ARS MATHEMATICA CONTEMPORANEA

Volume 16, Number 2, Spring/Summer 2019, Pages 277–641

Covered by:
Mathematical Reviews
zbMATH (formerly Zentralblatt MATH)
COBISS
SCOPUS
Science Citation Index-Expanded (SCIE)
Web of Science
ISI Alerting Service
Current Contents/Physical, Chemical & Earth Sciences (CC/PC & ES)
dblp computer science bibliography

The University of Primorska

The Society of Mathematicians, Physicists and Astronomers of Slovenia
The Institute of Mathematics, Physics and Mechanics
The Slovenian Discrete and Applied Mathematics Society

The publication is partially supported by the Slovenian Research Agency from the Call for co-financing of scientific periodical publications.

Petra Šparl Award 2020: Call for Nominations

The Petra Šparl Award was established in 2017 to recognise (in each even-numbered year) the best paper published recently by a young woman mathematician in one of the two journals *Ars Mathematica Contemporanea* (AMC) and *The Art of Discrete and Applied Mathematics* (ADAM).

The award is named in memory of Dr Petra Šparl, a talented woman mathematician with promising future who worked in graph theory and combinatorics, but died mid-career in 2016 after a battle with cancer.

This award consists of a certificate with the recipient's name, and an invitation to give a lecture at the Mathematics Colloquium at the University of Primorska, and to give lectures at the University of Maribor and University of Ljubljana.

The first award was made in May 2018 to Dr Monika Pilśniak (AGH University, Kraków, Poland), for her work on the distinguishing index, in *Ars Mathematica Contemporanea* **13** (2017), 259–274.

The Petra Šparl Award Committee is now calling for nominations for the second award.

Eligibility: Each nominee must be a woman author or co-author of a paper published either in AMC or ADAM in the last five years, who was at most 40 years old at the time of the paper's first submission.

Nomination format: Each nomination should specify the following:

(a) the name, birth-date and affiliation of the candidate;

(b) the title and other bibliographic details of the paper for which the award is recommended;

(c) reasons why the candidate's contribution to the paper is worthy of the award, in at most 500 words; and

(d) names and email addresses of one or two referees who could be consulted with regard to the quality of the paper.

Procedure: Nominations should be submitted by email to any one of the three members of the Petra Šparl Award Committee (see below), by 31 August 2019.

Award Committee:

- Marston Conder, `m.conder@auckland.ac.nz`
- Asia Ivić Weiss, `weiss@mathstat.yorku.ca`
- Aleksander Malnič, `aleksander.malnic@guest.arnes.si`

Marston Conder, Asia Ivić Weiss and Aleksander Malnič
Members of the 2020 Petra Šparl Award Committee

Contents

Volume 16, Number 2, Spring/Summer 2019, Pages 277–641

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 277–298
https://doi.org/10.26493/1855-3974.1601.e75
(Also available at http://amc-journal.eu)

Smallest snarks with oddness 4 and cyclic connectivity 4 have order 44*

Jan Goedgebeur
Department of Applied Mathematics, Computer Science & Statistics, Ghent University,
Krijgslaan 281-S9, 9000 Ghent, Belgium and
Computer Science Department, University of Mons,
Place du Parc 20, 7000 Mons, Belgium

Edita Máčajová , Martin Škoviera
Department of Computer Science, Comenius University, 842 48 Bratislava, Slovakia

Received 14 February 2018, accepted 22 September 2018, published online 6 January 2019

Abstract

The family of snarks – connected bridgeless cubic graphs that cannot be 3-edge-coloured – is well-known as a potential source of counterexamples to several important and long-standing conjectures in graph theory. These include the cycle double cover conjecture, Tutte's 5-flow conjecture, Fulkerson's conjecture, and several others. One way of approaching these conjectures is through the study of structural properties of snarks and construction of small examples with given properties. In this paper we deal with the problem of determining the smallest order of a nontrivial snark (that is, one which is cyclically 4-edge-connected and has girth at least 5) of oddness at least 4. Using a combination of structural analysis with extensive computations we prove that the smallest order of a snark with oddness at least 4 and cyclic connectivity 4 is 44. Formerly it was known that such a snark must have at least 38 vertices and one such snark on 44 vertices was constructed by Lukoťka, Máčajová, Mazák and Škoviera in 2015. The proof requires determining all cyclically 4-edge-connected snarks on 36 vertices, which extends the previously compiled list of all such snarks up to 34 vertices. As a by-product, we use this new list to test the validity of several conjectures where snarks can be smallest counterexamples.

Keywords: Cubic graph, cyclic connectivity, edge-colouring, snark, oddness, computation.

Math. Subj. Class.: 05C15, 05C21, 05C30, 05C40, 05C75, 68R10

*The first author was supported by a Postdoctoral Fellowship of the Research Foundation Flanders (FWO). The second and the third author were partially supported by VEGA 1/0876/16 and by APVV-15-0220. Most computations for this work were carried out using the Stevin Supercomputer Infrastructure at Ghent University.

E-mail addresses: jan.goedgebeur@ugent.be (Jan Goedgebeur), macajova@dcs.fmph.uniba.sk (Edita Máčajová), skoviera@dcs.fmph.uniba.sk (Martin Škoviera)

1 Introduction

Snarks are an interesting, important, but somewhat mysterious family of cubic graphs whose characteristic property is that their edges cannot be properly coloured with three colours. Very little is known about the nature of snarks because the reasons which cause the absence of 3-edge-colourability in cubic graphs are not well understood. Snarks are also difficult to find because almost all cubic graphs are hamiltonian and hence 3-edge-colourable [44]. On the other hand, deciding whether a cubic graph is 3-edge-colourable or not is NP-complete [26], implying that the family of snarks is sufficiently rich.

The importance of snarks resides mainly in the fact that many difficult conjectures in graph theory, such as Tutte's 5-flow conjecture or the cycle double cover conjecture, would be proved in general if they could be established for snarks [29, 30]. While most of these problems are trivial for 3-edge-colourable graphs, and exceedingly difficult for snarks in general, they often become tractable for snarks that are in a certain sense close to being 3-edge-colourable.

There exist a number of measures of uncolourability of cubic graphs (see [16] for a recent survey). Among them, the smallest number of odd circuits in a 2-factor of a cubic graph, known as *oddness*, has received the widest attention. Note that the oddness of a cubic graph is an even integer which equals zero precisely when the graph is 3-edge-colourable. It is known, for example, that the 5-flow conjecture and the Fan-Raspaud conjecture are true for cubic graphs of oddness at most two [30, 37], while the cycle double cover conjecture is known to hold for cubic graphs of oddness at most 4 [24, 27]. Snarks with large oddness thus still remain potential counterexamples to these conjectures and therefore merit further study.

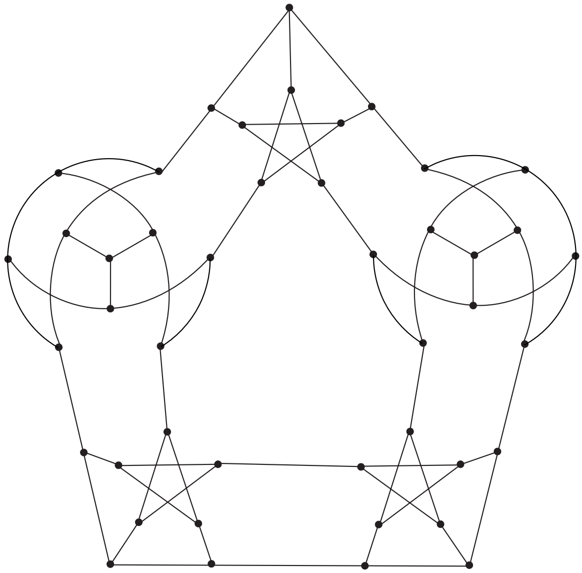


Figure 1: The smallest known nontrivial snark with oddness ≥ 4.

Several authors have provided constructions of infinite families of snarks with increasing oddness, see, for example, [25, 33, 35, 49]. Most of them focus on snarks with cyclic connectivity at least 4 and girth at least 5, because snarks that lack these two properties can be easily reduced to smaller snarks. We call such snarks *nontrivial*. All currently available constructions indicate that snarks of oddness greater than 2 are extremely rare. From [7, Observation 4.10] it follows that there exist no nontrivial snarks of oddness greater than 2 on up to 36 vertices. The smallest known example of a nontrivial snark with oddness at least 4 has 44 vertices and its oddness equals 4. It was constructed by Lukot'ka et al. in [35], superseding an earlier construction of Hägglund [25] on 46 vertices; it is shown in Figure 1 in a form different from the one displayed in [35]. In [35, Theorem 12] it is also shown that if we allow trivial snarks, the smallest one with oddness greater than 2 has 28 vertices and oddness 4. As explained in [22, 34], there are exactly three such snarks, one with cyclic connectivity 3 and two with cyclic connectivity 2. (The latter result rectifies the false claim made in [35] that there are only two snarks of oddness 4 on 28 vertices.)

The aim of the present paper is to prove the following result.

Theorem 1.1. *The smallest number of vertices of a snark with cyclic connectivity* 4 *and oddness at least* 4 *is* 44*. The girth of each such snark is at least* 5.

This theorem bridges the gap between the order 36 up to which all nontrivial snarks have been generated (and none of oddness greater than 2 was found [7]) and the order 44 where an example of oddness 4 has been constructed [35]. Since generating all nontrivial snarks beyond 36 vertices seems currently infeasible, it would be hardly possible to find a smallest nontrivial snark with oddness at least 4 by employing computational force alone. On the other hand, the current state-of-the-art in the area of snarks, with constructions significantly prevailing over structural theorems, does not provide sufficient tools for a purely theoretical proof of our theorem. Our proof is therefore an inevitable combination of structural analysis of snarks with computations.

The proof consists of two steps. First we prove that every snark with oddness at least 4, cyclic connectivity 4, and minimum number of vertices can be decomposed into two smaller cyclically 4-edge-connected snarks G_1 and G_2 by removing a cycle-separating 4-edge-cut, adding at most two vertices to each of the components, and by restoring 3-regularity. Conversely, every such snark arises from two smaller cyclically 4-edge-connected snarks G_1 and G_2 by the reverse process. In the second step of the proof we computationally verify that no combination of G_1 and G_2 can result in a cyclically 4-edge-connected snark of oddness at least 4 on fewer than 44 vertices. This requires checking all suitable pairs of cyclically 4-edge-connected snarks on up to 36 vertices, including those that contain 4-cycles. Such snarks have been previously generated only up to order 34 [7], which is why we had to additionally generate all cyclically 4-edge-connected snarks on 36 vertices containing a 4-cycle. This took about 80 CPU years and yielded exactly $404\,899\,916$ cyclically 4-edge-connected snarks.

It is important to realise that Theorem 1.1 does not yet determine the order of a smallest nontrivial snark with oddness at least 4. The reason is that it does not exclude the existence of cyclically 5-connected snarks with oddness at least 4 on fewer than 44 vertices. However, the smallest currently known cyclically 5-edge-connected snark with oddness at least 4 has 76 vertices (see Steffen [49, Theorem 2.3]), which indicates that a cyclically 5-edge-connected snark with oddness at least 4 on fewer than 44 vertices either does not exist or will be very difficult to find.

Our paper is organised as follows. Section 2 provides the necessary background material for the proof of Theorem 1.1 and for the results that precede it, in particular for the decomposition theorems proved in Section 3. In Section 4 we employ these decomposition theorems to prove Theorem 1.1. We further discuss this theorem in Section 5 where we also pose two related problems. In the final section we report about the tests which we have performed on the set of all cyclically 4-edge-connected snarks of order 36 concerning the validity of several interesting conjectures in graph theory, such as the dominating cycle conjecture, the total colouring conjecture, and the Petersen colouring conjecture.

We will continue our investigation of the smallest snarks with oddness at least 4 and cyclic connectivity 4 in the sequel of this paper [23]. We will display a set of 31 such snarks, analyse their properties, and prove that they constitute the complete set of snarks with oddness at least 4 and cyclic connectivity 4 on 44 vertices.

2 Preliminaries

2.1 Graphs and multipoles

All graphs in this paper are finite. For the sake of completeness, we have to permit graphs containing multiple edges or loops, although these features will in most cases be excluded by the imposed connectivity or colouring restrictions.

Besides graphs we also consider graph-like structures, called *multipoles*, that may contain dangling edges and even isolated edges. Multipoles serve as a convenient tool for constructing larger graphs from smaller building blocks. They also naturally arise as a result of severing one or several edges of a graph, in particular edges forming an edge-cut. In this paper all multipoles will be cubic (3-valent).

Every edge of a multipole has two ends and each end can, but need not, be incident with a vertex. An edge which has both ends incident with a vertex is called *proper*. If one end of an edge is incident with a vertex and the other is not, then the edge is called a *dangling edge* and, if neither end of an edge is incident with a vertex, it is called an *isolated edge*. An end of an edge that is not incident with a vertex is called a *semiedge*. A multipole with k semiedges is called a *k-pole*. Two semiedges s and t of a multipole can be joined to produce an edge $s * t$ connecting the end-vertices of the corresponding dangling edges. Given two k-poles M and N with semiedges $s_1, \dots, s_k$ and $t_1, \dots, t_k$, respectively, we define their *complete junction* $M * N$ to be the graph obtained by performing the junctions $s_i * t_i$ for each $i \in \{1, \dots, k\}$. A *partial junction* is defined in a similar way except that a proper subset of semiedges of M is joined to semiedges of N. Partial junctions can be used to construct larger multipoles from smaller ones. In either case, whenever a junction of two multipoles is to be performed, we assume that their semiedges are assigned a fixed order. For a more detailed formal development of concepts related to multipoles we refer the reader, for example, to [15, 36] or [13].

2.2 Cyclic connectivity

Let G be a connected graph. An *edge-cut* of a graph G, or just a *cut* for short, is any set S of edges of G such that $G - S$ is disconnected. An edge-cut is said to be *trivial* if it consists of all edges incident with one vertex, and *nontrivial* otherwise. An important kind of an edge-cut is a cocycle, which arises by taking a set of vertices or an induced subgraph H of G and letting S to be the set $\delta_G(H)$ of all edges with exactly one end in H. We omit the subscript G whenever G is clear from the context.

An edge-cut is said to be *cycle-separating* if at least two components of $G - S$ contain cycles. We say that a connected graph G is *cyclically k-edge-connected* if no set of fewer than k edges is cycle-separating in G. The *cyclic connectivity* of G, denoted by $\zeta(G)$, is the largest number $k \leq \beta(G)$, where $\beta(G) = |E(G)| - |V(G)| + 1$ is the cycle rank of G, for which G is cyclically k-connected (cf. [41, 43]).

It is not difficult to see that for a cubic graph G with $\zeta(G) \leq 3$ the value $\zeta(G)$ coincides with the usual vertex-connectivity or edge-connectivity of G. Thus cyclic connectivity in cubic graphs is a natural extension of the common versions of connectivity (which unlike cyclic connectivity are bounded above by 3). Another useful observation is that the value of cyclic connectivity remains invariant under subdivisions and adjoining new vertices of degree 1.

The following well-known result [41, 43] relates $\zeta(G)$ to the length of a shortest cycle in G, denoted by $g(G)$ and called the *girth* of G.

Proposition 2.1. *For every connected cubic graph G we have $\zeta(G) \leq g(G)$.*

Let us observe that in a connected cubic graph every edge-cut S consisting of independent edges is cycle-separating: indeed the minimum valency of $G - S$ is 2, so each component of $G - S$ contains a cycle. Conversely, a cycle-separating edge-cut of minimum size is easily seen to be independent; moreover, $G - S$ has precisely two components, called *cyclic parts* or *fragments*. A fragment minimal under inclusion will be called an *atom*. A *nontrivial atom* is any atom different from a shortest cycle.

The following two propositions provide useful tools in handling cyclic connectivity. The first of them follows easily by mathematical induction. For the latter we refer the reader to [41, Proposition 4 and Theorem 11].

Lemma 2.2. *Let H be a connected acyclic subgraph of a cubic graph separated from the rest by a k-edge-cut. Then H has $k - 2$ vertices.*

Proposition 2.3. *Let G be a connected cubic graph. The following statements hold:*

(i) Every fragment of G is connected, and every atom is 2-connected. Moreover, if $\zeta(G) \geq 3$, then every fragment is 2-connected.

(ii) If A is a nontrivial atom of G, then $\zeta(A) > \zeta(G)/2$.

In the present paper we focus on cyclically 4-edge-connected cubic graphs, in particular on those with cyclic connectivity exactly 4. From the results mentioned earlier it follows that a cyclically 4-edge-connected cubic graph has no bridges and no 2-edge-cuts. Furthermore, every 3-edge-cut separates a single vertex, and every 4-edge-cut which is not cycle-separating consists of the four edges adjacent to some edge.

An important method of constructing cyclically 4-edge-connected cubic graphs from smaller ones applies the following operation which we call an I-extension. In a cubic graph G take two edges e and f, subdivide each of e and f with a new vertex v_e and v_f, respectively, and by add a new edge between v_e and v_f. The resulting graph, denoted by $G(e, f)$ is said to be obtained by an I-*extension* across e and f. It is not difficult to see that if G is cyclically 4-edge-connected and e and f are non-adjacent edges of G, then so is $G(e, f)$.

A well-known theorem of Fontet [19] and Wormald [51] states that all cyclically 4-edge-connected cubic graphs can be obtained from the complete graph K_4 and the cube Q_3 by repeatedly applying I-extensions to pairs of non-adjacent edges. However, I-extensions are also useful for constructing cubic graphs in general. For example, in [8] all connected cubic graphs up to 32 vertices have been generated by using I-extensions as main construction operation.

For more information on cyclic connectivity the reader may wish to consult [41].

2.3 Edge-colourings

A k*-edge-colouring* of a graph G is a mapping $\phi\colon E(G) \to \mathbf{C}$ where $\mathbf{C}$ is a set of k colours. If all pairs of adjacent edges receive distinct colours, ϕ is said to be *proper*; otherwise it is called *improper*. Graphs with loops do not admit proper edge-colourings because of the self-adjacency of loops. Since we are mainly interested in proper colourings, the adjective "proper" will usually be dropped. For multipoles, edge-colourings are defined similarly; that is to say, each edge receives a colour irrespectively of the fact whether it is, or it is not, incident with a vertex.

The result of Shannon [47] implies that every loopless cubic graph, and hence every loopless cubic multipole, can be properly coloured with four colours, see also [32]. In the study of snarks it is often convenient to take the set of colours $\mathbf{C}$ to be the set $\mathbb{Z}_2 \times \mathbb{Z}_2 = \{(0,0),(0,1),(1,0),(1,1)\}$ where $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$ are identified with 0, 1, 2, and 3, respectively. We say that a multipole is *colourable* if it admits a 3-edge-colouring and *uncolourable* otherwise. For a 3-edge-colouring of a cubic graph or a cubic multipole we use the colour-set $\mathbf{C} = \{1,2,3\}$ because such a colouring is in fact a nowhere-zero $\mathbb{Z}_2 \times \mathbb{Z}_2$-flow. This means that for every vertex v the sum of colours incident with v, the *outflow* at v, equals 0 in $\mathbb{Z}_2 \times \mathbb{Z}_2$. The following fundamental result [5, 14] is a direct consequence of this fact.

Theorem 2.4 (Parity Lemma). *Let M be a k-pole endowed with a proper 3-edge-colouring with colours 1, 2, and 3. If the set of all semi-edges contains k_i edges of colour i for $i \in \{1,2,3\}$, then*

$$k_1 \equiv k_2 \equiv k_3 \equiv k \pmod 2.$$

Now let M be a loopless cubic multipole that cannot be properly 3-edge-coloured. Then M has a proper 4-edge-colouring with colours from the set $\mathbf{C} = \mathbb{Z}_2 \times \mathbb{Z}_2$. Such a colouring will not be a $\mathbb{Z}_2 \times \mathbb{Z}_2$-flow anymore since every vertex incident with an edge coloured 0 will have a non-zero outflow. It is natural to require the colour 0 to be used as little as possible, that is, to require the set of edges coloured 0 to be the minimum-size colour class. Such a 4-edge-colouring will be called *minimum*. In a minimum 4-edge-colouring of M every edge e coloured 0 must be adjacent to edges of all three non-zero colours; in particular, e must be a proper edge. It follows that exactly one colour around e appears twice.

By summing the outflows at vertices incident with edges coloured 0 we obtain the following useful result due to Fouquet [20, Theorem 1] and Steffen [48, Lemma 2.2].

Theorem 2.5. *Let ϕ be a minimum 4-edge-colouring of a loopless cubic multipole M with m edges coloured 0, and for $i \in \{1,2,3\}$ let m_i denote the number of those edges coloured 0 that are adjacent to two edges coloured i. Then*

$$m_1 \equiv m_2 \equiv m_3 \equiv m \pmod 2.$$

We finish the discussion of colourings with the definition of the standard recolouring tool, a Kempe chain. Let M be a cubic multipole whose edges have been properly coloured with colours from the set $\{0,1,2,3\} = \mathbb{Z}_2 \times \mathbb{Z}_2$. For any two distinct colours $i, j \in \{1,2,3\}$ we define an *i-j-Kempe chain P* to be a non-extendable walk that alternates the edges with colours i and j. Clearly, P is either an even circuit, or is a path that ends with either a semiedge or with a vertex incident with an edge coloured 0. It is easy to see that switching the colours i and j on P gives rise to a new proper 4-edge-colouring of M. Furthermore, if the original colouring was a minimum 4-edge-colouring, so is the new one.

2.4 Snarks

A snark is, essentially, a nontrivial cubic graph that has no 3-edge-colouring. Precise definitions vary depending on what is to be considered "nontrivial". In many papers, especially those dealing with snark constructions, snarks are required to be cyclically 4-edge-connected and have girth at least 5; see for example [12, 16]. However, in [9, 25] the girth requirement is dropped, demanding snarks to be cyclically 4-edge-connected but allowing them to have 4-cycles.

Another group of papers, especially those dealing with the structural analysis of snarks, adopts the widest possible definition of a snark, permitting all kinds of trivial features such as triangles, digons and even bridges; see, for example [11, 13, 42]. In this paper, our usage of the term snark agrees with the latter group: we define a *snark* to be a connected cubic graph that cannot be 3-edge-coloured.

This paper deals with snarks that are far from being 3-edge-colourable. Two measures of uncolourability will be prominent in this paper. The *oddness* $\omega(G)$ of a bridgeless cubic graph G is the smallest number of odd circuits in a 2-factor of G. The *resistance* $\rho(G)$ of a cubic graph G is the smallest number of edges of G which have to be removed in order to obtain a colourable graph. Obviously, if G is colourable, then $\omega(G) = \rho(G) = 0$. If G is uncolourable, then both $\omega(G) \geq 2$ and $\rho(G) \geq 2$. Furthermore, $\rho(G) \leq \omega(G)$ for every bridgeless cubic graph G.

The following lemma is due to Steffen [48].

Lemma 2.6. *Let G be a bridgeless cubic graph. Then $\rho(G) = 2$ if and only if $\omega(G) = 2$.*

One of the methods of constructing snarks from smaller ones uses I-extensions (cf. Subsection 2.2). The following result from [42] tells us when an I-extension of a snark is again a snark.

Lemma 2.7. *Let G be a snark and e and f be distinct edges of G. Then $G(e,f)$ is a snark if and only if the graph $G - \{e,f\}$ is uncolourable.*

Another method of constructing snarks is based on extending multipoles to cubic graphs, see [13]. If the multipole in question is uncolourable, it can be extended to a snark simply by restoring 3-regularity. We are therefore interested in extending colourable multipoles. For $k \geq 2$, we say that a k-pole M *extends* to a snark if there exists a colourable multipole N such that $M * N$ is a snark. The graph $M * N$ is called a snark *extension* of M.

Given a k-pole M with semiedges $e_1, e_2, \ldots, e_k$, we define its *colouring set* to be the following set of k-tuples:

$$\mathrm{Col}(M) = \{\phi(e_1)\phi(e_2)\ldots\phi(e_k) : \phi \text{ is a 3-edge-colouring of } M\}.$$

Note that the set $\mathrm{Col}(M)$ depends on the ordering in which the semiedges are listed. We therefore implicitly assume that such an ordering is given. As the colourings "inside" a multipole can usually be ignored, we define two multipoles M and N to be *colour-equivalent* if $\mathrm{Col}(M) = \mathrm{Col}(N)$.

Any colouring of a colourable multipole can be changed to a different colouring by permuting the set of colours. The particular colour of a semiedge is therefore not important, it is only important whether it equals or differs from the colour of any other semiedge. By saying this we actually define the *type* of a colouring ϕ of a multipole M: it is the lexicographically smallest sequence of colours assigned to the semiedges of M which can be obtained from ϕ by permuting the colours.

By the Parity Lemma (Theorem 2.4), each colouring of a 4-pole has one of the following types: 1111, 1122, 1212, and 1221. Observe that every colourable 4-pole admits at least two different types of colourings. Indeed, we can start with any colouring and switch the colours along an arbitrary Kempe chain to obtain a colouring of another type. Colourable 4-poles thus can have two, three, or four different types of colourings. Those attaining exactly two types are particularly important for the study of snarks; we call them *colour-open* 4-poles, as opposed to *colour-closed* multipoles discussed in more detail in [42].

The following result appears in [13].

Proposition 2.8. *A colourable* 4*-pole extends to a snark if and only if it is colour-open.*

A 4-pole M will be called *isochromatic* if its semiedges can be partitioned into two pairs such that in every colouring of M the semiedges within each pair are coloured with the same colour. A 4-pole M will be called *heterochromatic* of its semiedges can be partitioned into two pairs such that in every colouring of M the semiedges within each pair are coloured with distinct colours. The pairs of semiedges of an isochromatic or a heterochromatic 4-pole mentioned above will be called *couples*.

Note that the 4-pole C_4 obtained from a 4-cycle by attaching one dangling edge to every vertex is colour-closed, and hence neither isochromatic nor heterochromatic. Indeed, with respect to a cyclic ordering of its semiedges it admits colourings of three types, namely 1111, 1122, and 1221 (but not 1212). In particular, if a snark G contains a 4-cycle C, then, as is well-known, $G - V(C)$ stays uncolourable.

The following two results are proved in [13]:

Proposition 2.9. *Every colour-open* 4*-pole is either isochromatic or heterochromatic, but not both. Moreover, it is isochromatic if and only if it admits a colouring of type* 1111*.*

Proposition 2.10. *Every colour-open* 4*-pole can be extended to a snark by adding at most two vertices, and such an extension is unique. A heterochromatic multipole extends by joining the semiedges within each couple, that is, by adding no new vertex. An isochromatic multipole extends by attaching the semiedges of each couple to a new vertex, and by connecting these two vertices with a new edge.*

Colour-open 4-poles can be combined to form larger 4-poles from smaller ones by employing partial junctions: we take two 4-poles M and N, choose two semiedges in each of them, and perform the individual junctions. In general, such a junction need not respect the structure of the couples of the 4-poles participating in the operation. In this manner it may happen that, for example, a partial junction of two heterochromatic 4-poles results in an isochromatic dipole or in a heterochromatic dipole. In Theorem 3.5, one of our decomposition theorems, partial junctions of 4-poles will occur in the reverse direction.

3 Decomposition theorems

The aim of this section is to show that every snark with oddness at least 4, cyclic connectivity 4, and minimum number of vertices can be decomposed into two smaller cyclically 4-edge-connected snarks G_1 and G_2 by removing a cycle-separating 4-edge-cut, adding at most two vertices to each of the components, and by restoring 3-regularity. This will be proved in two steps – Theorem 3.3 and Theorem 3.5.

Theorem 3.3 is a decomposition theorem for cyclically 4-edge-connected cubic graphs proved in 1988 by Andersen et al. [2, Lemma 7]. Roughly speaking, it states that every cubic graph G whose cyclic connectivity equals 4 can be decomposed into two smaller cyclically 4-edge-connected cubic graphs G_1 and G_2 by removing a cycle-separating 4-edge-cut, adding two vertices to each of the components, and by restoring 3-regularity. Our proof is different from the one in [2] and provides useful insights into the problem. For instance, it offers the possibility to determine conditions under which it is feasible to extend a 4-pole to a cyclically 4-edge-connected cubic graph by adding two isolated edges rather than by adding two new vertices.

Theorem 3.5 deals with a particular situation where the cyclically 4-edge-connected cubic graph G in question is a snark. As explained in the previous section, every snark containing a cycle-separating 4-edge-cut that leaves a colour-open component can be decomposed into two smaller snarks G_1 and G_2 by removing the cut, adding *at most two vertices* to each of the components, and by restoring 3-regularity. Unfortunately, G_1 or G_2 are not guaranteed to be cyclically 4-edge-connected because snark extensions forced by the colourings need not coincide with those forced by the cyclic connectivity (see Example 3.1 below). Moreover, Proposition 2.10 suggests that restoring 3-regularity by adding no new vertices, that is, by joining pairs of the four 2-valent vertices to each other in one of the components, may be necessary in order for G_1 or G_2 to be a snark. If this is the case, Theorem 3.3 cannot be applied. Nevertheless, Theorem 3.5 shows that if G is a smallest nontrivial snark with oddness at least 4, then we can form G_1 and G_2 in such a way that they indeed will be cyclically 4-edge-connected snarks.

Example 3.1. We give an example of a cyclically 4-edge-connected snark in which a decomposition along a given cycle-separating 4-edge-cut forces one of the resulting smaller snarks to have cyclic connectivity smaller than 4. To construct such a snark take the Petersen graph and form a 4-pole H of order 10 by severing two non-adjacent edges and a 4-pole I of order 8 by removing two adjacent vertices. It is easy to see that H is heterochromatic with couples being formed by the semiedges obtained from the same edge, and I is isochromatic with couples formed by the semiedges formerly incident with the same vertex. Let us create a cubic graph G by arranging two copies of H and one copy of I into a cycle, and by performing junctions that respect the structure of the couples. The partial junction of two copies of H contained in G, denoted by H^2, is again a heterochromatic 4-pole, so G is a junction of an isochromatic 4-pole I with a heterochromatic 4-pole, and therefore a snark. Furthermore, the cyclic connectivity of G equals 4. Let us decompose G by removing from G the 4-edge-cut S separating I from H^2 and by completing each of the components to a snark. Proposition 2.10 implies that I can be completed to a copy G' of the Petersen graph while H^2 extends to a snark G'' of order 20 by joining the semiedges within each couple, that is, by adding no new vertex. The same Proposition states that the decomposition of G into G' and G'' is uniquely determined by S. However, G'' has a cycle-separating 2-edge-cut connecting the two copies of H contained in it. Therefore the low connectivity of G'' is unavoidable.

We proceed to Theorem 3.3. It requires one auxiliary result about comparable cuts. Let S and T be two edge-cuts in a graph G. Let us denote the two components of $G - S$ by H_1 and H_2 and those of $G - T$ by K_1 and K_2. The cuts S and T are called *comparable* if $H_i \subseteq K_j$ or $K_j \subseteq H_i$ for some $i, j \in \{1, 2\}$.

Lemma 3.2. *Let G be a cyclically 4-edge-connected cubic graph and let K be a component arising from the removal of a cycle-separating 4-edge-cut from G. Then any two nontrivial 2-edge-cuts in K are comparable, or K is a 4-cycle.*

Proof. Let S be the cycle-separating 4-edge-cut that separates K from the rest of G, and let $A = \{a_1, a_2, a_3, a_4\}$ be the set of the vertices of K incident with an edge from S. Since S is independent, the vertices of A are pairwise distinct. Proposition 2.3 (i) implies that K is 2-connected. It follows that for every nontrivial 2-edge-cut Q in K the graph $K - Q$ consists of two components, each containing exactly two vertices of A.

Let R and T be two nontrivial 2-edge-cuts in K. Denote the components of $K - R$ by X_1 and X_2, and those of $K - T$ by Y_1 and Y_2. Observe that the subgraphs $X_i \cap Y_j$ for $i, j \in \{1, 2\}$ need not all be non-empty. Let a be the number of edges between $X_1 \cap Y_1$ and $X_1 \cap Y_2$, b the number of edges between $X_1 \cap Y_1$ and $X_2 \cap Y_2$, c the number of edges between $X_1 \cap Y_1$ and $X_2 \cap Y_1$, d the number of edges between $X_1 \cap Y_2$ and $X_2 \cap Y_1$, e the number of edges between $X_2 \cap Y_1$ to $X_2 \cap Y_2$, and finally f the number of edges between $X_1 \cap Y_2$ and $X_2 \cap Y_2$; see Figure 2.

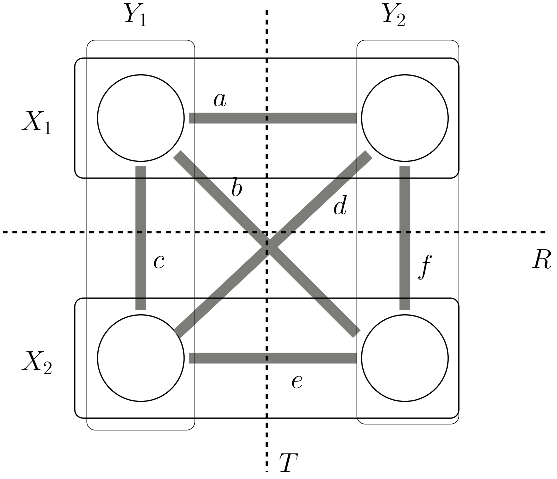


Figure 2: Crossing edge-cuts R and T.

If at least one of the sets $X_1 \cap Y_1$, $X_1 \cap Y_2$, $X_2 \cap Y_1$, and $X_2 \cap Y_2$ is empty, then the definition of comparable cuts directly implies that the cuts R and T are comparable, as required. Thus we can assume that all the subgraphs $X_i \cap Y_j$ are nonempty. Our aim is to show that in this case K is a 4-cycle. We start by showing that each of the subgraphs $X_1 \cap Y_1$, $X_1 \cap Y_2$, $X_2 \cap Y_1$, and $X_2 \cap Y_2$ contains exactly one element of A. Suppose that one of them, say $X_1 \cap Y_1$, contains no vertex from A. Since both R and T separate the vertices from A in such a way that both components contain two vertices from A, we deduce that both $X_1 \cap Y_2$ and $X_2 \cap Y_1$ contain two vertices from A each, while $X_2 \cap Y_2$ contains no vertex from A. Now $|\delta_K(X_1 \cap Y_1)| = a + b + c \geq 3$, because G has no bridges and no 2-edge-cuts. Further, since $X_1 \cap Y_2$ contains two vertices from A and G is cyclically 4-edge-connected, we see that $|\delta_K(X_1 \cap Y_2)| = a + d + f \geq 2$. However, R is a 2-cut, so $b + c + d + f = 2$. Therefore $2a \geq 3$ and hence $a \geq 2$. Similarly, $e \geq 2$. But then $|T| \geq a + e \geq 4$, which contradicts the fact that T is a 2-edge-cut. Thus all the subgraphs $X_i \cap Y_j$ contain an element of A, which in turn implies that each $X_i \cap X_j$ contains exactly one vertex from A.

To finish the proof we show that $a = c = e = f = 1$ and $b = d = 0$. Suppose that $a = 2$. Since T is a 2-edge-cut, we have that $b = d = e = 0$. Now $c + d + e \geq 2$ and $b + e + f \geq 2$ because G is 3-edge-connected, so $c \geq 2$ and $f \geq 2$, and hence $|R| \geq c + f \geq 4$, a contradiction. Thus $a \leq 1$. Similarly, we can derive that $c \leq 1$, $e \leq 1$, and $f \leq 1$. If $b = 2$, then $a = c = d = e = f = 0$ implying that G has a bridge, which is absurd. Hence $b \leq 1$ and similarly $d \leq 1$. Suppose that $a = 0$. As G is 3-edge-connected, we have $2 \leq a + b + c = b + c \leq 2$ and similarly $2 \leq a + d + f = d + f \leq 2$. It follows that that $b = c = d = f = 1$ and hence $|R| = b + c + d + f = 4$, which contradicts the fact that R is a 2-cut. Therefore $a = 1$ and similarly $c = e = f = 1$, which also implies that $b = d = 0$. Finally, every subgraph $X_i \cap Y_j$ has $|\delta_G(X_i \cap Y_j)| = 3$, so $X_i \cap Y_j$ is acyclic and therefore, by Lemma 2.2, a single vertex. In other words, K is a 4-cycle. This completes the proof. □

We are ready to prove the decomposition theorem of Andresen et al. [2].

Theorem 3.3. *Let G be a cyclically 4-edge-connected cubic graph with a cycle-separating 4-edge-cut whose removal leaves components G_1 and G_2. Then each of G_1 and G_2 can be extended to a cyclically 4-edge-connected cubic graph by adding two adjacent vertices and restoring 3-regularity.*

Proof. It suffices to prove the statement for G_1. If G_1 is a 4-cycle, we can easily extend it to the complete bipartite graph $K_{3,3}$ which is cyclically 4-edge-connected, as required. We therefore assume that G_1 is not a 4-cycle. Let $A = \{a_1, a_2, a_3, a_4\}$ be the set of vertices of G_1 incident with an edge of $\delta_G(G_1)$. By Lemma 3.2, every 2-edge-cut in G_1 separates the vertices of A into the same two 2-element sets, say $\{a_1, a_2\}$ from $\{a_3, a_4\}$. We extend G_1 to a cyclically 4-edge-connected cubic graph $\tilde{G}_1$ as follows. Let us take two new vertices x_1 and x_2 and construct $\tilde{G}_1$ from G_1 by adding to G_1 the edges x_1x_2, x_1a_1, x_1a_3, x_2a_2,

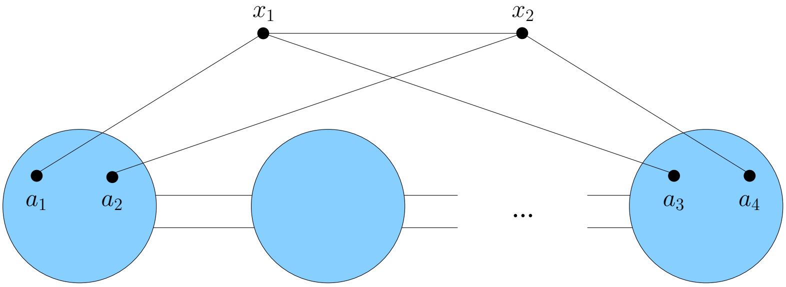


Figure 3: Extending G_1 to $\tilde{G}_1$.

and x_2a_4, see Figure 3. We now verify that $\tilde{G}_1$ is indeed cyclically 4-edge-connected.

Suppose to the contrary that $\tilde{G}_1$ is not cyclically 4-edge-connected. Then $\tilde{G}_1$ has a minimum-size cycle-separating edge-cut F such that $|F| < 4$. Let H_1 and H_2 be the components of $G_1 - F$. The cut F cannot consist entirely of edges of $G_1 \cup \delta_G(G_1)$ for otherwise F would be a cycle-separating edge-cut of G of size smaller than 4. Therefore the edge x_1x_2 is contained in F. Since F is an independent cut, the edges x_1a_1, x_1a_3, x_2a_2, and x_2a_4 do not belong to F. This in turn implies that a_1 and a_3 belong to one component of $\tilde{G}_1 - F$ while a_2 and a_4 belong to the other component of $\tilde{G}_1 - F$; without loss of generality, let a_1 and a_3 belong to H_1. Since G_1 contains no bridge, there exist edges e_1 and e_2 in G_1 such that $F = \{x_1x_2, e_1, e_2\}$. But then $\{e_1, e_2\}$ is a 2-edge cut in G_1 that separates the set $\{a_1, a_3\}$ from $\{a_2, a_4\}$, which is a contradiction. This completes the proof. □

Before proving the second main result of this section we need the following fact.

Proposition 3.4. *Let G be a cubic graph with a cycle-separating 4-edge-cut whose removal leaves components G_1 and G_2. If both G_1 and G_2 are 3-edge-colourable, then $\omega(G) \leq 2$.*

Proof. Assume that both G_1 and G_2 are 3-edge-colourable. If G is 3-edge-colourable, then $\omega(G) = 0$. Therefore we may assume that G is not 3-edge-colourable. In this situation G_1 admits two types of colourings and G_2 admits the other two types of colourings. One of them, say G_1 has a colouring ϕ_1 of the type 1111; by Proposition 2.9, G_1 is isochromatic and G_2 is heterochromatic. Parity Lemma (Theorem 2.4) implies that if we take an arbitrary 3-edge-colouring ϕ_2 of G_2, then exactly two colours occur on the dangling edges of G_2. Let e and f be any two of the dangling edges that receive the same colour. Then, after permuting the colours in G_1, if necessary, ϕ_1 and ϕ_2 can be easily combined to a 3-edge-colouring of $G - \{e, f\}$. This shows that $\rho(G) = 2$ and therefore $\omega(G) = 2$. □

Now we are in position to prove our second decomposition theorem.

Theorem 3.5. *Let G be a snark with oddness at least 4, cyclic connectivity 4, and minimum number of vertices. Then G contains a cycle-separating 4-edge-cut S such that both components of $G - S$ can be extended to a cyclically 4-edge-connected snark by adding at most two vertices.*

In fact, we prove the following stronger and more detailed result which will also be needed in our next paper [23].

Theorem 3.6. *Let G be a snark with oddness at least 4, cyclic connectivity 4, and minimum number of vertices. Let S be a cycle-separating 4-edge-cut in G whose removal leaves components G_1 and G_2. Then, up to permutation of the index set $\{1, 2\}$, exactly one of the following occurs.*

(i) Both G_1 and G_2 are uncolourable, in which case each of them can be extended to a cyclically 4-edge-connected snark by adding two vertices.

(ii) G_1 is uncolourable and G_2 is heterochromatic, in which case G_1 can be extended to a cyclically 4-edge-connected snark by adding two vertices, and G_2 can be extended to a cyclically 4-edge-connected snark by adding two isolated edges.

(iii) G_1 is uncolourable and G_2 is isochromatic, in which case G_1 can be extended to a cyclically 4-edge-connected snark by adding two vertices, and G_2 can be extended to a cyclically 4-edge-connected snark by adding two vertices, except possibly $\zeta(G_2) = 2$. In the latter case, G_2 is a partial junction of two colour-open 4-poles, which may be isochromatic or heterochromatic in any combination.

Proof. Let G be a snark with $\omega(G) \geq 4$, $\zeta(G) = 4$, and with minimum number of vertices. Let $S = \{s_1, s_2, s_3, s_4\}$ be an arbitrary fixed cycle-separating 4-edge-cut in G, and let G_1 and G_2 be the components of $G - S$. According to Proposition 3.4, at least one of G_1 and G_2 is uncolourable. If both G_1 and G_2 are uncolourable, we can extend each of them to a cyclically 4-edge-connected snark by applying Theorem 3.3, establishing (i). For the rest of the proof we may therefore assume that G_2 is colourable and G_1 is not. Again, G_1 can be extended to a cyclically 4-edge-connected snark by Theorem 3.3. Let $\tilde{G}_1$ be an extension of G_1 to a cyclically 4-edge-connected snark by adding two adjacent vertices y_1 and y_2 according to Theorem 3.3. Without loss of generality we may assume that the vertex y_1 is incident with the edges s_1 and s_2 while y_2 is incident with s_3 and s_4.

As regards G_2, we prove that either (ii) or (iii) holds. Our first step in this direction is showing that G_2 can be extended to a snark. In view of Proposition 2.8, this amounts to verifying that G_2 is colour-open.

Claim 1. *The 4-pole G_2 is colour-open.*

Proof of Claim 1. Suppose to the contrary that G_2 is not colour-open. This means that it has at least three types of colourings. Since G is a smallest cyclically 4-edge-connected snark with oddness at least 4 and $\tilde{G}_1$ is a cyclically 4-edge-connected snark with fewer vertices than G, we infer that $\omega(\tilde{G}_1) = 2$. By Lemma 2.6, there exist two nonadjacent edges e_1 and e_2 in $\tilde{G}_1$ such that $\tilde{G}_1 - \{e_1, e_2\}$ is colourable. Equivalently, by Lemma 2.7, the cubic graph $\tilde{G}_1(e_1, e_2)$ is colourable.

We claim that the edge y_1y_2 is one of e_1 and e_2. Suppose not. Then both e_1 and e_2 have at least one end-vertex in G_1. As mentioned, $\tilde{G}_1(e_1, e_2)$ is a colourable cubic graph. Hence $G_1(e_1, e_2)$ is a colourable 4-pole, and therefore it has at least two types of colourings. Since G_2 has at least three of the four types, both $G_1(e_1, e_2)$ and G_2 admit colourings of the same type. These colourings can be combined into a colouring of $G(e_1, e_2)$, implying that $G - \{e_1, e_2\}$ is also colourable. However, from Lemma 2.6 we get that $\omega(G) = 2$, which is a contradiction proving that one of e_1 and e_2 coincides with y_1y_2.

Assuming that $y_1y_2 = e_1$, let us consider a minimum 4-edge-colouring ϕ_1 of $\tilde{G}_1$ where e_1 and e_2 are the only edges of $\tilde{G}_1$ coloured 0. Theorem 2.5 implies that there exist a unique non-zero colour that is repeated at both e_1 and e_2. Without loss of generality we may assume that the repeated colour is 1 and that $\phi_1(s_1) = \phi_1(s_3) = 1$, $\phi_1(s_2) = 2$, and $\phi_1(s_4) = 3$. In this situation, G_2 cannot have a colouring of type 1212 for otherwise we could combine this colouring with ϕ_1 to produce a 3-edge-colouring of $G - \{e_2, s_4\}$, which is impossible since $\omega(G) \geq 4$. Therefore G_2 has colourings of all the remaining three types 1111, 1122, and 1221.

Consider the 1-2-Kempe chain P in $\tilde{G}_1$ with respect to the colouring ϕ_1 beginning at the vertex y_2. Clearly, the other end of P must be the end-vertex of e_2 incident with edges of colours 1, 3, and 0. If P does not pass through the vertex y_1, we switch the colours on P producing a 4-edge-colouring ϕ_1' of $\tilde{G}_1$ where $\phi_1'(s_1) = 1$, $\phi_1'(s_2) = \phi_1'(s_3) = 2$, and $\phi_1'(s_4) = 3$. However, ϕ_1' can be combined with a colouring of G_2 of type 1221 to obtain a 3-edge-colouring of $G - \{e_2, s_4\}$, which is impossible since $\omega(G) \geq 4$. If P passes through y_1, we switch the colours only on the segment P_0 between y_2 and y_1, producing an improper colouring ϕ_1'' of $\tilde{G}_1$ with y_1 being its only faulty vertex. Depending on whether P_0 ends with an edge coloured 1 or 2 we get $\phi_1''(s_1) = \phi_1''(s_2) = \phi_1''(s_3) = 2$ and $\phi_1''(s_4) = 3$, or $\phi_1''(s_1) = \phi_1''(s_2) = 1$, $\phi_1''(s_3) = 2$, and $\phi_1''(s_4) = 3$. In the latter case we can combine ϕ_1'' with a colouring of G_2 of type 1122, producing a 3-edge-colouring of $G - \{e_2, s_4\}$. In the former case we first interchange the colours 1 and 2 on G_1 and then combine the resulting colouring with a colouring of G_2 of type 1111, again producing a 3-edge-colouring of $G - \{e_2, s_4\}$. Since $\omega(G) \geq 4$, in both cases we have reached a contradiction. This establishes Claim 1. □

Proposition 2.10 now implies that G_2 can be extended to a snark $\bar{G}_2$ by adding at most two vertices. Recall that such an extension is unique up to isomorphism and depends only on whether G_2 is isochromatic or heterochromatic. We discuss these two cases separately.

Case 1. *G_2 is isochromatic.* First note that in this case $\bar{G}_2$ arises from G_2 by adding two new vertices x_1 and x_2 joined by an edge and by attaching each of the new vertices to the semiedges in the same couple. From Proposition 2.3 (i) we get that $\zeta(G_2) \geq 2$. If $\zeta(G_2) \geq 4$, then the same obviously holds for $\bar{G}_2$. Assume that $\zeta(G_2) = 3$, and let A denote the set of end-vertices in G_2 of the edges of the edge-cut S. Note that $|A| = 4$ because S is independent. Since $\zeta(G) = 4$, every cycle separating 3-edge-cut R in G_2 has the property that each component of $G_2 - R$ contains at least one vertex from A. This readily implies that $\zeta(\bar{G}_2) \geq 4$ and establishes the statement (iii) whenever $\zeta(G_2) \geq 3$. It remains to consider the case where $\zeta(G_2) = 2$. Let U be a cycle-separating 2-edge-cut in G_2 and let Q_1 and Q_2 be the components of $G_2 - U$. Since G is cyclically 4-edge-connected, each Q_i contains exactly two vertices from A and thus both Q_1 and Q_2 are 4-poles. Each Q_i is colourable because any 3-edge-colouring of G_2 provides one for Q_i. Furthermore, each Q_i is colour-open, because G_2 and hence also Q_i has an extension to $\bar{G}_2$. Thus G_2 is a partial junction of two colour-open 4-poles. It is not difficult to show that an isochromatic 4-pole can arise from a partial junction of any combination of isochromatic and heterochromatic 4-poles, as claimed.

Case 2. *G_2 is heterochromatic.* In this case G_2 arises from a snark by severing two independent edges. Suppose to the contrary that $\bar{G}_2$ is not cyclically 4-edge-connected. Then G_2 has at least twelve vertices, because there is only one 2-edge-connected snark of order less than twelve – the Petersen graph – and its cyclic connectivity equals 5. Let us take a heterochromatic 4-pole H of order 10 obtained from the Petersen graph and substitute G_2 in G with H, creating a new cubic graph G'. Clearly, G' is a snark of order smaller than G. To derive a final contradiction with the minimality of G we show that G' is cyclically 4-edge-connected and has oddness at least 4.

Claim 2. $\omega(G') \geq 4$.

Proof of Claim 2. Suppose to the contrary that $\omega(G') < 4$. Since G_1 is uncolourable and $G_1 \subseteq G'$, we infer that $\omega(G') = 2$ which in turn implies that $\rho(G') = 2$. Therefore there exist edges e_1 and e_2 in G' such that $G' - \{e_1, e_2\}$ is colourable. In other words, G' has a minimum 4-edge-colouring ψ where e_1 and e_2 are the only edges of G' coloured 0.

Since G_1 is uncolourable, at least one of e_1 and e_2 must have both end-vertices in G_1. Without loss of generality assume that at e_1 has both end-vertices in G_1. If e_2 had at least one end-vertex in G_1, we could take a 3-edge-colouring of $G' - \{e_1, e_2\}$, remove H and reinstate G_2 coloured in such a way that the edges in $S - \{e_2\}$ receive the same colours from G_2 as they did from H; this is possible since G_2 and H are colour-equivalent. However, in this way we would produce a 3-edge-colouring of $G - \{e_1, e_2\}$, contrary to the assumption that $\omega(G) = 4$. Therefore e_2 has both ends in H.

Since H is heterochromatic, the edges of S can be partitioned into couples such that for every 3-edge-colouring of H the colours of both edges within a couple are always different. Let $\{s_i, s_j\}$ and $\{s_k, s_l\}$ be the couples of H. Further, since ψ is a minimum 4-edge-colouring of G', all three non-zero colours are present on the edges adjacent to each of e_i, one of the colours being represented twice. By Theorem 2.5, the same colour occurs twice at both e_1 and e_2, say colour 1. If we regard ψ as a $\mathbb{Z}_2 \times \mathbb{Z}_2$-valuation and sum the outflows from vertices of G_1 we see that the flow through S equals $\psi(s_1) + \psi(s_2) + \psi(s_3) + \psi(s_4) = 1$. Hence, the distribution of colours in the couples of S, the set $\{\{\psi(s_i), \psi(s_j)\}, \{\psi(s_k), \psi(s_l)\}\}$, must have one of the following four forms:

$$D_1 = \{\{1,1\}, \{2,3\}\},$$
$$D_2 = \{\{1,2\}, \{1,3\}\},$$
$$D_3 = \{\{2,2\}, \{2,3\}\},$$
$$D_4 = \{\{2,3\}, \{3,3\}\}.$$

We now concentrate on the restriction of ψ to G_1 and show that it can be modified to a 4-edge-colouring λ of G_1 with distribution either D_2 or D_3. If the colouring ψ of G' has distribution D_4, we can simply interchange the colours 2 and 3 to obtain the distribution D_3. Assume that ψ has distribution D_1. Let us consider the unique end-vertex u of e_1 in G_1 such that the edges incident with u receive colours 1, 3, and 0 from ψ. The 1-2-Kempe chain P starting at u ends with a vertex incident with e_2, which means that P traverses S. Let s be the first edge of S that belongs to P. If $\psi(s) = 1$, then the desired 4-edge-colouring λ of G_1 with distribution D_2 can be obtained by the Kempe switch on the segment of P that ends with s and by a subsequent permutation of colours interchanging 1 and 2. If $\psi(s) = 2$, then a 4-edge-colouring of G_1 with distribution D_3 can be obtained similarly. In both cases, e_1 is the only edge coloured 0 under λ.

If λ has distribution D_2, then λ and a 3-edge-colouring of H of type 1212 can be combined to a 3-edge-colouring of $G' - \{e_1, s_l\}$. However, as observed earlier, by removing H and reinstating G_2 we could produce a 3-edge-colouring of $G - \{e_1, s_l\}$, which is impossible because $\omega(G) \geq 4$. If λ has distribution D_3, we can similarly combine λ with a 3-edge-colouring of H of type 1221 to a 3-edge-colouring of $G' - \{e_1, s_i\}$ which is impossible for the same reason. This contradiction completes the proof of Claim 2. □

Claim 3. $\zeta(G') = 4$.

Proof of Claim 3. Suppose to the contrary that $\zeta(G') < 4$. Let S' be a minimum size cycle-separating edge-cut in G'. If all the edges of S' had at least one end vertex in G_1, then S' would be a cycle-separating cut also in G, which is impossible. Therefore at least one edge of S' has both ends in H, which means that S' intersects H. Since H is connected, we conclude that $S'_H = S' \cap E(H)$ is an edge-cut of H. Note that S'_H is an independent set of edges, so S'_H must be a cycle-separating edge-cut in H. Recall, however, that H arises from the Petersen graph by severing two independent edges e and f. It follows that $S'_H \cup \{e, f\}$ is a cycle-separating edge-cut in the Petersen graph. Hence, $|S'_H \cup \{e, f\}| \geq 5$, and consequently $3 \leq |S'_H| \leq |S'| \leq 3$. This shows that $S'_H = S'$ and therefore S' is completely contained in H; in particular $S' \cap S = \emptyset$. Because S' is an edge-cut of the entire G', all the edges of S must join G_1 to the same component of $H - S'$. On the other hand, the Petersen graph is cyclically 5-edge-connected, therefore both e and f have end-vertices in different components of $H - S'$. The way how G' was constructed from G now implies that the set of end-vertices of S in H coincides with the set of end-vertices of e and f. Therefore S has an end-vertex in each component of $H - S'$, contradicting the previous observation. This contradiction establishes Claim 3. □

Claim 2 and Claim 3 combined provide a final contradiction with the choice of G, which concludes the proof. □

We proceed to proving our second decomposition theorem.

Proof of Theorem 3.5. Let G be a snark with oddness at least 4, cyclic connectivity 4, and minimum number of vertices. If G contains a cycle-separating 4-edge-cut whose removal leaves either two uncolourable components or one uncolourable component and one heterochromatic component, then the conclusion follows directly from Theorem 3.6 (i) or (ii), respectively. Otherwise one of the components is uncolourable and the other one, denoted by G_2, is isochromatic. In this case, G_2 contains a subgraph K which is an atom, possibly $K = G_2$. Clearly, K is colourable and $\delta_G(K)$ is a cycle-separating 4-edge-cut. If K is heterochromatic, then the conclusion again follows from Theorem 3.6 (ii). Therefore we may assume that K is isochromatic. Since $4 = \zeta(G) < 5 \leq g(G)$, we see that K is a non-trivial atom and from Proposition 2.3 (ii) we infer that $\zeta(K) \geq 3$. Applying statement (iii) of Theorem 3.6 with $S = \delta_G(K)$ we finally get the desired result. □

4 Main result

We are now ready to prove our main result.

Theorem 1.1. *The smallest number of vertices of a snark with cyclic connectivity* 4 *and oddness at least* 4 *is* 44*. The girth of each such snark is at least* 5.

Proof. Let G be a snark with oddness at least 4, cyclic connectivity 4, and minimum order. We first prove that G has girth at least 5. By Proposition 2.1, the girth of G is at least 4. Suppose to the contrary that G contains a 4-cycle C, and let S be the edge-cut separating C from the rest of G. Since S is cycle-separating, it has to satisfy one of the statements (i) – (iii) of Theorem 3.6. In the notation of Theorem 3.6, C necessarily plays the role of G_2, because it is colourable. In particular, S does not satisfy (i). However, S satisfies neither (ii) because G_2 is not heterochromatic, nor (iii) since G_2 is not isochromatic. Thus we have reached a contradiction proving that the girth of G is at least 5.

In Figure 1 we have displayed a snark with oddness at least 4, cyclic connectivity 4 on 44 vertices. It remains to show that there are no snarks of oddness at least 4 and cyclic connectivity 4 with fewer than 44 vertices.

Our main tool is Theorem 3.5. It implies that every snark with oddness at least 4, cyclic connectivity 4, and minimum number of vertices can be obtained from two smaller cyclically 4-edge-connected snarks G_1 and G_2 by the following process:

- Form a 4-pole H_i from each G_i by either removing two adjacent vertices or two nonadjacent edges and by retaining the dangling edges.

- Construct a cubic graph G by identifying the dangling edges of H_1 with those of H_2 after possibly applying a permutation to the dangling edges of H_1 or H_2.

Any graph G obtained in this manner will be called a 4-*join* of G_1 and G_2. Note that the well-known operation of a *dot product* of snarks [1, 28] is a special case of a 4-join.

We proceed to proving that every snark with cyclic connectivity 4 on at most 42 vertices has oddness 2. If G is a snark with cyclic connectivity 4 on at most 42 vertices, then by Theorem 3.5 it contains a cycle-separating 4-edge-cut S such that both components K_1 and K_2 of $G - S$ can be extended to snarks G_1 and G_2, respectively, by adding at most two vertices; in other words, G is a 4-join of G_1 and G_2. Clearly, $|V(K_1)| + |V(K_2)| = |V(G)| \leq 42$. Assuming that $|V(K_1)| \leq |V(K_2)|$ we see that $|V(K_1)| \geq 8$, because the smallest cyclically 4-edge-connected snark has 10 vertices, and hence $|V(K_2)| \leq 34$. Therefore both G_1 and G_2 have order at least 10 and at most 36.

Let $\mathcal{S}_n$ denote the set of all pairwise non-isomorphic cyclically 4-edge-connected snarks of order not exceeding n. To finish the proof it remains to show that every 4-join of two snarks from $\mathcal{S}_{36}$ with at most 42 vertices has oddness 2. Unfortunately, verification of this statement in a purely theoretical way is far beyond currently available methods. The final step of our proof has been therefore performed by a computer.

We have written a program which applies a 4-join in all possible ways to two given input graphs and have applied this program to the complete list of snarks from the set $\mathcal{S}_{36}$. More specifically, given an arbitrary pair of input graphs, the program removes in all possible ways either two adjacent vertices or two nonadjacent edges from each of the graphs (retaining the dangling edges) and then identifies the dangling edges from the first graph in the pair with the dangling edges of the second graph, again in all possible (i.e., $4! = 24$) ways. We also use the *nauty* library [39, 40] to determine the orbits of edges and edge pairs in the input graphs, so the program only removes two adjacent vertices or two nonadjacent edges once from every orbit of edges or edge pairs, respectively. The resulting graphs can still contain isomorphic copies, therefore we also use *nauty* to compute a canonical labelling of the graphs and remove the isomorphic copies.

Until now, only the set $\mathcal{S}_{34}$ has been known; it was determined by Brinkmann, Hägglund, Markström, and the first author [7] in 2013 and was shown to contain exactly 27 205 766 snarks. Using the program *snarkhunter* [7, 8] we have been able to generate all cyclically 4-edge-connected snarks on 36 vertices, thereby completing the determination of $\mathcal{S}_{36}$. This took about 80 CPU years and yielded exactly 404 899 916 such graphs. The size of $\mathcal{S}_{36}$ thus totals to 432 105 682 graphs. (The new list of snarks can be downloaded from the *House of Graphs* [6] at `http://hog.grinvin.org/Snarks`.)

Finally, we have performed all possible 4-joins of two snarks from $\mathcal{S}_{36}$ that produce a snark with at most 42 vertices and checked their oddness. This computation required approximately 75 CPU days. We have used two independent programs to compute the oddness of the resulting graphs (the source code of these programs can be obtained from [21]) and in each case the results of both programs were in complete agreement. No snark of oddness greater than 2 among them was found, which completes the proof of Theorem 1.1. □

5 Remarks and open problems

We have applied the 4-join operation to all valid pairs of snarks from $\mathcal{S}_{36}$ to construct cyclically 4-edge-connected snarks on 44 vertices and checked their oddness. In this manner we have produced 31 cyclically 4-edge-connected snarks of oddness 4, including the one from

Figure 1, all of them having girth 5. The most symmetric of them is shown in Figure 4. We will describe and analyse these 31 snarks in the sequel of this paper [23], where we also prove that they constitute a complete list of all snarks with oddness at least 4, cyclic connectivity 4, and minimum number of vertices.

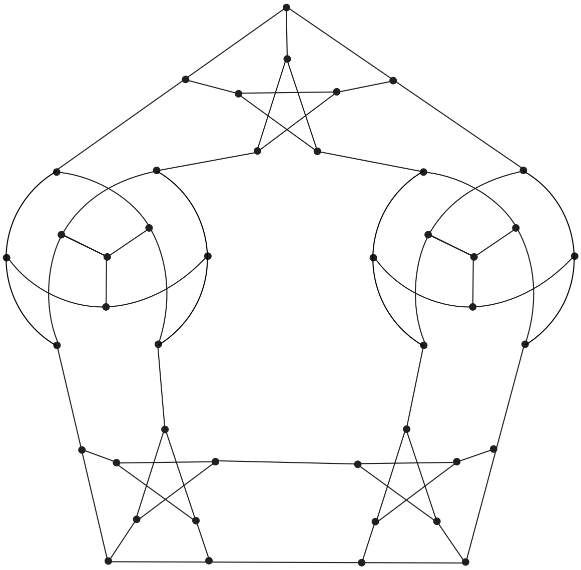


Figure 4: The most symmetric nontrivial snark of oddness 4 on 44 vertices.

As we have already mentioned in Introduction (Section 1), Theorem 1.1 does not yet determine the smallest order of a nontrivial snark with oddness 4, because there might exist snarks with oddness at least 4 of order 38, 40, or 42 with cyclic connectivity greater than 4. Furthermore, it is not immediately clear why a snark G with $\omega(G) \geq m$ and minimum order should have oddness exactly m. This situation suggests two natural problems which require the following definition: Given integers $\omega \geq 2$ and $k \geq 2$, let $m(\omega, k)$ denote the minimum order of a cyclically k-edge-connected snark with oddness at least ω. For example, one has $m(2,2) = m(2,3) = m(2,4) = m(2,5) = 10$ as exemplified by the Petersen graph, and $m(2,6) = 28$ as exemplified by the Isaacs flower snark J_7. The values $m(2,k)$ for $k \geq 7$ are not known, however the well-known conjecture of Jaeger and Swart [31] that there are no cyclically 7-edge-connected snarks would imply that these values are not defined. For $\omega = 4$, Lukot'ka et al. [35, Theorem 12] showed that $m(4,2) = m(4,3) = 28$. The value $m(4,4)$ remains unknown although our Theorem 1.1 seems to suggest that $m(4,4) = 44$.

Problem 5.1. Determine the value $m(4,4)$.

Our second problem asks whether the function $m(\omega, k)$ is monotonous in both coordinates.

Problem 5.2. Is it true that $m(\omega+1, k) \geq m(\omega, k)$ and $m(\omega, k+1) \geq m(\omega, k)$ whenever the involved values are defined?

6 Testing conjectures

After having generated all snarks from the set $\mathcal{S}_{34}$ and those from $\mathcal{S}_{36}$ that have girth at least 5, Brinkmann et al. [7] tested the validity of several important conjectures whose minimal counterexamples, provided that they exist, must be snarks. For most of the considered conjectures the potential minimal counterexamples are proven to be nontrivial snarks, that is, those with cyclic connectivity at least 4 and girth at least 5. Nevertheless, in some cases the girth condition has not been established. Therefore it appears reasonable to check the validity of such conjectures on the set $\mathcal{S}_{36} \setminus \mathcal{S}_{34}$ of all cyclically 4-edge-connected snarks of order 36. We have performed these tests and arrived at the conclusions discussed below; for more details on the conjectures we refer the reader to [7].

A *dominating* circuit in a graph G is a circuit C such that every edge of G has an end-vertex on C. Fleischner [17] made the following conjecture on dominating cycles.

Conjecture 6.1 (Dominating circuit conjecture)**.** *Every cyclically* 4*-edge-connected snark has a dominating circuit.*

The dominating circuit conjecture exists in several different forms (see, for example, [3, 18]) and is equivalent to a number of other seemingly unrelated conjectures such as the Matthews-Sumner conjecture about the hamiltonicity of claw-free graphs [38]. For more information on these conjectures see [10].

Our tests have resulted in the following claim.

Claim 6.2. *Conjecture 6.1 has no counterexample on* 36 *or fewer vertices.*

The *total chromatic number* of a graph G is the minimum number of colours required to colour the vertices and the edges of G in such a way that adjacent vertices and edges have different colours and no vertex has the same colour as its incident edges. The total colouring conjecture [4, 50] suggests that the total chromatic number of every graph with maximum degree Δ is either $\Delta + 1$ or $\Delta + 2$. For cubic graphs this conjecture is known to be true by a result of Rosenfeld [45], therefore the total chromatic number of a cubic graph is either 4 or 5. Cavicchioli et al. [12, Problem 5.1] asked for a smallest nontrivial snark with total chromatic number 5. Brinkmann et al. [7] showed that such a snark must have at least 38 vertices. Sasaki et al. [46] displayed examples of snarks with connectivity 2 or 3 whose total chromatic number is 5 and asked [46, Question 2] for the order of a smallest cyclically 4-edge-connected snark with total chromatic number 5. Brinkmann et al. [9] constructed cyclically 4-edge-connected snarks with girth 4 and total chromatic number 5 for each even order greater than or equal to 40. Our next claim shows that the value asked for by Sasaki et al. is either 38 or 40.

Claim 6.3. *All cyclically* 4*-edge-connected snarks with at most* 36 *vertices have total chromatic number* 4*.*

The following conjecture was made by Jaeger [30] and is known as the *Petersen colouring conjecture*. If true, this conjecture would imply several other profound conjectures, in particular, the 5-cycle double cover conjecture and the Fulkerson conjecture.

Conjecture 6.4 (Petersen colouring conjecture)**.** *Every bridgeless cubic graph G admits a colouring of its edges using the edges of the Petersen graph as colours in such a way that any three mutually adjacent edges of G are coloured with three mutually adjacent edges of the Petersen graph.*

It is easy to see that the smallest counterexample to this conjecture must be a cyclically 4-edge-connected snark. Brinkmann et al. [7] showed that the smallest counterexample to the Petersen colouring conjecture must have order at least 36. Here we improve the latter value to 38.

Claim 6.5. *Conjecture 6.4 has no counterexamples on* 36 *or fewer vertices.*

References

[1] G. M. Adelson-Velsky and V. K. Titov, On edge 4-chromatic cubic graphs (in Russian), in: V. K. Zakharov, V. P. Kozyrev, K. A. Rybnikov, V. N. Sachkov, V. E. Stepanov and V. E. Tarakanov (eds.), *Voprosy Kibernetiki, Proceedings of the Seminar on Combinatorial Mathematics (Moscow State University, Moscow, January 27 – 29, 1971)*, Moscow, 1973 pp. 5–14.

[2] L. D. Andersen, H. Fleischner and B. Jackson, Removable edges in cyclically 4-edge-connected cubic graphs, *Graphs Combin.* **4** (1988), 1–21, doi:10.1007/bf01864149.

[3] P. Ash and B. Jackson, Dominating cycles in bipartite graphs, in: J. A. Bondy and U. S. R. Murty (eds.), *Progress in Graph Theory*, Academic Press, Toronto, Ontario, 1984 pp. 81–87, proceedings of the conference on combinatorics held at the University of Waterloo, Waterloo, Ontario, 1982.

[4] M. Behzad, G. Chartrand and J. K. Cooper, Jr., The colour numbers of complete graphs, *J. London Math. Soc.* **42** (1967), 226–228, doi:10.1112/jlms/s1-42.1.226.

[5] D. Blanuša, Problem četiriju boja, *Glasnik Mat. Fiz. Astr. Ser. II* **1** (1946), 31–42.

[6] G. Brinkmann, K. Coolsaet, J. Goedgebeur and H. Mélot, House of Graphs: a database of interesting graphs, *Discrete Appl. Math.* **161** (2013), 311–314, doi:10.1016/j.dam.2012.07.018, available at `http://hog.grinvin.org/`.

[7] G. Brinkmann, J. Goedgebeur, J. Hägglund and K. Markström, Generation and properties of snarks, *J. Comb. Theory Ser. B* **103** (2013), 468–488, doi:10.1016/j.jctb.2013.05.001.

[8] G. Brinkmann, J. Goedgebeur and B. D. McKay, Generation of cubic graphs, *Discrete Math. Theor. Comput. Sci.* **13** (2011), 69–80, `https://www.dmtcs.org/dmtcs-ojs/index.php/dmtcs/article/view/1801/0.html`.

[9] G. Brinkmann, M. Preissmann and D. Sasaki, Snarks with total chromatic number 5, *Discrete Math. Theor. Comput. Sci.* **17** (2015), 369–382, `https://www.dmtcs.org/dmtcs-ojs/index.php/dmtcs/article/view/2567.1.html`.

[10] H. Broersma, G. Fijavž, T. Kaiser, R. Kužel, Z. Ryjáček and P. Vrána, Contractible subgraphs, Thomassen's conjecture and the dominating cycle conjecture for snarks, *Discrete Math.* **308** (2008), 6064–6077, doi:10.1016/j.disc.2007.11.026.

[11] P. J. Cameron, A. G. Chetwynd and J. J. Watkins, Decomposition of snarks, *J. Graph Theory* **11** (1987), 13–19, doi:10.1002/jgt.3190110104.

[12] A. Cavicchioli, T. E. Murgolo, B. Ruini and F. Spaggiari, Special classes of snarks, *Acta Appl. Math.* **76** (2003), 57–88, doi:10.1023/a:1022864000162.

[13] M. Chladný and M. Škoviera, Factorisation of snarks, *Electron. J. Combin.* **17** (2010), #R32, `https://www.combinatorics.org/ojs/index.php/eljc/article/view/v17i1r32`.

[14] B. Descartes, Network-colourings, *Math. Gaz.* **32** (1948), 67–69, doi:10.2307/3610702.

[15] M. A. Fiol, A Boolean algebra approach to the construction of snarks, in: Y. Alavi, G. Chartrand, O. R. Oellermann and A. J. Schwenk (eds.), *Graph Theory, Combinatorics, and Applications, Volume 1*, John Wiley & Sons, New York, Wiley-Interscience Publication, 1991 pp.

493–524, proceedings of the Sixth Quadrennial International Conference on the Theory and Applications of Graphs held at Western Michigan University, Kalamazoo, Michigan, May 30 – June 3, 1988.

[16] M. A. Fiol, G. Mazzuoccolo and E. Steffen, On measures of edge-uncolorability of cubic graphs: A brief survey and some new results, 2017, `arXiv:1702.07156 [math.CO]`.

[17] H. Fleischner, Cycle decompositions, 2-coverings, removable cycles, and the four-color disease, in: J. A. Bondy and U. S. R. Murty (eds.), *Progress in Graph Theory*, Academic Press, Toronto, Ontario, 1984 pp. 233–246, proceedings of the conference on combinatorics held at the University of Waterloo, Waterloo, Ontario, 1982.

[18] H. Fleischner and M. Kochol, A note about the dominating circuit conjecture, *Discrete Math.* **259** (2002), 307–309, doi:10.1016/s0012-365x(02)00588-5.

[19] M. Fontet, Graphes 4-essentiels, *C. R. Acad. Sci. Paris Ser. A* **287** (1978), 289–290.

[20] J.-L. Fouquet, Graphes cubiques d'indice chromatique quatre, *Ann. Discrete Math.* **9** (1980), 23–28, doi:10.1016/s0167-5060(08)70028-1.

[21] J. Goedgebeur, Source code of two programs to compute the oddness of a graph, `http://caagt.ugent.be/oddness/`.

[22] J. Goedgebeur, On the smallest snarks with oddness 4 and connectivity 2, *Electron. J. Combin.* **25** (2018), #P2.15, `https://www.combinatorics.org/ojs/index.php/eljc/article/view/v25i2p15`.

[23] J. Goedgebeur, E. Máčajová and M. Škoviera, The smallest nontrivial snarks of oddness 4, in prepration.

[24] R. Häggkvist and S. McGuinness, Double covers of cubic graphs with oddness 4, *J. Comb. Theory Ser. B* **93** (2005), 251–277, doi:10.1016/j.jctb.2004.11.003.

[25] J. Hägglund, On snarks that are far from being 3-edge colorable, *Electron. J. Combin.* **23** (2016), #P2.6, `https://www.combinatorics.org/ojs/index.php/eljc/article/view/v23i2p6`.

[26] I. Holyer, The NP-completeness of edge-coloring, *SIAM J. Comput.* **10** (1981), 718–720, doi: 10.1137/0210055.

[27] A. Huck and M. Kochol, Five cycle double covers of some cubic graphs, *J. Comb. Theory Ser. B* **64** (1995), 119–125, doi:10.1006/jctb.1995.1029.

[28] R. Isaacs, Infinite families of nontrivial trivalent graphs which are not Tait colorable, *Amer. Math. Monthly* **82** (1975), 221–239, doi:10.2307/2319844.

[29] F. Jaeger, A survey of the cycle double cover conjecture, in: B. R. Alspach and C. D. Godsil (eds.), *Cycles in Graphs*, North-Holland, Amsterdam, volume 115 of *North-Holland Mathematics Studies*, 1985 pp. 1–12, doi:10.1016/s0304-0208(08)72993-1, papers from the workshop held at Simon Fraser University, Burnaby, British Columbia, July 5 – August 20, 1982.

[30] F. Jaeger, Nowhere-zero flow problems, in: L. W. Beineke and R. J. Wilson (eds.), *Selected Topics in Graph Theory, Volume 3*, Academic Press, San Diego, California, pp. 71–95, 1988.

[31] F. Jaeger and T. Swart, Problem session, *Ann. Discrete Math.* **9** (1980), 304–305, doi:10.1016/s0167-5060(08)70086-4.

[32] E. L. Johnson, A proof of 4-coloring the edges of a cubic graph, *Amer. Math. Monthly* **73** (1966), 52–55, doi:10.2307/2313923.

[33] M. Kochol, Superposition and constructions of graphs without nowhere-zero k-flows, *European J. Combin.* **23** (2002), 281–306, doi:10.1006/eujc.2001.0563.

[34] R. Lukoťka, E. Máčajová, J. Mazák and M. Škoviera, Erratum to "Small snarks with large oddness", in preparation.

[35] R. Lukoťka, E. Máčajová, J. Mazák and M. Škoviera, Small snarks with large oddness, *Electron. J. Combin.* **22** (2015), #P1.51, `https://www.combinatorics.org/ojs/index.php/eljc/article/view/v22i1p51`.

[36] E. Máčajová and M. Škoviera, Irreducible snarks of given order and cyclic connectivity, *Discrete Math.* **306** (2006), 779–791, doi:10.1016/j.disc.2006.02.003.

[37] E. Máčajová and M. Škoviera, Sparsely intersecting perfect matchings in cubic graphs, *Combinatorica* **34** (2014), 61–94, doi:10.1007/s00493-014-2550-4.

[38] M. M. Matthews and D. P. Sumner, Hamiltonian results in $K_{1,3}$-free graphs, *J. Graph Theory* **8** (1984), 139–146, doi:10.1002/jgt.3190080116.

[39] B. D. McKay, *nauty User's Guide (Version 1.5)*, Technical Report TR-CS-90-02, Department of Computer Science, Australian National University, 1990, the latest version of the software is available at `http://cs.anu.edu.au/~bdm/nauty`.

[40] B. D. McKay and A. Piperno, Practical graph isomorphism, II, *J. Symbolic Comput.* **60** (2014), 94–112, doi:10.1016/j.jsc.2013.09.003.

[41] R. Nedela and M. Škoviera, Atoms of cyclic connectivity in cubic graphs, *Math. Slovaca* **45** (1995), 481–499.

[42] R. Nedela and M. Škoviera, Decompositions and reductions of snarks, *J. Graph Theory* **22** (1996), 253–279, doi:10.1002/(sici)1097-0118(199607)22:3<253::aid-jgt6>3.0.co;2-l.

[43] N. Robertson, Minimal cyclic-4-connected graphs, *Trans. Amer. Math. Soc.* **284** (1984), 665–687, doi:10.2307/1999101.

[44] R. W. Robinson and N. C. Wormald, Almost all cubic graphs are Hamiltonian, *Random Structures Algorithms* **3** (1992), 117–125, doi:10.1002/rsa.3240030202.

[45] M. Rosenfeld, On the total coloring of certain graphs, *Israel J. Math.* **9** (1971), 396–402, doi:10.1007/bf02771690.

[46] D. Sasaki, S. Dantas, C. M. H. de Figueiredo and M. Preissmann, The hunting of a snark with total chromatic number 5, *Discrete Appl. Math.* **164** (2014), 470–481, doi:10.1016/j.dam.2013.04.006.

[47] C. E. Shannon, A theorem on coloring the lines of a network, *J. Math. Physics* **28** (1949), 148–151, doi:10.1002/sapm1949281148.

[48] E. Steffen, Classification and characterizations of snarks, *Discrete Math.* **188** (1998), 183–203, doi:10.1016/s0012-365x(97)00255-0.

[49] E. Steffen, Measurements of edge-uncolorability, *Discrete Math.* **280** (2004), 191–214, doi:10.1016/j.disc.2003.05.005.

[50] V. G. Vizing, On an estimate of the chromatic class of a p-graph (in Russian), *Diskret. Analiz* **3** (1964), 25–30.

[51] N. C. Wormald, Classifying k-connected cubic graphs, in: A. F. Horadam and W. D. Wallis (eds.), *Combinatorial Mathematics VI*, Springer, Berlin, volume 748 of *Lecture Notes in Mathematics*, pp. 199–206, 1979, doi:10.1007/bfb0102696, proceedings of the Sixth Australian Conference held at the University of New England, Armidale, August, 1978.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 299–317
https://doi.org/10.26493/1855-3974.1164.2f7
(Also available at http://amc-journal.eu)

Order-chain polytopes*

Takayuki Hibi

Department of Pure and Applied Mathematics, Graduate School of Information Science and Technology, Osaka University, Suita, Osaka 565-0871, Japan

Nan Li

Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Teresa Xueshan Li †

School of Mathematics and Statistics, Southwest University, Chongqing 400715, PR China

Li Li Mu

School of Mathematics, Liaoning Normal University, Dalian 116029, PR China

Akiyoshi Tsuchiya

Department of Pure and Applied Mathematics, Graduate School of Information Science and Technology, Osaka University, Suita, Osaka 565-0871, Japan

Received 18 July 2016, accepted 28 November 2018, published online 6 January 2019

Abstract

Given two families X and Y of integral polytopes with nice combinatorial and algebraic properties, a natural way to generate a new class of polytopes is to take the intersection $\mathcal{P} = \mathcal{P}_1 \cap \mathcal{P}_2$, where $\mathcal{P}_1 \in X$, $\mathcal{P}_2 \in Y$. Two basic questions then arise: 1) when $\mathcal{P}$ is integral and 2) whether $\mathcal{P}$ inherits the "old type" from $\mathcal{P}_1, \mathcal{P}_2$ or has a "new type", that is, whether $\mathcal{P}$ is unimodularly equivalent to a polytope in $X \cup Y$ or not. In this paper, we focus on the families of order polytopes and chain polytopes. Following the above framework, we create a new class of polytopes which are named order-chain polytopes. When studying

*This work was initiated when the third author and the fourth author were visiting the MIT Department of Mathematics. These two authors would like to thank Professor Richard Stanley for many helpful discussions. This work was supported by the China Scholarship Council, the National Science Foundation of China (Grant No. 11601440, 11701249), the Natural Science Foundation of Chongqing (Grant No. cstc2016jcyjA0245) and Fundamental Research Funds for Central Universities (Grant No. XDJK2018C075).

†Corresponding author.

their volumes, we discover a natural relation with Ehrenborg and Mahajan's results on maximizing descent statistics.

Keywords: Poset, order-chain polytope, unimodular equivalence.

Math. Subj. Class.: 52B05, 52B20

1 Introduction

This paper was motivated by the following two questions about intersecting two integral polytopes $\mathcal{P}_1$ and $\mathcal{P}_2$, which come from two given families X and Y of polytopes respectively:

1) when the intersection $\mathcal{P} = \mathcal{P}_1 \cap \mathcal{P}_2$ is integral and
2) whether $\mathcal{P}$ inherits the "old type" from $\mathcal{P}_1, \mathcal{P}_2$ or has a "new type", that is, whether $\mathcal{P}$ is unimodularly equivalent to a polytope in $X \cup Y$ or not.

Usually, we shall start with those families X and Y of polytopes which have nice combinatorial and algebraic properties. In this paper, we focus on the families of order polytopes and chain polytopes. Instead of considering the intersection of an arbitrary d-dimensional order polytope and an arbitrary d-dimensional chain polytope, we will consider the intersection of an order polytope $\mathcal{O}(P')$ and a chain polytope $\mathcal{C}(P'')$, both of which arise from weak subposets P', P'' of a given poset. The resulting polytope is called an order-chain polytope, which generalizes both order polytope and chain polytope.

The order polytope $\mathcal{O}(P)$ as well as the chain polytope $\mathcal{C}(P)$ arising from a finite partially ordered set P has been studied by many authors from viewpoints of both combinatorics and commutative algebra. Especially, in [16], the combinatorial structures of order polytopes and chain polytopes are explicitly discussed. Furthermore, in [9], the natural question when the order polytope $\mathcal{O}(P)$ and the chain polytope $\mathcal{C}(P)$ are unimodularly equivalent is solved completely. It follows from [5] and [8] that the toric ring ([7, p. 37]) of $\mathcal{O}(P)$ and that of $\mathcal{C}(P)$ are algebras with straightening laws ([6, p. 124]) on finite distributive lattices. Thus in particular the toric ideal ([7, p. 35]) of each of $\mathcal{O}(P)$ and $\mathcal{C}(P)$ possesses a squarefree quadratic initial ideal ([7, p. 10]) and possesses a regular unimodular triangulation ([7, p. 254]) arising from a flag complex. Furthermore, toric rings of order polytopes naturally appear in algebraic geometry (e.g., [2]) and in representation theory (e.g., [18]).

We begin by introducing some basic notation and terminology. Given a convex polytope $\mathcal{P} \subset \mathbb{R}^d$, a *facet hyperplane* of $\mathcal{P} \subset \mathbb{R}^d$ is defined to be a hyperplane in $\mathbb{R}^d$ which contains a facet of $\mathcal{P}$. If

$$H = \{(x_1, x_2, \dots, x_d) \in \mathbb{R}^d : a_1x_1 + a_2x_2 + \cdots + a_dx_d - b = 0\},$$

where each a_i and b belong to $\mathbb{R}$, is a hyperplane of $\mathbb{R}^d$ and $v = (y_1, y_2, \dots, y_d) \in \mathbb{R}^d$, then we set

$$H(v) = a_1y_1 + a_2y_2 + \cdots + a_dy_d - b.$$

E-mail addresses: hibi@math.sci.osaka-u.ac.jp (Takayuki Hibi), amenda860111@gmail.com (Nan Li), pmgb@swu.edu.cn (Teresa Xueshan Li), lly-mu@hotmail.com (Li Li Mu), a-tsuchiya@ist.osaka-u.ac.jp (Akiyoshi Tsuchiya)

Let $(P, \preccurlyeq)$ be a finite partially ordered set (*poset*, for short) on $[d] = \{1, \ldots, d\}$. For each subset $S \subseteq P$, we define $\rho(S) = \sum_{i \in S} \mathbf{e}_i$, where $\mathbf{e}_1, \ldots, \mathbf{e}_d$ are the canonical unit coordinate vectors of $\mathbb{R}^d$. In particular $\rho(\emptyset) = (0, 0, \ldots, 0)$, the origin of $\mathbb{R}^d$. A subset I of P is an *order ideal* of P if $i \in I$, $j \in [d]$ together with $j \preccurlyeq i$ in P imply $j \in I$. An *antichain* of P is a subset A of P such that any two elements in A are incomparable. We say that j *covers* i if $i \prec j$ and there is no $k \in P$ such that $i \prec k \prec j$. A chain $j_1 \prec j_2 \prec \cdots \prec j_s$ is *saturated* if j_q covers j_{q-1} for $1 < q \leq s$, and it is called a *maximal chain* if, moreover, j_1 is a minimal element and j_s is a maximal element of P. A poset can be represented with its Hasse diagram, in which each cover relation $i \prec j$ corresponds to an edge denoted by $e = \{i, j\}$. For a finite poset P, we let $c(P)$, $m_\star(P)$ and $m^\star(P)$ denote the number of maximal chains, the number of minimal elements and the number of maximal elements of P, respectively. We denote by $E(P)$ the set of edges in the Hasse diagram of P.

In [16], Stanley introduced two convex polytopes arising from a finite poset, the order polytope and the chain polytope. Following [9], we employ slightly different definitions. Given a finite poset $(P, \preccurlyeq)$ on $[d]$, the *order polytope* $\mathcal{O}(P)$ is defined to be the convex polytope consisting of those $(x_1, \ldots, x_d) \in \mathbb{R}^d$ such that

(1) $0 \leq x_i \leq 1$ for $1 \leq i \leq d$;

(2) $x_i \geq x_j$ if $i \preccurlyeq j$ in P.

The *chain polytope* $\mathcal{C}(P)$ of P is defined to be the convex polytope consisting of those $(x_1, \ldots, x_d) \in \mathbb{R}^d$ such that

(1) $x_i \geq 0$ for $1 \leq i \leq d$;

(2) $x_{i_1} + \cdots + x_{i_k} \leq 1$ for every maximal chain $i_1 \prec \cdots \prec i_k$ of P.

Recall (see [16] for details) that there is a close connection between the combinatorial structure of P and the geometric structures of $\mathcal{O}(P)$ and $\mathcal{C}(P)$. For instance, the following connections are not hard to prove:

- The number $f_{d-1}(\mathcal{O}(p))$ of facets of $\mathcal{O}(P)$ is equal to $m_\star(P) + m^\star(P) + |E(P)|$. Equivalently, if we let $\hat{P} = P \cup \{\hat{0}, \hat{1}\}$ be the poset obtained from P by adjoining a minimum element $\hat{0}$ and a maximum element $\hat{1}$, then we have $f_{d-1}(\mathcal{O}(P)) = |E(\hat{P})|$.
- The number $f_{d-1}(\mathcal{C}(P))$ of facets of $\mathcal{C}(P)$ is equal to $d + c(P)$.
- The vertices of $\mathcal{O}(P)$ are exactly those $\rho(I)$ for which I is an order ideal of P, and the vertices of $\mathcal{C}(P)$ are exactly those $\rho(A)$ for which A is an antichain of P. Since it is well known that order ideals of P are in one-to-one correspondence with antichains of P, the order polytope $\mathcal{O}(P)$ and the chain polytope $\mathcal{C}(P)$ have the same number of vertices.

Let P be a finite poset, we define an *edge partition* of P to be a map

$$\ell \colon E(P) \longrightarrow \{o, c\}.$$

Equivalently, an edge partition of P is an ordered pair

$$(oE(P), cE(P))$$

of subsets of $E(P)$ such that $oE(P) \cup cE(P) = E(P)$ and $oE(P) \cap cE(P) = \emptyset$. An edge partition ℓ is called *proper* if $oE(P) \neq \emptyset$ and $cE(P) \neq \emptyset$.

Suppose that $(P, \preccurlyeq)$ is a poset on $[d]$ with an edge partition $\ell = (oE(P), cE(P))$. Let P'_ℓ and P''_ℓ denote the d-element weak subposets of P with cover relations given by the edge sets $oE(P)$ and $cE(P)$ respectively. Here by a weak subposet of P, we mean a subset Q of elements of P and a partial ordering $\preccurlyeq^*$ of Q such that if $x \preccurlyeq^* y$ in Q, then $x \preccurlyeq y$ in P. The *order-chain polytope* $\mathcal{OC}_\ell(P)$ with respect to the edge partition ℓ of P is defined to be the convex polytope

$$\mathcal{O}(P'_\ell) \cap \mathcal{C}(P''_\ell)$$

in $\mathbb{R}^d$. Clearly the notion of order-chain polytope is a natural generalization of both order polytope and chain polytope of a finite poset.

For example, let P be the chain $1 \prec 2 \prec \cdots \prec 7$ with

$$oE(P) = \{\{1,2\},\{4,5\},\{5,6\}\}, \qquad cE(P) = \{\{2,3\},\{3,4\},\{6,7\}\}.$$

Then P'_ℓ is the disjoint union of the following four chains:

$$1 \prec 2, \quad 3, \quad 4 \prec 5 \prec 6, \quad 7$$

and P''_ℓ is the disjoint union of

$$1, \quad 2 \prec 3 \prec 4, \quad 5 \quad \text{and} \quad 6 \prec 7.$$

Hence the order-chain polytope $\mathcal{OC}_\ell(P)$ is the convex polytope consisting of those $(x_1, \ldots, x_7) \in \mathbb{R}^7$ such that

(1) $0 \leq x_i \leq 1$ for $1 \leq i \leq 7$;

(2) $x_1 \geq x_2,\ x_4 \geq x_5 \geq x_6$;

(3) $x_2 + x_3 + x_4 \leq 1,\ x_6 + x_7 \leq 1$.

It should be noted that, for any poset P on $[d]$ and any edge partition ℓ of P, the dimension of the order-chain polytope $\mathcal{OC}_\ell(P)$ is equal to d. In fact, let $x = (1/d, \ldots, 1/d) \in \mathbb{R}^d$, clearly, we have $x \in \mathcal{OC}_\ell(P)$. If P'_ℓ is an antichain, then $\mathcal{O}(P'_\ell)$ is the d-cube $[0,1]^d$. In this case, $\mathcal{OC}_\ell(P)$ is exactly the same as the chain polytope $\mathcal{C}(P)$ and so is d-dimensional. If P'_ℓ is not an antichain, then P''_ℓ is not a d-element chain. In this case, $x \in \partial\mathcal{O}(P'_\ell)$ and $x \in \mathcal{C}(P''_\ell) \setminus \partial\mathcal{C}(P''_\ell)$, since no facet hyperplane of $\mathcal{C}(P''_\ell)$ contains x. In this case, we can find a ball $B_d(x)$ centered at x such that $B_d(x) \subset \mathcal{C}(P''_\ell) \setminus \partial\mathcal{C}(P''_\ell)$. Keeping in mind that x belongs to the boundary of $\mathcal{O}(P'_\ell)$, we deduce that $B_d(x) \cap (\mathcal{O}(P'_\ell) \setminus \partial\mathcal{O}(P'_\ell)) \neq \emptyset$. It follows that $(\mathcal{O}(P'_\ell) \setminus \partial\mathcal{O}(P'_\ell)) \cap (\mathcal{C}(P''_\ell) \setminus \partial\mathcal{C}(P''_\ell)) \neq \emptyset$, as desired.

Recall that an integral convex polytope (a convex polytope is *integral* if all of its vertices have integer coordinates) is called *compressed* ([15]) if all of its "pulling triangulations" are unimodular. Equivalently, a compressed polytope is an integral convex polytope any of whose reverse lexicographic initial ideals are squarefree ([17]). It follows from [13, Theorem 1.1] that all order polytopes and all chain polytopes are compressed. Hence the intersection of an order polytope and a chain polytope is compressed if it is integral. In particular every integral order-chain polytope is compressed. It then follows that every integral order-chain polytope possesses a unimodular triangulation and is normal ([12]).

Then one of the natural question, which we study in Section 2, is when an order-chain polytope is integral. We call an edge partition ℓ of a finite poset P *integral* if the order-chain polytope $\mathcal{OC}_\ell(P)$ is integral. We show that every edge partition of a finite poset P is integral if and only if P is cycle-free. Here by a cycle-free poset P we mean that the Hasse diagram of P is a cycle-free graph (i.e., an unoriented graph that does not have cycles). Furthermore, we prove that every poset P with $|E(P)| \geq 2$ possesses at least one proper integral edge partition.

In Section 3, we consider the problem when an integral order-chain polytope is unimodularly equivalent to either an order polytope or a chain polytope. This problem is related to the work [9], in which the authors characterize all finite posets P such that $\mathcal{O}(P)$ and $\mathcal{C}(P)$ are unimodularly equivalent. We show that if P is either a disjoint union of chains or a zigzag poset, then the order-chain polytope $\mathcal{OC}_\ell(P)$, with respect to each edge partition ℓ of P, is unimodularly equivalent to the chain polytope of some poset (Theorem 3.3 and Theorem 3.4). On the other hand, for each positive integer $d \geq 6$, we find a d-dimensional integral order-chain polytope which is not unimodularly equivalent to any chain polytope nor order polytope. This means that the notion of order-chain polytope is a nontrivial generalization of order polytope or chain polytope.

We conclude the present paper with an observation on the volume of order-chain polytopes in Section 4. An interesting question is to find an edge partition ℓ of a poset P which maximizes the volume of $\mathcal{OC}_\ell(P)$. In general, it seems to be very difficult to find a complete answer. We shall discuss the case when P is a chain on $[d]$, which involves Ehrenborg and Mahajan's problem (see [3]) of maximizing the descent statistics over certain family of subsets.

2 Integral order-chain polytopes

In this section, we consider the problem when an order-chain polytope is integral. We shall prove that every edge partition of a poset P is integral if and only if P is cycle-free. We also prove that every finite poset P with $|E(P)| \geq 2$ has at least one proper integral edge partition.

Theorem 2.1. *Let P be a finite poset. Then every edge partition of P is integral if and only if P is a cycle-free poset.*

Proof. Suppose that each edge partition ℓ of P is integral. If the Hasse diagram of P has a cycle C, then it is easy to find a non-integral edge partition. In fact, let $e = \{i, j\}$ be an arbitrary edge from C and $\ell = (E(P) \setminus \{e\}, \{e\})$. We now show that ℓ is not integral. To this end, let I be the connected component of the Hasse diagram of P'_ℓ which contains i and j and let $v = (v_1, v_2, \ldots, v_d) \in \mathbb{R}^d$ with

$$v_k = \begin{cases} \frac{1}{2}, & \text{if } k \in I \\ 0, & \text{otherwise.} \end{cases}$$

Then it is easy to see that

$$v = \bigcap_{\{p,q\} \in E(I)} H_{pq} \bigcap_{t \notin I} H_t \bigcap H_{ij},$$

where

$$\begin{aligned} H_{pq} &= \{(x_1, x_2, \dots, x_d) \mid x_p = x_q\} \text{ for } e = \{p, q\} \in E(I) \\ H_t &= \{(x_1, x_2, \dots, x_d) \mid x_t = 0\} \text{ for } t \notin I \\ H_{ij} &= \{(x_1, x_2, \dots, x_d) \mid x_i + x_j = 1\} \end{aligned}$$

are all facet hyperplanes of $\mathcal{OC}_\ell(P)$. So we deduce that v is a vertex of $\mathcal{OC}_\ell(P)$, and ℓ is not integral.

Conversely, suppose that P is a cycle-free poset on $[d]$ and ℓ is an edge partition of P. If $v = (a_1, a_2, \dots, a_d)$ is a vertex of $\mathcal{OC}_\ell(P)$, then we can find d independent facet hyperplanes of $\mathcal{OC}_\ell(P)$ such that

$$v = \left(\bigcap_{i=1}^{d-m} H'_i\right) \cap \left(\bigcap_{j=1}^{m} H''_j\right), \tag{2.1}$$

where $m = \dim\left(\bigcap_{i=1}^{d-m} H'_i\right)$, each H'_i is a facet hyperplane of $\mathcal{O}(P'_\ell)$ and each H''_j is a facet hyperplane of $\mathcal{C}(P''_\ell)$ which corresponds to a chain C_j of length ≥ 2 in P''_ℓ. By [16, Theorem 2.1], there is a set partition $\pi = \{B_1, B_2, \dots, B_{m+1}\}$ of $[d]$ such that $B_1, B_2, \dots, B_m$ are connected as subposets of P'_ℓ, $B_{m+1} = \{i \in [d] : a_i = 0 \text{ or } 1\}$ and

$$\bigcap_{i=1}^{d-m} H'_i = \{(x_1, x_2, \dots, x_d) \mid x_i = x_j \text{ if } \{i, j\} \subseteq B_k \text{ for some } 1 \leq k \leq m, \\ \text{and } x_r = a_r \text{ if } r \in B_{m+1}\}.$$

Let $B_{m+1} = \{r_1, r_2, \dots, r_s\}$ and for $1 \leq k \leq m$, let b_k denote the same values of all $a'_i s, i \in B_k$. Then it suffices to show that each b_k is an integer. Keeping in mind the assumption that the Hasse diagram of P is cycle-free, we find that $|C_i \cap B_j| \leq 1$ for $1 \leq i, j \leq m$. For $1 \leq i, j \leq m$, let

$$c_{ij} = \begin{cases} 1, & \text{if } |C_i \cap B_j| = 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

and for $1 \leq i \leq m, 1 \leq j \leq s$, let

$$d_{i,m+j} = \begin{cases} 1, & \text{if } r_j \in C_i \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

By (2.1), $(b_1, b_2, \dots, b_m, a_{r_1}, a_{r_2}, \dots, a_{r_s})$ must be the unique solution of the following linear system:

$$\begin{cases} \sum_{j=1}^{m} c_{ij} y_j + \sum_{j=m+1}^{m+s} d_{ij} y_j = 1, \quad 1 \leq i \leq m \\ y_{m+1} = a_{r_1} \\ y_{m+2} = a_{r_2} \\ \quad \vdots \\ y_{m+s} = a_{r_s}. \end{cases} \tag{2.4}$$

Now it suffices to show that the determinant of the coefficient matrix

$$A = \begin{pmatrix} c_{11} & \cdots & c_{1m} & d_{1,m+1} & \cdots & d_{1,m+s} \\ \vdots & & \vdots & \vdots & & \vdots \\ c_{m1} & \cdots & c_{mm} & d_{m,m+1} & \cdots & d_{m,m+s} \\ 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{pmatrix} \tag{2.5}$$

is equal to 1 or -1. Now construct a bipartite graph G with vertex set

$$\{B_1, B_2, \dots, B_m, C_1, C_2, \dots, C_m\},$$

and edge set

$$\{\{B_i, C_j\} \mid 1 \leq i, j \leq m, |B_i \cap C_j| = 1\}.$$

Let

$$C = \begin{pmatrix} c_{11} & \cdots & c_{1m} \\ \vdots & & \vdots \\ c_{m1} & \cdots & c_{mm} \end{pmatrix}.$$

Then we have

$$\det(C) = \sum_{\sigma \in \mathfrak{S}_m} \operatorname{sign}(\sigma) c_{1\sigma_1} \cdots c_{m\sigma_m}. \tag{2.6}$$

Clearly, each nonzero term in (2.6) corresponds to a perfect matching in the graph G. Since the Hasse diagram of P is cycle-free, the graph G must be a cycle-free bipartite graph, which means that there is at most one perfect matching in G. So we have $\det(C) = 0, 1$ or -1. Note that the linear equations (2.4) has unique solution $(b_1, b_2, \dots, b_m, a_{r_1}, \dots, a_{r_s})$. Then we find that $\det(C) = \pm 1$. It follows that each b_i is an integer. So the vertex v of $\mathcal{OC}_\ell(P)$ is integral. □

For a general finite poset P with $|E(P)| \geq 2$, the following theorem indicates that there exists at least one proper integral edge partition.

Theorem 2.2. *Suppose that P is a finite poset. Let $\operatorname{Min}(P)$ denote the set of all minimal elements in P. For $S \subseteq \operatorname{Min}(P)$, let $E_S(P)$ denote the set of all edges in $E(P)$ which are incident to some elements in S. Then the edge partition*

$$\ell = (E(P) \setminus E_S(P), E_S(P))$$

is integral.

Proof. Suppose that v is a vertex of $\mathcal{OC}_\ell(P)$. Then v can be represented as intersection of d independent facet hyperplanes, as in (2.1). Keeping the notation in the proof of Theorem 2.1, we can deduce that $|C_i| = 2$ and $|B_i \cap C_j| \leq 1$ for $1 \leq i, j \leq m$. So we can construct in the same way two matrices A and C as those in the proof of Theorem 2.1. Then, we can construct a graph G with vertex set $\{B_1, B_2, \dots, B_m, r_1, r_2, \dots, r_s\}$ and edge set determined by $C_1, C_2, \dots, C_m$. More precisely, $\{B_i, B_j\}$ is an edge of G if and only if there exists $1 \leq k \leq m$ such that $C_k = \{i', j'\}$ for some $i' \in B_i, j' \in B_j$, and $\{B_i, r_j\}$ is an edge of G if and only if there exists $1 \leq k \leq m$ such that $C_k = \{r_j, i'\}$ for some $i' \in B_i$. Obviously, G is a bipartite graph with bipartition $(\mathcal{B}_1, \mathcal{B}_2)$, where

$$\mathcal{B}_2 = \{B_j : 1 \leq j \leq m,\ B_j = \{k\} \text{ for some } k \in S\} \cup \{r_t :\ 1 \leq t \leq s,\ r_t \in S\},$$

and

$$\mathcal{B}_1 = \{B_1, B_2, \dots, B_m, r_1, r_2, \dots, r_s\} \setminus \mathcal{B}_2.$$

Moreover, by the construction of the graph G, its incidence matrix is

$$\begin{pmatrix} c_{11} & \cdots & c_{1m} & d_{1,m+1} & \cdots & d_{1,m+s} \\ \vdots & & \vdots & \vdots & & \vdots \\ c_{m1} & \cdots & c_{mm} & d_{m,m+1} & \cdots & d_{m,m+s} \end{pmatrix}$$

where $c_{ij}, d_{i,m+j}$ are defined in (2.2) and in (2.3) respectively. A well known fact shows that the incidence matrix of any bipartite graph is totally unimodular (a matrix A is *totally unimodular* if every square submatrix has determinant 0, +1, or −1). So the submatrix C has determinant $0, 1$ or -1. This completes the proof. □

Example 2.3. By Theorem 2.1, if the Hasse diagram of P has a cycle, then there exists at least one non-integral edge partition ℓ.

(1) For example, let P denote the poset whose Hasse diagram is a 4-cycle (see Figure 1) and let $E_1 = \{\{1,2\}, \{2,4\}, \{3,4\}\}$. Then the edge partition $\ell_1 = (E_1, \{1,3\})$ is non-integral, since $v = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$ is a vertex of $\mathcal{OC}_{\ell_1}(P)$ given by

$$\begin{cases} x_1 = x_2 = x_4 = x_3 \\ x_1 + x_3 = 1. \end{cases}$$

However, it is easy to see that the edge partition $\ell_2 = (\{1,3\}, E_1)$ is integral. So we find that the complementary edge partition $\ell^c = (cE(P), oE(P))$ of an integral edge partition $\ell = (oE(P), cE(P))$ is not necessarily integral.

(2) For any poset P whose Hasse diagram is a cycle and any edge partition ℓ of P, it is not hard to show that all coordinates of each vertex of $\mathcal{OC}_\ell(P)$ are 0, 1 or $\frac{1}{2}$.

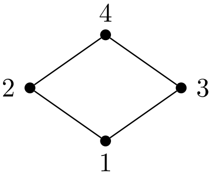


Figure 1: Hasse diagram of poset P from Example 2.3.

3 Unimodular equivalence

In this section, we shall compare the newly constructed order-chain polytopes with some known polytopes. Specifically, we will focus on integral order-chain polytopes and consider their unimodular equivalence relation with order polytopes or chain polytopes.

Recall (see, for example, [9]) that a $d \times d$ integral matrix U is *unimodular* if $\det(U) = \pm 1$. A map $\varphi\colon \mathbb{R}^d \to \mathbb{R}^d$ is a *unimodular transformation* if there exist a $d \times d$ unimodular matrix U and an integral vector $w \in \mathbb{Z}^d$ such that $\varphi(v) = vU + w$. Two integral polytopes $\mathcal{P}$ and $\mathcal{Q}$ in $\mathbb{R}^d$ are *unimodularly equivalent* if there exists a unimodular transformation $\varphi\colon \mathbb{R}^d \to \mathbb{R}^d$ such that $\mathcal{Q} = \varphi(\mathcal{P})$. Much of the importance of unimodular equivalence arises from the fact that combinatorial type and Ehrhart polynomial of an integral polytope are invariant modulo unimodular equivalence. For instance, classification of polytopes with certain properties (modulo unimodular equivalence) has gained some attentions recently (see, for example, [1, 10, 11]).

We shall use the ideas in the proof of the following theorem due to Hibi and Li [9].

Theorem 3.1 ([9, Theorem 1.3]). *The order polytope $\mathcal{O}(P)$ and the chain polytope $\mathcal{C}(P)$ of a finite poset P are unimodularly equivalent if and only if the poset shown in Figure 2 does not appear as a subposet of P.*

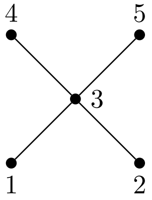


Figure 2: The "forbidden" poset from Theorem 3.1.

Definition 3.2. A poset P on $[d]$ is said to be a zigzag poset if its cover relations are given by

$$1 \prec \cdots \prec i_1 \succ i_1 + 1 \succ \cdots \succ i_2 \prec i_2 + 1 \prec \cdots \prec i_3 \succ \cdots \succ i_k \prec i_k + 1 \prec \cdots \prec d$$

for some $0 \le i_1 < i_2 < \cdots < i_k \le d$.

Theorem 3.3. *Suppose that P is a disjoint union of chains. Then for any edge partition ℓ, the order-chain polytope $\mathcal{OC}_\ell(P)$ is unimodularly equivalent to a chain polytope $\mathcal{C}(Q)$, where Q is a disjoint union of zigzag posets.*

Proof. We firstly assume that P is a chain:

$$1 \prec 2 \prec 3 \prec \cdots \prec d.$$

and ℓ is an edge partition of P given by:

$$\begin{array}{ll}
o\colon & 1 \prec 2 \prec \cdots \prec i_1 \\
c\colon & i_1 \prec i_1+1 \prec \cdots \prec i_2 \\
o\colon & i_2 \prec i_2+1 \prec \cdots \prec i_3 \\
 & \vdots \\
c\colon & i_{t-1} \prec i_{t-1}+1 \prec \cdots \prec i_t \\
o\colon & i_t \prec i_t+1 \prec \cdots \prec i_{t+1} \\
 & \vdots \\
c\colon & i_{k-1} \prec i_{k-1}+1 \prec \cdots \prec i_k = d,
\end{array}$$

where $1 \leq i_1 < i_2 < \cdots < i_{k-1} \leq i_k = d$. Then the order-chain polytope $\mathcal{OC}_\ell(P)$ is given by

$$\begin{cases}
x_1 \geq x_2 \geq \cdots \geq x_{i_1} \\
x_{i_1} + x_{i_1+1} + \cdots + x_{i_2} \leq 1 \\
x_{i_2} \geq x_{i_2+1} \geq \cdots \geq x_{i_3} \\
\qquad \vdots \\
x_{i_{t-1}} + x_{i_{t-1}+1} + \cdots + x_{i_t} \leq 1 \\
x_{i_t} \geq x_{i_t+1} \geq \cdots \geq x_{i_{t+1}} \\
\qquad \vdots \\
x_{i_{k-1}} + x_{i_{k-1}+1} + \cdots + x_d \leq 1 \\
0 \leq x_i \leq 1, \quad 1 \leq i \leq d.
\end{cases} \tag{3.1}$$

Now define a map $\varphi\colon \mathbb{R}^d \to \mathbb{R}^d$ as follows:

(1) if i is a maximal element in P'_ℓ, then let $x'_i = x_i$;

(2) if i is not a maximal element in P'_ℓ, then $\{i, i+1\}$ must be an edge in the Hasse diagram of P'_ℓ. Let $x'_i = x_i - x_{i+1}$.

Let $\varphi(x_1, x_2, \ldots, x_d) = (x'_1, x'_2, \ldots, x'_d)$.

Now it is easy to show that φ is a unimodular transformation. Moreover, the system (3.1) is transformed into:

$$\begin{cases}
x'_1 + x'_2 + \cdots + x'_{i_1} \leq 1 \\
x'_{i_1} + x'_{i_1+1} + \cdots + x'_{i_2} + x'_{i_2+1} + \cdots + x'_{i_3} \leq 1 \\
\qquad \vdots \\
x'_{i_{t-1}} + x'_{i_{t-1}+1} + \cdots + x'_{i_t} + x'_{i_t+1} + \cdots + x'_{i_{t+1}} \leq 1 \\
\qquad \vdots \\
x'_{i_{k-1}} + x'_{i_{k-1}+1} + \cdots + x'_d \leq 1 \\
0 \leq x'_i \leq 1, \quad 1 \leq i \leq d.
\end{cases}$$

Obviously, this system corresponds to the chain polytope $\mathcal{C}(Q)$ for the zigzag poset Q:

$$1 \prec 2 \prec \cdots \prec i_1 \succ i_1 + 1 \succ \cdots \succ i_2 \succ i_2 + 1 \succ \cdots \succ i_3 \prec \cdots$$

or the dual zigzag poset Q^*:

$$1 \succ 2 \succ \cdots \succ i_1 \prec i_1 + 1 \prec \cdots \prec i_2 \prec i_2 + 1 \prec \cdots \prec i_3 \succ \cdots$$

So we deduce that $\mathcal{OC}_\ell(P)$ is unimodularly equivalent to the chain polytope of some zigzag poset.

Now we continue to prove the general case that P is a disjoint union of k chains:

$$P = C_1 \uplus C_2 \uplus \cdots \uplus C_k.$$

Since

$$\mathcal{O}(P \uplus Q) = \mathcal{O}(P) \times \mathcal{O}(Q) \quad \text{and} \quad \mathcal{C}(P \uplus Q) = \mathcal{C}(P) \times \mathcal{C}(Q),$$

we have

$$\begin{aligned} \mathcal{OC}_\ell(P \uplus Q) &= \mathcal{O}((P \uplus Q)'_\ell) \cap \mathcal{C}((P \uplus Q)''_\ell) \\ &= \mathcal{O}(P'_\ell \uplus Q'_\ell) \cap \mathcal{C}(P''_\ell \uplus Q''_\ell) \\ &= [\mathcal{O}(P'_\ell) \times \mathcal{O}(Q'_\ell)] \cap [\mathcal{C}(P''_\ell) \times \mathcal{C}(Q''_\ell)] \\ &= [\mathcal{O}(P'_\ell) \cap \mathcal{C}(P''_\ell)] \times [\mathcal{O}(Q'_\ell) \cap \mathcal{C}(Q''_\ell)] \\ &= \mathcal{OC}_\ell(P) \times \mathcal{OC}_\ell(Q). \end{aligned} \tag{3.2}$$

Hence we conclude that

$$\begin{aligned} \mathcal{OC}_\ell(C_1 \uplus \cdots \uplus C_k) &= \mathcal{OC}_\ell(C_1) \times \cdots \times \mathcal{OC}_\ell(C_k) \\ &\overset{\varphi_1 \times \cdots \times \varphi_k}{\cong} \mathcal{C}(Q_1) \times \cdots \times \mathcal{C}(Q_k) \\ &= \mathcal{C}(Q_1 \uplus \cdots \uplus Q_k), \end{aligned}$$

where Q_i are zigzag posets. □

Similarly, we can modify the proof of Theorem 3.3 slightly to get the following result:

Theorem 3.4. *Suppose that P is a finite zigzag poset. Then for any edge partition ℓ, the order-chain polytope $\mathcal{OC}_\ell(P)$ is unimodularly equivalent to a chain polytope $\mathcal{C}(Q)$ for some zigzag poset Q.*

Proof. Suppose that P is a zigzag poset on $[d]$ and ℓ is an edge partition of P. Define a map $\varphi\colon \mathbb{R}^d \to \mathbb{R}^d$ as follows:

(1) if i is covered by at most one element in P'_ℓ, let

$$x'_i = \begin{cases} x_i, & \text{if } i \text{ is a maximal element in } P'_\ell \\ x_i - x_j, & \text{if } i \text{ is covered by } j \text{ in } P'_\ell \quad (j = i-1 \text{ or } i+1). \end{cases}$$

(2) if i is covered by both $i-1$ and $i+1$ in P'_ℓ, let

$$x'_i = 1 - x_i.$$

Let

$$\varphi(x_1, x_2, \ldots, x_d) = (x'_1, x'_2, \ldots, x'_d).$$

It is not hard to show that φ is the desired unimodular transformation. □

The following example shows that not every order-chain polytope $\mathcal{OC}_\ell(P)$ of a cycle-free poset P is unimodularly equivalent to some chain polytope.

Example 3.5. Let P be the poset shown in Figure 2 with an edge partition

$$\ell = (\{\{1,3\},\{3,4\},\{3,5\}\},\{2,3\}).$$

Let

$$\varphi(x_1, x_2, x_3, x_4, x_5) = (x_1, 1 - x_2, x_3, x_4, x_5).$$

It is obvious that φ is a unimodular transformation and $\varphi(\mathcal{OC}_\ell(P)) = \mathcal{O}(P)$. However, by checking all 63 different non-isomorphic posets with 5 elements, we find that $\mathcal{O}(P)$ is not equivalent to any chain polytope.

Furthermore, for any $d \geq 6$, we shall find an integral order-chain polytope in $\mathbb{R}^d$ which is not unimodularly equivalent to any chain polytope or order polytope. To this end, we need the following lemma.

Lemma 3.6.

(1) None of the chain polytopes of finite posets on $[d]$ possesses $d+4$ vertices and $d+7$ facets.

(2) None of the order polytopes of finite posets on $[d]$ possesses $d+4$ vertices and $d+7$ facets.

Proof. (1) Assume, by contradiction, that P is a finite poset on $[d]$ such that $\mathcal{C}(P)$ has $d+4$ vertices and $d+7$ facets. Since the vertices of $\mathcal{C}(P)$ are those $\rho(A)$ for which A is an antichain of P, we can deduce that P possesses exactly $d+4$ antichains. Keeping in mind that $\emptyset, \{1\}, \ldots, \{d\}$ are antichains of P, we find that there is no antichain A in P with $|A| \geq 3$. Otherwise, the number of antichains of P is at least $d+5$. It then follows that there are exactly three 2-element antichains in P. We need to consider the following four cases:

(i) Let, say, $\{1,2\},\{1,3\},\{1,4\}$ be the 2-element antichains of P. Then the maximal chains of P are $P \setminus \{1\}$ and $P \setminus \{2,3,4\}$.

(ii) Let, say, $\{1,2\},\{1,3\},\{2,4\}$ be the 2-element antichains of P. Then the maximal chains of P are $P \setminus \{1,2\}$, $P \setminus \{1,4\}$ and $P \setminus \{2,3\}$.

(iii) Let, say, $\{1,2\},\{1,3\},\{4,5\}$ be the 2-element antichains of P. Then the maximal chains of P are $P \setminus \{1,4\}$, $P \setminus \{1,5\}$, $P \setminus \{2,3,4\}$ and $P \setminus \{2,3,5\}$.

(iv) Let, say, $\{1,2\},\{3,4\},\{5,6\}$ be the 2-element antichains of P. It can be shown easily that P possesses exactly eight maximal chains.

Recall that the number of facets of $\mathcal{C}(P)$ is equal to $d+c(P)$, it follows from the assumption that there are exactly 7 maximal chains in P, which is a contradiction. As a result, none of the chain polytopes $\mathcal{C}(P)$ of a finite poset P on $[d]$ with $d+4$ vertices can possess $d+7$ facets, as desired.

(2) Let P be a finite poset on $[d]$ and suppose that the number of vertices of $\mathcal{O}(P)$ is $d+4$ and the number of facets of $\mathcal{O}(P)$ is $d+7$. Since the number of vertices of $\mathcal{O}(P)$ and that of $\mathcal{C}(P)$ coincide, it follows from the proof of (a) that there is no antichain A in P with $|A| \geq 3$ and that P includes exactly three 2-element antichains. On the other hand, it is known [9, Corollary 1.2] that the number of facets of $\mathcal{O}(P)$ is less than or equal to that of $\mathcal{C}(P)$. Hence the number of maximal chains of P is at least 7. Thus, by using the argument in the proof of (a), we can assume that the antichains of P are $\{1,2\}, \{3,4\}$ and $\{5,6\}$. Then, it is easy to prove that the number $|E(\hat{P})|$ of edges in the Hasse diagram of $\hat{P} = P \cup \{\hat{0}, \hat{1}\}$ is at most $d+6$. So we deduce that the number of facets of $\mathcal{O}(P)$ is at most $d+6$, a contradiction with the assumption. □

We remark that, by modifying the argument of the statement (1) in Lemma 3.6, we can prove directly that the order polytope of Example 3.5 cannot be unimodularly equivalent to any chain polytope.

Example 3.7. Let P be the finite poset shown in Figure 3. Let ℓ be the edge partition with $oE(P) = \{\{3,5\},\{3,6\}\}$ and $cE(P) = E(P) \setminus oE(P)$. Then it is easy to verify that $\mathcal{OC}_\ell(P)$ is an integral polytope with 10 vertices and 13 facets. (Since the number of facets of the order-chain polytope is small, we can compute this by hand. Of course, we can also compute this by using the software polymake [4].) So it follows from Lemma 3.6 that the integral order-chain polytope $\mathcal{OC}_\ell(P)$ cannot be unimodularly equivalent to any order polytope or any chain polytope.

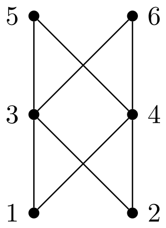


Figure 3: Poset P from Example 3.7.

In fact, for any $d > 6$, let P_d be the poset shown in Figure 4 and let ℓ be the edge partition with

$$oE(P_d) = \{\{3,5\},\{3,6\},\{5,7\},\{6,7\},\{7,8\},\ldots,\{d-1,d\}\}.$$

It is easy to see that the order-chain polytope $\mathcal{OC}_\ell(P_d)$ has $d+4$ vertices and $d+7$ facets. Therefore $\mathcal{OC}_\ell(P_d)$ cannot be unimodularly equivalent to any order polytope or any chain polytope.

Recall that Example 3.5 shows that there is an order polytope which is not unimodularly equivalent to any chain polytope. To conclude this section, we will prove that, for each $d \geq 9$, there exists a finite poset P on $[d]$ for which the chain polytope $\mathcal{C}(P)$ cannot be unimodularly equivalent to any order polytope.

Recall that, for a finite poset P on $[d]$, we have

$$f_{d-1}(\mathcal{O}(P)) = m_\star(P) + m^\star(P) + |E(P)|$$

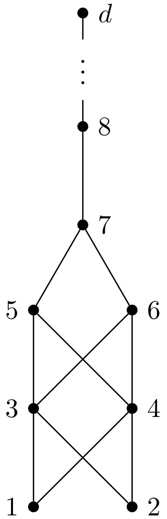


Figure 4: Poset P_d from Example 3.7.

and

$$f_{d-1}(\mathcal{C}(P)) = d + c(P),$$

To present our results, we firstly discuss upper bounds for $f_{d-1}(\mathcal{O}(P))$ and $f_{d-1}(\mathcal{C}(P))$. By [9, Theorem 2.1], if $d \leq 4$, then $\mathcal{O}(P)$ and $\mathcal{C}(P)$ are unimodularly equivalent and $f_{d-1}(\mathcal{O}(P)) = f_{d-1}(\mathcal{C}(P)) \leq 2d$. Moreover, for each $1 \leq d \leq 4$, there exists a finite poset P on $[d]$ with $f_{d-1}(\mathcal{O}(P)) = f_{d-1}(\mathcal{C}(P)) = 2d$.

Lemma 3.8. *Let $d \geq 5$ and P be a finite poset on $[d]$. Then*

$$f_{d-1}(\mathcal{O}(P)) \leq \left\lfloor \frac{d+1}{2} \right\rfloor \left(d - \left\lfloor \frac{d+1}{2} \right\rfloor \right) + d \tag{3.3}$$

and

$$f_{d-1}(\mathcal{C}(P)) \leq \begin{cases} 3^k + d, & d = 3k \\ 4 \cdot 3^{k-1} + d, & d = 3k+1 \\ 2 \cdot 3^k + d, & d = 3k+2. \end{cases} \tag{3.4}$$

Furthermore, both upper bounds for $f_{d-1}(\mathcal{O}(P))$ and $f_{d-1}(\mathcal{C}(P))$ are tight.

Proof. **(Order polytope)** Let $d = 4$. Since the right-hand side of (3.3) is equal to $2d$ $(= 8)$, the inequality (3.3) also holds for $d = 4$. Let $d \geq 5$ and P be a finite poset on $[d]$. We will prove (3.3) by induction on d. Suppose that 1 is a minimal element of P and let a be the number of elements in P which cover 1.

If $a = 0$, then $\mathcal{O}(P) = \mathcal{O}(P \setminus \{1\}) \times [0,1]$ and so

$$\begin{aligned} f_{d-1}(\mathcal{O}(P)) &= f_{d-2}(\mathcal{O}(P \setminus \{1\})) + 2 \\ &\leq \left\lfloor \frac{d}{2} \right\rfloor \left(d - 1 - \left\lfloor \frac{d}{2} \right\rfloor \right) + d - 1 + 2 \\ &\leq \left\lfloor \frac{d+1}{2} \right\rfloor \left(d - \left\lfloor \frac{d+1}{2} \right\rfloor \right) + d. \end{aligned}$$

If $1 \leq a \leq \lfloor d/2 \rfloor$, then from the facts that $|E(P \setminus \{1\})| = |E(P)| - a$, $m_\star(P \setminus \{1\}) \geq m_\star(P) - 1$ and $m^\star(P \setminus \{1\}) = m^\star(P)$, we have

$$\begin{aligned} f_{d-1}(\mathcal{O}(P)) &= m^\star(P) + m_\star(P) + |E(P)| \\ &\leq m^\star(P \setminus \{1\}) + m_\star(P \setminus \{1\}) + 1 + |E(P \setminus \{1\})| + a \\ &\leq \left\lfloor \frac{d}{2} \right\rfloor \left(d - 1 - \left\lfloor \frac{d}{2} \right\rfloor \right) + (d-1) + \left\lfloor \frac{d}{2} \right\rfloor + 1 \\ &\leq \left\lfloor \frac{d+1}{2} \right\rfloor \left(d - \left\lfloor \frac{d+1}{2} \right\rfloor \right) + d. \end{aligned}$$

Now we consider the case $\lfloor d/2 \rfloor + 1 \leq a \leq d - 1$. Let, say, 2 be an element of P which covers 1. Since the set of the elements of P which cover 1 is an antichain of P, it follows that $|E(P \setminus \{2\})| \geq |E(P)| - (d-a)$, $m_\star(P \setminus \{2\}) \geq m_\star(P)$ and $m^\star(P \setminus \{2\}) \geq m^\star(P) - 1$. Hence

$$\begin{aligned} f_{d-1}(\mathcal{O}(P)) &= m^\star(P) + m_\star(P) + |E(P)| \\ &\leq m^\star(P \setminus \{2\}) + 1 + m_\star(P \setminus \{2\}) + |E(P \setminus \{2\})| + (d-a) \\ &\leq \left\lfloor \frac{d}{2} \right\rfloor \left(d - 1 - \left\lfloor \frac{d}{2} \right\rfloor \right) + (d-1) + \left(d - \left\lfloor \frac{d}{2} \right\rfloor - 1 \right) + 1 \\ &\leq \left\lfloor \frac{d+1}{2} \right\rfloor \left(d - \left\lfloor \frac{d+1}{2} \right\rfloor \right) + d. \end{aligned}$$

Therefore, the inequality (3.3) holds. We proceed to show that this upper bound for $f_{d-1}(\mathcal{O}(P))$ is tight. In fact, let P be the finite poset P on $[d]$ with

$$E(P) = \left\{ \{i,j\} \in [d] \times [d] : 1 \leq i \leq \left\lfloor \frac{d+1}{2} \right\rfloor, \left\lfloor \frac{d+1}{2} \right\rfloor + 1 \leq j \leq d \right\}.$$

Clearly, we have

$$f_{d-1}(\mathcal{O}(P)) = \left\lfloor \frac{d+1}{2} \right\rfloor \left(d - \left\lfloor \frac{d+1}{2} \right\rfloor \right) + d.$$

(Chain polytope) Let $d \geq 5$. Let P_1 be a finite poset on $[d]$ and M_1 the set of minimal elements of P_1. If P_1 is an antichain, then $f_{d-1}(\mathcal{C}(P_1)) = 2d$. Suppose that P_1 is not an antichain. Let $P_2 = P_1 \setminus M_1$ and M_2 be the set of minimal elements of P_2. In general, if P_i is not an antichain and M_i is the set of minimal element of P_i, then we set $P_{i+1} = P_i \setminus M_i$. By continuing this construction, we can get an integer $r \geq 1$ such that each of the $P_1, \ldots, P_{r-1}$ is not an antichain and that P_r is an antichain. Let P be the finite poset on $[d]$ such that $i_1 \prec i_2 \prec \cdots \prec i_r$ if $i_j \in M_j$ for $1 \leq j \leq r$. One has $c(P_1) \leq c(P) = |M_1| \cdots |M_r|$. For any integer $d \geq 5$, let

$$M(d) = \max \left\{ \prod_{i=1}^{r} m_i : 1 \leq r \leq d,\, m_1 + m_2 + \cdots + m_r = d,\, m_i \in \mathbb{N}^+ \right\}.$$

Then the desired inequalities (3.4) follows immediately from the following claim:

$$M(d) = \begin{cases} 3^k, & d = 3k \\ 4 \cdot 3^{k-1}, & d = 3k+1 \\ 2 \cdot 3^k, & d = 3k+2. \end{cases} \tag{3.5}$$

So it suffices to prove this claim. Since for any integer $m \geq 4$,

$$m \leq \left\lfloor \frac{m+1}{2} \right\rfloor \left(m - \left\lfloor \frac{m+1}{2} \right\rfloor \right),$$

we can assume that, to maximize the product $\prod_{i=1}^{r} m_i$, all parts $m_i \leq 3$. We can also assume without loss of generality that there are at most two m_is that are equal to 2, since $2^3 < 3^2$. Then the claim (3.5) follows immediately.

Finally, for each $d \geq 5$, the existence of a finite poset P on $[d]$ for which the equality holds in (3.4) follows easily from the above argument. □

Remark 3.9. The special case $d = 1976$ of claim (3.5) is exactly the problem 4 in the International Mathematical Olympiad (IMO) in 1976, where the maximum value of a product of positive integers summing up to 1976 is asked for. The answer is $2 \cdot 3^{658}$ since $1976 = 3 \cdot 658 + 2$.

A routine computation shows that, for each $1 \leq d \leq 8$, the right-hand side of (3.3) coincides with that of (3.4) and that, for each $d \geq 9$, the right-hand side of (3.3) is strictly less than that of (3.4). Hence

Corollary 3.10. *For each $d \geq 9$, there exists a finite poset P on $[d]$ for which the chain polytope $\mathcal{C}(P)$ cannot be unimodularly equivalent to any order polytope.*

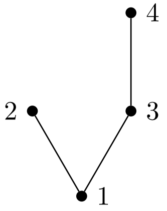
4 Volumes of $\mathcal{OC}_\ell(P)$

Given a poset P on $[d]$, Corollary 4.2 in [16] shows that the volumes of $\mathcal{O}(P)$ and $\mathcal{C}(P)$ are given by

$$V(\mathcal{O}(P)) = V(\mathcal{C}(P)) = \frac{e(P)}{d!},$$

where $e(P)$ is the number of linear extensions of P. (Recall that a linear extension of P is a permutation $\pi = \pi_1\pi_2\cdots\pi_d$ of $[d]$ such that $\pi^{-1}(i) < \pi^{-1}(j)$ if $i \prec j$ in P.)

For order-chain polytopes, different edge partitions usually give rise to polytopes with different volumes. For example, let P be the poset as follows:



It is easy to see that
$$V(\mathcal{O}(P)) = V(\mathcal{C}(P)) = \frac{3}{4!}.$$
Let
$$\ell = (\{1,2\},\{\{1,3\},\{3,4\}\}), \quad \ell' = (\{\{1,2\},\{1,3\}\},\{3,4\}),$$
then we have
$$V(\mathcal{OC}_\ell(P)) = \frac{1}{4!} \quad \text{and} \quad V(\mathcal{OC}_{\ell'}(P)) = \frac{5}{4!}.$$
Hence one has the following inequality:
$$V(\mathcal{OC}_\ell(P)) < V(\mathcal{O}(P)) = V(\mathcal{C}(P)) < V(\mathcal{OC}_{\ell'}(P)).$$

It should be noted that, for an arbitrary poset P, we can not always find edge partitions such that this inequality holds. For example, if P is a chain, then there is no edge partition ℓ such that $V(\mathcal{OC}_\ell(P)) < V(\mathcal{O}(P)) = V(\mathcal{C}(P))$. Then a natural question is to ask which edge partition ℓ gives rise to an order-chain polytope with maximum volume. It seems very difficult to solve this problem in general case. In this section, we consider the special case when P is a chain P on $[d]$. We transform it to a problem of maximizing descent statistics over certain family of subsets. For references on this topic, we refer the reader to [3] and [14].

Let P be a chain on $[d]$. By the proof of Theorem 3.3, for an edge partition ℓ of P, the order-chain polytope $\mathcal{OC}_\ell(P)$ is unimodularly equivalent to a chain polytope $\mathcal{C}(P_1)$, where P_1 is a zigzag poset such that all maximal chains, except the first one (containing 1) and the last one (containing d), consist of at least three elements. So we have
$$V(\mathcal{OC}_\ell(P)) = V(\mathcal{C}(P_1)) = \frac{e(P_1)}{d!}.$$
Conversely, for such a zigzag poset P_1, it is easy to find an edge partition ℓ of P such that $\mathcal{OC}_\ell(P)$ is unimodularly equivalent to $\mathcal{C}(P_1)$. Denote by $\mathcal{Z}(d)$ the set of such zigzag posets P_1 on $[d]$. Thus, to compute the maximum volume over all order-chain polytopes of the chain P, it suffices to compute the maximum number of linear extensions for all zigzag posets $P_1 \in \mathcal{Z}(d)$. Next we shall represent this problem as a problem of maximizing descent statistic over a certain class of subsets. To this end, we recall some notions and basic facts. Given a permutation $\pi = \pi_1\pi_2\cdots\pi_d$, let $\mathrm{Des}(\pi)$ denote its descent set $\{i \in [d-1] : \pi_i > \pi_{i+1}\}$. For $S \subseteq [d-1]$, define the descent statistic $\beta(S)$ to be the number of permutations of $[d]$ with descent set S. Note that there is an obvious bijection between zigzag posets on $[d]$ and subsets of $[d-1]$ given by
$$S\colon P \mapsto \{j \in [d-1] : j \succ j+1\}.$$

Moreover, a permutation $\pi = \pi_1\pi_2\cdots\pi_d$ of $[d]$ is a linear extension of P if and only if $\mathrm{Des}(\pi^{-1}) = S(P)$. Let $\mathcal{F}(d) = S(\mathcal{Z}(d))$. Then we can transform the problem of maximizing volume of order-chain polytopes of a d-chain to the problem of maximizing the descent statistic $\beta(S)$, where S ranges over $\mathcal{F}(d)$.

Observe that $\beta(S) = \beta(\bar{S})$, where $\bar{S} = [d-1] \setminus S$. Following [3], we will encode both S and $\bar{S}$ by a list $L = (l_1, l_2, \ldots, l_k)$ of positive integers such that $l_1 + l_2 + \cdots + l_k = d - 1$. Given $S \subseteq [d-1]$, a run of S is a set $R \subseteq [d-1]$ of consecutive integers of maximal cardinality such that $R \subseteq S$ or $R \subseteq \bar{S}$. For example, if $d = 10$, then the set $S = \{1, 2, 5, 8, 9\}$ has 5 runs: $\{1, 2\}$, $\{3, 4\}$, $\{5\}$, $\{6, 7\}$, $\{8, 9\}$. Suppose that S has k runs $R_1, R_2, \ldots, R_k$ with $|R_i| = l_i$, let $L(S) = (l_1, l_2, \ldots, l_k)$.

Lemma 4.1. *Suppose that $S \subseteq [d-1]$ and $L(S) = (l_1, l_2, \ldots, l_k)$. Then $S \in \mathcal{F}(d)$ if and only if $l_i \geq 2$ for all $2 \leq i \leq k-1$.*

Proof. The lemma follows immediately from the fact that $\mathcal{Z}(d)$ consists of zigzag posets P such that all maximal chains in P, except the first one (containing 1) and the last one (containing d), contain at least three elements. □

Denote by F_d the dth Fibonacci number. By Lemma 4.1, it is easy to see that $|\mathcal{F}(d)| = 2F_d$ for $d \geq 2$. Based on computer evidences, we conjectured the following results about maximizing descent statistic over $\mathcal{F}(d)$, which in fact[1] is a special case of Theorem 6.1 in [3].

Proposition 4.2. *Suppose that $d \geq 2$ and $S \subseteq [d-1]$.*

(1) If $d = 2m$ and

$$L(S) = (1, \underbrace{2, 2, \ldots, 2}_{m-1}) \quad \text{or} \quad L(S) = (\underbrace{2, 2, \ldots, 2}_{m-1}, 1),$$

then $\beta(T) \leq \beta(S)$ for any $T \in \mathcal{F}(d)$.

(2) If $d = 2m+1$ and

$$L(S) = (1, \underbrace{2, 2, \ldots, 2}_{m-1}, 1),$$

then $\beta(T) \leq \beta(S)$ for any $T \in \mathcal{F}(d)$.

Equivalently, by the proof of Theorem 3.3, we have

Proposition 4.3. *Let P be a chain on $[d]$. Then the alternating edge partition $\ell = (oE(P), cE(P))$ with*

$$oE(P) = \begin{cases} \{\{1,2\},\{3,4\},\ldots,\{d-1,d\}\}, & \text{if } d \text{ is even} \\ \{\{1,2\},\{3,4\},\ldots,\{d-2,d-1\}\}, & \text{otherwise} \end{cases}$$

gives rise to an order-chain polytope $\mathcal{OC}_\ell(P)$ with maximum volume.

[1] We thank Joe Gallian and Mitchell Lee for bringing [3, Theorem 6.1] to our attention.

References

[1] M. Blanco and F. Santos, Lattice 3-polytopes with few lattice points, *SIAM J. Discrete Math.* **30** (2016), 669–686, doi:10.1137/15m1014450.

[2] J. Brown and V. Lakshmibai, Singular loci of Bruhat-Hibi toric varieties, *J. Algebra* **319** (2008), 4759–4779, doi:10.1016/j.jalgebra.2007.10.033.

[3] R. Ehrenborg and S. Mahajan, Maximizing the descent statistic, *Ann. Comb.* **2** (1998), 111–129, doi:10.1007/bf01608482.

[4] E. Gawrilow and M. Joswig, `polymake`: a framework for analyzing convex polytopes, in: G. Kalai and G. M. Ziegler (eds.), *Polytopes — Combinatorics and Computation*, Birkhäuser, Basel, volume 29 of *DMV Seminar*, pp. 43–73, 2000, doi:10.1007/978-3-0348-8438-9_2, including papers from the DMV-Seminar "Polytopes and Optimization" held in Oberwolfach, November 1997.

[5] T. Hibi, Distributive lattices, affine semigroup rings and algebras with straightening laws, in: M. Nagata and H. Matsumura (eds.), *Commutative Algebra and Combinatorics*, North-Holland, Amsterdam, volume 11 of *Advanced Studies in Pure Mathematics*, pp. 93–109, 1987, papers from the U.S.-Japan Joint Seminar held in Kyoto, August 25 – 31, 1985.

[6] T. Hibi, *Algebraic Combinatorics on Convex Polytopes*, Carslaw Publications, Glebe, New South Wales, 1992.

[7] T. Hibi (ed.), *Gröbner Bases: Statistics and Software Systems*, Springer, Tokyo, 2013, doi:10.1007/978-4-431-54574-3.

[8] T. Hibi and N. Li, Chain polytopes and algebras with straightening laws, *Acta Math. Vietnam.* **40** (2015), 447–452, doi:10.1007/s40306-015-0115-2.

[9] T. Hibi and N. Li, Unimodular equivalence of order and chain polytopes, *Math. Scand.* **118** (2016), 5–12, doi:10.7146/math.scand.a-23291.

[10] T. Hibi and A. Tsuchiya, Classification of lattice polytopes with small volumes, 2018, `arXiv:1708.00413 [math.CO]`.

[11] F. Liu, On positivity of Ehrhart polynomials, 2018, `arXiv:1711.09962 [math.CO]`.

[12] H. Ohsugi and T. Hibi, Normal polytopes arising from finite graphs, *J. Algebra* **207** (1998), 409–426, doi:10.1006/jabr.1998.7476.

[13] H. Ohsugi and T. Hibi, Convex polytopes all of whose reverse lexicographic initial ideals are squarefree, *Proc. Amer. Math. Soc.* **129** (2001), 2541–2546, doi:10.1090/s0002-9939-01-05853-1.

[14] B. E. Sagan, Y.-N. Yeh and G. M. Ziegler, Maximizing Möbius functions on subsets of Boolean algebras, *Discrete Math.* **126** (1994), 293–311, doi:10.1016/0012-365x(94)90273-9.

[15] R. P. Stanley, Decompositions of rational convex polytopes, *Ann. Discrete Math.* **6** (1980), 333–342, doi:10.1016/s0167-5060(08)70717-9.

[16] R. P. Stanley, Two poset polytopes, *Discrete Comput. Geom.* **1** (1986), 9–23, doi:10.1007/bf02187680.

[17] B. Sturmfels, *Gröbner Bases and Convex Polytopes*, volume 8 of *University Lecture Series*, American Mathematical Society, Providence, RI, 1996.

[18] Y. Wang, Sign Hibi cones and the anti-row iterated Pieri algebras for the general linear groups, *J. Algebra* **410** (2014), 355–392, doi:10.1016/j.jalgebra.2014.01.039.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 319–329
https://doi.org/10.26493/1855-3974.1436.7a2
(Also available at http://amc-journal.eu)

Comparing topologies on linearly recursive sequences*

Laiachi El Kaoutit [†]
Universidad de Granada, Departamento de Álgebra and *IEMath-Granada,*
Facultad de Educación, Econonía y Tecnología de Ceuta,
Cortadura del Valle, s/n. E-51001 Ceuta, Spain

Paolo Saracco [‡]
Université Libre de Bruxelles, Département de Mathématique,
Boulevard du Triomphe, B-1050 Brussels, Belgium

Received 29 June 2017, accepted 24 March 2018, published online 9 January 2019

Abstract

The space of linearly recursive sequences of complex numbers admits two distinguished topologies. Namely, the adic topology induced by the ideal of those sequences whose first term is 0 and the topology induced from the Krull topology on the space of complex power series via a suitable embedding. We show that these topologies are not equivalent.

Keywords: Linearly recursive sequences, adic topologies, power series, Hopf algebras.

Math. Subj. Class.: 13J05, 40A05, 16W70, 13J10, 54A10, 16W80

1 Introduction

A linearly recursive sequence of complex numbers is a sequence of elements of $\mathbb{C}$ which satisfies a recurrence relation with constant coefficients. These sequences arise widely in mathematics and have been studied extensively and from different points of view, see [12]

[*] This paper was written while P. Saracco was member of the “National Group for Algebraic and Geometric Structures, and their Applications” (GNSAGA – INdAM). His stay as visiting researcher at the campus of Ceuta of the University of Granada was financially supported by IEMath-GR. Research supported by grant PRX16/00108, and Spanish Ministerio de Economía y Competitividad and the European Union FEDER, grant MTM2016-77033-P.

The authors thank A. Ardizzoni for the fruitful (coffees and) discussions. The first author would like to thank all the members of the dipartimento di Matematica “Giuseppe Peano” for providing a very fruitful working ambience.

[†] Home page: `https://www.ugr.es/~kaoutit/`

[‡] Home page: `https://sites.google.com/site/paolosaracco/`

E-mail addresses: kaoutit@ugr.es (Laiachi El Kaoutit), paolo.saracco@ulb.ac.be (Paolo Saracco)

for a survey on the subject. Classically they are related with formal power series, in the sense that a sequence $(s_n)_{n\geq 0}$ is linearly recursive if and only if its generating function $\sum_{n\geq 0} s_n Z^n$ is a rational function $p(Z)/q(Z)$, where $p(Z), q(Z) \in \mathbb{C}[Z]$ and $q(0) \neq 0$. Nevertheless, few topological properties seems to be known. For instance, it is known that the space $\mathcal{L}in(\mathbb{C})$ of all linearly recursive sequences of any order forms an augmented algebra under the Hurwitz product, with augmentation given by the projection on the 0-th component. As such, it comes endowed with a natural topology, which is the adic topology induced by the kernel J of this augmentation. Besides, there is a monomorphism of algebras which assigns every linearly recursive sequence $(s_n)_{n\geq 0}$ to the power series $\sum_{n\geq 0} (s_n/n!)\, Z^n$. Through this monomorphism, the algebra of linearly recursive sequences can be considered as a subalgebra of $\mathbb{C}[[Z]]$ and, as such, it inherits another natural topology, namely the one induced by the Krull topology on $\mathbb{C}[[Z]]$ (the adic topology induced by the unique maximal ideal $\mathfrak{m}$ of $\mathbb{C}[[Z]]$, which is also the augmentation ideal induced by the evaluation at 0).

These two topologies are very close. Namely, up to the embedding above, one can see that $J = \mathfrak{m} \cap \mathcal{L}in(\mathbb{C})$, so that the adic topology is finer than the induced one. A natural question which arises is if these are equivalent or not. Notice that, since the finiteness hypotheses are not fulfilled, Artin-Rees Lemma fails to be applied in this context. In fact, in this note we will compare these two topologies and we will give a negative answer to the previous question: the J-adic topology is strictly finer than the induced one.

As a by-product, we will see that the adic completion of $\mathcal{L}in(\mathbb{C})$ is larger than its completion with respect to the induced topology, which in fact can be identified with $\mathbb{C}[[Z]]$ itself, in the sense that we will provide a split surjective morphism

$$\widehat{\mathcal{L}in(\mathbb{C})} \to \mathbb{C}[[Z]].$$

Our approach will take advantage of the fact that, from an algebraic point of view, linearly recursive sequences may be identified with the finite (or continuous) dual of the algebra $\mathbb{C}[X]$ of polynomial functions of the additive affine algebraic $\mathbb{C}$-group, and that the formal power series can be considered as its full linear dual $\mathbb{C}[X]^*$.

The main motivation behind this note comes from studying the completion of the finite dual Hopf algebra of the universal enveloping algebra of a finite-dimensional complex Lie algebra.

2 The space of linearly recursive sequences and Hurwitz's product

We assume to work over the field $\mathbb{C}$ of complex numbers. However, it will be clear that this choice is not restrictive as the results will hold as well if we consider any algebraically closed field $\Bbbk$ of characteristic 0 instead. An augmented complex algebra A, is an algebra endowed with a morphism of algebras $A \to \mathbb{C}$, called the *augmentation*.

All vector spaces, algebras and coalgebras are assumed to be over $\mathbb{C}$. The unadorned tensor product $\otimes$ denotes the tensor product over $\mathbb{C}$. All maps are assumed to be at least $\mathbb{C}$-linear. For every vector space V, the $\mathbb{C}$-linear dual of V is $V^* = \mathrm{Hom}_{\mathbb{C}}(V, \mathbb{C})$ (i.e., the vector space of all linear forms from V to $\mathbb{C}$). Given a coalgebra C, for dealing with the comultiplication of an element $x \in C$ we will resort to the Sweedler's Sigma notation $\Delta(x) = \sum x_1 \otimes x_2$.

Consider the vector space $\mathbb{C}^{\mathbb{N}}$ of all sequences $(z_n)_{n\geq 0}$ of complex numbers. A sequence $(z_n)_{n\geq 0} \in \mathbb{C}^{\mathbb{N}}$ is said to be *linearly recursive* if there exists a family of constant coefficients $c_1, \ldots, c_r \in \mathbb{C}$, $r \geq 1$, such that

$$z_n = c_1 z_{n-1} + c_2 z_{n-2} + \cdots + c_r z_{n-r} \qquad \text{for all } n \geq r.$$

Denote by $\mathcal{L}in(\mathbb{C}) \subseteq \mathbb{C}^{\mathbb{N}}$ the vector subspace of all linearly recursive sequences.

Then the study of the algebraic and/or topological properties of the vector space $\mathcal{L}in(\mathbb{C})$ depends heavily on which product we are choosing on the vector space $\mathbb{C}^{\mathbb{N}}$, since the latter can be endowed with at least two algebra structures, as the subsequent Lemma 2.1 entails.

Lemma 2.1. *The assignment* $\Phi\colon \mathbb{C}^{\mathbb{N}} \to \mathbb{C}[X]^*$ *given by*

$$\big[\Phi\big((z_n)_{n\geq 0}\big)\big](X^m) = z_m \qquad \textit{for all } m \geq 0$$

is an isomorphism of vector spaces.

Next we recall how the vector space $\mathcal{L}in(\mathbb{C})$ can be endowed with a Hopf algebra structure, by using the Hurwitz's product. Recall first that $\mathbb{C}[X]$ is in fact a Hopf algebra, as it can be identified with the algebra of polynomial functions on the affine complex line $\mathbb{A}^1_{\mathbb{C}} = \mathbb{C}$, viewed as an algebraic group with the sum. Comultiplication, counit and antipode are the algebra morphisms induced by the assignments

$$\Delta(X) = X \otimes 1 + 1 \otimes X, \qquad \varepsilon(X) = 0, \qquad S(X) = -X.$$

From this it follows that $\mathbb{C}[X]^*$ is an augmented algebra under the *convolution product*

$$(f * g)\left(X^n\right) = \sum_{k=0}^{n} \binom{n}{k} f\left(X^k\right) g\left(X^{n-k}\right) \qquad \text{for all } n \geq 0. \tag{2.1}$$

The unit of $\mathbb{C}[X]^*$ is the counit ε of $\mathbb{C}[X]$. The augmentation ε_* is given by evaluation at 1.

As a consequence, the vector space $\mathbb{C}^{\mathbb{N}}$ turns out to be an augmented algebra as well in such a way that Φ becomes an algebra isomorphism. The product of this algebra is the so-called *Hurwitz's product*

$$(z_n)_{n\geq 0} * (u_n)_{n\geq 0} = \left(\sum_{k=0}^{n} \binom{n}{k} z_k\, u_{n-k}\right)_{n\geq 0}. \tag{2.2}$$

The unit is the sequence $(z_n)_{n\geq 0}$ with $z_0 = 1$ and $z_n = 0$ for all $n \geq 1$. The augmentation is given by the projection on the 0-th component. The vector space of all sequences $\mathbb{C}^{\mathbb{N}}$ endowed with this algebra structure will be denoted by $\mathcal{H}\mathbb{C}^{\mathbb{N}}$.

Recall also that given a Hopf algebra as $\mathbb{C}[X]$, we may consider its *finite dual*[1] Hopf algebra $\mathbb{C}[X]^\circ$. This is the vector subspace of all linear maps which vanish on a finite-codimensional ideal (i.e., one that leads to a finite-dimensional quotient algebra). Here we will focus only on the case of our interest and we refer to [4, Chapter 9] and [9, Chapter VI] for a more extended treatment.

[1] In the literature, it appears also under the names *Sweedler dual* or *continuous dual*, where continuity is with respect to the linear topology whose neighbourhood base at 0 consists exactly of the finite-codimensional ideals, see e.g. [6, §3].

Lemma 2.2. *Given the Hopf algebra $\mathbb{C}[X]$, the set*

$$\mathbb{C}[X]^\circ = \{f \in \mathbb{C}[X]^* \mid \ker(f) \supseteq I, \textit{ for } I \textit{ a non-zero ideal of } \mathbb{C}[X]\}$$

is an augmented subalgebra of $\mathbb{C}[X]^*$ which is also a Hopf algebra. The augmentation $\varepsilon_\circ$ is given by the restriction of ε_*. The comultiplication on $\mathbb{C}[X]^\circ$ is defined in such a way that for $f \in \mathbb{C}[X]^\circ$, we have

$$\Delta_\circ(f) = \sum f_1 \otimes f_2 \iff \Big(f(pq) = \sum f_1(p)f_2(q), \textit{ for all } p, q \in \mathbb{C}[X]\Big). \tag{2.3}$$

The antipode is given by pre-composing with the one of $\mathbb{C}[X]$, i.e., $S_\circ(f) = f \circ S$ for all $f \in \mathbb{C}[X]^\circ$.

Since the algebra $\mathbb{C}[X]$ is a principal ideal domain, it turns out that the space of linearly recursive sequences $\mathcal{L}in(\mathbb{C})$ can be identified with $\mathbb{C}[X]^\circ$ via the isomorphism Φ, whence it becomes an augmented subalgebra of $\mathcal{H}\mathbb{C}^{\mathbb{N}}$ and a Hopf algebra. For further reading, we refer the interested reader to [3, 6, 10] and [11, Chapter 2].

3 Two filtrations on the space of linearly recursive sequences

In this and in the next section, we will implicitly make use of the identifications $\mathcal{H}\mathbb{C}^{\mathbb{N}} = \mathbb{C}[X]^*$ and $\mathcal{L}in(\mathbb{C}) = \mathbb{C}[X]^\circ$, via the isomorphism of algebras Φ of Lemma 2.1.

As augmented algebras, both $\mathbb{C}[X]^*$ and $\mathbb{C}[X]^\circ$ inherit a natural filtration. Namely, if we let $I := \ker(\varepsilon_*)$ and $J := \ker(\varepsilon_\circ)$ be their *augmentation ideals*, then we can consider $\mathbb{C}[X]^*$ and $\mathbb{C}[X]^\circ$ as filtered with the adic filtrations $F_n\left(\mathbb{C}[X]^*\right) = I^n$ and $F_n\left(\mathbb{C}[X]^\circ\right) = J^n, n \geq 0$.

Moreover, $\mathbb{C}[X]^\circ$ inherits a filtration $F'_n\left(\mathbb{C}[X]^\circ\right) = I^n \cap \mathbb{C}[X]^\circ$ induced from the canonical inclusion $\mathbb{C}[X]^\circ \subseteq \mathbb{C}[X]^*$ as well and it is clear that $F_n\left(\mathbb{C}[X]^\circ\right) \subseteq F'_n\left(\mathbb{C}[X]^\circ\right)$. Hence, the J-adic filtration on $\mathbb{C}[X]^\circ$ is finer than the induced one. As we will show in this section, it is in fact strictly finer.

For every $\lambda \in \mathbb{C}$ we define $\phi_\lambda \colon \mathbb{C}[X] \to \mathbb{C}$ to be the algebra map such that $\phi_\lambda(X) = \lambda$. The set $G_a := \mathsf{Alg}_{\mathbb{C}}\left(\mathbb{C}[X], \mathbb{C}\right) = \{\phi_\lambda \mid \lambda \in \mathbb{C}\}$ is a group with group structure given by

$$\phi_\lambda \cdot \phi_{\lambda'} := \phi_\lambda * \phi_{\lambda'} = \phi_{\lambda+\lambda'}, \qquad e_{G_a} := \varepsilon = \phi_0, \qquad (\phi_\lambda)^{-1} := \phi_\lambda \circ S = \phi_{-\lambda}.$$

Lemma 3.1 ([4, Example 9.1.7])**.** *Denote by ξ the distinguished element in $\mathbb{C}[X]^*$ which satisfies $\xi(X^n) = \delta_{n,1}$ for all $n \geq 0$ (Kronecker's delta). Then the convolution product* (2.1) *induces an isomorphism of Hopf algebras*

$$\Psi \colon \mathbb{C}[\xi] \otimes \mathbb{C}G_a \longrightarrow \mathbb{C}[X]^\circ, \quad \Big(\xi^n \otimes \phi_\lambda \longmapsto \xi^n * \phi_\lambda\Big), \tag{3.1}$$

where $\mathbb{C}G_a$ is the group algebra on G_a and $\mathbb{C}[\xi]$ is a polynomial Hopf algebra as in Section 2.

We denote by

$$\vartheta \colon \mathbb{C}[\xi] \xhookrightarrow{\Psi} \mathbb{C}[X]^\circ \xhookrightarrow{\iota} \mathbb{C}[X]^* \tag{3.2}$$

the algebra monomorphism induced by Ψ.

Remark 3.2. It is worthy to point out that Lemma 3.1 is a particular instance of the renowned *Cartier-Gabriel-Kostant-Milnor-Moore Theorem*, which states that for a cocommutative Hopf $\Bbbk$-algebra H over an algebraically closed field $\Bbbk$ of characteristic zero, the multiplication in H induces an isomorphism of Hopf algebras $U\left(P\left(H\right)\right) \# \mathbb{C}G\left(H\right) \cong H$, where the left-hand side is endowed with the smash product algebra structure (see [4, Corollary 5.6.4 and Theorem 5.6.5], [9, Theorems 8.1.5, 13.0.1 and §13.1] and [8, Theorem 15.3.4]).

Denote by $\varepsilon_a\colon \mathbb{C}G_a \to \mathbb{C}$ the counit of the group algebra, which acts via $\varepsilon_a(\phi_\lambda) = \phi_\lambda(1) = 1$ for all $\lambda \in \mathbb{C}$, and by $\varepsilon_\xi\colon \mathbb{C}[\xi] \to \mathbb{C}$ the counit of the polynomial algebra in ξ defined by $\varepsilon_\xi(\xi) = 0$. These maps are in fact the restrictions of the counit $\varepsilon_\circ\colon \mathbb{C}[X]^\circ \to \mathbb{C}$ to the vector subspaces of $\mathbb{C}[X]^\circ$ generated by G_a and $\{\xi^n \mid n \geq 0\}$, respectively. Thus, up to the isomorphism Ψ of equation (3.1), we have $\varepsilon_\circ = \varepsilon_\xi \otimes \varepsilon_a$.

Lemma 3.3. *The isomorphism* Ψ *of* (3.1) *induces an isomorphism of vector spaces*

$$\mathbb{C}\bar{\xi} \oplus \frac{\ker(\varepsilon_a)}{\ker(\varepsilon_a)^2} \cong \frac{J}{J^2},$$

where $\bar{\xi} = \xi + \langle \xi^2 \rangle$ *in the quotient* $\langle \xi \rangle / \langle \xi^2 \rangle$.

Proof. First of all, as Ψ is an isomorphism of Hopf algebras, it induces an isomorphism of vector spaces between J/J^2 and $\ker(\varepsilon_\xi \otimes \varepsilon_a)/\ker(\varepsilon_\xi \otimes \varepsilon_a)^2$. Set $K := \ker(\varepsilon_\xi \otimes \varepsilon_a)$. The family of assignments

$$\begin{aligned} \frac{\langle \xi^k \rangle}{\langle \xi^{k+1} \rangle} \otimes \frac{\ker(\varepsilon_a)^h}{\ker(\varepsilon_a)^{h+1}} &\longrightarrow \frac{K^n}{K^{n+1}} \\ \left(\xi^k + \langle \xi^{k+1} \rangle\right) \otimes \left(x + \ker(\varepsilon_a)^{h+1}\right) &\longmapsto \left(\xi^k \otimes x\right) + K^{n+1} \end{aligned}$$

for $h, k \geq 0$ and $n = h + k$ induces a graded isomorphism of graded vector spaces

$$\mathsf{gr}(\mathbb{C}[\xi]) \otimes \mathsf{gr}(\mathbb{C}G_a) \cong \mathsf{gr}\big(\mathbb{C}[\xi] \otimes \mathbb{C}G_a\big),$$

see e.g. [5, Lemma VIII.2]. In particular, the degree 1 component of this together with Ψ induce the stated isomorphism

$$\mathbb{C}\bar{\xi} \oplus \frac{\ker(\varepsilon_a)}{\ker(\varepsilon_a)^2} \cong \left(\frac{\langle \xi \rangle}{\langle \xi^2 \rangle} \otimes \frac{\mathbb{C}G_a}{\ker(\varepsilon_a)}\right) \oplus \left(\frac{\mathbb{C}[\xi]}{\langle \xi \rangle} \otimes \frac{\ker(\varepsilon_a)}{\ker(\varepsilon_a)^2}\right) \cong \frac{K}{K^2} \cong \frac{J}{J^2}. \qquad \square$$

The key fact is that the quotient $\ker(\varepsilon_a)/\ker(\varepsilon_a)^2$ does not vanish, as we will show in the subsequent lemma. To this aim, recall that there is an algebra isomorphism

$$\Theta\colon \mathbb{C}[X]^* \longrightarrow \mathbb{C}[[Z]], \quad \Big(f \longmapsto \sum_{k\geq 0} f(e_k) Z^k\Big), \tag{3.3}$$

where $e_k = X^k/k!$ for all $k \geq 0$. Notice that $\Theta \circ \vartheta(\xi) = Z$, where ϑ is the morphism given in (3.2).

Lemma 3.4. *The element* $\phi_1 - \varepsilon + \ker(\varepsilon_a)^2$ *in the quotient* $\ker(\varepsilon_a)/\ker(\varepsilon_a)^2$ *is non-zero.*

Proof. Assume by contradiction that $\phi_1 - \varepsilon \in \ker(\varepsilon_a)^2$. By applying Ψ, this implies that $\phi_1 - \varepsilon \in J^2$, whence $\phi_1 - \varepsilon \in I^2$ in $\mathbb{C}[X]^*$. Since Θ induces a bijection between I^n and $\langle Z^n \rangle \subseteq \mathbb{C}[[Z]]$ for all $n \geq 1$, claiming that $\phi_1 - \varepsilon \in I^2$ in $\mathbb{C}[X]^*$ would imply that $\sum_{k\geq 1} Z^k/k! \in \langle Z^2 \rangle$, which is a contradiction. Thus, $\phi_1 - \varepsilon \notin \ker(\varepsilon_a)^2$. □

It follows from Lemma 3.3 and Lemma 3.4 that the elements $\xi + J^2$ and $\phi_1 - \varepsilon + J^2$ are linearly independent in J/J^2. In particular, $\phi_1 - \varepsilon - \xi \notin J^2$. However, $\phi_1 - \varepsilon - \xi$ as an element of $\mathbb{C}[X]^*$ maps $e_0 = 1$ and $e_1 = X$ to 0 and it maps $e_n = X^n/n!$ to $1/n!$ for all $n \geq 2$. Hence $\phi_1 - \varepsilon - \xi = \xi^2 * h_{(2)} \in I^2$, where for every $k \geq 0$

$$h_{(k)}(e_n) := \frac{1}{(n+k)!} \quad \text{for all } n \geq 0.$$

Indeed,

$$\left(\xi^2 * h_{(2)}\right)(e_n) = \sum_{i+j=n} \xi^2(e_i) h_{(2)}(e_j) = \begin{cases} 0 & n = 0,1 \\ \frac{1}{n!} & n \geq 2 \end{cases}$$

This shows that $\phi_1 - \varepsilon - \xi$ is an element in $\mathbb{C}[X]^\circ \cap I^2$ but not in J^2, so that $J^2 \subsetneq \mathbb{C}[X]^\circ \cap I^2$. Now, by induction one may see that for every $n \geq 1$ the element

$$\phi_1 - \left(\sum_{k=0}^{n-1} \frac{1}{k!} \xi^k \right) = \xi^n * h_{(n)} \in I^n \cap \mathbb{C}[X]^\circ \tag{3.4}$$

does not belong to J^n, so that the two filtrations do not coincide. We point out that, under the isomorphism $\mathbb{C}[X]^* \cong \mathbb{C}[[Z]]$ of equation (3.3), the element of equation (3.4) corresponds to

$$\mathsf{exp}(Z) - \left(\sum_{k=0}^{n-1} \frac{1}{k!} Z^k \right) = Z^n \cdot \left(\sum_{k\geq 0} \frac{1}{(n+k)!} Z^k \right).$$

Summing up, we have shown that the J-adic filtration on $\mathbb{C}[X]^\circ$ is strictly finer than the filtration induced from the inclusion $\iota \colon \mathbb{C}[X]^\circ \to \mathbb{C}[X]^*$.

4 Comparing the two topologies on $\mathbb{C}[X]^\circ$

Recall that a filtration on an algebra naturally induces on it a linear topology, whose neighbourhood base at 0 is given exactly by the elements of the filtration (see for example [1, III.49, Example 3] or [5, §I, Chapter D]). Furthermore, given an algebra A endowed with the $\mathfrak{m}$-adic filtration associated to an ideal $\mathfrak{m} \subseteq A$, the *completion* of A with respect to the linear topology induced by this filtration is, by definition, $\widehat{A} = \varprojlim_n (A/\mathfrak{m}^n)$, i.e., the projective limit of the projective system $A/\mathfrak{m}^n$ with the obvious projection maps $A/\mathfrak{m}^n \twoheadrightarrow A/\mathfrak{m}^m$ for $n \geq m$. An algebra A is said to be *Hausdorff and complete* if the canonical map $A \to \widehat{A}$ is an isomorphism. For further details, we refer to [5, §II, Chapter D].

Example 4.1. For every $n \geq 0$, there is a linear isomorphism between $\mathbb{C}[X]^*/I^{n+1}$ and the linear dual of the vector subspace $\mathbb{C}[X]_{\leq n} \subseteq \mathbb{C}[X]$ of all polynomials of degree up to n. These in turn induce an isomorphism

$$\widehat{\mathbb{C}[X]^*} = \varprojlim_n \left(\frac{\mathbb{C}[X]^*}{I^{n+1}} \right) \cong \mathbb{C}[X]^* \tag{4.1}$$

by which we conclude that $\mathbb{C}[X]^*$ is complete with respect to the I-adic topology.

Remark 4.2. Let us consider again the algebra monomorphism $\vartheta\colon \mathbb{C}[\xi] \to \mathbb{C}[X]^*$ of equation (3.2). Since $\vartheta(\xi) \in \ker(\varepsilon_*) = I$, we have that ϑ is a filtered morphism of filtered algebras and so we may consider its completion $\widehat{\vartheta} : \widehat{\mathbb{C}[\xi]} \to \widehat{\mathbb{C}[X]^*} \cong \mathbb{C}[X]^*$. Therefore, up to the canonical identification $\mathbb{C}[[\xi]] = \mathbb{C}[[Z]]$, the map $\widehat{\vartheta}$ turns out to be the inverse of Θ. A useful consequence of this is that every element $g \in \mathbb{C}[X]^*$ can be written as

$$g = \sum_{k \geq 0} g(e_k)\xi^k, \tag{4.2}$$

where as before $e_k = X^k/k!$ for all $k \geq 0$. By the right-hand side of equation (4.2), we mean the image in $\mathbb{C}[X]^*$ of the element

$$\left(\sum_{k=0}^{n} g(e_k)\xi^k + I^{n+1}\right)_{n \geq 0} = \lim_{n \to \infty} \left(\sum_{k=0}^{n} g(e_k)\xi^k\right)$$

via the isomorphism (4.1). Since $\xi^i\,(e_j) = \delta_{i,j}$ for all $i, j \geq 0$, given any $p = \sum_{i=0}^{t} p_i e_i \in \mathbb{C}[X]$ the sequence $\left(\sum_{k=0}^{n} g(e_k)\xi^k\right)(p)$, $n \geq 0$, eventually becomes constant and it equals the element $\sum_{i=0}^{t} p_i g(e_i) = g(p)$. In light of this interpretation, $I^n = \langle \xi^n \rangle$ for all $n \geq 0$, in the algebra $\mathbb{C}[X]^*$.

We already know from Section 3 that the J-adic filtration on $\mathbb{C}[X]^\circ$ does not coincide with the one induced by the inclusion $\mathbb{C}[X]^\circ \subseteq \mathbb{C}[X]^*$. Nevertheless, the topologies they induce may still be equivalent ones (that is, the two filtrations may be equivalent). Our next aim is to show that these topologies are not even equivalent, by showing that the J-adic completion of $\mathbb{C}[X]^\circ$ is not homeomorphic to $\mathbb{C}[X]^*$ via the completion of the inclusion map $\iota\colon \mathbb{C}[X]^\circ \to \mathbb{C}[X]^*$.

Remark 4.3. It is worthy to mention that $\mathbb{C}[X]^\circ$ is dense in $\mathbb{C}[X]^*$ with respect to the finite topology on $\mathbb{C}[X]^*$ (the one induced by the product topology on $\mathbb{C}^{\mathbb{C}[X]}$), see for instance [2, Exercise 1.5.21]. On the other hand, since for every $f \in \mathbb{C}[X]^*$ and for all $n \geq 0$, we have that $f + \langle \xi^n \rangle = \mathcal{O}\left(f; e_0, e_1, \ldots, e_{n-1}\right)$, the space of linear maps which coincide with f on $e_0, e_1, \ldots, e_{n-1}$, it turns out that the I-adic topology on $\mathbb{C}[X]^*$ is coarser then the linear one. It follows then that $\mathbb{C}[X]^\circ \subseteq \mathbb{C}[X]^*$ is dense with respect to the I-adic topology as well and hence one may check that

$$\varprojlim_{n} \left(\frac{\mathbb{C}[X]^\circ}{\mathbb{C}[X]^\circ \cap I^n}\right) \cong \varprojlim_{n} \left(\frac{\mathbb{C}[X]^*}{I^n}\right) \cong \mathbb{C}[X]^*.$$

Now, consider the completion $\widehat{\psi}\colon \mathbb{C}[[\xi]] \to \widehat{\mathbb{C}[X]^\circ}$, where ψ is the filtered monomorphism of algebras given in (3.2). In view of Remark 4.2, one shows that $\widehat{\iota} \circ \widehat{\psi} = \widehat{\vartheta}$. Therefore, $\widehat{\iota}$ is a split epimorphism, as $\widehat{\vartheta}$ is an homeomorphism whose inverse is Θ.

Remark 4.4. In fact $\mathbb{C}[X]^*$ is a complete Hopf algebra in the sense of [7, Appendix A] and $\widehat{\iota}\colon \widehat{\mathbb{C}[X]^\circ} \to \mathbb{C}[X]^*$ becomes an effective epimorphism of complete Hopf algebras (see [7, Proposition 2.19, page 274]).

The subsequent proposition gives conditions under which $\widehat{\iota}$ becomes an homeomorphism.

Proposition 4.5. *The following assertions are equivalent:*

(1) the canonical map $\widehat{\iota}\colon \widehat{\mathbb{C}[X]^\circ} \to \mathbb{C}[X]^*$ *is injective;*

(2) the canonical map $\widehat{\iota}\colon \widehat{\mathbb{C}[X]^\circ} \to \mathbb{C}[X]^*$ *is an homeomorphism;*

(3) the J*-adic and the induced filtrations on* $\mathbb{C}[X]^\circ$ *coincide;*

(4) the J*-adic and the induced topologies on* $\mathbb{C}[X]^\circ$ *are equivalent.*

Proof. We already observed that $\widehat{\psi} \circ \Theta$ is a continuous section of $\widehat{\iota}$. Thus, if $\widehat{\iota}$ injective then it will be bijective with inverse $\widehat{\psi} \circ \Theta$, and so an homeomorphism. This proves the implication (1) ⇒ (2). To show that (2) ⇒ (3), let us denote by

$$F_n\left(\widehat{\mathbb{C}[X]^\circ}\right) = \ker\left(\widehat{\mathbb{C}[X]^\circ} \to \mathbb{C}[X]^\circ/J^n\right)$$

the canonical filtration on $\widehat{\mathbb{C}[X]^\circ}$. If $\widehat{\iota}$ is an homeomorphism, then its inverse has to be $\widehat{\psi} \circ \Theta$. As a consequence, we obtain the second of the following chain of isomorphisms

$$\frac{\mathbb{C}[X]^\circ}{J^n} \cong \frac{\widehat{\mathbb{C}[X]^\circ}}{F_n\left(\widehat{\mathbb{C}[X]^\circ}\right)} \cong \frac{\mathbb{C}[X]^*}{I^n},$$

for every $n \geq 1$. Their composition sends

$$f + J^n \in \mathbb{C}[X]^\circ/J^n \quad \text{to} \quad \iota(f) + I^n \in \mathbb{C}[X]^*/I^n,$$

which shows that $J^n = I^n \cap \mathbb{C}[X]^\circ$. Thus the two filtrations coincide. Since the implication (3) ⇒ (4) is clear, let us show that (4) ⇒ (1). Saying that the two topologies are equivalent, implies that every J^n (which is open in the J-adic topology) has to be open in the induced topology as well. In particular, it has to contain an element of the neighbourhood base of 0. Therefore, we may assume that for every $n \geq 0$, there exists $m \geq n$ such that $I^m \cap \mathbb{C}[X]^\circ \subseteq J^n$. Given $(f_n + J^n)_{n\geq 0}$ an element in the kernel of $\widehat{\iota}$, we have that $f_n \in I^n$ for every $n \geq 0$. This implies that for every $n \geq 0$, there exists $m \geq n$ such that

$$f_n + J^n = f_m + J^n \in (I^m \cap \mathbb{C}[X]^\circ) + J^n = J^n,$$

which means that $(f_n + J^n)_{n\geq 0} = 0$ and this settles the proof. □

In conclusion, it follows from the result of Section 3 that none of the equivalent conditions in Proposition 4.5 holds, as the two filtrations do not coincide. An explicit non-zero element which lies in the kernel of $\widehat{\iota}$ is exactly the one coming from equation (3.4). Indeed, on the one hand

$$\left(\phi_1 - \sum_{k=0}^{n} \frac{1}{k!}\xi^k + J^{n+1}\right)_{n\geq 0} \in \widehat{\mathbb{C}[X]^\circ}$$

is non-zero, but on the other hand a direct check shows that

$$\widehat{\iota}\left(\left(\phi_1 - \sum_{k=0}^{n} \frac{1}{k!}\xi^k + J^{n+1}\right)_{n\geq 0}\right) = \left(\phi_1 - \sum_{k=0}^{n} \frac{1}{k!}\xi^k + I^{n+1}\right)_{n\geq 0} = 0$$

in $\mathbb{C}[X]^*$.

Remark 4.6. Observe that an element $\left(f_n + J^{n+1}\right)_{n\geq 0}$ in $\widehat{\mathbb{C}[X]^\circ}$ can be considered as the formal limit $\lim_{n\to\infty}(f_n)$ of the Cauchy sequence $\{f_n \mid n \geq 0\}$ in $\mathbb{C}[X]^\circ$ with the J-adic topology. The element $\left(\phi_1 + J^{n+1}\right)_{n\geq 0}$ can be identified with ϕ_1 itself, as limit of a constant sequence. On the other hand the element $\left(\sum_{k=0}^{n}\xi^k/k! + J^{n+1}\right)_{n\geq 0}$ can be considered as the limit $\lim_{n\to\infty}\left(\sum_{k=0}^{n}\xi^k/k!\right)$. As we already noticed, ϕ_1 is associated with the exponential function, in the sense that its power series expansion in $\mathbb{C}[X]^*$ is $\sum_{k\geq 0}\xi^k/k! = \mathsf{exp}(\xi)$. However, it follows from what we showed that in $\widehat{\mathbb{C}[X]^\circ}$ the Cauchy sequence $\left\{\sum_{k=0}^{n}\xi^k/k! \mid n \geq 0\right\}$ does not converge to ϕ_1.

5 Final remarks

As we mentioned in the introduction, linearly recursive sequences have already been studied deeply as "rational" power series. What we plan to do in this section is to provide a possible explanation of why the topological richness expounded in the previous sections didn't enter the picture before and to provide an overview of the different interpretations of these sequences.

The commutative diagram of algebras in (5.1) summarizes the state of the art. Therein, $\mathbb{C}^{\mathbb{N}}$ is endowed with the algebra structure given by the product $(x_n)_{n\geq 0}\,(y_n)_{n\geq 0} = \left(\sum_{k=0}^{n} x_k y_{n-k}\right)_{n\geq 0}$.

$$\begin{tikzcd}
\mathcal{L}in(\mathbb{C}) \arrow[rr, "\cong"] \arrow[dr, hook] & & \mathbb{C}[X]^\circ \arrow[dr, hook, "\iota"] & \\
& \mathcal{H}\mathbb{C}^{\mathbb{N}} \arrow[rr, "\Phi"] \arrow[d, "\zeta"', "\cong"] & & \mathbb{C}[X]^* \arrow[d, "\Theta", "\cong"'] \\
& \mathbb{C}^{\mathbb{N}} \arrow[rr, "\Omega"] & & \mathbb{C}[[Z]] \\
\mathcal{L}in(\mathbb{C}) \arrow[ur, hook] \arrow[rr, "\omega", "\cong"'] & & \mathbb{C}[Z]_{\langle Z\rangle} \arrow[ur, hook] &
\end{tikzcd} \tag{5.1}$$

The isomorphism Ω sends any sequence $(x_n)_{n\geq 0}$ to the power series $\sum_{n\geq 0} x_n Z^n$, while the isomorphism ζ sends a sequence $(z_n)_{n\geq 0}$ to the sequence $(z_n/n!)_{n\geq 0}$. The algebra $\mathbb{C}[Z]_{\langle Z\rangle}$ denotes the localization of $\mathbb{C}[Z]$ at the maximal ideal $\langle Z\rangle$, that is, the set of fractions $p(Z)/q(Z)$ with $q(0) \neq 0$. Lastly, the isomorphism ω is induced by the restriction of Ω to $\mathcal{L}in(\mathbb{C})$ and it is given as follows. For a sequence $(a_n)_{n\geq 0}$ in $\mathcal{L}in(\mathbb{C})$, let $c_r = 1, c_{r-1}, \ldots, c_0 \in \mathbb{C}$, $r \geq 1$ be the family of constant coefficients such that

$$a_{l+r} + c_{r-1}a_{l+r-1} + c_{r-2}a_{l+r-2} + \cdots + c_0 a_l = 0, \qquad \text{for all } l \geq 0.$$

If we consider

$$q(Z) = \sum_{i=0}^{r} c_{r-i} Z^i \quad \text{and} \quad p(Z) = \sum_{j=0}^{r-1}\left(\sum_{i=0}^{j} c_{r-i} a_{j-i}\right) Z^j,$$

then

$$q(Z)\left(\sum_{n\geq 0} a_n Z^n\right) = p(Z).$$

Thus, ω acts via

$$\omega((a_n)_{n\geq 0}) := p(Z)/q(Z) \in \mathbb{C}[Z]_{\langle Z\rangle}.$$

As one can realize from diagram (5.1), there are essentially two linear topologies which can be induced on $\mathcal{L}in(\mathbb{C})$: one from $\mathcal{H}\mathbb{C}^{\mathbb{N}}$, which we denote by $\mathcal{T}_{\mathcal{H}}$, and the other from $\mathbb{C}^{\mathbb{N}}$, which we denote by $\mathcal{T}$. Apart from these, $\mathcal{L}in(\mathbb{C})$ has its own two adic topologies given by the ideals $\mathcal{I} := \mathcal{L}in(\mathbb{C}) \cap \mathfrak{a}$, where $\mathfrak{a}$ is the augmentation ideal of $\mathcal{H}\mathbb{C}^{\mathbb{N}}$, and $\mathcal{J} := \mathcal{L}in(\mathbb{C}) \cap \mathfrak{b}$, where $\mathfrak{b}$ is the augmentation ideal of $\mathbb{C}^{\mathbb{N}}$.

It follows from the definitions that the $\mathcal{I}$-adic topology on $\mathcal{L}in(\mathbb{C})$ is finer than $\mathcal{T}_{\mathcal{H}}$ and the $\mathcal{J}$-adic one is finer than $\mathcal{T}$. On the one hand, in view of the previous sections, the $\mathcal{I}$-adic topology is in fact strictly finer than $\mathcal{T}_{\mathcal{H}}$. On the other hand, however, one can show that the $\mathcal{J}$-adic topology turns out to be equivalent to $\mathcal{T}$, since it is known that $\widehat{\mathbb{C}[Z]_{\langle Z\rangle}}$ is homeomorphic to $\widehat{\mathbb{C}[Z]} \cong \mathbb{C}[[Z]]$, and this may be the reason why topologies on $\mathcal{L}in(\mathbb{C})$ weren't analysed before.

Finally, comparing the topologies $\mathcal{T}$ and $\mathcal{T}_{\mathcal{H}}$ on $\mathcal{L}in(\mathbb{C})$ seems to be more involved. Apparently it is possible that these are different. However, it is not clear to us how to show, for instance, that any open neighbourhood of the form $\mathcal{L}in(\mathbb{C}) \cap \mathfrak{a}^n$ (the product is that of $\mathbb{C}^{\mathbb{N}}$) is not contained in some open $\mathcal{L}in(\mathbb{C}) \cap \mathfrak{b}^m$ (the product now is in $\mathcal{H}\mathbb{C}^{\mathbb{N}}$). What is clear instead is that the isomorphism ζ does not map linearly recursive sequences in $\mathcal{H}\mathbb{C}^{\mathbb{N}}$ to linearly recursive sequences in $\mathbb{C}^{\mathbb{N}}$.

References

[1] N. Bourbaki, *Topologie générale: Chapitres 1 à 4*, Éléments de mathématique, Hermann, Paris, 1971.

[2] S. Dăscălescu, C. Năstăsescu and Ş. Raianu, *Hopf Algebras: An Introduction*, volume 235 of *Monographs and Textbooks in Pure and Applied Mathematics*, Marcel Dekker, New York, 2001.

[3] C. A. Futia, E. F. Müller and E. J. Taft, Bialgebras of recursive sequences and combinatorial identities, *Adv. Appl. Math.* **28** (2002), 203–230, doi:10.1006/aama.2001.0778.

[4] S. Montgomery, *Hopf Algebras and Their Actions on Rings*, volume 82 of *CBMS Regional Conference Series in Mathematics*, American Mathematical Society, Providence, Rhode Island, 1993, doi:10.1090/cbms/082.

[5] C. Năstăsescu and F. van Oystaeyen, *Graded Ring Theory*, volume 28 of *North-Holland Mathematical Library*, North-Holland, Amsterdam, 1982.

[6] B. Peterson and E. J. Taft, The Hopf algebra of linearly recursive sequences, *Aequationes Math.* **20** (1980), 1–17, doi:10.1007/bf02190488.

[7] D. Quillen, Rational homotopy theory, *Ann. Math.* **90** (1969), 205–295, doi:10.2307/1970725.

[8] D. E. Radford, *Hopf Algebras*, volume 49 of *Series on Knots and Everything*, World Scientific, Hackensack, New Jersey, 2012, doi:10.1142/8055.

[9] M. E. Sweedler, *Hopf Algebras*, Mathematics Lecture Note Series, W. A. Benjamin, New York, 1969.

[10] E. J. Taft, Algebraic aspects of linearly recursive sequences, in: J. Bergen and S. Montgomery (eds.), *Advances in Hopf Algebras*, Marcel Dekker, New York, volume 158 of *Lecture Notes in Pure and Applied Mathematics*, 1994 pp. 299–317, papers from the NSF-CBMS Conference on Hopf Algebras and their Actions on Rings held at DePaul University, Chicago, Illinois, August 10 – 14, 1992.

[11] R. G. Underwood, *Fundamentals of Hopf Algebras*, Universitext, Springer, Cham, 2015, doi: 10.1007/978-3-319-18991-8.

[12] A. J. van der Poorten, Some facts that should be better known, especially about rational functions, in: R. A. Mollin (ed.), *Number Theory and Applications*, Kluwer Academic Publishers, Dordrecht, volume 265 of *NATO Advanced Science Institutes Series C: Mathematical and Physical Sciences*, 1989 pp. 497–528, proceedings of the NATO Advanced Study Institute held in Banff, Alberta, April 27 – May 5, 1988.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 331–348
https://doi.org/10.26493/1855-3974.1427.75c
(Also available at http://amc-journal.eu)

Intrinsic linking with linking numbers of specified divisibility

Christopher Tuffley
School of Fundamental Sciences, Massey University,
Private Bag 11 222, Palmerston North 4442, New Zealand

Received 20 June 2017, accepted 29 November 2018, published online 16 January 2019

Abstract

Let n, q and r be positive integers, and let K_N^n be the n-skeleton of an $(N-1)$-simplex. We show that for N sufficiently large every embedding of K_N^n in $\mathbb{R}^{2n+1}$ contains a link consisting of r disjoint n-spheres, such that every pairwise linking number is a nonzero multiple of q. This result is new in the classical case $n = 1$ (graphs embedded in $\mathbb{R}^3$) as well as the higher dimensional cases $n \geq 2$; and since it implies the existence of an r-component link with all pairwise linking numbers at least q in absolute value, it also extends a result of Flapan et al. from $n = 1$ to higher dimensions. Additionally, for $r = 2$ we obtain an improved upper bound on the number of vertices required to force a two-component link with linking number a nonzero multiple of q. Our new bound has growth $O(nq^2)$, in contrast to the previous bound of growth $O(\sqrt{n}4^n q^{n+2})$.

Keywords: Intrinsic linking, complete n-complex, Ramsey theory.

Math. Subj. Class.: 57Q45, 57M25, 57M15

1 Introduction

In the early 1980s Sachs [11] and Conway and Gordon [1] proved that every embedding of the complete graph K_6 in $\mathbb{R}^3$ contains a pair of disjoint cycles that form a nontrivial link, and Conway and Gordon also showed that every embedding of K_7 in $\mathbb{R}^3$ contains a nontrivial knot. These facts are expressed by saying that K_6 is *intrinsically linked*, and K_7 is *intrinsically knotted.* Since then, a number of authors have shown that embeddings of larger complete graphs necessarily exhibit more complex linking behaviour, such as non-split many-component links [4, 6]; two component links with linking number large in absolute value [2]; and two component links with linking number a nonzero multiple of a given integer [5, 6]. Embeddings of larger complete graphs must also exhibit more

E-mail address: c.tuffley@massey.ac.nz (Christopher Tuffley)

complicated knotting behaviour, such as knots with second Conway co-efficient large in absolute value [2].

Such *Ramsey-type results* for intrinsic linking can also be shown to hold in higher dimensions. Let K_N^n be the n-skeleton of an $(N-1)$-simplex, which we call the *complete n-complex on N vertices*. Then K_{2n+4}^n is intrinsically linked, in the sense that every embedding in $\mathbb{R}^{2n+1}$ contains a pair of disjoint n-spheres that have nonzero linking number [10, 12]; and moreover, the linking results described above can all be extended to embeddings of sufficiently large complete n-complexes in $\mathbb{R}^{2n+1}$ [13].

Flapan, Mellor and Naimi [3, Theorem 1] have shown that intrinsic linking of graphs is arbitrarily complex, in the following sense: Given positive integers r and α, every embedding of a sufficiently large complete graph in $\mathbb{R}^3$ contains an r-component link in which the linking number of each pair of components is at least α in absolute value. The main goal of this paper is to prove an analogue of this result in all dimensions, with the condition on the magnitude of the linking numbers replaced by a divisibility condition instead. Namely, we show that, given positive integers r and q, every embedding of a sufficiently large complete n-complex in $\mathbb{R}^{2n+1}$ contains a link consisting of r disjoint n-spheres, in which all pairwise linking numbers are nonzero multiples of q.

This result is new in the classical case $n = 1$ as well as the higher dimensional cases $n \geq 2$. Since a nonzero multiple of q has magnitude at least q, it also extends the Flapan-Mellor-Naimi result to $n \geq 2$. The techniques used to prove it draw heavily on those of Flapan, Mellor and Naimi (for the construction of many-component links with all pairwise linking numbers nonzero), as well as those of our previous paper [13] (for intrinsic linking with $n \geq 2$, and constructing links with linking numbers divisible by q). By refining a technique from [13] we also obtain a vastly improved upper bound on the number of vertices required in the case $r = 2$. Our new bound has growth $O(nq^2)$, in contrast to the previous best bound [13, Theorem 1.4] of growth $O(\sqrt{n}4^n q^{n+2})$.

We note that Flapan, Mellor and Naimi [3, Theorem 2] further show that intrinsic linking of complete graphs is arbitrarily complex in an even stronger sense: one can additionally require that the second co-efficient of the Conway polynomial of each component has absolute value at least α as well. As an integral measure of the complexity of a knot, the second Conway co-efficient may be regarded as the natural analogue of the pairwise linking number, viewed as an integral measure of the complexity of a two-component link. By Hoste [7, Lemma 2.1(i)] the Conway polynomial $\nabla_{\mathcal{L}}(z)$ of an oriented r-component link $\mathcal{L} = K_1 \cup K_2 \cup \cdots \cup K_r$ has the form

$$\nabla_{\mathcal{L}}(z) = z^{r-1}[a_0 + a_1 z^2 + \cdots + a_m z^{2m}],$$

and by the second Conway co-efficient we mean the co-efficient a_1. When $\mathcal{L} = K_1$ is a knot we have $a_0 = 1$ (Kauffman [9, Proposition 4.1], or see Hoste [7, Lemma 2.1(iii)]), so a_1 is the first nontrivial co-efficient of $\nabla_{\mathcal{L}}(z)$; and when $\mathcal{L} = K_1 \cup K_2$ is a two-component link we have $a_0 = \ell k(K_1, K_2)$ (Hoste [7, Lemma 2.1(iv)]), so here it is the linking number that is the first nontrivial co-efficient of $\nabla_{\mathcal{L}}(z)$. Moreover, for a knot K the mod two reduction of a_1 is equal to the Arf invariant of K (Kauffman [9, Section 4(a)], or see [7, Lemma 2.1(iii)]), so the linking number and the second Conway co-efficient may both be regarded as integral lifts of the mod two invariants used to establish the first results in intrinsic knotting and linking: the intrinsic linking of K_6 is proved by considering a sum of pairwise linking numbers mod two, and the intrinsic knotting of K_7 is proved by considering the sum of the Arf invariants of the Hamiltonian cycles in an embedding of K_7

in $\mathbb{R}^3$ [1].

We do not consider knotting of the components in this paper. This is chiefly for reasons of dimension: knotting of n-spheres occurs in $\mathbb{R}^{n+2}$, whereas linking of n-spheres occurs in $\mathbb{R}^{2n+1}$, so the only dimension in which we can consider intrinsic knotting and linking of n-complexes simultaneously is the classical case $n = 1$. We have not given this case separate consideration, instead giving uniform arguments that work for all n. To our knowledge there are at present no known divisibility results for intrinsic knotting, and we pose the following question:

Question 1.1. Let $q \geq 2$ be a positive integer. Does there exist N such that every embedding of K_N in $\mathbb{R}^3$ contains a knot with second Conway co-efficient a nonzero multiple of q?

Hoste [8] shows that the first Conway co-efficient a_0 of an r-component oriented link $\mathcal{L}$ is equal to any cofactor of a certain matrix of pairwise linking numbers associated with $\mathcal{L}$. It then follows from Theorem 1.3 below that for N sufficiently large every embedding of K_N^n in $\mathbb{R}^{2n+1}$ contains a non-split r-component link satisfying $a_0 \equiv 0 \pmod{q}$. As a strengthening of Theorem 1.3, we might additionally ask that a_0 be nonzero, motivating the following question:

Question 1.2. Let n, q and r be positive integers, with $q \geq 2$ and $r \geq 3$. Does there exist N such that every embedding of K_N^n in $\mathbb{R}^{2n+1}$ contains an r-component link with first Conway co-efficient a nonzero multiple of q?

We conjecture that the answer to both questions above is yes.

1.1 Statement of results

Throughout this paper, an *r-component link* means r disjoint oriented n-spheres embedded in $\mathbb{R}^{2n+1}$. Given a 2-component link $L_1 \cup L_2$ we will write $\ell k(L_1, L_2)$ for their linking number, and $\ell k_2(L_1, L_2)$ for their linking number mod two. For $\{i, j\} = \{1, 2\}$ the integral linking number is given by the homology class $[L_i]$ in $H_n(\mathbb{R}^{2n+1} - L_j; \mathbb{Z}) \cong \mathbb{Z}$.

Our main result is as follows:

Theorem 1.3. *Let n, q and r be positive integers, with $r \geq 2$. For N sufficiently large every embedding of K_N^n in $\mathbb{R}^{2n+1}$ contains an r-component link $L_1 \cup \cdots \cup L_r$ such that, for every $i \neq j$, $\ell k(L_i, L_j)$ is a nonzero multiple of q.*

Since every nonzero multiple of q has absolute value at least q, Theorem 1.3 immediately gives us the following extension of Theorem 1 of Flapan et al. [3] to higher dimensions:

Corollary 1.4. *Let n, λ and r be positive integers, with $r \geq 2$. For N sufficiently large every embedding of K_N^n in $\mathbb{R}^{2n+1}$ contains an r-component link $L_1 \cup \cdots \cup L_r$ such that, for every $i \neq j$, $|\ell k(L_i, L_j)| \geq \lambda$.*

The $r = 2$ case of Theorem 1.3 is proved as Theorem 1.4 of [13], with an upper bound of growth $O(\sqrt{n}4^n q^{n+2})$ on the number of vertices required. We re-prove this result with a greatly improved bound with growth $O(nq^2)$:

Theorem 1.5. *For $r = 2$, the conclusion of Theorem 1.3 holds for*

$$N \geq \kappa_n(q) = \begin{cases} 24q^2, & n = 1, \\ 4q^2(2n+4) + n + \left\lceil \frac{4q^2-2}{n} \right\rceil + 1, & n \geq 2. \end{cases}$$

In other words, every embedding of $K^n_{\kappa_n(q)}$ in $\mathbb{R}^{2n+1}$ contains a two component link $L_1 \cup L_2$ such that the linking number $\ell k(L_1, L_2)$ is a nonzero multiple of q.

We note that the bound of Theorem 1.5 is equal to the best known upper bound on the number of vertices required to force the existence of a generalised key ring with q keys (see Flapan et al. [3, Lemma 1] for the case $n = 1$ (although they don't state the bound explicitly), and Tuffley [13, Theorem 1.2] for $n \geq 2$).

1.2 Overview

As is the case with most Ramsey-type results on intrinsic linking, Theorems 1.3 and 1.5 are proved by using the connect sum operation to combine simpler links into more complicated ones. To achieve the divisibility condition we will require the building block components to be "large", in the sense that they all contain two copies of a fixed suitably triangulated disc. The triangulation will not only need to have many n-simplices, but must also have a combinatorial structure analogous to a path in a graph. Accordingly, we call such a triangulated disc an *n-path*. We give a precise definition of a path in Section 2, and then re-establish a number of known results on intrinsic linking to show that we can require the necessary components to be large in this sense.

The bulk of the work required to prove Theorem 1.3 is done in Proposition 3.1, which forms the main technical lemma of the paper. Section 3 is devoted to the proof of this. The proposition plays the role of Flapan, Mellor and Naimi's Lemma 2, and the statement and proof are heavily modelled on theirs, making modifications as needed for it to work in all dimensions and achieve the divisibility condition. From an arithmetic standpoint, realising the divisibility condition largely boils down to repeatedly applying the following simple number-theoretic observation, used by both Fleming [5] and Tuffley [13]:

Let $\ell_1, \ell_2, \ldots, \ell_q$ be integers. Then there exist $0 \leq a < b \leq q$ such that

$$\sum_{i=a+1}^{b} \ell_i \equiv 0 \pmod{q}.$$

The work then is in achieving this sum topologically, with the integers involved being linking numbers with respect to some fixed sphere S. Paths and generalised key rings (links in which one component has nonzero linking number with all the others) play crucial roles in this.

With Proposition 3.1 established it is a relatively simple matter to prove Theorem 1.3, and we do this in Section 4. The underlying argument is essentially that of Flapan, Mellor and Naimi's proof of their Theorem 1, using our Proposition 3.1 in place of their Lemma 2, and with some additional considerations to ensure that the building block components are sufficiently large, in the sense described above.

Finally, we turn our attention to the two component case in Section 5, and establish the improved bound of Theorem 1.5. This is done by simply improving the construction of the building block link used in our original proof [13, Theorem 1.4] of this result. This building block is a generalised key ring with q keys that are all sufficiently large, and our original approach was to obtain this by working with a subdivision of K_N^n. By taking the subdivision fine enough, we could ensure that each key contained the required pair of paths. However, Lemma 5.1 gives us a simple way to enlarge the keys of an existing key ring, thereby eliminating the need to subdivide. This by itself dramatically reduces the number of vertices required. By additionally "recycling" vertices left over from earlier stages of the construction, we show that we can in fact do this using no more vertices than were needed to construct the initial key ring with q keys, reducing the number of vertices still further.

1.3 Some notation and terminology

The combinatorial structure of a link with many components is usefully described by its *linking pattern:*

Definition 1.6 (Flapan et al. [3, Definitions 1 and 2])**.** Given a link $\mathcal{L}$, the *linking pattern* of $\mathcal{L}$ is the graph with vertices the components of $\mathcal{L}$, and an edge between two components K and L if and only if $\ell k(K, L) \neq 0$. The *mod* 2 *linking pattern* of $\mathcal{L}$ is the graph with vertices the components of $\mathcal{L}$, and an edge between two components K and L if and only if $\ell k_2(K, L) \neq 0$.

An $(r + 1)$-component link $R \cup L_1 \cup \cdots \cup L_r$ is a *generalised key ring* with ring R and keys $L_1, \ldots, L_r$ if its linking pattern contains the star on $r + 1$ vertices as a subgraph, with R as the central vertex. Thus, the components L_i all link R, just like the keys on a key ring. The link is referred to as a "generalised" key ring to reflect the fact that the keys may link each other, which is not typically the case with the kinds of key rings we carry on our persons.

The linking numbers between components of two disjoint many-component links are conveniently collected into a *linking matrix* as follows:

Definition 1.7. Given disjoint ordered oriented links $\mathcal{J} = J_1 \cup \cdots \cup J_s$, $\mathcal{L} = L_1 \cup \cdots \cup L_t$, we define their *linking matrix* $\ell k(\mathcal{J}, \mathcal{L})$ to be the $s \times t$ matrix with (i, j)-entry $\ell k(J_i, L_j)$.

We will say that a matrix A is *positive* if all entries of A are positive, and *nonvanishing* if every entry of A is nonzero.

2 Constructing links with large components

A common strategy in proving Ramsey-type results for intrinsic linking is to start with a link with many components and relatively simple linking behaviour, and combine some of the components to form a link with fewer components but more complicated linking behaviour. Our arguments to prove Theorem 1.3 will require that the building block linking components are "large" in a suitable sense. Thus, in this section we re-establish a number of known results on intrinsic linking to prove the existence of links with large components.

In the classical one-dimensional case (graphs embedded in $\mathbb{R}^3$) we will simply require our components to have sufficiently many vertices (equivalently, sufficiently many edges). In principle, no additional work is required in this case, because we could simply take a sufficiently large complete graph and subdivide each edge into a suitably long path, as is done in Flapan [2]. The combinatorics of triangulated n-spheres are more complicated for $n \geq 2$, however, and it will not be sufficient to simply work with spheres with many vertices or n-simplices. Instead, we will additionally require our components to be large in the following sense, where D is chosen in advance:

Definition 2.1. Let D be an n-dimensional triangulated disc. A triangulated n-sphere is *large with respect to* D or D*-large* if it contains two disjoint oppositely oriented copies of D.

When it comes time to prove Theorem 1.3 we will choose D so that it has a triangulation of the following form:

Definition 2.2. Let D be an n-dimensional triangulated disc with ℓ n-simplices. Then D is a *path of length* ℓ if its n-simplices may be labelled $\Delta_1, \ldots, \Delta_\ell$ such that

$$D_{ab} = \bigcup_{i=a}^{b} \Delta_i$$

is a disc for any $1 \leq a \leq b \leq \ell$.

For $n = 1$ this definition co-incides with the usual meaning of a path in a graph. To construct a path for $n \geq 2$ we may start with ℓ n-simplices $\Delta_1, \ldots, \Delta_\ell$, and choose distinct $(n-1)$-simplices γ_i, δ_i belonging to Δ_i. Choose simplicial isomorphisms $\phi_i \colon \delta_i \to \gamma_{i+1}$ for $1 \leq i \leq \ell - 1$, and glue the Δ_i according to the ϕ_i. The result is a disc D^n, and the triangulation $D^n = \Delta_1 \cup \cdots \cup \Delta_\ell$ satisfies Definition 2.2 by construction. In Lemma 2.6 of [13] it is shown that a disc constructed in this way has $\ell + n$ vertices, and the number of $(n-1)$-simplices in ∂D^n is $\ell(n-1) + 2$. We note that for $n \geq 2$ a path does not necessarily have this form: for instance, for $n = 2$ the triangulation of a regular n-gon by radii may be given the structure of a path.

We begin by establishing the existence of D-large n-spheres with arbitrarily many additional n-simplices. For convenience, we let $\sigma_n(D, m)$ be the minimal number of vertices of a triangulated sphere satisfying the conditions of the following lemma.

Lemma 2.3. *Let D be a triangulated disc, and let m be a positive integer. There is a triangulation of S^n that contains two disjoint oppositely oriented copies of D, together with at least m additional n-simplices.*

Proof. Consider $D \times I$. If $V = \{v_0, \ldots, v_N\}$ is the vertex set of D, then $D \times I$ has a triangulation with vertex set $V \times \{0, 1\}$, and simplices of the form

$$\delta_j = [(v_{i_0}, 0), \ldots, (v_{i_j}, 0), (v_{i_j}, 1), \ldots, (v_{i_k}, 1)]$$

for $0 \leq j \leq k$ and each k-simplex $\delta = [v_{i_0}, \ldots, v_{i_k}]$ of D with $i_0 < i_1 < \cdots < i_k$. As a first pass we let $S = \partial(D \times I)$ with the induced triangulation.

The n-sphere S contains two disjoint copies of D, namely $D \times \{0\}$ and $D \times \{1\}$, and they are oppositely oriented because they are exchanged by reflection in the equator $\partial D \times \{\frac{1}{2}\}$. Suppose that ∂D contains a total of t simplices of dimension $n - 1$. Each contributes a total of n simplices of dimension n to $\partial D \times I$, so S has a total of nt additional n-simplices. If $nt \geq m$ does not hold then let S' be a triangulated n-sphere with at least $m - nt + 1$ simplices of dimension n (such a triangulated sphere certainly exists, for example by taking the boundary of a sufficiently long $(n+1)$-path, as constructed above).

Choose n-simplices δ and δ' belonging to $\partial D \times I$ and S', respectively, and form the connected sum of S and S' by gluing the discs $S - \delta$ and $S' - \delta'$ along their boundaries. The resulting sphere satisfies the conditions given in the conclusion of the lemma. □

We now use Lemma 2.3 to prove the existence of generalised key rings with large rings. To do this we require the following slight strengthening of Lemma 3.2 of [13], which is in turn an extension of Lemma 1 of Flapan et al. [3] to all dimensions.

Lemma 2.4. *Let D be a triangulated disc. Suppose that K_N^n is embedded in $\mathbb{R}^{2n+1}$ such that it contains a link*

$$L \cup J_1 \cup \cdots \cup J_{m^2} \cup X_1 \cup \cdots \cup X_{m^2},$$

where $\ell k_2(J_i, X_i) = 1$ for all i, and L contains two disjoint oppositely oriented copies of D and at least m^2 additional n-simplices. Then there is an n-sphere Z in K_N^n with all its vertices on $L \cup J_1 \cup \cdots \cup J_{m^2}$, and an index set I with $|I| \geq \frac{m}{2}$, such that $\ell k_2(Z, X_j) = 1$ for all $j \in I$ and Z contains two disjoint oppositely oriented copies of D.

In Lemma 3.2 of [13] we require only that L has at least m^2 n-simplices. Thus, the difference between the two results is the stronger condition that L contains the two copies of D and a further m^2 n-simplices, and the additional conclusion that Z contains two disjoint oppositely oriented copies of D. To prove the stronger form it is only necessary to observe that in proving the original result we can ensure that the copies of D in L end up in Z.

Proof. The first step in the proof of [13, Lemma 3.2] is to construct an n-sphere $\mathcal{S}$ with all its vertices on $L \cup J_1 \cup \cdots \cup J_{m^2}$, and meeting each sphere J_i in an n-simplex δ_i. This is done by choosing a distinct n-simplex δ_i' belonging to L for each $i = 1, \ldots, m^2$, and applying [13, Corollary 2.2] to obtain a sphere $Q_i \subseteq K_N^n$ with all its vertices on $\delta_i \cup \delta_i'$, and meeting J_i in δ_i and L in δ_i'. The sphere $\mathcal{S}$ is then constructed from L and the Q_i by omitting the interiors of the discs δ_i'. Thus, we can ensure that $\mathcal{S}$ contains two disjoint oppositely oriented copies of D by choosing the δ_i' from among the m^2 additional n-simplices of L, leaving the copies of D intact.

At the final step in the proof of [13, Lemma 3.2], the required sphere Z is constructed from $\mathcal{S}$ and a (possibly empty) subset of the J_i, by omitting the interiors of the corresponding n-simplices δ_i. Therefore, since $\mathcal{S}$ contains the required copies of D, we are guaranteed that Z does too. □

Corollary 2.5. *Let D be a triangulated disc, and r a positive integer. For N sufficiently large, every embedding of K_N^n in $\mathbb{R}^{2n+1}$ contains an $(r+1)$-component link $R \cup L_1 \cup \cdots \cup L_r$ such that $\ell k_2(R, L_i) = 1$ for all i, and R contains two disjoint oppositely oriented copies of D. It suffices to take*

$$N \geq \kappa_n(D, r) = 4r^2(2n+4) + \sigma_n(D, 4r^2).$$

Proof. Given an embedding of $K_{\kappa_n(D,r)}^n$ in $\mathbb{R}^{2n+1}$, choose $4r^2$ disjoint copies of K_{2n+4}^n contained in the embedding, together with a copy of $K_{\sigma_n(D,4r^2)}^n$. By Taniyama [12] the ith copy of K_{2n+4}^n contains a 2-component link $J_i \cup X_i$ such that $\ell k_2(J_i, X_i) = 1$, and the copy of $K_{\sigma_n(D,4r^2)}^n$ contains a triangulated sphere L that contains two disjoint oppositely oriented copies of D and at least $4r^2$ additional n-simplices. The result now follows by applying Lemma 2.4 with $m = 2r$ to the link

$$L \cup J_1 \cup \cdots \cup J_{4r^2} \cup X_1 \cup \cdots \cup X_{4r^2}. \qquad \square$$

Finally, we extend Proposition 1 of Flapan et al. [3] to higher dimensions, with the additional conclusion that all components are large with respect to a chosen triangulated disc D. This result serves as the base case for the inductive argument proving Theorem 1.3 in Section 4.

Proposition 2.6. *Let D be a triangulated disc, and let r be a positive integer. For N sufficiently large, every embedding of K_N^n in $\mathbb{R}^{2n+1}$ contains a $2r$-component link*

$$J_1 \cup \cdots \cup J_r \cup L_1 \cup \cdots \cup L_r,$$

such that $\ell k_2(J_i, L_j)$ is nonzero for all i and j, and each component contains two disjoint oppositely oriented copies of D.

The link given by this result has mod two linking pattern containing the complete bipartite graph $K_{r,r}$, because each component J_i has nonzero mod 2 linking number with each component L_j. The argument to prove the existence of such a link is exactly that of Flapan et al.'s proof of their Proposition 1, and the extension to higher dimensions already follows from our paper [13]: as noted in Section 1.2.2 of [13] their Proposition 1 is a purely combinatorial argument that depends only on their Lemma 1 and the existence of generalised key rings, and these are generalised to higher dimensions in [13]. So the work to be done here is to ensure that each component contains copies of the disc D.

For $n = 1$ this already follows from Flapan et al.'s Proposition 1, because we may simply subdivide each edge of a sufficiently large complete graph into paths of length ℓ. A similar approach could be taken in higher dimensions, using the subdivisions of K_N^n constructed in [13], but this introduces many unnecessary vertices. We give a simpler argument that doesn't make use of subdivision, and requires far fewer vertices.

Proof. Following Flapan et al. [3] let $m = \frac{(4r)^{2^r}}{4}$, and let

$$N = m\kappa_n(D, r) + r\sigma_n(D, m).$$

Then K_N^n contains m copies of $K_{\kappa_n(D,r)}^n$ and r copies of $K_{\sigma_n(D,m)}^n$, all disjoint from one another. Given an embedding of K_N^n in $\mathbb{R}^{2n+1}$, by Corollary 2.5 the ith copy of $K_{\kappa_n(D,r)}^n$ contains a generalised key ring

$$R_i \cup J_{i1} \cup \cdots \cup J_{ir}$$

such that the ring R_i is D-large; and the jth copy of $K_{\sigma_n(D,m)}^n$ contains a D-large sphere L_j that contains at least m additional n-simplices.

Apply Lemma 2.4 to the link

$$L_1 \cup J_{11} \cup \cdots \cup J_{m1} \cup R_1 \cup \cdots \cup R_m.$$

This yields a D-large sphere Z_1 with all its vertices on $L_1 \cup J_{11} \cup \cdots \cup J_{m1}$, and an index set I_1 with $|I_1| \geq \frac{\sqrt{m}}{2} = \frac{(4r)^{2^{r-1}}}{4} = m_1$, such that $\ell k_2(Z_1, R_i) = 1$ for all $i \in I_1$. Suppose now that for some $1 \leq k < r$ we have constructed D-large spheres $Z_1, \ldots, Z_k$ and an index set I_k such that

(1) all vertices of Z_j lie on $L_j \cup J_{1j} \cup \cdots \cup J_{mj}$ for $1 \leq j \leq k$;

(2) $|I_k| \geq m_k = \frac{(4r)^{2^{r-k}}}{4}$;

(3) $\ell k_2(Z_j, R_i) = 1$ for all $1 \leq j \leq k$ and $i \in I_k$.

Applying Lemma 2.4 to the link

$$L_{k+1} \cup \left(\bigcup_{i \in I_k} J_{i(k+1)} \right) \cup \left(\bigcup_{i \in I_k} R_i \right)$$

we obtain a D-large sphere Z_{k+1} with all its vertices on $L_{k+1} \cup J_{1(k+1)} \cup \cdots \cup J_{m(k+1)}$, and an index set $I_{k+1} \subseteq I_k$ with $|I_{k+1}| \geq \frac{\sqrt{m_k}}{2} = \frac{(4r)^{2^{r-k-1}}}{4} = m_{k+1}$, such that $\ell k_2(Z_{k+1}, R_i) = 1$ for all $i \in I_{k+1}$. This gives us D-large spheres $Z_1, \ldots, Z_{k+1}$ and an index set I_{k+1} such that conditions (1) – (3) hold with k replaced by $k+1$, so by induction there are D-large spheres $Z_1, \ldots, Z_r$ and an index set I_r such that they hold for $k = r$. Since $m_r = \frac{(4r)^{2^{r-r}}}{4} = r$, the first $2r$ components of

$$Z_1 \cup \cdots \cup Z_r \cup \left(\bigcup_{i \in I_r} R_i \right)$$

are the required link. □

3 The main technical lemma

This section is dedicated to proving the following analogue of Lemma 2 of Flapan et al. [3], which forms the main technical lemma of this paper:

Proposition 3.1 (Main technical lemma). *Let $q \in \mathbb{N}$. Suppose that K_N^n is embedded in $\mathbb{R}^{2n+1}$ such that it contains a link with oriented components $J_1, \ldots, J_A$, $L_1, \ldots, L_B$, $X_1, \ldots, X_S$ and $Y_1, \ldots, Y_T$ satisfying*

(1) $A \geq 2^S q^{S+T}$;

(2) $B \geq 3^S 2^T (S+T) q^{S+T}$;

(3) $\ell k(J_a, X_s)$ is nonzero for all a and s;

(4) $\ell k(L_b, Y_t)$ is nonzero for all b and t; and

(5) each component J_a, L_b contains two disjoint oppositely oriented copies of a fixed path $\mathcal{D}$ of length $\lambda \geq (2q)^{S+T}$.

Then K_N^n contains an n-sphere Z with all its vertices on $J_1 \cup \cdots \cup J_A \cup L_1 \cup \cdots \cup L_B$ such that, for each s and t, $\ell k(Z, X_s)$ and $\ell k(Z, Y_t)$ are nonzero multiples of q.

We note that the hypotheses of our Proposition 3.1 are much stronger than the hypotheses of Flapan et al.'s Lemma 2: we require A and B to be much greater, and we have the additional hypothesis (5) that the components J_a, L_b are large with respect to a certain path. This is to be expected, since our conclusion is strictly stronger than theirs: any nonzero multiple of q is necessarily at least q in magnitude.

Before proving Proposition 3.1 we first establish the following lemma on sums of vectors in $\mathbb{R}^d$, which we will use in the proof.

Lemma 3.2. *Let* $\mathbf{f} \in \mathbb{R}^d$ *be a vector with all entries nonzero, and for* $i = 0, \ldots, N$ *let* $\mathbf{v}_i \in \mathbb{R}^d$. *If* $N \geq 2^d$ *then there exist* $0 \leq j < k \leq N$ *such that every entry of* $\mathbf{f} + \mathbf{v}_k - \mathbf{v}_j$ *is nonzero.*

Proof. The proof is by induction on d. In the base case $d = 1$, suppose that $N \geq 2$. If either $f + v_1 - v_0$ or $f + v_2 - v_1$ is nonzero then we are done, and otherwise

$$f + v_2 - v_0 = (f + v_2 - v_1) + (f + v_1 - v_0) - f = -f \neq 0.$$

Thus the lemma holds in the base case $d = 1$.

Suppose now that the lemma holds for some $d \geq 1$, and let $\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_N$ be $N + 1 \geq 2^{d+1} + 1$ vectors in $\mathbb{R}^{d+1}$. We claim that there is $N' \geq 2^d$ and $N' + 1$ indices $0 \leq i_0 < i_1 < \cdots < i_{N'} \leq N$ such that, for any $0 \leq j < k \leq N'$, the $(d + 1)$th entry of $\mathbf{f} + \mathbf{v}_{i_k} - \mathbf{v}_{i_j}$ is nonzero. The inductive step will then follow by applying the inductive hypothesis to the first d entries of $\mathbf{f}$ and $\mathbf{v}_{i_0}, \ldots, \mathbf{v}_{i_{N'}}$.

Write $x^{(i)}$ for the ith entry of $\mathbf{x} \in \mathbb{R}^m$. To prove the claim we consider the graph with vertex set $\{0, 1, \ldots, N\}$, and an edge between j and k if $j < k$ and the difference $v_k^{(d+1)} - v_j^{(d+1)}$ is equal to the forbidden value $-f^{(d+1)}$. Now observe that for any path $(i_0, i_1, \ldots, i_m)$ in this graph we have

$$v_{i_m}^{(d+1)} - v_{i_0}^{(d+1)} = \sum_{j=1}^{m} [v_{i_j}^{(d+1)} - v_{i_{j-1}}^{(d+1)}] = -f^{(d+1)} \sum_{j=1}^{m-1} \operatorname{sign}(i_j - i_{j-1}).$$

In particular, if the path is a cycle then $i_m = i_0$, and it follows that

$$f^{d+1} \sum_{j=1}^{m-1} \operatorname{sign}(i_{j+1} - i_j) = 0.$$

Since $f^{(d+1)}$ is nonzero by hypothesis the sum must be zero, and since each term is ±1, for this to occur it must involve an even number of terms. Thus any cycle must be of even length, and it follows that our graph is bipartite.

Colour the vertices black and white in such a way that there is no edge between vertices of the same colour, and let $0 \leq i_0 < i_1 < \cdots < i_{N'} \leq N$ be the vertices belonging to the larger colour class. Then $N' + 1 \geq \lceil (N + 1)/2 \rceil \geq \lceil (2^{d+1} + 1)/2 \rceil = 2^d + 1$, and for any $0 \leq j < k \leq N'$ we have $f^{(d+1)} + v_{i_k}^{(d+1)} - v_{i_j}^{(d+1)} \neq 0$, as required. Lemma 3.2 now follows by our discussion above. □

Proof of Proposition 3.1. Let

$$\mathcal{J} = J_1 \cup \cdots \cup J_A, \qquad \mathcal{X} = X_1 \cup \cdots \cup X_S,$$
$$\mathcal{L} = L_1 \cup \cdots \cup L_B, \qquad \mathcal{Y} = Y_1 \cup \cdots \cup Y_T.$$

Following Flapan et al. [3], we begin by replacing the links $\mathcal{J}$ and $\mathcal{L}$ with sublinks $\mathcal{J}'$, $\mathcal{L}''$ for which we have some control over the signs of the entries of the linking matrices $\ell k(\mathcal{J}', \mathcal{X})$, $\ell k(\mathcal{L}'', \mathcal{Y})$ and $\ell k(\mathcal{L}'', \mathcal{X})$. To do this, we first consider the patterns of signs of the entries of the vectors $\ell k(J_a, \mathcal{X})$. Since these vectors have S entries, and all are nonzero, there are 2^S possibilities for the patterns of signs (positive and negative) in each one. It follows that we can choose at least $A/2^S \geq q^{S+T}$ of them that all have the same pattern of signs. Moreover, after reversing the orientation of some components of $\mathcal{X}$ if necessary, we may assume that these signs are all positive. Thus, setting $\mathcal{J}' = J_1 \cup \cdots \cup J_{q^{S+T}}$, we may assume without loss of generality that the linking matrix $\ell k(\mathcal{J}', \mathcal{X})$ is positive.

Applying the same argument to the vectors $\ell k(L_b, \mathcal{Y})$, we obtain a sublink $\mathcal{L}'$ of $\mathcal{L}$ with at least $3^S(S+T)q^{S+T}$ components such that the linking matrix $\ell k(\mathcal{L}', \mathcal{Y})$ is positive. We now consider the patterns of signs (positive, negative or zero) of the vectors $\ell k(L_b, \mathcal{X})$ for L_b a component of $\mathcal{L}'$. There are now 3^S possibilities for these patterns, so we may choose at least $(S+T)q^{S+T}$ components that have the same pattern. Setting $\mathcal{L}'' = L_1 \cup \cdots \cup L_{(S+T)q^{S+T}}$ we may therefore assume without loss of generality that the linking matrix $\ell k(\mathcal{L}'', \mathcal{Y})$ is positive, and that each column of $\ell k(\mathcal{L}'', \mathcal{X})$ is either positive, negative, or zero. From now on we restrict our attention to the sublinks $\mathcal{J}'$ and $\mathcal{L}''$ of $\mathcal{J}$ and $\mathcal{L}$.

Our next goal is to construct a sublink $\mathcal{Z} = Z_1 \cup \cdots \cup Z_C$ of $\mathcal{J}' \cup \mathcal{L}''$ such that every entry of

$$\mathbf{z} = \sum_{c=1}^{C} \ell k(Z_c, \mathcal{X} \cup \mathcal{Y})$$

is a nonzero multiple of q. At the final step we will obtain the required n-sphere Z as a connect sum of the components of $\mathcal{Z}$. To this end we begin by considering the sums

$$\mathbf{j}_\alpha = \sum_{a=1}^{\alpha} \ell k(J_a, \mathcal{X} \cup \mathcal{Y})$$

modulo q for $1 \leq \alpha \leq q^{S+T}$. Each vector $\mathbf{j}_\alpha$ has $S+T$ entries, so there are q^{S+T} possibilities when considered mod q. Since we have q^{S+T} vectors in total, by the pigeonhole principle we can either find one that is zero modulo q, or two that are equal modulo q. In either case, there are integers $0 \leq \alpha_0 < \alpha_1 \leq q^{S+T}$ such that the vector

$$\mathbf{j} = \sum_{a=\alpha_0+1}^{\alpha_1} \ell k(J_a, \mathcal{X} \cup \mathcal{Y})$$

is zero modulo q. Moreover, the first S entries of $\mathbf{j}$ are given by $\sum_{a=\alpha_0+1}^{\alpha_1} \ell k(J_a, \mathcal{X})$, and are therefore nonzero, because the vector $\ell k(J_a, \mathcal{X})$ is positive for each a. We will use $J_{\alpha_0+1} \cup \cdots \cup J_{\alpha_1}$ as the first $\alpha_1 - \alpha_0$ components of $\mathcal{Z}$.

We now consider the sums

$$\sum_{b=1}^{\beta} \ell k(L_b, \mathcal{X} \cup \mathcal{Y})$$

modulo q for $1 \leq \beta \leq (S+T)q^{S+T}$. Since there are again q^{S+T} possibilities mod q, and we have $(S+T)q^{S+T}$ sums in total, we can either find $S+T$ of them that are zero mod q, or $S+T+1$ of them that are identical mod q. In either case, there are integers $0 \leq \beta_0 < \beta_1 < \cdots < \beta_{S+T} \leq (S+T)q^{S+T}$ such that the vectors

$$\boldsymbol{\ell}_i = \sum_{b=\beta_0+1}^{\beta_i} \ell k(L_b, \mathcal{X} \cup \mathcal{Y})$$

are zero modulo q. Any additional components of $\mathcal{Z}$ will be chosen from among $L_{\beta_0+1} \cup L_{\beta_0+2} \cup \cdots \cup L_{\beta_{S+T}}$.

To choose the remaining components of $\mathcal{Z}$ we consider the sequence of $S+T+1$ vectors $\mathbf{j}, \mathbf{j}+\boldsymbol{\ell}_1, \dots, \mathbf{j}+\boldsymbol{\ell}_{S+T}$. From above these vectors are all zero when considered modulo q, and we claim that it is possible to choose at least one of them that is nonvanishing when considered as an integer vector. To see this, consider first the $(S+t)$-entries for some $1 \leq t \leq T$, which are given by

$$\begin{aligned} j^{(S+t)} &= \sum_{a=\alpha_0+1}^{\alpha_1} \ell k(J_a, Y_t), \\ (j+\ell_i)^{(S+t)} &= \sum_{a=\alpha_0+1}^{\alpha_1} \ell k(J_a, Y_t) + \sum_{b=\beta_0+1}^{\beta_i} \ell k(L_b, Y_t). \end{aligned}$$

Since the linking matrix $\ell k(\mathcal{L}'', \mathcal{Y})$ is positive these form a strictly increasing sequence, and consequently the $(S+t)$-entry vanishes for at most one of our $S+T+1$ vectors.

Next, consider the s-entries for some $1 \leq s \leq S$, which are given by

$$\begin{aligned} j^{(s)} &= \sum_{a=\alpha_0+1}^{\alpha_1} \ell k(J_a, X_s), \\ (j+\ell_i)^{(s)} &= \sum_{a=\alpha_0+1}^{\alpha_1} \ell k(J_a, X_s) + \sum_{b=\beta_0+1}^{\beta_i} \ell k(L_b, X_s). \end{aligned}$$

Recall that the first sum is positive, and that each column of the linking matrix $\ell k(\mathcal{L}'', \mathcal{X})$ is either positive, negative, or zero. It follows that the above sequence of integers is either constant (in which case it is positive), or it is strictly increasing or strictly decreasing. In any case we again conclude that the s-entry vanishes for at most one of our $S+T+1$ vectors. Thus there are at most $S+T$ vectors for which one of the entries vanishes, and so there is at least one for which no entry vanishes, proving the claim. We may then set

$$\begin{aligned} \mathcal{Z} &= Z_1 \cup \cdots \cup Z_C \\ &= \begin{cases} J_{\alpha_0+1} \cup \cdots \cup J_{\alpha_1} & \text{if } \mathbf{j} \text{ is nonvanishing, or} \\ J_{\alpha_0+1} \cup \cdots \cup J_{\alpha_1} \cup L_{\beta_0+1} \cup \cdots \cup L_{\beta_i} & \text{if } \mathbf{j}+\boldsymbol{\ell}_i \text{ is nonvanishing.} \end{cases} \end{aligned}$$

With this choice of $\mathcal{Z}$, every entry of

$$\mathbf{z}_0 = \sum_{c=1}^{C} \ell k(Z_c, \mathcal{X} \cup \mathcal{Y})$$

is a nonzero multiple of q, as required.

Our final task is to obtain the required n-sphere as a suitable connect sum of the components of $\mathcal{Z}$. To do this we will inductively construct oriented spheres $F_1, \dots, F_{C-1}$ such that, for each $1 \leq \gamma \leq C-1$,

(a) the vertices of F_γ lie on $Z_\gamma \cup Z_{\gamma+1}$ (and so F_γ is disjoint from $\mathcal{X}$, $\mathcal{Y}$, and the rest of $\mathcal{Z}$);

(b) $F_{\gamma-1} \cap Z_\gamma$ and $F_\gamma \cap Z_\gamma$ are disjoint discs, each of which is oppositely oriented by Z_γ and $F_{\gamma-1}$ or F_γ;

(c) every entry of the vector

$$\mathbf{z}_\gamma = \mathbf{z}_0 + \sum_{i=1}^{\gamma} \ell k(F_i, \mathcal{X} \cup \mathcal{Y})$$

is a nonzero multiple of q.

We will then obtain the required sphere Z from the union of $\mathcal{Z}$ and the F_c by omitting the interiors of the discs $F_c \cap Z_c$ and $F_c \cap Z_{c+1}$. Conditions (a) and (b) imply that F_c and $F_{c'}$ are disjoint for all c and c', and it follows that Z is a connect sum of spheres, and hence itself a sphere. Moreover, as a chain we have $Z = \sum_{c=1}^{C} Z_c + \sum_{c=1}^{C-1} F_c$, so

$$\ell k(Z, \mathcal{X} \cup \mathcal{Y}) = \mathbf{z}_0 + \sum_{c=1}^{C-1} \ell k(F_c, \mathcal{X} \cup \mathcal{Y}),$$

and by condition (c) every entry of this vector is a nonvanishing multiple of q.

The underlying technique for constructing the spheres F_c comes from the proof of Theorem 1.4 of Tuffley [13], but additional work is required to ensure that condition (c) is satisfied. By hypothesis (5) each sphere Z_c contains two disjoint copies of the path $\mathcal{D}$, one of each orientation. We begin by labelling these D_c and D'_c in such a way that there is an orientation reversing simplicial isomorphism $\phi_c \colon D_c \to D'_{c+1}$. This may be done inductively: first label the copies of $\mathcal{D}$ contained in Z_1 arbitrarily, and then once D_c and D'_c have been chosen, choose D_{c+1} and D'_{c+1} so that D'_{c+1} is oppositely oriented to D_c. We will choose the spheres F_c so that the following strengthened form of condition (a) holds for $1 \le \gamma \le C - 1$:

(a′) the vertices of F_γ lie on $D_\gamma \cup D'_{\gamma+1}$.

This condition serves to ensure that $F_{\gamma-1} \cap Z_\gamma$ and $F_\gamma \cap Z_\gamma$ are disjoint, as required by condition (b).

Suppose that for some $0 \le c < C - 1$ the spheres $F_1, \dots, F_c$ have been constructed so that conditions (a′), (b) and (c) hold for $0 \le \gamma \le c$. When $c = 0$ conditions (a′) and (b) are empty, and condition (c) is that every entry of $\mathbf{z}_0$ is a nonzero multiple of q, so we may take $c = 0$ as our base case. Let $\Delta_1, \dots, \Delta_\lambda$ be a labelling of the n-simplices of the path D_{c+1} as in Definition 2.2, and for $1 \le \ell \le \lambda$ let P_ℓ be the oriented sphere satisfying

$$P_\ell \cap Z_{c+1} = \Delta_\ell, \qquad\qquad P_\ell \cap Z_{c+2} = \phi_{c+1}(\Delta_\ell)$$

that results from applying Corollary 2.2 of Tuffley [13] to the pairs (Z_{c+1}, D_{c+1}) and (Z_{c+2}, D'_{c+2}). The vertices of these spheres all lie on $D_{c+1} \cup D'_{c+2}$, and for any $1 \le \mu \le \nu \le \lambda$, the chain $\sum_{\ell=\mu}^{\nu} P_\ell$ represents a sphere meeting D_{c+1} in the disc $\bigcup_{\ell=\mu}^{\nu} \Delta_\ell$, and D'_{c+2} in the disc $\bigcup_{\ell=\mu}^{\nu} \phi_{c+1}(\Delta_\ell)$.

For $1 \le \ell \le \lambda$ we consider the sums

$$\sum_{i=1}^{\ell} \ell k(P_i, \mathcal{X} \cup \mathcal{Y})$$

modulo q. As above there are q^{S+T} possibilities for these modulo q, and we have $\lambda \ge 2^{S+T} q^{S+T}$ of them, so we can either find 2^{S+T} of them that are identically zero mod q, or

$2^{S+T}+1$ of them that are equal mod q. In either case there are integers $0 \leq \mu_0 < \mu_1 < \cdots < \mu_{2^{S+T}}$ such that the vectors

$$\mathbf{p}_j = \sum_{i=\mu_0+1}^{\mu_j} \ell k(P_i, \mathcal{X} \cup \mathcal{Y})$$

are identically zero mod q for $1 \leq j \leq 2^{S+T}$.

Set $\mathbf{p}_0 = \mathbf{0}$, and apply Lemma 3.2 to the vectors $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_{2^{S+T}} \in \mathbb{R}^{S+T}$ with $\mathbf{f} = \mathbf{z}_c$. This yields indices $0 \leq j < k \leq 2^{S+T}$ such that no entry of

$$\mathbf{z}_c + \mathbf{p}_k - \mathbf{p}_j = \mathbf{z}_c + \sum_{i=\mu_j+1}^{\mu_k} \ell k(P_i, \mathcal{X} \cup \mathcal{Y})$$

is zero. Moreover, the vectors $\mathbf{z}_c$, $\mathbf{p}_j$ and $\mathbf{p}_k$ are all identically zero mod q, so every entry of $\mathbf{z}_c + \mathbf{p}_k - \mathbf{p}_j$ is a nonzero multiple of q.

Let $F_{c+1} = \sum_{i=\mu_j+1}^{\mu_k} P_i$. Then F_{c+1} represents an n-sphere with all its vertices on $Z_{c+1} \cup Z_{c+2}$, and meeting Z_{c+1} and Z_{c+2} in the discs

$$F_{c+1} \cap Z_{c+1} = \bigcup_{i=\mu_j+1}^{\mu_k} \Delta_i \subseteq D_{c+1}, \quad F_{c+1} \cap Z_{c+2} = \phi_{c+1}\left(\bigcup_{i=\mu_j+1}^{\mu_k} \Delta_i\right) \subseteq D'_{c+2}.$$

The construction of Corollary 2.2 of Tuffley [13] ensures that these discs are oppositely oriented by F_{c+1} and $Z_{c+1} \cup Z_{c+2}$, so conditions (a′) and (b) are satisfied; and with this choice of F_{c+1} we have $\mathbf{z}_{c+1} = \mathbf{z}_c + \mathbf{p}_k - \mathbf{p}_j$, so condition (c) is too. This completes the inductive step, and we now obtain the required sphere Z as described above. □

4 Proof of Theorem 1.3

We are now in a position to prove our main result, Theorem 1.3. The strategy is that of Flapan et al.'s proof of their Theorem 1.

Proof of Theorem 1.3. Following Flapan et al. [3], for each $u, v \in \mathbb{N}$ let $H(u,v)$ denote the complete $(u+2)$-partite graph with parts P_1 and P_2 containing v vertices each, and parts $Q_1, \ldots, Q_u$ containing a single vertex each. We will prove by induction on u that for every $u \geq 0$ and $v, \ell \geq 1$, for N sufficiently large every embedding of K_N^n in $\mathbb{R}^{2n+1}$ contains a link $\mathcal{L}$ such that

(L1) the linking pattern of $\mathcal{L}$ contains the graph $H(u,v)$;

(L2) the linking number between any two distinct components in $Q_1 \cup \cdots \cup Q_u$ is a nonzero multiple of q; and

(L3) every component in $P_1 \cup P_2$ contains disjoint oppositely oriented copies of a path D of length at least ℓ.

For simplicity, we will say that a link $\mathcal{L}$ satisfying conditions (L1) – (L3) with the given parameter values satisfies property (u, v, ℓ).

The base case $u = 0$ follows from Proposition 2.6 with $r = v$, by choosing D to be a path of length ℓ. Suppose then that the claim holds for some $u \geq 0$. Given $v, \ell \geq 0$, let

$$\begin{aligned} S &= v, \\ T &= u + v, \\ A = B &= 2^T 3^S (S+T) q^{S+T} \geq 2^S q^{S+T}, \\ \lambda &= \max\{\ell, (2q)^{S+T}\}, \end{aligned}$$

and let $w = S + A = S + B$. By our inductive hypothesis, for N sufficiently large every embedding of K_N^n in $\mathbb{R}^{2n+1}$ contains a link $\mathcal{L}$ satisfying property (u, w, λ). We will show that every such embedding also contains a link $\mathcal{L}'$ satisfying property $(u+1, v, \ell)$.

Given an embedding of K_N^n in $\mathbb{R}^{2n+1}$ and a link $\mathcal{L}$ contained in it satisfying property (u, w, λ), label the components of $\mathcal{L}$ such that

$$\begin{aligned} P_1 &= \{X_1, \ldots, X_S, L_1, \ldots, L_B\}, \\ P_2 &= \{Y_1, \ldots, Y_S, J_1, \ldots, J_A\}, \end{aligned}$$

and $Q_i = \{Y_{v+i}\}$ for $1 \leq i \leq u$. Then all linking numbers $\ell k(J_a, X_s)$ and $\ell k(L_b, Y_t)$ are nonzero by (L1), and every component J_a, L_b contains two disjoint copies of a path $\mathcal{D}$ of length at least $\lambda \geq (2q)^{S+T}$, by (L3). So we may apply Proposition 3.1 to $\mathcal{L}$ to obtain a sphere Z with all its vertices on $J_1 \cup \cdots \cup J_A \cup L_1 \cup \cdots \cup L_B$ and linking every component X_s, Y_t with linking number a nonzero multiple of q. Let

$$\begin{aligned} \mathcal{L}' &= X_1 \cup \cdots \cup X_S \cup Y_1 \cup \cdots \cup Y_T \cup Z \\ &= X_1 \cup \cdots \cup X_v \cup Y_1 \cup \cdots \cup Y_{u+v} \cup Z, \end{aligned}$$

and partition the components as $P_1' \cup P_2' \cup Q_1' \cup \cdots \cup Q_{u+1}'$ such that

$$\begin{aligned} P_1' &= \{X_1, \ldots, X_v\}, \\ P_1' &= \{Y_1, \ldots, Y_v\}, \end{aligned}$$

and

$$Q_i' = \begin{cases} \{Y_{v+i}\} & 1 \leq i \leq u, \\ \{Z\} & i = u+1. \end{cases}$$

Then with respect to this partition the linking pattern of $\mathcal{L}'$ contains the graph $H(u+1, v)$; any two components in $Q_1' \cup \cdots \cup Q_{u+1}'$ have linking number a nonzero multiple of q; and every component in $P_1 \cup P_2$ contains a copy of $\mathcal{D}$, which is a path of length at least $\lambda \geq \ell$. So $\mathcal{L}'$ satisfies property $(u+1, v, \ell)$, completing the inductive step. By (L2) the result now follows by restricting attention to $Q_1 \cup \cdots \cup Q_u$, with $u = r$. □

5 The two component case

We now turn to the two component case, and establish the improved bound of Theorem 1.5.

From the proof of [13, Theorem 1.4] it suffices to prove every embedding of $K^n_{\kappa_n(q)}$ contains a generalised key ring with q keys each large with respect to a path D of length q. The approach of [13] was to work with a subdivision of K_N^n, in which each n-simplex was subdivided into q^n simplices. This is a fairly extravagant approach, since only $2q$

n-simplices from each component are used to form the required paths. The reduction in the number of vertices required comes from Lemma 5.1, which gives us a simple and economical way to enlarge the keys of an existing generalised key ring. A further modest saving comes from "recycling" some of the vertices leftover from the construction of the initial key ring.

Lemma 5.1. *Let K_N^n be embedded in $\mathbb{R}^{2n+1}$ such that it contains a link $X \cup Y$ with $\ell k(X, Y) \neq 0$. Let D be a triangulated n-disc with d vertices, and suppose that V is a set of $2d - (n+1)$ vertices of K_N^n disjoint from $X \cup Y$. Then K_N^n contains a D-large sphere Z with all its vertices on $Y \cup V$ such that $\ell k(X, Z) \neq 0$.*

The result also holds with all linking numbers calculated mod 2.

Proof. Choose an n-simplex Δ belonging to Y, and let $S = \partial(D \times I)$ with the triangulation with $2d$ vertices from the proof of Lemma 2.3. Then $\Delta \cup V$ contains a total of $(n+1)+(2d-(n+1)) = 2d$ vertices, so we may embed S in K_N^n such that all vertices of S lie on $\Delta \cup V$ and Δ is an n-simplex of $\partial D \times I$. Orient S such that Δ receives opposite orientations from S and Y, and consider the chains S and $T = S + Y$. Both represent D-large n-spheres with all their vertices on $Y \cup V$, and the linking numbers $\ell k(X, S)$, $\ell k(X, T)$ cannot both be zero because in the homology group $H_n(\mathbb{R}^{2n+1} - X)$ we have

$$[T] - [S] = [S + Y] - [S] = [Y] \neq 0. \tag{5.1}$$

We may therefore choose one of S and T to be Z so that $\ell k(X, Z) \neq 0$.

If $\ell k_2(X, Y) \neq 0$ then equation (5.1) holds in $H_n(\mathbb{R}^{2n+1} - X; \mathbb{Z}/2\mathbb{Z})$, and we may again choose Z to be one of S and T so that $\ell k_2(X, Z) \neq 0$. □

Corollary 5.2. *Let q be a positive integer. Then every embedding of $K^n_{\kappa_n(q)}$ in $\mathbb{R}^{2n+1}$ contains a generalised key ring in which each key is large with respect to a path D of length q.*

Proof. By [13, Theorem 1.2] every embedding of $K^n_{\kappa_n(q)}$ in $\mathbb{R}^{2n+1}$ contains a generalised key ring $\mathcal{L}$ with q keys. This link is constructed by applying [13, Lemma 3.2] (the extension of [3, Lemma 1] to higher dimensions) to a link

$$L \cup J_1 \cup \cdots \cup J_{4q^2} \cup K_1 \cup \cdots \cup K_{4q^2},$$

in which $\ell k_2(J_i, K_i)$ is nonzero for all i, and each component J_i, K_i is the boundary of an $(n+1)$-simplex. This yields an n-sphere R with all vertices on $L \cup J_1 \cup \cdots \cup J_{4q^2}$ and linking at least q of the K_i, which forms the ring of the generalised key ring. Let $K_{i_1}, \ldots, K_{i_q}$ be the keys.

Recall that a path D of length q can be constructed using as few as $d = q + n$ vertices. Since only q of the K_i are components of $\mathcal{L}$ this leaves at least $(4q^2 - q)(n + 2) = q(4q - 1)(n + 2)$ vertices of $K^n_{\kappa_n(q)}$ that do not belong to $\mathcal{L}$. Observe that

$$(4q - 1)(n + 2) = (4q - 1)n + 8q - 2 \geq 2n + 2q = 2d > 2d - (n + 1).$$

The spare vertices are therefore more than enough to apply Lemma 5.1 q times to R and each key K_{i_j} in turn, replacing K_{i_j} with a D-large sphere Z_j that still links R. Then

$$R \cup Z_1 \cup \cdots \cup Z_q$$

is the desired link. □

For completeness' sake we sketch the steps needed to prove Theorem 1.5 from this point. For any missing details see the proof of [13, Theorem 1.4], or the corresponding step of the proof of Proposition 3.1.

Proof of Theorem 1.5. By Corollary 5.2, every embedding of $K^n_{\kappa_n(q)}$ in $\mathbb{R}^{2n+1}$ contains a generalised key ring $R \cup Z_1 \cup \cdots \cup Z_q$ such that each key Z_i is large with respect to a path D of length q. Orient the Z_i so that all linking numbers with R are positive. Working in the homology group $H_n(\mathbb{R}^{2n+1} - R; \mathbb{Z})$, let $1 \leq a \leq b \leq q$ be such that

$$\sum_{i=a}^{b} [Z_i] \equiv 0 \pmod{q},$$

and note that this sum is positive. From now on we restrict our attention to the spheres $Z_a, \ldots, Z_b$.

If $a = b$ we are done. Otherwise, we use the fact that each component Z_i is D-large to construct oriented spheres $F_a, \ldots, F_{b-1}$ such that, for $a \leq i \leq b-1$,

(a) the vertices of F_i lie on $Z_i \cup Z_{i+1}$ (and so F_i is disjoint from R and the rest of the Z_j);

(b) $F_{i-1} \cap Z_i$ and $F_i \cap Z_i$ are disjoint discs, each of which is oppositely oriented by Z_i and F_{i-1} or F_i;

(c) the linking number $\ell k(R, F_i)$ is zero mod q.

The construction of the F_i is identical to that of the corresponding spheres in Proposition 3.1, except that the simpler condition (c) means we only require D to have length q, and the spheres can all be constructed simultaneously instead of inductively. Now if $\ell k(R, F_i)$ is nonzero for some i then $R \cup F_i$ is the required link; and otherwise, we let Z be the connect sum of $Z_a, \ldots, Z_b, F_a, \ldots, F_{b-1}$ obtained by omitting the interiors of the discs $F_i \cap Z_i$ and $F_i \cap Z_{i+1}$ for each i. Then Z is an n-sphere, and in $H_n(\mathbb{R}^{2n+1} - R)$ we have

$$[Z] = \sum_{i=a}^{b} [Z_i] + \sum_{i=a}^{b-1} [F_i] = \sum_{i=a}^{b} [Z_i],$$

which is a nonzero multiple of q. □

References

[1] J. H. Conway and C. McA. Gordon, Knots and links in spatial graphs, *J. Graph Theory* **7** (1983), 445–453, doi:10.1002/jgt.3190070410.

[2] E. Flapan, Intrinsic knotting and linking of complete graphs, *Algebr. Geom. Topol.* **2** (2002), 371–380, doi:10.2140/agt.2002.2.371.

[3] E. Flapan, B. Mellor and R. Naimi, Intrinsic linking and knotting are arbitrarily complex, *Fund. Math.* **201** (2008), 131–148, doi:10.4064/fm201-2-3.

[4] E. Flapan, J. Pommersheim, J. Foisy and R. Naimi, Intrinsically n-linked graphs, *J. Knot Theory Ramifications* **10** (2001), 1143–1154, doi:10.1142/s0218216501001360.

[5] T. Fleming, Intrinsically linked graphs with knotted components, *J. Knot Theory Ramifications* **21** (2012), 1250065 (10 pages), doi:10.1142/s0218216512500654.

[6] T. Fleming and A. Diesl, Intrinsically linked graphs and even linking number, *Algebr. Geom. Topol.* **5** (2005), 1419–1432, doi:10.2140/agt.2005.5.1419.

[7] J. Hoste, The Arf invariant of a totally proper link, *Topology Appl.* **18** (1984), 163–177, doi: 10.1016/0166-8641(84)90008-7.

[8] J. Hoste, The first coefficient of the Conway polynomial, *Proc. Amer. Math. Soc.* **95** (1985), 299–302, doi:10.2307/2044531.

[9] L. H. Kauffman, The Conway polynomial, *Topology* **20** (1981), 101–108, doi:10.1016/0040-9383(81)90017-3.

[10] L. Lovász and A. Schrijver, A Borsuk theorem for antipodal links and a spectral characterization of linklessly embeddable graphs, *Proc. Amer. Math. Soc.* **126** (1998), 1275–1285, doi: 10.1090/s0002-9939-98-04244-0.

[11] H. Sachs, On a spatial analogue of Kuratowski's theorem on planar graphs—an open problem, in: M. Borowiecki, J. W. Kennedy and M. M. Sysło (eds.), *Graph Theory*, Springer, Berlin, volume 1018 of *Lecture Notes in Mathematics*, pp. 230–241, 1983, doi:10.1007/bfb0071633, proceedings of a conference held in Łagów, February 10 – 13, 1981.

[12] K. Taniyama, Higher dimensional links in a simplicial complex embedded in a sphere, *Pacific J. Math.* **194** (2000), 465–467, doi:10.2140/pjm.2000.194.465.

[13] C. Tuffley, Some Ramsey-type results on intrinsic linking of n-complexes, *Algebr. Geom. Topol.* **13** (2013), 1579–1612, doi:10.2140/agt.2013.13.1579.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 349–358
https://doi.org/10.26493/1855-3974.1485.0b1
(Also available at http://amc-journal.eu)

Decomposition method related to saturated hyperball packings

Jenő Szirmai

Budapest University of Technology and Economics, Institute of Mathematics, Department of Geometry, H-1521 Budapest, Hungary

Received 14 September 2017, accepted 28 November 2018, published online 22 January 2019

Abstract

In this paper we study the problem of hyperball (hypersphere) packings in 3-dimensional hyperbolic space. We introduce a new definition of the non-compact saturated ball packings with generalized balls (horoballs, hyperballs) and describe to each saturated hyperball packing, a new procedure to get a decomposition of 3-dimensional hyperbolic space $\mathbb{H}^3$ into truncated tetrahedra. Therefore, in order to get a density upper bound for hyperball packings, it is sufficient to determine the density upper bound of hyperball packings in truncated simplices.

Keywords: Hyperbolic geometry, hyperball packings, Dirichlet-Voronoi cell, packing density, Coxeter tilings.

Math. Subj. Class.: 52C17, 52C22, 52B15

1 Introduction

In n-dimensional hyperbolic space $\mathbb{H}^n$ ($n \geq 2$) there are 3 kinds of generalized “balls” (spheres): the usual balls (spheres), horoballs (horospheres) and hyperballs (hyperspheres).

The classical problems of ball packings and coverings with congruent *generalized balls* of hyperbolic spaces $\mathbb{H}^n$ are extensively discussed in the literature, however there are several essential open questions e.g.:

1. What are the optimal ball packing and covering configurations of *usual spheres* and what are their densities ($n \geq 3$) (see [1, 5, 7, 12])?
2. The monotonicity of the density related to the Böröczky type ball configurations depending on the radius of the congruent balls ($n \geq 4$) (see [4, 10]).

E-mail address: szirmai@math.bme.hu (Jenő Szirmai)

3. What are the optimal horoball packing and covering configurations and what are their densities allowing horoballs in different types ($n \geq 4$) (see [3, 8, 9])?
4. What are the optimal packing and covering arrangements using non-compact balls (horoballs and hyperballs) and what are their densities? These are the so-called hyp-hor packings and coverings (see [21]).
5. What are the optimal hyperball packing and covering configurations and what are their densities ($n \geq 3$)?

In this paper we study the 5th question related to saturated, congruent hyperball packings in 3-dimensional hyperbolic space $\mathbb{H}^3$.

In the hyperbolic plane $\mathbb{H}^2$ the universal upper bound of the hypercycle packing density is $\frac{3}{\pi}$, proved by I. Vermes in [24] and the universal lower bound of the hypercycle covering density is $\frac{\sqrt{12}}{\pi}$ determined by I. Vermes in [25].

In [15] and [16] we studied the regular prism tilings (simply truncated Coxeter orthoscheme tilings) and the corresponding optimal hyperball packings in $\mathbb{H}^n$ $(n = 3, 4)$ and we extended the method developed in the former paper [20] to 5-dimensional hyperbolic space. Moreover, their metric data and their densities have been determined. In paper [19] we studied the n-dimensional hyperbolic regular prism honeycombs and the corresponding coverings by congruent hyperballs and we determined their least dense covering densities. Furthermore, we formulated conjectures for the candidates of the least dense hyperball covering by congruent hyperballs in the 3- and 5-dimensional hyperbolic space $(n \in \mathbb{N}, 3 \leq n \leq 5)$.

In [22] we discussed congruent and non-congruent hyperball (hypersphere) packings of the truncated regular tetrahedron tilings. These are derived from the Coxeter simplex tilings $\{p, 3, 3\}$ $(7 \leq p \in \mathbb{N})$ and $\{5, 3, 3, 3, 3\}$ in 3- and 5-dimensional hyperbolic space. We determined the densest hyperball packing arrangement and its density with congruent hyperballs in $\mathbb{H}^5$ and determined the smallest density upper bounds of non-congruent hyperball packings generated by the above tilings in $\mathbb{H}^n$ $(n = 3, 5)$.

In [21] we deal with packings derived by horo- and hyperballs (briefly hyp-hor packings) in n-dimensional hyperbolic spaces $\mathbb{H}^n$ $(n = 2, 3)$ which form a new class of the classical packing problems. We constructed in the 2- and 3-dimensional hyperbolic spaces hyp-hor packings that are generated by complete Coxeter tilings of degree 1 i.e. the fundamental domains of these tilings are simple frustum orthoschemes and we determined their densest packing configurations and their densities. We proved using also numerical approximation methods that in the hyperbolic plane $(n = 2)$ the density of the above hyp-hor packings arbitrarily approximate the universal upper bound of the hypercycle or horocycle packing density $\frac{3}{\pi}$ and in $\mathbb{H}^3$ the optimal configuration belongs to the $\{7, 3, 6\}$ Coxeter tiling with density ≈ 0.83267. Furthermore, we analyzed the hyp-hor packings in truncated orthoschemes $\{p, 3, 6\}$ $(6 < p < 7,\ p \in \mathbb{R})$ whose density function is attained its maximum for a parameter which lies in the interval $[6.05, 6.06]$ and the densities for parameters lying in this interval are larger that ≈ 0.85397. That means that these locally optimal hyp-hor configurations provide larger densities that the Böröczky-Florian density upper bound (≈ 0.85328) for ball and horoball packings but these hyp-hor packing configurations can not be extended to the entirety of hyperbolic space $\mathbb{H}^3$.

In [23] we studied a large class of hyperball packings in $\mathbb{H}^3$ that can be derived from truncated tetrahedron tilings. In order to get a density upper bound for the above hyperball packings, it is sufficient to determine this density upper bound locally, e.g. in truncated tetrahedra. Thus, we proved that if the truncated tetrahedron is regular, then the density of the densest packing is ≈ 0.86338. This is larger than the Böröczky-Florian density upper bound for balls and horoballs but our locally optimal hyperball packing configuration cannot be extended to the entirety of $\mathbb{H}^3$. However, we described a hyperball packing construction, by the regular truncated tetrahedron tiling under the extended Coxeter group $\{3,3,7\}$ with maximal density ≈ 0.82251.

Recently, (to the best of author's knowledge) the candidates for the densest hyperball (hypersphere) packings in the $3,4$ and 5-dimensional hyperbolic space $\mathbb{H}^n$ are derived by the regular prism tilings that have been published in papers [15, 16] and [20].

In this paper we study hyperball (hypersphere) packings in 3*-dimensional hyperbolic space. We develope a decomposition algorithm that for each saturated hyperball packing provides a decomposition of* $\mathbb{H}^3$ *into truncated tetrahedra. Therefore, in order to get a density upper bound for hyperball packings, it is sufficient to determine the density upper bound of hyperball packings in truncated simplices.*

2 Projective model and saturated hyperball packings in $\mathbb{H}^3$

We use for $\mathbb{H}^3$ (and analogously for $\mathbb{H}^n$, $n \geq 3$) the projective model in the Lorentz space $\mathbb{E}^{1,3}$ that denotes the real vector space $\mathbf{V}^4$ equipped with the bilinear form of signature $(1,3)$,

$$\langle \mathbf{x}, \mathbf{y} \rangle = -x^0y^0 + x^1y^1 + x^2y^2 + x^3y^3,$$

where the non-zero vectors

$$\mathbf{x} = (x^0, x^1, x^2, x^3) \in \mathbf{V}^4 \quad \text{and} \quad \mathbf{y} = (y^0, y^1, y^2, y^3) \in \mathbf{V}^4,$$

are determined up to real factors, for representing points of $\mathcal{P}^n(\mathbb{R})$. Then $\mathbb{H}^3$ can be interpreted as the interior of the quadric $Q = \{(\mathbf{x}) \in \mathcal{P}^3 \mid \langle \mathbf{x}, \mathbf{x} \rangle = 0\} =: \partial\mathbb{H}^3$ in the real projective space $\mathcal{P}^3(\mathbf{V}^4, \boldsymbol{V}_4)$ (here $\boldsymbol{V}_4$ is the dual space of $\mathbf{V}^4$). Namely, for an interior point $\mathbf{y}$ holds $\langle \mathbf{y}, \mathbf{y} \rangle < 0$.

Points of the boundary $\partial\mathbb{H}^3$ in $\mathcal{P}^3$ are called points at infinity, or at the absolute of $\mathbb{H}^3$. Points lying outside $\partial\mathbb{H}^3$ are said to be outer points of $\mathbb{H}^3$ relative to Q. Let $(\mathbf{x}) \in \mathcal{P}^3$, a point $(\mathbf{y}) \in \mathcal{P}^3$ is said to be conjugate to $(\mathbf{x})$ relative to Q if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ holds. The set of all points which are conjugate to $(\mathbf{x})$ form a projective (polar) hyperplane $pol(\mathbf{x}) := \{(\mathbf{y}) \in \mathcal{P}^3 \mid \langle \mathbf{x}, \mathbf{y} \rangle = 0\}$. Thus, the quadric Q induces a bijection (linear polarity $\mathbf{V}^4 \to \boldsymbol{V}_4$) from the points of $\mathcal{P}^3$ onto their polar hyperplanes.

Point $X(\mathbf{x})$ and hyperplane $\alpha(\boldsymbol{a})$ are incident if $\mathbf{x}\boldsymbol{a} = 0$ ($\mathbf{x} \in \mathbf{V}^4 \setminus \{\mathbf{0}\}$, $\boldsymbol{a} \in \boldsymbol{V}_4 \setminus \{\mathbf{0}\}$).

The hypersphere (or equidistance surface) is a quadratic surface at a constant distance from a plane (base plane) in both halfspaces. The infinite body of the hypersphere, containing the base plane, is called hyperball.

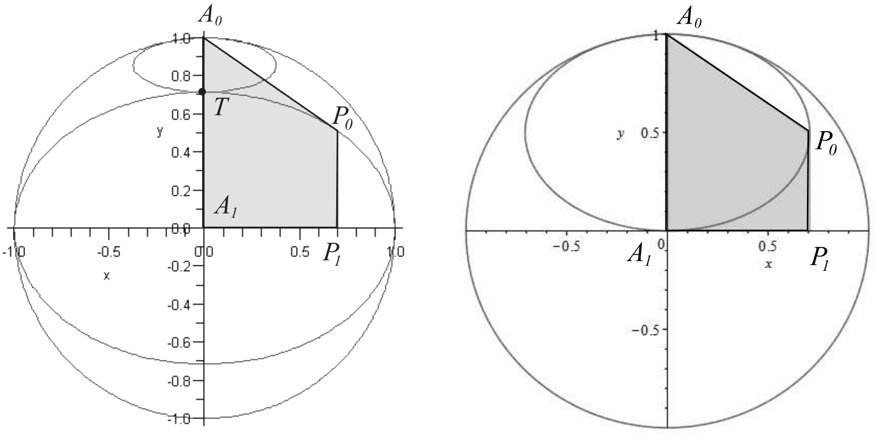
The *half hyperball* with distance h to a base plane β is denoted by $\mathcal{H}^h_+$. The volume of a bounded hyperball piece $\mathcal{H}^h_+(\mathcal{A})$, delimited by a 2-polygon $\mathcal{A} \subset \beta$, and its prism orthogonal to β, can be determined by the classical formula (2.1) of J. Bolyai [2].

$$\mathrm{Vol}(\mathcal{H}^h_+(\mathcal{A})) = \frac{1}{4}\,\mathrm{Area}(\mathcal{A})\left[k \sinh\frac{2h}{k} + 2h\right] \tag{2.1}$$

The constant $k = \sqrt{\frac{-1}{K}}$ is the natural length unit in $\mathbb{H}^3$, where K denotes the constant negative sectional curvature. In the following we may assume that $k = 1$.

Let $\mathcal{B}^h$ be a hyperball packing in $\mathbb{H}^3$ with congruent hyperballs of height h.

The notion of *saturated packing* follows from that fact that the density of any packing can be improved by adding further packing elements as long as there is sufficient room to do so. However, we usually apply this notion for packings with congruent elements. Now, *we modify the classical definition of saturated packing* for non-compact ball packings with generalized balls (horoballs, hyperballs) in n-dimensional hyperbolic space $\mathbb{H}^n$ ($n \geq 2$ integer parameter):



(a) (b)

Figure 1: (a) Saturated hyp-hor packing, at present $a = 0.7$. (b) Saturated horocycle packing with parameter $a = \frac{1}{\sqrt{2}}$.

Definition 2.1. A ball packing with non-compact balls (horoballs or/and hyperballs) in $\mathbb{H}^n$ is saturated if no new non-compact ball can be added to it.

We illustrate the meaning of the above definition by 2-dimensional Coxeter tilings given by the Coxeter symbol $[\infty]$ (see Figure 1), which are denoted by $\mathcal{T}_a$. The fundamental domain of $\mathcal{T}_a$ is a Lambert quadrilateral $A_0A_1P_0P_1$ (see [21]) that is denoted by $\mathcal{F}_a$. It is derived by the truncation of the orthoscheme $A_0A_1A_2$ by the polar line π of the outer vertex A_2. The other initial principal vertex A_0 of the orthoscheme is lying on the absolute quadric of the Beltrami-Cayley-Klein model.

The images of $\mathcal{F}_a$ under reflections on its sides fill the hyperbolic plane $\mathbb{H}^2$ without overlap. The tilings $\mathcal{T}_a$ contain a free parameter $0 < a < 1$, $a \in \mathbb{R}$. The polar straight line of A_2 is π and $\pi \cap A_0A_2 = P_0$, $\pi \cap A_1A_2 = P_1$. If we fix the parameter a then a optimal hypercycle tiling can be derived from the mentioned Coxeter tiling (see Figure 1(a)) but here there are sufficient rooms to add horocycles with centre A_0 and with centres at the images of A_0. This saturated *hyp-hor packing* (packing with horo- and hyperballs) is illustrated in Figure 1(a). The Figure 1(b) shows a saturated horocycle packing belonging to the same Coxeter tiling.

To obtain hyperball (hypersphere) packing bounds it obviously suffices to study saturated hyperball packings (using the above definition) and in what follows we assume that all packings are saturated unless otherwise stated.

3 Decomposition into truncated tetrahedra

We take the set of hyperballs $\{\mathcal{H}_i^h\}$ of a saturated hyperball packing $\mathcal{B}^h$ (see Definition 2.1). Their base planes are denoted by β_i. Thus, in a saturated hyperball packing the distance between two ultraparallel base planes $d(\beta_i, \beta_j)$ is at least $2h$ (where for the natural indices holds $i < j$ and d is the hyperbolic distance function).

In this section we describe a procedure to get a decomposition of 3-dimensional hyperbolic space $\mathbb{H}^3$ into truncated tetrahedra corresponding to a given saturated hyperball packing.

Step 1. The notion of the radical plane (or power plane) of two Euclidean spheres can be extended to the hyperspheres. The radical plane (or power plane) of two non-intersecting hyperspheres is the locus of points at which tangents drawn to both hyperspheres have the same length (so these points have equal power with respect to the two non-intersecting hyperspheres). If the two non-intersecting hyperspheres are congruent also in Euclidean sense in the model then their radical plane coincides with their "Euclidean symmetry plane" and any two congruent hypersphere can be transformed into such an hypersphere arrangement.

Using the radical planes of the hyperballs $\mathcal{H}_i^h$, similarly to the Euclidean space, can be constructed the unique Dirichlet-Voronoi (in short D-V) decomposition of $\mathbb{H}^3$ to the given hyperball packing $\mathcal{B}^h$. Now, the D-V cells are infinite hyperbolic polyhedra containing the corresponding hyperball, and its vertices are proper points of $\mathbb{H}^3$. We note here (it is easy to see), that a vertex of any D-V cell cannot be outer or boundary point of $\mathbb{H}^3$ relative to Q, because the hyperball packing $\mathcal{B}^h$ is saturated by the Definition 2.1.

Step 2. We consider an arbitrary *proper* vertex $P \in \mathbb{H}^3$ of the above D-V decomposition and the hyperballs $\mathcal{H}_i^h(P)$ whose D-V cells meet at P. The base planes of the hyperballs $\mathcal{H}_i^h(P)$ are denoted by $\beta_i(P)$, and these planes determine a non-compact polyhedron $\mathcal{D}^i(P)$ with the intersection of their halfspaces containing the vertex P. Moreover, denote $A_1, A_2, A_3, \dots$ the outer vertices of $\mathcal{D}^i(P)$ and cut off $\mathcal{D}^i(P)$ with the polar planes $\alpha_j(P)$ of its outer vertices A_j. Thus, we obtain a convex compact polyhedron $\mathcal{D}(P)$. This is bounded by the base planes $\beta_i(P)$ and "polar planes" $\alpha_j(P)$. Applying this procedure for all vertices of the above Dirichlet-Voronoi decomposition, we obtain an other decomposition of $\mathbb{H}^3$ into convex polyhedra.

Step 3. We consider $\mathcal{D}(P)$ as a tile of the above decomposition. The planes from the finite set of base planes $\{\beta_i(P)\}$ are called adjacent if there is a vertex A_s of $\mathcal{D}^i(P)$ that lies on each of the above plane. We consider non-adjacent planes $\beta_{k_1}(P), \beta_{k_2}(P), \beta_{k_3}(P), \dots,$ $\beta_{k_m}(P) \in \{\beta_i(P)\}$ ($k_l \in \mathbb{N}^+$, $l = 1, 2, 3, \dots, m$) that have an outer point of intersection denoted by $A_{k_1 \cdots k_m}$. Let $N_{\mathcal{D}(P)} \in \mathbb{N}$ denote the *finite* number of the outer points $A_{k_1 \cdots k_m}$ related to $\mathcal{D}(P)$. It is clear, that its minimum is 0 if $\mathcal{D}^i(P)$ is tetrahedron. The polar plane $\alpha_{k_1 \cdots k_m}$ of $A_{k_1 \cdots k_m}$ is orthogonal to planes $\beta_{k_1}(P), \beta_{k_2}(P), \dots, \beta_{k_m}(P)$ (thus, it contains their poles $B_{k_1}, B_{k_2}, \dots, B_{k_m}$) and divides $\mathcal{D}(P)$ into two convex polyhedra $\mathcal{D}_1(P)$ and $\mathcal{D}_2(P)$.

Step 4. If $N_{\mathcal{D}_1(P)} \neq 0$ and $N_{\mathcal{D}_2(P)} \neq 0$ then $N_{\mathcal{D}_1(P)} < N_{\mathcal{D}(P)}$ and $N_{\mathcal{D}_2(P)} < N_{\mathcal{D}(P)}$ then we apply the Step 3 for polyhedra $\mathcal{D}_i(P)$, $i \in \{1, 2\}$.

Step 5. If $N_{\mathcal{D}_i(P)} \neq 0$ or $N_{\mathcal{D}_j(P)} = 0$ ($i \neq j$, $i, j \in \{1, 2\}$) then we consider the polyhedron $\mathcal{D}_i(P)$ where $N_{\mathcal{D}_i(P)} = N_{\mathcal{D}(P)} - 1$ because the vertex $A_{k_1 \cdots k_m}$ is left out and apply the Step 3.

Step 6. If $N_{\mathcal{D}_1(P)} = 0$ and $N_{\mathcal{D}_2(P)} = 0$ then the procedure is over for $\mathcal{D}(P)$. We continue the procedure with the next cell.

Step 7. It is clear, that the above plane $\alpha_{k_1\cdots k_m}$ intersects every hyperball $\mathcal{H}^h_j(P)$ $(j = k_1, \dots, k_m)$.

Lemma 3.1. *The plane $\alpha_{k_1\cdots k_m}$ of $A_{k_1\cdots k_m}$ does not intersect the hyperballs $\mathcal{H}^h_s(P)$ where $A_{k_1\cdots k_m} \notin \beta_s(P)$.*

Proof. Let $\mathcal{H}^h_s(P)$ $(A_{k_1\cdots k_m} \notin \beta_s(P))$ be an arbitrary hyperball corresponding to $\mathcal{D}(P)$ with base plane $\beta_s(P)$ whose pole is denoted by B_s. The common perpendicular σ of the planes $\alpha_{k_1\cdots k_m}$ and $\beta_s(P)$ is the line through the point $A_{k_1\cdots k_m}$ and B_s. We take a plane κ containing the above common perpendicular, and its intersections with $\alpha_{k_1\cdots k_m}$ and $\mathcal{H}^h_s(P)$ are denoted by ϕ and η. We obtain the plane arrangement illustrated in Figure 2 which coincides with the configuration that is investigated in [24]. There I. Vermes noticed that the straight line $\phi = \alpha_{k_1\cdots k_m} \cap \kappa$ does not intersect the hypercycle $\eta = \mathcal{H}^h_s(P) \cap \kappa$. The plane $\alpha_{k_1\cdots k_m}$ and the hyperball $\mathcal{H}^h_s(P)$ can be generated by rotation of ϕ and η about the common perpendicular σ; therefore, they are disjoint. □

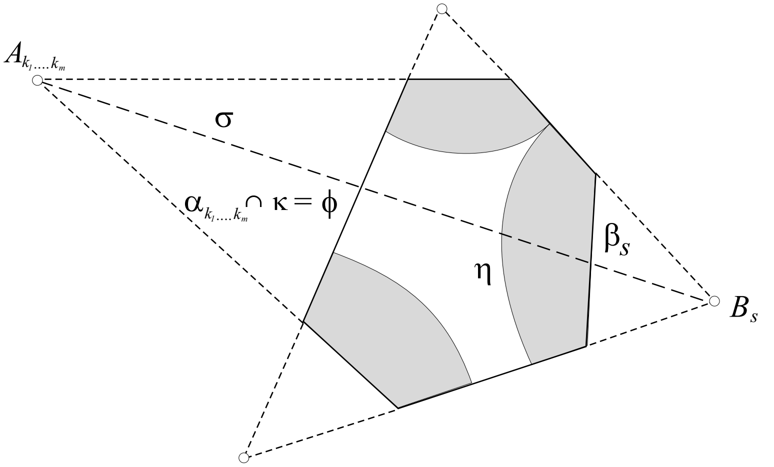


Figure 2: The plane κ and its intersections with $\mathcal{D}(P)$ and $\mathcal{H}^h_s(P)$.

Step 8. We have seen in Steps 3, 4, 5 and 6 that the number of the outer vertices $A_{k_1\cdots k_m}$ of any polyhedron obtained after the cutting process is less than the original one, and we have proven in Step 7 that the original hyperballs form packings in the new polyhedra $\mathcal{D}_1(P)$ and $\mathcal{D}_2(P)$, as well. We continue the cutting procedure described in Step 3 for both polyhedra $\mathcal{D}_1(P)$ and $\mathcal{D}_2(P)$. If a derived polyhedron is a truncated tetrahedron then the cutting procedure does not give new polyhedra, thus the procedure will not be continued. Finally, after a *finite number of cuttings* we get a decomposition of $\mathcal{D}(P)$ into truncated tetrahedra, and in any truncated tetrahedron the corresponding congruent hyperballs from $\{\mathcal{H}^h_i\}$ form a packing. Moreover, we apply the above method for the further cells.

Finally we get the following:

Theorem 3.2. *The above described algorithm provides for each congruent saturated hyperball packing a decomposition of* $\mathbb{H}^3$ *into truncated tetrahedra.* □

The above procedure is illustrated for regular octahedron tilings derived by the regular prism tilings with Coxeter-Schläfli symbol $\{p, 3, 4\}$, $6 < p \in \mathbb{N}$. These Coxeter tilings and the corresponding hyperball packings are investigated in [15]. In this situation the convex polyhedron $\mathcal{D}(P)$ is a truncated octahedron (see Figure 3) whose vertices B_i $(i = 1, 2, 3, 4, 5, 6)$ are outer points and the octahedron is cut off with their polar planes β_i. These planes are the base planes of the hyperballs $\mathcal{H}_i^h$. We can assume that the centre of the octahedron coincides with the centre of the model.

First, we choose three non-adjacent base planes $\beta_2, \beta_3, \beta_4$. Their common point, denoted by A_{234} and its polar plane α_{234} are determined by points B_2, B_3, B_4 containing the centre P as well. Then we consider the non-adjacent base planes $\beta_2, \beta_4, \beta_5$ and the polar plane α_{245} of their common point A_{245}. It is clear that the points B_2, B_4, B_5 lie in the plane α_{245} (see Figure 3).

By the above two "cuttings" we get the decomposition of $\mathcal{D}(P)$ into truncated simplices.

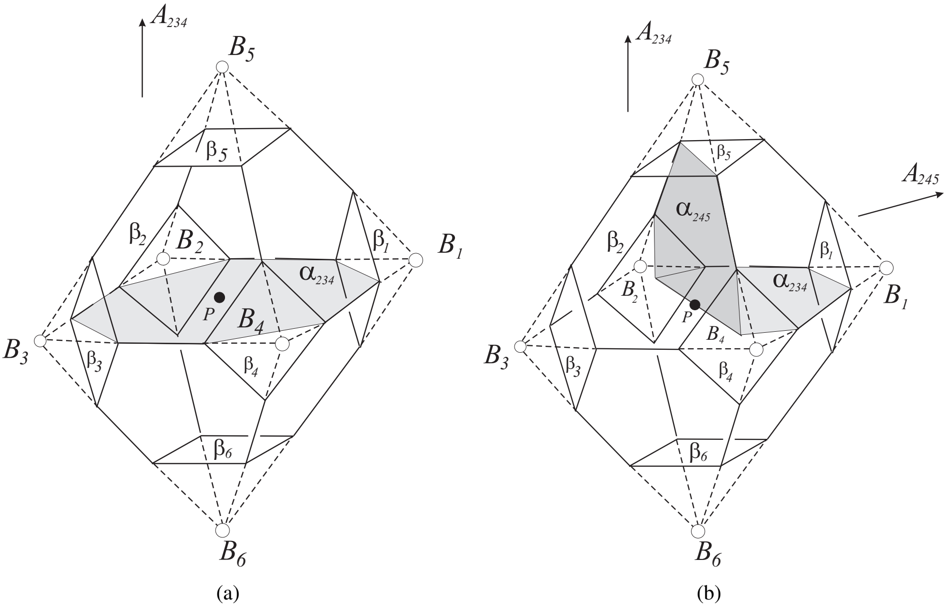


Figure 3: Truncated octahedron tiling derived from the regular prism tilings with Coxeter-Schläfli symbol $\{p, 3, 4\}$ and its decomposition into truncated tetrahedra.

Remark 3.3.

1. If we try to define the density of system of sets in hyperbolic space as we did in Euclidean space, i.e. by the limiting value of the density with respect to a sphere $C(r)$ of radius r with a fixed centre O. But since for a fixed value of h the volume

of spherical shell $C(r+h) - C(r)$ is the same order of magnitude as the volume of $C(r)$, the argument used in Euclidean space to prove that the limiting value is independent of the choice of O is does not work in hyperbolic space. Therefore the definition of packing density is crucial in hyperbolic spaces $\mathbb{H}^n$ as shown by K. Böröczky [3], for nice examples also see [6, 14]. The most widely accepted notion of packing density considers the local densities of balls with respect to their Dirichlet-Voronoi cells (cf. [3] and [7]), but in our cases these cells are infinite hyperbolic polyhedra. The other possibility: the packing density δ can be defined (see [15, 20, 24, 25]) as the reciprocal of the ratio of the volume of a fundamental domain for the symmetry group of a tiling to the volume of the ball pieces contained in the fundamental domain ($\delta < 1$). Similarly is defined the covering density $\Delta > 1$. In order to determine an upper bound for the density of congruent hyperball packings in $\overline{\mathbb{H}}^n$ we used an extended notion of such local density. Therefore, we had to construct a decomposition of $\mathbb{H}^n$ into compact cells to define local density to a given hyperball packing and these corresponding cells are (not absolutely congruent) truncated tetrahedra (see the above algorithm and [23]).

2. From the above section it follows that, to each saturated hyperball packing $\mathcal{B}^h$ of hyperballs $\mathcal{H}_i^h$ there is a decomposition of $\mathbb{H}^3$ into truncated tetrahedra. Therefore, in order to get a density upper bound for hyperball packings, it is sufficient to determine the density upper bound of hyperball packings in truncated simplices. We observed in [23] that some extremal properties of hyperball packings naturally belong to the regular truncated tetrahedron (or simplex, in general, see Lemma 3.2 and Lemma 3.3 in [23]). Therefore, we studied hyperball packings in regular truncated tetrahedra, and prove that if the truncated tetrahedron is regular, then the density of the densest packing is ≈ 0.86338 (see Theorem 5.1 in [23]). However, these hyperball packing configurations are only locally optimal, and cannot be extended to the whole space $\mathbb{H}^3$. Moreover, we showed that the densest known hyperball packing, dually related to the regular prism tilings, introduced in [15], can be realized by a regular truncated tetrahedron tiling with density ≈ 0.82251.

3. In [22] we discussed the problem of congruent and non-congruent hyperball (hypersphere) packings to each truncated regular tetrahedron tiling. These are derived from the Coxeter simplex tilings $\{p,3,3\}$ and $\{5,3,3,3,3\}$ in the 3- and 5-dimensional hyperbolic space. We determined the densest hyperball packing arrangement and its density with congruent hyperballs in $\mathbb{H}^5$ (≈ 0.50514) and determined the smallest density upper bounds of non-congruent hyperball packings generated by the above tilings: in $\mathbb{H}^3$ (≈ 0.82251); in $\mathbb{H}^5$ (≈ 0.50514).

The question of finding the densest hyperball packings and horoball packings with horoballs of different types in the n-dimensional hyperbolic spaces $n \geq 3$ has not been settled yet either (see e.g. [8, 9, 13, 23]).

Optimal sphere packings in other homogeneous Thurston geometries represent another huge class of open mathematical problems. For these non-Euclidean geometries only very few results are known (e.g. [17, 18]). Detailed studies are the objective of ongoing research. The applications of the above projective method seem to be interesting in (non-Euclidean) crystallography as well, a topic of much current interest.

References

[1] K. Bezdek, Sphere packings revisited, *European J. Combin.* **27** (2006), 864–883, doi:10.1016/j.ejc.2005.05.001.

[2] J. Bolyai, Appendix: Scientiam spatii absolute veram exhibens, in: F. Bolyai (ed.), *Tentamen juventutem studiosam in elementa matheseos purae, elementaris ac sublimioris, methodo intuitiva, evidentiaque huic propria, introducendi, Tomus primus*, Maros Vásárhelyini: J. & S. Kali, 1832.

[3] K. Böröczky, Packing of spheres in spaces of constant curvature, *Acta Math. Acad. Sci. Hungar.* **32** (1978), 243–261, doi:10.1007/bf01902361.

[4] K. Böröczky and A. Florian, Über die dichteste Kugelpackung im hyperbolischen Raum, *Acta Math. Acad. Sci. Hungar.* **15** (1964), 237–245, doi:10.1007/bf01897041.

[5] G. Fejes Tóth, G. Kuperberg and W. Kuperberg, Highly saturated packings and reduced coverings, *Monatsh. Math.* **125** (1998), 127–145, doi:10.1007/bf01332823.

[6] G. Fejes Tóth and W. Kuperberg, Packing and covering with convex sets, in: P. M. Gruber and J. M. Wills (eds.), *Handbook of Convex Geometry, Volume B*, North-Holland, Amsterdam, pp. 799–860, 1993, doi:10.1016/c2009-0-15706-9.

[7] R. Kellerhals, Ball packings in spaces of constant curvature and the simplicial density function, *J. Reine Angew. Math.* **494** (1998), 189–203, doi:10.1515/crll.1998.006.

[8] R. T. Kozma and J. Szirmai, Optimally dense packings for fully asymptotic Coxeter tilings by horoballs of different types, *Monatsh. Math.* **168** (2012), 27–47, doi:10.1007/s00605-012-0393-x.

[9] R. T. Kozma and J. Szirmai, New lower bound for the optimal ball packing density in hyperbolic 4-space, *Discrete Comput. Geom.* **53** (2015), 182–198, doi:10.1007/s00454-014-9634-1.

[10] T. H. Marshall, Asymptotic volume formulae and hyperbolic ball packing, *Ann. Acad. Sci. Fenn. Math.* **24** (1999), 31–43, `http://www.acadsci.fi/mathematica/Vol24/marshall.html`.

[11] E. Molnár, The projective interpretation of the eight 3-dimensional homogeneous geometries, *Beiträge Algebra Geom.* **38** (1997), 261–288, `https://www.emis.de/journals/BAG/vol.38/no.2/8.html`.

[12] E. Molnár and J. Szirmai, Top dense hyperbolic ball packings and coverings for complete Coxeter orthoscheme groups, *Publ. Inst. Math. (Beograd)* **103** (2018), 129–146, doi:10.2298/pim1817129m.

[13] L. Németh, On the hyperbolic Pascal pyramid, *Beiträge Algebra Geom.* **57** (2016), 913–927, doi:10.1007/s13366-016-0293-7.

[14] C. Radin, The symmetry of optimally dense packings, in: A. Prékopa and E. Molnár (eds.), *Non-Euclidean Geometries*, Springer, New York, volume 581 of *Mathematics and Its Applications (New York)*, 2006 pp. 197–207, doi:10.1007/0-387-29555-0_10, papers from the International Conference on Hyperbolic Geometry held in Budapest, July 6 – 12, 2002.

[15] J. Szirmai, The regular p-gonal prism tilings and their optimal hyperball packings in the hyperbolic 3-space, *Acta Math. Hungar.* **111** (2006), 65–76, doi:10.1007/s10474-006-0034-8.

[16] J. Szirmai, The regular prism tilings and their optimal hyperball packings in the hyperbolic n-space, *Publ. Math. Debrecen* **69** (2006), 195–207.

[17] J. Szirmai, A candidate for the densest packing with equal balls in Thurston geometries, *Beiträge Algebra Geom.* **55** (2014), 441–452, doi:10.1007/s13366-013-0158-2.

[18] J. Szirmai, Simply transitive geodesic ball packings to $\mathbf{S}^2 \times \mathbf{R}$ space groups generated by glide reflections, *Ann. Mat. Pura Appl.* **193** (2014), 1201–1211, doi:10.1007/s10231-013-0324-z.

[19] J. Szirmai, The least dense hyperball covering of regular prism tilings in hyperbolic n-space, *Ann. Mat. Pura Appl.* **195** (2016), 235–248, doi:10.1007/s10231-014-0460-0.

[20] J. Szirmai, The optimal hyperball packings related to the smallest compact arithmetic 5-orbifolds, *Kragujevac J. Math.* **40** (2016), 260–270, doi:10.5937/kgjmath1602260s.

[21] J. Szirmai, Packings with horo- and hyperballs generated by simple frustum orthoschemes, *Acta Math. Hungar.* **152** (2017), 365–382, doi:10.1007/s10474-017-0728-0.

[22] J. Szirmai, Density upper bound for congruent and non-congruent hyperball packings generated by truncated regular simplex tilings, *Rend. Circ. Mat. Palermo* **67** (2018), 307–322, doi:10.1007/s12215-017-0316-8.

[23] J. Szirmai, Hyperball packings in hyperbolic 3-space, *Mat. Vesnik* **70** (2018), 211–221, `http://www.vesnik.math.rs/vol/mv18303.pdf`.

[24] I. Vermes, Ausfüllungen der hyperbolischen Ebene durch kongruente Hyperzykelbereiche, *Period. Math. Hungar.* **10** (1979), 217–229, doi:10.1007/bf02020020.

[25] I. Vermes, Über reguläre Überdeckungen der Bolyai-Lobatschewskischen Ebene durch kongruente Hyperzykelbereiche, *Period. Polytech. Mech. Engrg.* **25** (1981), 249–261, `https://pp.bme.hu/me/article/view/5842`.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 359–375
https://doi.org/10.26493/1855-3974.1525.7f3
(Also available at http://amc-journal.eu)

A characterization of graphs with disjoint total dominating sets*

Michael A. Henning [†]
*Department of Pure and Applied Mathematics, University of Johannesburg,
Auckland Park, 2006 South Africa*

Iztok Peterin [‡]
*Faculty of Electrical Engineering and Computer Science, University of Maribor,
Koroška 46, 2000 Maribor, Slovenia*

Received 8 November 2017, accepted 11 November 2018, published online 27 January 2019

Abstract

A set S of vertices in a graph G is a total dominating set of G if every vertex is adjacent to a vertex in S. A fundamental problem in total domination theory in graphs is to determine which graphs have two disjoint total dominating sets. In this paper, we solve this problem by providing a constructive characterization of the graphs that have two disjoint total dominating sets. Our characterization gives an entirely new description of graphs with two disjoint total dominating sets and places them in another context, developing them from four base graphs and applies a sequence of operations from seventeen operations that are independent and necessary to produce all such graphs. We show that every graph with two disjoint total dominating sets can be constructed using this method.

Keywords: Total domination number, disjoint total dominating sets.

Math. Subj. Class.: 05C69

[*] The authors express their sincere thanks to the referees for their meticulous and thorough reading of the paper, and for their very helpful comments which improved the exposition and clarity of the revised version of the paper. In particular, we thank one of the reviewers for suggesting to us Claim 4.9 and Claim 4.10 which greatly simplified the original proof.

[†] Research supported in part by the South African National Research Foundation and the University of Johannesburg.

[‡] This author is also affiliated with Institute of Mathematics, Physics and Mechanics, Jadranska 19, 1000 Ljubljana, Slovenia. Research supported in part by the Slovenian Research Agency by the projects No. P1-0297 and No. J1-9109.

E-mail addresses: mahenning@uj.ac.za (Michael A. Henning), iztok.peterin@um.si (Iztok Peterin)

1 Introduction

A *dominating set* of a graph G is a set S of vertices of G such that every vertex not in S has a neighbor in S, where two vertices are neighbors if they are adjacent. A *total dominating set* of a graph G with no isolated vertex is a set S of vertices such that every vertex in G has a neighbor in S. Domination and its variations in graphs are now well studied. The literature on this subject has been surveyed and detailed in the two books by Haynes, Hedetniemi, and Slater [6, 7]. For a recent book on total domination in graphs we refer the reader to [13]. A survey of total domination in graphs can also be found in [9].

A classical result in domination theory, due to Ore [14] in 1962, is that every graph with no isolated vertex has two disjoint dominating sets. However, it is not the case that every graph with no isolated vertex can be partitioned into a dominating set and a total dominating set. Henning and Southey [11] showed that every connected graph with minimum degree at least two that is not a cycle on five vertices has a disjoint dominating set and a total dominating set. Further, in [12] they present a constructive characterization of connected graphs of order at least 4 that have a disjoint dominating set and a total dominating set. Disjoint dominating and total dominating sets in graphs are studied further, for example, in [10]. A characterization of graphs with disjoint dominating and paired-dominating sets is characterized in [15].

It remains, however, an outstanding problem to determine which graphs have two disjoint total dominating sets. Zelinka [16] in 1989 showed that no minimum degree condition in a graph is sufficient to guarantee that there exist two disjoint total dominating sets in the graph. Heggernes and Telle [8] showed that the decision problem to decide for a given graph G if it has two disjoint total dominating sets is NP-complete, even for bipartite graphs. Sufficient conditions for a graph to have two disjoint total dominating sets were obtained by Delgado, Desormeaux, and Haynes [4], but the authors in [4] were not able to characterize such graphs. Cubic graphs that have two disjoint total dominating sets were recently studied by Desormeaux, Henning and Haynes [5]. In particular, they show that cubic graphs that are 5-chordal or claw-free (we do not define these concepts here) can be partitioned into two total dominating sets.

The *total domatic number* $tdom(G)$ of G is the maximum number of disjoint total dominating sets [3]. This can also be considered as a coloring of the vertices such that every vertex has a neighbor of every color (and has been called the *coupon coloring problem* [2]). Recent work on the total domatic number can be found, for example, in [1, 5]. The fundamental problem in total domination theory in graphs of determining which graphs have two disjoint total dominating sets can be phrased as follows: Determine which graphs G satisfy $tdom(G) \geq 2$. We call a graph a *TDP-graph* (standing for "total dominating partitionable graph") if its vertex set can be partitioned into two total dominating sets; that is, a graph G is a TDP-graph if and only if $tdom(G) \geq 2$.

In this paper, we provide a constructive characterization of the graphs that have two disjoint total dominating sets, or, equivalently, a characterization of the TDP-graphs. We describe a procedure to build TDP-graphs in terms of a 2-coloring of the vertices that indicate the role each vertex plays in the sets associated with the two disjoint total dominating sets. We show that the resulting family we construct, starting from four initial base graphs and applying one of seventeen operations to extend graphs in the family to larger graphs, is precisely the class of all TDP-graphs.

Our characterization provides a method for creating the TDP-graphs using a finite set of precise operations. The construction places the TDP-graphs in another context, developing them from four base graphs and applying a sequence of operations from seventeen operations that are independent and necessary to produce a TDP-graph; that is, we show that this method produces precisely the family of TDP-graphs in that every graph generated by this method/algorithm is a TDP-graph and further every TDP-graph can be constructed in this way.

We remark that this procedure does not solve the decision problem to decide if a given graph has two disjoint total dominating sets in polynomial time. If one follows the steps in the proof of Section 4, one does indeed obtain an algorithm for this decision problem. However, this algorithm is far from polynomial time complexity. In particular, the first step of this algorithm is to discard some edges in order to obtain so-called sparse TDP-graph. Unfortunately, the proof does not provide those edges and this already spoils the time complexity.

1.1 Notation

For notation and graph theory terminology we generally follow [13]. All graphs in this paper are finite and simple, without loops or multiple edges. The *order* of a graph G is denoted by $n(G) = |V(G)|$, and the *size* of G by $m(G) = |E(G)|$. We denote the *degree* of a vertex v in the graph G by $d_G(v)$. The maximum (minimum) degree among the vertices of G is denoted by $\Delta(G)$ ($\delta(G)$, respectively). The *open neighborhood* of v is $N_G(v) = \{u \in V(G) \mid uv \in E(G)\}$. For a set $S \subseteq V(G)$, its *open neighborhood* is the set $N_G(S) = \bigcup_{v \in S} N_G(v)$. For subsets X and Y of vertices of G, we denote the set of edges with one end in X and the other end in Y by $[X, Y]$. For a set $S \subseteq V(G)$, the subgraph induced by S is denoted by $G[S]$. Further, the subgraph of G obtained from G by deleting all vertices in S and all edges incident with vertices in S is denoted by $G - S$; that is, $G - S = G[V(G) \setminus S]$. If $S = \{v\}$, we simply denote $G - \{v\}$ by $G - v$.

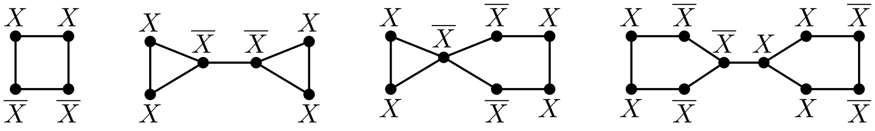
The *distance* between two vertices u and v in G, denoted $d_G(u, v)$, is the minimum length of a (u, v)-path in G. By W_{uv} we denote the set of all vertices of G which are closer to u than to v; that is, $W_{uv} = \{w \mid d_G(w, u) < d_G(w, v)\}$. Symmetrically, W_{vu} is defined. A *block* of a graph G is a maximal connected subgraph of G which has no cut-vertex of its own. A block containing exactly one cut-vertex of G is called an *end-block*. It is well known that any two different blocks of a graph have at most one vertex in common, namely a cut-vertex. Furthermore, a connected graph with at least one cut-vertex has at least two end-blocks. Let X denote the set of cut-vertices of a connected graph G and let Y denote the set of its blocks. The *block graph* of G is a bipartite graph with partite sets X and Y in which a vertex $x \in X$ is adjacent to a vertex $y \in Y$ if x is a vertex of the block y. It is well-known that the block graph of any connected graph is a tree.

A *walk* is a finite, alternating sequence of vertices and edges in which each edge of the sequence joins the vertex that precedes it in the sequence to the vertex that follows it in the sequence. A *non-backtracking walk* is a walk with the additional constraint that no two consecutive edges on the walk are repeated.

Let u be a cut-vertex of a graph G. Let H_1 and H_2 be two vertex disjoint subgraphs of $G - u$ that contain all the components of $G - u$, where each of H_1 and H_2 contain at least one component of $G - u$. We call H_1 and H_2 the *associated subgraphs* of $G - u$. For $i \in [2]$, we denote by H_i^u the subgraph of G induced by $V(H_i) \cup \{u\}$. Further, the vertex in H_1^u named u we rename u', and the vertex in H_2^u named u we rename u'' in order to distinguish between u, u' and u''. We use the standard notation $[k] = \{1, \ldots, k\}$.

2 The graph family $\mathcal{G}$

In this section, we construct a graph family $\mathcal{G}$ such that every graph in the family has two disjoint total dominating sets. First, we define a 2-*coloring* of a graph G as a partition $S = (S_R, S_B)$ of $V(G)$. The *color* of a vertex v, denoted $\mathrm{color}(v)$, is the letter $X \in \{R, B\}$ such that $v \in S_X$, where "R" and "B" here stand for red and blue, respectively. Thus, our 2-coloring of G is a coloring of the vertices of G, one color to each vertex, using the colors red and blue. We denote by $\overline{X}$ the letter $\overline{X} \in \{R, B\} \setminus \{X\}$, and we call $\overline{X}$ the color different from X. Thus, if $X = R$, then $\overline{X} = B$ while if $X = B$, then $\overline{X} = R$. We denote by (G, S) a graph G with a given 2-coloring S. Our aim is to describe a procedure to build TDP-graphs in terms of 2-colorings. For $i \in [4]$, by a 2-*colored* G_i, we shall mean the graph G_i and its associated 2-coloring shown in Figure 1. Further, we call each 2-colored G_i a 2-*colored base graph*.



(a) G_1 (b) G_2 (c) G_3 (d) G_4

Figure 1: The four 2-colored base graphs G_1, G_2, G_3, G_4.

Let $\mathcal{G}$ be the minimum family of 2-colored graphs that:

(i) contains the four 2-colored base graphs; and

(ii) is closed under the seventeen operations $\mathcal{O}_1$ through to $\mathcal{O}_{17}$ listed below, which extend a 2-colored graph (G', S') to a new 2-colored graph (G, S).

In Figures 2 – 7, the vertices of G' are colored black and the new vertices of G are colored white for illustrative purposes, even though the actual colors of the vertices are indicated by the letters X and $\overline{X}$.

Operation $\mathcal{O}_1$: (G, S) is obtained from (G', S') by adding an edge between two nonadjacent vertices of the same color. See the upper diagram of Figure 2.

Operation $\mathcal{O}_2$: (G, S) is obtained from (G', S') by adding an edge between two nonadjacent vertices of different color. See the lower diagram of Figure 2.

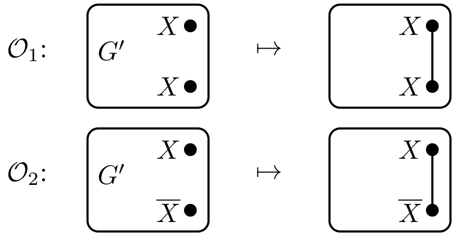


Figure 2: The operations $\mathcal{O}_1$ and $\mathcal{O}_2$.

Operation $\mathcal{O}_3$: If u and v are distinct vertices of different color from (G', S'), then (G, S) is obtained from (G', S') by adding a new vertex of any color adjacent to both u and v. See the left diagram in the upper part of Figure 3.

Operation $\mathcal{O}_4$: If u and v are distinct vertices of the same color from (G', S'), then (G, S) is obtained from (G', S') by adding adjacent vertices x and y and edges ux and vy with $\text{color}(x) = \text{color}(y) \neq \text{color}(u)$. See the middle diagram in the upper part of Figure 3.

Operation $\mathcal{O}_5$: If u and v are distinct vertices of different color from (G', S'), then (G, S) is obtained from (G', S') by adding adjacent vertices x and y and edges ux and vy with $\text{color}(x) = \text{color}(u) \neq \text{color}(y)$. See the right diagram in the upper part of Figure 3.

Operation $\mathcal{O}_6$: If u and v are distinct vertices of the same color from (G', S'), then (G, S) is obtained from (G', S') by adding a path xyz with $\text{color}(y) = \text{color}(z) \neq \text{color}(x) = \text{color}(u)$ and adding edges ux and vz. See the left diagram in the lower part of Figure 3.

Operation $\mathcal{O}_7$: If u and v are distinct vertices of the same color from (G', S'), then (G, S) is obtained from (G', S') by adding a path $xyzw$ and edges ux and vw with $\text{color}(x) = \text{color}(w) = \text{color}(u) \neq \text{color}(y) = \text{color}(z)$. See the middle diagram in the lower part of Figure 3.

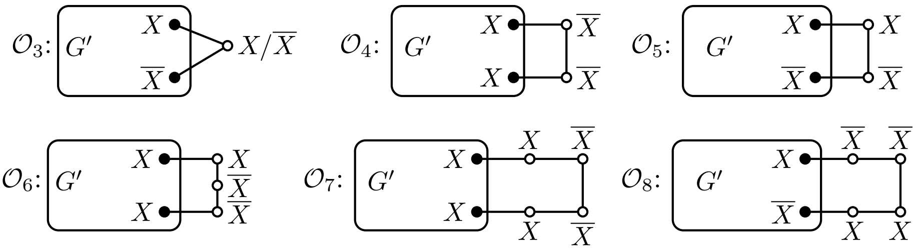


Figure 3: The operations $\mathcal{O}_3 - \mathcal{O}_8$.

Operation $\mathcal{O}_8$: If u and v are distinct vertices of different color from (G', S'), then (G, S) is obtained from (G', S') by adding a path $xyzw$ and edges ux and vw with $\text{color}(x) = \text{color}(y) = \text{color}(v) \neq \text{color}(z) = \text{color}(w)$. See the right diagram in the lower part of Figure 3.

Operation $\mathcal{O}_9$: If u and v are adjacent vertices of different color from (G', S'), then (G, S) is obtained from (G', S') by subdividing uv with four consecutive vertices x, y, z, w where x is adjacent to u and $\text{color}(u) = \text{color}(z) = \text{color}(w) \neq \text{color}(x) = \text{color}(y)$. See the upper diagram of Figure 4.

Operation $\mathcal{O}_{10}$: If u and v are adjacent vertices of the same color from (G', S'), then (G, S) is obtained from (G', S') by subdividing uv with four consecutive vertices x, y, z, w where x is adjacent to u and $\text{color}(u) = \text{color}(x) = \text{color}(w) \neq \text{color}(y) = \text{color}(z)$. See the lower diagram of Figure 4.

Operation $\mathcal{O}_{11}$: If v is a vertex from (G', S'), then (G, S) is obtained from (G', S') by adding an edge xy together with the edges vx and vy where $\text{color}(x) = \text{color}(y) \neq \text{color}(v)$. See the left diagram of Figure 5.

Operation $\mathcal{O}_{12}$: If v is a vertex from (G', S'), then (G, S) is obtained from (G', S') by

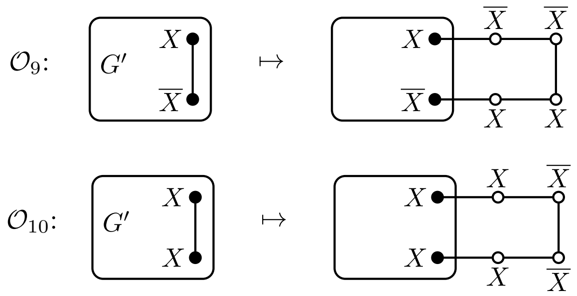


Figure 4: The operations $\mathcal{O}_9$ and $\mathcal{O}_{10}$.

adding a path xyz together with the edges vx and vz where $\text{color}(x) = \text{color}(y) \neq \text{color}(z) = \text{color}(v)$. See the middle diagram of Figure 5.

Operation $\mathcal{O}_{13}$: If v is a vertex from (G', S'), then (G, S) is obtained from (G', S') by adding a path $xyzw$ together with the edges vx and vw where $\text{color}(x) = \text{color}(w) = \text{color}(v) \neq \text{color}(y) = \text{color}(z)$. See the right diagram of Figure 5.

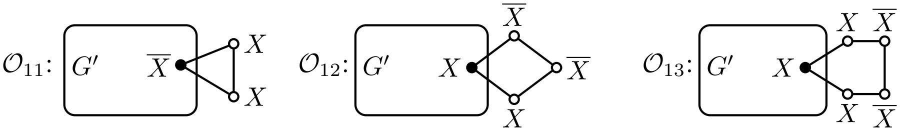


Figure 5: The operations $\mathcal{O}_{11}$, $\mathcal{O}_{12}$ and $\mathcal{O}_{13}$.

Operation $\mathcal{O}_{14}$: If v is a vertex from (G', S'), then (G, S) is obtained from (G', S') by adding a 3-cycle, $xyzx$, together with the edge vx where $\text{color}(x) = \text{color}(v) \neq \text{color}(y) = \text{color}(z)$. See the left diagram of Figure 6.

Operation $\mathcal{O}_{15}$: If v is a vertex from (G', S') of any color, then (G, S) is obtained from (G', S') by adding a 4-cycle, $xyzwx$, together with the edge vx where $\text{color}(x) = \text{color}(y) \neq \text{color}(z) = \text{color}(w)$. See the middle diagram of Figure 6, where the notation $X/\overline{X}$ means that the vertex can have any color.

Operation $\mathcal{O}_{16}$: If v is a vertex from (G', S'), then (G, S) is obtained from (G', S') by adding a 5-cycle, $xyzwtx$, together with the edge vx where $\text{color}(x) = \text{color}(y) = \text{color}(t) \neq \text{color}(z) = \text{color}(w) = \text{color}(v)$. See the right diagram of Figure 6.

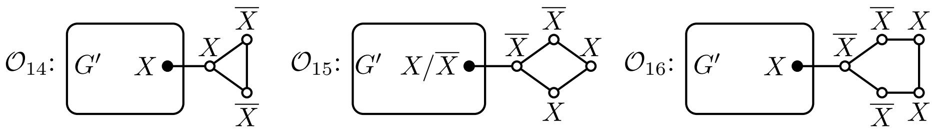


Figure 6: The operations $\mathcal{O}_{14}$, $\mathcal{O}_{15}$ and $\mathcal{O}_{16}$.

Operation $\mathcal{O}_{17}$: If u is a cut-vertex from (G', S') with associated subgraphs H_1^u and H_2^u, and in $N_{H_1^u}(u')$ there exists a vertex of the same color as u and in $N_{H_2^u}(u'')$ there exists a vertex of different color as u, then (G,S) is obtained from H_1^u and H_2^u by adding a new vertex v and the edges $u'v$ and vu''. The color of all vertices from H_1^u remains the same as in G', $\text{color}(v) = \text{color}(u'') \neq \text{color}(u') = \text{color}(u)$ and the color of all vertices from H_2^u is exchanged with respect to their color in G'. See the diagram of Figure 7, where the notation $\overline{A}$ means that the color of all vertices from the set A in (G',S') is changed in (G,S).

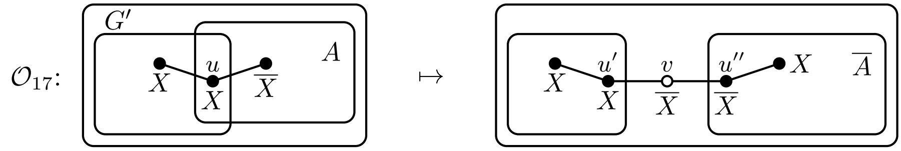


Figure 7: The operation $\mathcal{O}_{17}$.

We remark that, by definition, all operations $\mathcal{O}_3$ to $\mathcal{O}_{17}$ produce new vertices. Further, exactly one new vertex created in each of the operations $\mathcal{O}_{14}$ to $\mathcal{O}_{16}$ has degree 3, and all other new vertices created using operations $\mathcal{O}_3$ to $\mathcal{O}_{17}$ have degree 2 in G. In operations $\mathcal{O}_{11}$ to $\mathcal{O}_{13}$, if the selected vertex v from (G',S') is a cut-vertex of G' it is also a cut-vertex in G, while if v is not a cut-vertex of G' it becomes a cut-vertex in G. Moreover all operations from $\mathcal{O}_{14}$ to $\mathcal{O}_{17}$ produce new cut vertices. In this sense all operations, except $\mathcal{O}_1$ and $\mathcal{O}_2$, can be viewed as base operations which build the sparse skeleton of TDP-graphs, while $\mathcal{O}_1$ and $\mathcal{O}_2$ fill this skeleton with additional edges. This is also the main idea of the proof. First to discard all edges which are there by one of the operations $\mathcal{O}_1$ and $\mathcal{O}_2$, and then study the resulting vertices of degree two.

Lemma 2.1. *If $(G,S) \in \mathcal{G}$ for some 2-coloring $S = (S_R, S_B)$, then G is a TDP-graph. Further, $S = (S_R, S_B)$ is a partition of $V(G)$ into two total dominating sets of G.*

Proof. We proceed by induction on the number, $k \geq 0$, of operations $\mathcal{O}_1$ through $\mathcal{O}_{17}$ used to construct a 2-colored graph $(G,S) \in \mathcal{G}$. If $k = 0$, then (G,S) is one of the four 2-colored base graphs illustrated in Figure 1, and one can readily observe that G is a TDP-graph and $S = (S_R, S_B)$ is a partition of $V(G)$ into two total dominating sets of G. This establishes the base case. Let $k \geq 1$ and suppose that every 2-colored graph $(G',S') \in \mathcal{G}$ that can be constructed using fewer than k operations satisfies the desired result.

Let $(G,S) \in \mathcal{G}$ be a 2-colored graph that can be built from one of the 2-colored base graphs by a sequence of k operations $\mathcal{O}_1 - \mathcal{O}_{17}$. Let $\mathcal{O}_j$ be the last operation of that sequence where $j \in [17]$, and let (G',S') be the graph obtained from the same 2-colored base graph with the same sequence as that used to construct (G,S) but without applying the last operation $\mathcal{O}_j$. Thus, $(G',S') \in \mathcal{G}$ can be constructed using fewer than k operations.

By the induction hypothesis, the graph G' is a TDP-graph and $S' = (S'_A, S'_B)$ is a partition of $V(G')$ into two total dominating sets of G'. If $j \in [2]$, then $S = S'$ and G is a TDP-graph since no new vertices were added. For $3 \leq j \leq 17$ it is a simple exercise to check from the color of the new vertices added to (G',S') when forming (G,S) that the operation $\mathcal{O}_j$ yields two disjoint total domination sets, namely S_R and S_B. Thus, G is a TDP-graph, and $S = (S_R, S_B)$ is a partition of $V(G)$ into two total dominating sets of G. □

3 Main result

Our main result is to provide a constructive characterization of the graphs that have two disjoint total dominating sets, or, equivalently, a characterization of the TDP-graphs.

We prove that the class of all TDP-graphs is precisely the family $\mathcal{G}$ constructed in Section 2. A proof of Theorem 3.1 is given in Section 4.

Theorem 3.1. *A graph G is a TDP-graph if and only if every component of (G, S) is in $\mathcal{G}$ for some 2-coloring S. Further, if $(G, S) \in \mathcal{G}$, then $S = (S_R, S_B)$ is a partition of $V(G)$ into two total dominating sets of G.*

4 Proof of Theorem 3.1

The sufficiency follows from Lemma 2.1. To prove the necessity, let G be a TDP-graph and let $S = (S_R, S_B)$ be a partition of $V(G)$ into two total dominating sets of G. We show that $(G, S) \in \mathcal{G}$ by induction on $m = |E(G)|$. Since G is a TDP-graph, we note that $\delta(G) \geq 2$, G has order $n \geq 4$, and $m \geq 4$. If $m = 4$, then necessarily $G \cong C_4$, and (G, S) is the 2-colored base graph G_1, and so $(G, S) \in \mathcal{G}$. This establishes the base case. Let $m \geq 5$ and assume that every TDP-graph G' of size less than m where $S' = (S'_R, S'_B)$ is a partition of $V(G')$ into two total dominating sets satisfies $(G', S') \in \mathcal{G}$.

Let G be a TDP-graph of order n and size m, and let $S = (S_R, S_B)$ be a partition of $V(G)$ into two total dominating sets of G. If G is disconnected, we apply the inductive hypothesis to each component of G to produce the desired result. Hence, we may assume that G is connected.

Our general strategy in what follows is to reduce the graph G to a TDP-graph G' of size less than m, apply the inductive hypothesis to G' to show that $(G', S') \in \mathcal{G}$, and then reconstruct the graph (G, S) from (G', S') by applying one of the operations $\mathcal{O}_x$, $x \in [17]$, to show that $(G, S) \in \mathcal{G}$. We state this formally, since we will frequently use the following statement.

Statement 4.1. *If G' is a TDP-graph of size less than m, where $S' = (S'_R, S'_B)$ is a partition of $V(G')$ into two total dominating sets, and (G, S) can be constructed from (G', S') by applying one of the operations $\mathcal{O}_x$, where $x \in [17]$, then $(G, S) \in \mathcal{G}$.*

We define three graphs G_R, G_B and G_{RB} associated with the graph G and the partition $S = (S_R, S_B)$. Let G_R and G_B be the subgraphs of G induced by the sets S_R and S_B, respectively, and so $G_R = G[S_R]$ and $G_B = G[S_B]$. Let G_{RB} be the (spanning) subgraph of G with $V(G_{RB}) = V(G)$ and $E(G_{RB}) = E(G) \setminus (E(G_R) \cup E(G_B))$.

Claim 4.2. *If some component of G_R, G_B or G_{RB} is not a star, then $(G, S) \in \mathcal{G}$.*

Proof. Suppose that there exists a component, C, of G_R, G_B or G_{RB} which is not a star. If C contains a cycle $v_1 \dots v_k v_1$, $k \geq 3$, then G can be obtained from $G' = G - v_1v_2$ by either applying operation $\mathcal{O}_1$ in the case when C is a component of G_R or G_B or by applying operation $\mathcal{O}_2$ in the case when C is a component of G_{RB}. If C contains no cycle, then C is a tree different from a star. Therefore, there exists a path $u_1u_2u_3u_4$ in C and G can be obtained from $G' = G - u_2u_3$ by either applying operation $\mathcal{O}_1$ in the case when C is a component of G_R or G_B or by applying operation $\mathcal{O}_2$ in the case when C is a component of G_{RB}. In all cases, since $S = (S_R, S_B)$ is a partition of $V(G)$ into two total dominating sets of G, the same partition $S' = S = (S_R, S_B)$ is a partition of $V(G')$ into two total dominating sets of G'. By the inductive hypothesis, $(G', S') \in \mathcal{G}$. We can obtain G from the same 2-colored base graph as G' and the same sequence of operations from $\mathcal{O}_1 - \mathcal{O}_{17}$ used to construct (G', S') by adding at the end of this sequence the operation $\mathcal{O}_1$ or $\mathcal{O}_2$. Hence $(G, S) \in \mathcal{G}$. □

By Claim 4.2, we may assume that every component of G_R, G_B or G_{RB} is a star, for otherwise the desired result follows. We call the resulting graph G a *sparse TDP-graph with associated partition* $S = (S_R, S_B)$.

We now partition the sets S_R and S_B in two different ways depending on the role that the vertices in S_R and S_B, respectively, play in the graphs G_R, G_B and G_{RB}. First, let $S_R = R_1 \cup R_2 \cup R_3$ and $S_B = B_1 \cup B_2 \cup B_3$ where

$$\begin{aligned} R_1 &= \{v \in S_R \mid d_{G_R}(v) \geq 2\} \\ R_2 &= \{v \in S_R \setminus R_1 \mid N_G(v) \cap R_1 \neq \emptyset\} \\ R_3 &= S_R \setminus (R_1 \cup R_2) \end{aligned}$$

and

$$\begin{aligned} B_1 &= \{v \in S_B \mid d_{G_B}(v) \geq 2\} \\ B_2 &= \{v \in S_B \setminus B_1 \mid N_G(v) \cap B_1 \neq \emptyset\} \\ B_3 &= S_B \setminus (B_1 \cup B_2). \end{aligned}$$

Next, we define a partition of $V(G) = V(G_{RB})$ as the union of the two partitions $S_R = R_1B \cup R_2B \cup R_3B$ and $S_B = RB_1 \cup RB_2 \cup RB_3$ where

$$\begin{aligned} R_1B &= \{v \in S_R \mid d_{G_{RB}}(v) \geq 2\} \\ R_2B &= \{v \in S_R \setminus R_1B \mid v \text{ has a neighbor in } G_{RB} \text{ that belongs to } RB_1\} \\ R_3B &= S_R \setminus (R_1B \cup R_2B) \end{aligned}$$

and

$$\begin{aligned} RB_1 &= \{v \in S_B \mid d_{G_{RB}}(v) \geq 2\} \\ RB_2 &= \{v \in S_B \setminus RB_1 \mid v \text{ has a neighbor in } G_{RB} \text{ that belongs to } R_1B\} \\ RB_3 &= S_B \setminus (RB_1 \cup RB_2). \end{aligned}$$

We note that every vertex in R_3 has degree 1 in G_R, and every vertex in R_3B has degree 1 in G_{RB}. Analogously, every vertex in B_3 and RB_3 has degree 1 in G_B and G_{RB}, respectively. In particular, vertices from $R_3 \cap R_3B$ and from $B_3 \cap RB_3$ have degree 2 in G. Further, the neighbor of a vertex from R_3 in G_R belongs to R_3, and, analogously, the neighbor of a vertex from B_3 in G_B belongs to B_3. We proceed further with the following series of structural properties of the graph G.

Claim 4.3. $\delta(G) = 2$.

Proof. Recall that G is a sparse TDP-graph with associated partition $S = (S_R, S_B)$. Thus, S_R and S_B are disjoint total dominating sets of G which form a partition of $V(G)$. Every vertex $v \in V(G)$ has at least one neighbor in S_R and at least one neighbor in S_B. Hence, $\delta(G) \geq 2$. Suppose, to the contrary, that $\delta(G) > 2$.

Suppose that $R_1B \neq \emptyset$ and let $v \in R_1B$. Let $v_1, \ldots, v_k$, where $k \geq 2$, be the neighbors of v in G_{RB}. By Claim 4.2 and the definition of the set RB_2, we note that for each $i \in [k]$, $v_i \in RB_2$ and the vertex v is the only neighbor of v_i that belongs to the set S_R. Further, since $d_G(v_i) > 2$, the vertex v_i has at least two neighbors in S_B. By Claim 4.2, every component of the graph G_B is a star, implying that no two neighbors of v are adjacent or have a common neighbor in G_B. Further, every neighbor of v_i in G different from v belongs to the set B_2, and has the vertex v_i as its only neighbor in G_B. Thus, the set B_2 contains at least $2k$ vertices at distance 2 from v in G.

For $i \in [k]$, let w_i denote an arbitrary neighbor of v_i in G_B, and so $w_i \in B_2$. Since $d_G(w_i) > 2$ and w_i has only one neighbor in S_B, namely the vertex v_i, we note that $w_i \in RB_1$ and therefore w_i has at least two neighbors in R_2B. By Claim 4.2 and the definition of the set R_2B, we note that every neighbor of w_i different from v_i belongs to the set R_2B. Further, each such neighbor of w_i has exactly one neighbor that belongs to the set S_B, namely the vertex w_i, and therefore has at least two neighbors in S_R by the minimum degree condition. By Claim 4.2, every component of the graph G_R is a star, and therefore two distinct vertices of degree at least 2 in G_R belong to different components of G_R. This implies that this subset R_2B of vertices in S_R contains at least $4k$ vertices.

By the minimum degree condition, these vertices in R_2B also belong to R_1 and each of them has at least two neighbors in R_2. Further, analogously as before, no two such vertices are the same, implying that this subset of R_2 contains at least $8k - 1$ vertices distinct from v, all of which belong to the set R_1B, noting that one of these vertices may possibly be the vertex v. By repeating this process for all these vertices we see that we have an infinite process with infinite growth, which is not possible in a finite graph G. Therefore, the set $R_1B = \emptyset$. Analogously, the set $RB_1 = \emptyset$. Therefore, R_2B and RB_2 are also empty.

We now consider a vertex $v \in R_3B$. By Claim 4.2, every component of the graph G_{RB} is a star, implying that the vertex v has exactly one neighbor in S_B and, by the minimum degree condition, at least two neighbors in S_R. Thus, $v \in R_1$ and each neighbor of v in S_R belong to R_2. Further, by Claim 4.2, each such neighbor of v in R_2 has degree 1 in G_R and, therefore, by the minimum degree condition, has at least two neighbors in S_B. Thus, every neighbor of v in R_2 belongs to the set R_1B, contradicting our earlier observation that the set R_1B is an empty set. This completes the proof of Claim 4.3. □

By Claim 4.3, every sparse TDP-graph has minimum degree 2. In particular, $\delta(G) = 2$. Let $D = \{v \in V(G) \mid d_G(v) = 2\}$.

Claim 4.4. *If a vertex in D is a cut-vertex of G, then $(G, S) \in \mathcal{G}$.*

Proof. Suppose that a vertex in D is a cut-vertex of G. Suppose firstly that D contains two adjacent vertices, x and y, that are both cut-vertices of G, and let $e = xy$. Let C_x and C_y be the components of $G - e$ which contain x and y, respectively. Further, let x' be the neighbor of x in C_x and let y' be the neighbor of y in C_y. We have two possibilities with respect to the color of the vertices x, x', y, y'. Either $\text{color}(x') = \text{color}(x) \neq \text{color}(y) = \text{color}(y')$ or $\text{color}(x') = \text{color}(y') \neq \text{color}(x) = \text{color}(y)$. In both cases, let G' be the graph obtained from $G - \{x, y\}$ by adding the edge $x'y'$, and changing the color of all vertices in $V(C_y) \setminus \{y\}$ while retaining the color of all vertices in $V(C_x) \setminus \{x\}$. Let $S' = (S'_R, S'_B)$ be the resulting partition of $V(G')$. We note that G' is a TDP-graph, where $S' = (S'_R, S'_B)$ is a partition of $V(G')$ into two total dominating sets and that x' and y' are cut vertices of G'. If x and x' have the same color in G, then we use Statement 4.1 with the operation $\mathcal{O}_{17}$

and the cut vertex y' to show that $(G, S) \in \mathcal{G}$, while if x and x' have different color in G, we use Statement 4.1 with the operation $\mathcal{O}_{17}$ and the cut vertex x'.

Thus, we may assume that no two adjacent vertices of D are both cut-vertices of G. Let v be a cut-vertex of G that belongs to D with neighbors u' and u''. Without loss of generality we may assume that $\text{color}(v) = \text{color}(u'') \neq \text{color}(u')$. Let $C_{u'}$ and $C_{u''}$ be the components of $G - v$ containing u' and u'', respectively. Since $S = (S_R, S_B)$ is a partition of $V(G)$ into two total dominating sets of G, there exists a neighbor of u' in $C_{u'}$ of the same color as u' and a neighbor of u'' in $C_{u''}$ whose color is different from that of u''. Let G' be the graph obtained from $G - v$ by identifying the vertices u' and u'' into one new vertex u, and joining u to every neighbor of u' and u''. Further, we assign to u the same color as that of u', while we change the color of all vertices in $V(C_{u''}) \setminus \{u''\}$ and retain the color of all vertices in $V(C_{u'}) \setminus \{u'\}$. Let $S' = (S'_R, S'_B)$ be the resulting partition of $V(G')$. We note that G' is a TDP-graph, where $S' = (S'_R, S'_B)$ is a partition of $V(G')$ into two total dominating sets. We now use Statement 4.1 with the operation $\mathcal{O}_{17}$ to show that $(G, S) \in \mathcal{G}$, where $H_1^u = C_{u'}$ and $H_2^u = C_{u''}$. □

By Claim 4.4, we may assume that no vertex in D is a cut-vertex of G, for otherwise the desired result follows. We note that every vertex in D has one neighbor in S_R and one neighbor in S_B. Further, every component in $G[D]$ is a path or a cycle.

Claim 4.5. *Let C be a component of $G[D]$. If C is a cycle or if C is a path of order at least 5 or if C is a path of order 4 and the ends of C do not have a common neighbor, then $(G, S) \in \mathcal{G}$.*

Proof. Suppose that C is a cycle. Since G is a connected TDP-graph, this implies that $G \cong C_n$ where $n \equiv 0 \pmod 4$. In this case, G can be obtained from the 2-colored base graph G_1 by repeated applications of operation $\mathcal{O}_9$ (or operation $\mathcal{O}_{10}$). Hence, we may assume that C is a path, for otherwise the desired result follows. Let C be the path $x_1 \dots x_k$, where $k \geq 4$. Let u be the neighbor of x_1 not on C. If $k \geq 5$, let $v = x_5$, while if $k = 4$, let v be the neighbor of x_4 not on C. By assumption, $u \neq v$. Let $X = \{x_1, x_2, x_3, x_4\}$.

Suppose first that $\text{color}(u) = \text{color}(x_1)$, implying that $\text{color}(x_2) = \text{color}(x_3) \neq \text{color}(x_4) = \text{color}(v) = \text{color}(x_1)$. If u and v are adjacent in G, let $G' = G - X$. In this case, the graph G' is a TDP-graph and we use Statement 4.1 with the operation $\mathcal{O}_7$ to show that $(G, S) \in \mathcal{G}$. If u and v are not adjacent in G, let G' be obtained from $G - X$ by adding the edge uv. Once again, the graph G' is a TDP-graph. We use Statement 4.1 with the operation $\mathcal{O}_{10}$ to show that $(G, S) \in \mathcal{G}$.

Suppose next that $\text{color}(u) \neq \text{color}(x_1)$, implying that $\text{color}(x_2) = \text{color}(v) \neq \text{color}(x_3) = \text{color}(x_4) = \text{color}(u)$. If u and v are adjacent in G, let $G' = G - X$. In this case, the graph G' is a TDP-graph and we use Statement 4.1 with the operation $\mathcal{O}_8$ to show that $(G, S) \in \mathcal{G}$. If u and v are not adjacent in G, let G' be obtained from $G - X$ by adding the edge uv. Once again, the graph G' is a TDP-graph. We use Statement 4.1 with the operation $\mathcal{O}_9$ to show that $(G, S) \in \mathcal{G}$. □

By Claim 4.5, we may assume that every component of $G[D]$ is a path-component of order at most 4, and that the ends of a path-component of $G[D]$ of order 4 have a common neighbor in G. In what follows we adopt the following notation. Let P be a path-component of $G[D]$, and so $P \cong P_k$ for some $k \in [4]$. Let P be the path $x_1 \dots x_k$, and let u and v be the vertices in G that do not belong to P and are adjacent to x_1 and x_k, respectively. We call u and v the vertices in $G - V(P)$ associated with the path P. By assumption, if $k = 4$, then $u = v$. We note that if $k = 1$, then $u \neq v$. We define next a good path-component.

Definition 4.6. A path-component P of $G[D]$ is a *good path-component* if $P \cong P_k$ where $k \in [3]$, and both u and v have neighbors of both colors in the graph $G^- = G - V(P)$, where u and v are the vertices in G^- associated with P.

Claim 4.7. *If $G[D]$ contains a good path-component, then $(G, S) \in \mathcal{G}$.*

Proof. Suppose that $G[D]$ contains a good path-component, $P\colon x_1 \dots x_k$. By definition, $k \in [3]$. Suppose that $k = 1$. Since P is a good path-component, the graph $G' = G - x_1$ is a TDP-graph. Furthermore, $\operatorname{color}(u) \neq \operatorname{color}(v)$ since G is a TDP-graph. We now use Statement 4.1 with the operation $\mathcal{O}_3$ to show that $(G, S) \in \mathcal{G}$.

Suppose that $k = 2$. Suppose that $\operatorname{color}(x_1) = \operatorname{color}(x_2)$. Then, $\operatorname{color}(u) \neq \operatorname{color}(x_1)$ and either $u = v$ or $u \neq v$ and $\operatorname{color}(u) = \operatorname{color}(v)$. In both cases, since P is a good path-component, the graph $G' = G - V(P)$ is a TDP-graph. If $u = v$, we use Statement 4.1 with the operation $\mathcal{O}_{11}$ to show that $(G, S) \in \mathcal{G}$, while if $u \neq v$, we use Statement 4.1 with the operation $\mathcal{O}_4$ to show that $(G, S) \in \mathcal{G}$. Suppose that $\operatorname{color}(x_1) \neq \operatorname{color}(x_2)$. Then, $\operatorname{color}(u) = \operatorname{color}(x_1)$ and $\operatorname{color}(v) = \operatorname{color}(x_2)$. Since P is a good path-component, the graph $G' = G - V(P)$ is a TDP-graph, and we use Statement 4.1 with the operation $\mathcal{O}_5$ to show that $(G, S) \in \mathcal{G}$.

Suppose that $k = 3$. Without loss of generality we may assume that $\operatorname{color}(x_1) \neq \operatorname{color}(x_2) = \operatorname{color}(x_3)$, implying that $\operatorname{color}(u) = \operatorname{color}(x_1)$ and either $u = v$ or $u \neq v$ and $\operatorname{color}(u) = \operatorname{color}(v)$. Since P is a good path-component, the graph $G' = G - V(P)$ is a TDP-graph. If $u = v$, we use Statement 4.1 with the operation $\mathcal{O}_{12}$ to show that $(G, S) \in \mathcal{G}$, while if $u \neq v$, we use Statement 4.1 with the operation $\mathcal{O}_6$ to show that $(G, S) \in \mathcal{G}$. □

By Claim 4.7, we may assume that G contains no good path-component, for otherwise the desired result follows. We define next an end-block path component of $G[D]$.

Definition 4.8. A path-component P of $G[D]$ with associated vertices u and v is an *end-block path component* of $G[D]$ if $u = v$.

We are now in a position to present the following property of non-backtracking walks in the graph G.

Claim 4.9. *Suppose that $W\colon w_1 w_2 \dots w_k$ is a non-backtracking walk in G and no vertex of W belongs to an end-block path component of $G[D]$. If w_2 is not the only neighbor of w_1 in G whose color is $\operatorname{color}(w_2)$, then w_{i-1} is the only neighbor of w_i in G whose color is $\operatorname{color}(w_{i-1})$ for all $i \in [k] \setminus \{1\}$.*

Proof. Since W is a non-backtracking walk in G, we note that no two consecutive edges on W are equal; that is, $w_{i-1} \neq w_{i+1}$ for all $i \in [k-1] \setminus \{1\}$. Suppose, to the contrary, that the claim is false. Let $\ell \geq 2$ be the smallest integer such that the vertex w_ℓ has a neighbor different from $w_{\ell-1}$ of the same color as $w_{\ell-1}$.

Claim 4.9.1. $\ell \geq 3$.

Proof. Renaming colors if necessary, we may assume that $\text{color}(w_1) = X$. By supposition, at least one neighbor, say v_1, of w_1 different from w_2 has the same color as w_2. Suppose firstly that $\text{color}(w_2) = X$. By supposition, $\text{color}(v_1) = X$. If w_2 has a neighbor, z_2 say, different from w_1, of color X, then either $v_1 = z_2$, in which case $v_1w_1w_2v_1$ is a 3-cycle in G_X, or $v_1 \neq z_2$, in which case $v_1w_1w_2z_2$ is a path P_4 in G_X. Both cases produce a contradiction. Suppose secondly that $\text{color}(w_2) = \overline{X}$. By supposition, $\text{color}(v_1) = \overline{X}$. If w_2 has a neighbor, z_2 say, different from w_1, of color X, then $v_1w_1w_2z_2$ is a path P_4 in G_{RB}, a contradiction. We deduce, therefore, that w_1 is the only neighbor of w_2 whose color is $\text{color}(w_1)$. Hence, $\ell \geq 3$. □

By Claim 4.9.1, we have that $\ell \geq 3$. Renaming colors if necessary, we may assume that $\text{color}(w_{\ell-1}) = X$. By supposition, the vertex w_ℓ has a neighbor, $v_{\ell+1}$ say, different from $w_{\ell-1}$ of the same color as $w_{\ell-1}$; that is, $\text{color}(v_{\ell+1}) = X$. Further since G is a TPD-graph, the vertex w_ℓ has a neighbor of color $\overline{X}$.

Claim 4.9.2. $d_G(w_{\ell-1}) = 2$.

Proof. Suppose that $d_G(w_{\ell-1}) \geq 3$. Let v_ℓ be a neighbor of $w_{\ell-1}$ different from $w_{\ell-2}$ and w_ℓ. Suppose that $\text{color}(w_{\ell-2}) = X$. By the minimality of ℓ, the vertex $w_{\ell-2}$ is the only neighbor of $w_{\ell-1}$ whose color is $\text{color}(w_{\ell-2})$; that is, all neighbors of $w_{\ell-1}$ different from $w_{\ell-2}$ must have color $\overline{X}$. In particular, $\text{color}(v_\ell) = \text{color}(w_\ell) = \overline{X}$. Hence, $v_\ell w_{\ell-1} w_\ell v_{\ell+1}$ is a path P_4 in G_{RB}, a contradiction. Hence, $\text{color}(w_{\ell-2}) = \overline{X}$. Thus, all neighbors of $w_{\ell-1}$ different from $w_{\ell-2}$ must have color X. In particular, $\text{color}(v_\ell) = \text{color}(w_\ell) = X$. If $v_\ell = v_{\ell+1}$, then $v_\ell w_{\ell-1} w_\ell v_\ell$ is a 3-cycle in G_X, a contradiction. If $v_\ell \neq v_{\ell+1}$, then $v_\ell w_{\ell-1} w_\ell v_{\ell+1}$ is a path P_4 in G_X, a contradiction. □

By Claim 4.9.2, the vertex $w_{\ell-1}$ has degree 2 in G; that is, $w_{\ell-1} \in D$. By supposition, the vertex w_1 has at least two neighbors whose color is $\text{color}(w_2)$ and at least one vertex whose color is different from $\text{color}(w_2)$. In particular, the vertex w_1 has degree at least 3 in G. Let $p \geq 1$ be the largest integer such that $d_G(w_p) \geq 3$ and $p \leq \ell-2$. Possibly, $p = \ell-2$. We now consider the path $P\colon w_{p+1} \dots w_{\ell-1}$ and note that P is a path-component in $G[D]$. If $w_p = w_\ell$, then P is an end-block path component of $G[D]$, contradicting the supposition that no vertex of W belongs to an end-block path component of $G[D]$. Hence, $w_p \neq w_\ell$ and the vertices w_p and w_ℓ associated with the path-component P in $G[D]$ are distinct vertices.

We now consider the graph $G^- = G - V(P)$. By our earlier observations, the vertex w_ℓ has neighbors of both colors in G^-. If $p = 1$, then the vertex w_p has neighbors of both colors in G^-. If $p \geq 2$, then by the minimality of ℓ the vertex w_p once again has neighbors of both colors in G^-. Thus the path P is a good-path component, contradicting our earlier assumption that G contains no good path-component. This completes the proof of Claim 4.9. □

Claim 4.10. *If G contains a cycle that is not an end-block of G, then $(G, S) \in \mathcal{G}$.*

Proof. Assume that some cycle C in G is not an end-block in G. Let P be a path-component of $G[D]$ with associated vertices u and v. Suppose firstly that $u = v$. Thus, P is an end-block path component of $G[D]$ and $C_P = G[V(P) \cup \{u\}]$ is a cycle in G. Further, C_P is an end-block of G with u as its cut-vertex in G. Suppose that $d_G(u) \geq 4$. We now consider the graph $G^- = G - V(P)$. By our earlier assumptions, no vertex in D is a cut-vertex of G, implying that G^- is a connected graph.

Claim 4.10.1. *The vertex u has neighbors of both colors in G^-.*

Proof. Suppose, to the contrary, that all neighbors of u in G^- have the same color. By supposition, there is a cycle C in G^- that contains no vertex that belongs to an end-block component of $G[D]$. Hence there exists a non-backtracking walk $W\colon w_1w_2 \dots w_k$ in G that starts at the vertex u, proceeds from u to C, goes all the way around C, and then returns to u, without entering any end-block path component of $G[D]$. We note that $k \geq 3$ and that $w_1 = w_k = u$. By our supposition that all neighbors of u in G^- have the same color, the vertex w_2 is not the only neighbor of w_1 in G whose color is $\mathrm{color}(w_2)$. By Claim 4.9, the vertex w_{k-1} is the only neighbor of w_k in G whose color is $\mathrm{color}(w_{k-1})$. This contradicts our supposition that all neighbors of u in G^- have the same color. □

By Claim 4.10.1, the vertex u has neighbors of both colors in G^-. Since G is a TDP-graph, this implies that the graph G^- is a TDP-graph. Hence, we can use Statement 4.1 with the operation $\mathcal{O}_{11}$ or $\mathcal{O}_{12}$ or $\mathcal{O}_{13}$, depending on the length of P, to show that $(G, S) \in \mathcal{G}$. We may therefore assume that $d_G(u) = 3$ (and still $u = v$), for otherwise $(G, S) \in \mathcal{G}$, as desired. Thus, the vertex u has degree 1 in G^-. Let x be the neighbor of u in G^-. By our earlier assumptions, no vertex in D is a cut-vertex of G. In particular, the cut-vertex x does not belong to D, and so $d_G(x) \geq 3$. We now consider the (connected) graph $G_u^- = G^- - u$ obtained from G^- by deleting the vertex u. Using analogous arguments as in the proof of Claim 4.10.1, the vertex x has neighbors of both colors in G_u^-. Hence, we can use Statement 4.1 with the operation $\mathcal{O}_{14}$ or $\mathcal{O}_{15}$ or $\mathcal{O}_{16}$, depending on the length of P, to show that $(G, S) \in \mathcal{G}$.

Suppose next that $u \neq v$. Using analogous arguments as in the proof of Claim 4.10.1, the vertices u and v each have neighbors of both colors in G_u^-. Thus the path P is a good-path component, contradicting our earlier assumption that G contains no good path-component. This completes the proof of Claim 4.10. □

By Claim 4.10, we may assume that every cycle in G is an end-block of G, for otherwise $(G, S) \in \mathcal{G}$ as desired. Every block of G that is not an end-block is a copy of K_2 consisting of a single edge. By our earlier assumptions, every cycle in G has length 3, 4 or 5. Let T^- be the graph obtained from G by deleting all vertices that belong to an end-block path component of $G[D]$. By our earlier assumptions, the graph T^- is a tree. In particular, every vertex of T^- is a cut-vertex of G. By our earlier assumptions, no vertex in D is a cut-vertex of G, implying that every vertex of D belongs to an end-block path component of $G[D]$. Hence, every vertex of D belongs to an end-block of G.

Claim 4.11. *If two cycles of G intersect, then $(G, S) \in \mathcal{G}$.*

Proof. Suppose that two different cycles C_1 and C_2 of G intersect. Since every cycle in G is an end-block of G, these two cycles intersect in exactly one common vertex, v say.

Claim 4.11.1. *If G contains exactly one cut-vertex, then (G, S) is a 2-colored base graph G_3.*

Proof. Suppose that G contains exactly one cut-vertex. Since the cut-vertices of G are precisely the vertices in the tree T^-, this implies that $V(T^-) = \{v\}$. Thus, every block of G is an end-block that contains the vertex u. Let C_1 be the cycle $vv_1v_2 \dots v_{k-1}v$ and let $\mathrm{color}(v) = X$, where $k \in \{3, 4, 5\}$. If $k = 3$, then $\mathrm{color}(v_1) = \mathrm{color}(v_2) = \overline{X}$. If $k = 4$, then $\mathrm{color}(v_2) = \overline{X}$ and, renaming v_1 and v_3, if necessary, we may assume that

$\text{color}(v_1) = X$ and $\text{color}(v_3) = \overline{X}$. If $k = 5$, then $\text{color}(v_2) = \text{color}(v_3) = \overline{X}$ and $\text{color}(v_1) = \text{color}(v_4) = X$.

Suppose that G contains an end-block, C say, that is a 4-cycle. If C' is an arbitrary end-block different from C, then $C' - v$ is a good path-component of $G[D]$, a contradiction. Hence, no end-block of G is a 4-cycle.

Thus, since G is a TDP-graph, at least one end-block is a 3-cycle and at least one end-block is a 5-cycle. Renaming the end-blocks if necessary, we may assume that C_1 is a 3-cycle and C_2 is a 5-cycle. These two cycles, together with their associated 2-colorings described above, form the 2-colored base graph G_3. If G contains at least three blocks and C' is an arbitrary end-block different from C_1 and C_2, then $C' - v$ is a good path-component of $G[D]$, a contradiction. Hence, G contains exactly two end-blocks, implying that (G, S) is the 2-colored base graph G_3. □

By Claim 4.11.1, we may assume that G contains at least two cut-vertices, for otherwise $(G, S) \in \mathcal{G}$ as desired. As observed earlier, the cut-vertices of G are precisely the vertices in the tree T^-. Let x be a neighbor of v in T^-. Renaming the cycle C_1 and C_2 and the vertex x if necessary, we may assume without loss of generality that the vertex v has a neighbor, y say, in C_1 such that $\text{color}(x) \neq \text{color}(y)$. We now consider the graph $G^- = G - (V(C_2) \setminus \{v\})$. Since G is a TDP-graph, this implies that the graph G^- is a TDP-graph. Hence, we can use Statement 4.1 with the operation $\mathcal{O}_{11}$ or $\mathcal{O}_{12}$ or $\mathcal{O}_{13}$, depending on the length of C_2, to show that $(G, S) \in \mathcal{G}$. □

By Claim 4.11, we may assume that no two cycles of G intersect, for otherwise $(G, S) \in \mathcal{G}$ as desired. The tree T^- therefore contains at least two vertices. Further, every leaf in T^- has degree 3 in G and belongs to exactly one end-block of G. Let $p_1 p_2 \dots p_k$ be a longest path in T^-. Necessarily, p_1 and p_k are both leaves in T^-. Since T^- contains no vertex of D, we note that every vertex in T^- has degree at least 3 in G. Let C_1 and C_k be the end-blocks in G that contain p_1 and p_k, respectively.

Claim 4.12. *If $k \in \{2, 3\}$, then $(G, S) \in \mathcal{G}$.*

Proof. Suppose firstly that $k = 2$. In this case, G is obtained from the two cycles C_1 and C_2 by adding the edge $p_1 p_2$. If C_1 is a 4-cycle, then the cycle C_1 together with its associated 2-coloring is the 2-colored base graph G_1. Starting with this 2-colored base graph G_1, we can use Statement 4.1 with the operation $\mathcal{O}_{14}$ or $\mathcal{O}_{15}$ or $\mathcal{O}_{16}$, depending on the length of C_2, to show that $(G, S) \in \mathcal{G}$. Analogously, if C_2 is a 4-cycle, $(G, S) \in \mathcal{G}$. Hence, we may assume that neither C_1 nor C_2 is a 4-cycle. With this assumption, if C_1 is a 3-cycle, then C_2 is also a 3-cycle noting that G is a TDP-graph. In this case, (G, S) is the 2-colored base graph G_2. If C_1 is a 5-cycle, then C_2 is also a 5-cycle. In this case, (G, S) is the 2-colored base graph G_4. Hence if $k = 2$, then $(G, S) \in \mathcal{G}$.

Suppose secondly that $k = 3$. We now consider the (connected) graph $G^- = G - V(C_1)$. We note that the vertex p_2 has degree at least 2 in G^-. If the vertex p_2 has neighbors of both colors in G^-, then G^- is a TPD-graph. In this case, we can use Statement 4.1 with the operation $\mathcal{O}_{14}$ or $\mathcal{O}_{15}$ or $\mathcal{O}_{16}$, depending on the length of C_1, to show that $(G, S) \in \mathcal{G}$. Hence we may assume that all neighbors of p_2 in G^- have the same color which is different to $\text{color}(p_1)$ (noting that G is a TPD-graph). This implies that the vertex p_2 has neighbors of both colors in the graph $G - V(C_2)$, and once again we can use Statement 4.1 with the operation $\mathcal{O}_{14}$ or $\mathcal{O}_{15}$ or $\mathcal{O}_{16}$, depending on the length of C_2, to show that $(G, S) \in \mathcal{G}$. □

By Claim 4.12, we may assume that $k \geq 4$, for otherwise $(G, S) \in \mathcal{G}$ as desired. We now consider the (connected) graph $G^- = G - V(C_1)$. If the vertex p_2 has neighbors of both colors in G^-, then as in the proof of Claim 4.12 we can use Statement 4.1 with the operation $\mathcal{O}_{14}$ or $\mathcal{O}_{15}$ or $\mathcal{O}_{16}$, depending on the length of C_1, to show that $(G, S) \in \mathcal{G}$. Hence we may assume that all neighbors of p_2 in G^- have the same color. We now consider the walk $p_2p_3 \dots p_k$. By assumption, p_3 is not the only neighbor of p_2 in G whose color is $\text{color}(p_3)$. By Claim 4.9, the vertex p_{k-2} is the only neighbor of p_{k-1} in G whose color is $\text{color}(p_{k-2})$. This implies that the vertex p_{k-1} has neighbors of both colors in the graph $G - V(C_2)$. Hence, $G - V(C_2)$ is a TPD-graph and we can use Statement 4.1 with the operation $\mathcal{O}_{14}$ or $\mathcal{O}_{15}$ or $\mathcal{O}_{16}$, depending on the length of C_2, to show that $(G, S) \in \mathcal{G}$. This completes the proof of Theorem 3.1. □

5 Closing remarks

We remark that although our characterization in Theorem 3.1 solves a long-standing problem in the theory of total domination in graphs which has been open for several decades, it remains a challenging problem to determine in polynomial time if a given graph is a TDP-graph even for some special graph classes. Our method cannot be used to decide if a given graph G is a TDP-graph in polynomial time. The reason for that is that we have no specified vertex partition together with G. Indeed, recognizing this class of graphs is known to be NP-complete (see [8]). However, we nonetheless believe that our constructive proof gives valuable insights into the problem and gives an entirely new description of TDP-graphs, placing them in another context.

We close with a short discussion about the independence of operations $\mathcal{O}_1$ to $\mathcal{O}_{17}$ in the class $\mathcal{G}$. For this purpose, we will construct small graphs in $\mathcal{G}$ from our 2-colored base graphs that cannot be built by any other construction in $\mathcal{G}$, thereby showing that operation $\mathcal{O}_i$ is independent for each $i \in [17]$. The independence of these seventeen operations used to build graphs in the family $\mathcal{G}$ show that none of them are redundant, and all are needed in the construction.

- Apply operation $\mathcal{O}_2$ on G_1 (to obtain the graph $K_4 - e$).
- Apply operation $\mathcal{O}_3$ on G_1 to obtain the house graph; that is, the graph obtained from a 5-cycle by adding an edge.
- Apply operation $\mathcal{O}_1$ once and operation $\mathcal{O}_2$ three times on the house graph to obtain K_5.
- Apply operation $\mathcal{O}_4$ to two nonadjacent vertices of degree 2 on G_2.
- The independence of operation $\mathcal{O}_x$, where $x \in \{5, 6, 11, 12, 13, 14, 15, 16\}$, can be seen by applying $\mathcal{O}_x$ once on G_1.
- The independence of operation $\mathcal{O}_x$, where $x \in \{7, 10\}$, can be seen by applying $\mathcal{O}_x$ once on adjacent vertices of degree 3 in G_2.
- The independence of operation $\mathcal{O}_x$, where $x \in \{8, 9\}$, can be seen by applying $\mathcal{O}_x$ once on adjacent vertices of degree 3 in G_4.
- Apply operation $\mathcal{O}_{17}$ once on the cut-vertex of G_3.

Hence, all seventeen operations are independent. Further, our proof of Theorem 3.1 shows that all seventeen operations are necessary to give our characterization of TDP-graphs.

References

[1] H. Aram, S. M. Sheikholeslami and L. Volkmann, On the total domatic number of regular graphs, *Trans. Comb.* **1** (2012), 45–51, doi:10.22108/toc.2012.760.

[2] B. Chen, J. H. Kim, M. Tait and J. Verstraete, On coupon colorings of graphs, *Discrete Appl. Math.* **193** (2015), 94–101, doi:10.1016/j.dam.2015.04.026.

[3] E. J. Cockayne, R. M. Dawes and S. T. Hedetniemi, Total domination in graphs, *Networks* **10** (1980), 211–219, doi:10.1002/net.3230100304.

[4] P. Delgado, W. J. Desormeaux and T. W. Haynes, Partitioning the vertices of a graph into two total dominating sets, *Quaest. Math.* **39** (2016), 863–873, doi:10.2989/16073606.2016.1188862.

[5] W. J. Desormeaux, T. W. Haynes and M. A. Henning, Partitioning the vertices of a cubic graph into two total dominating sets, *Discrete Appl. Math.* **223** (2017), 52–63, doi:10.1016/j.dam.2017.01.032.

[6] T. W. Haynes, S. T. Hedetniemi and P. J. Slater (eds.), *Domination in Graphs: Advanced Topics*, volume 209 of *Monographs and Textbooks in Pure and Applied Mathematics*, Marcel Dekker, New York, 1998.

[7] T. W. Haynes, S. T. Hedetniemi and P. J. Slater, *Fundamentals of Domination in Graphs*, volume 208 of *Monographs and Textbooks in Pure and Applied Mathematics*, Marcel Dekker, New York, 1998.

[8] P. Heggernes and J. A. Telle, Partitioning graphs into generalized dominating sets, *Nordic J. Comput.* **5** (1998), 128–142.

[9] M. A. Henning, A survey of selected recent results on total domination in graphs, *Discrete Math.* **309** (2009), 32–63, doi:10.1016/j.disc.2007.12.044.

[10] M. A. Henning, C. Löwenstein, D. Rautenbach and J. Southey, Disjoint dominating and total dominating sets in graphs, *Discrete Appl. Math.* **158** (2010), 1615–1623, doi:10.1016/j.dam.2010.06.004.

[11] M. A. Henning and J. Southey, A note on graphs with disjoint dominating and total dominating sets, *Ars Combin.* **89** (2008), 159–162.

[12] M. A. Henning and J. Southey, A characterization of graphs with disjoint dominating and total dominating sets, *Quaest. Math.* **32** (2009), 119–129, doi:10.2989/qm.2009.32.1.10.712.

[13] M. A. Henning and A. Yeo, *Total Domination in Graphs*, Springer Monographs in Mathematics, Springer, New York, 2013, doi:10.1007/978-1-4614-6525-6.

[14] O. Ore, *Theory of Graphs*, volume 38 of *American Mathematical Society Colloquium Publications*, American Mathematical Society, Providence, Rhode Island, 1962.

[15] J. Southey and M. A. Henning, A characterization of graphs with disjoint dominating and paired-dominating sets, *J. Comb. Optim.* **22** (2011), 217–234, doi:10.1007/s10878-009-9274-1.

[16] B. Zelinka, Total domatic number and degrees of vertices of a graph, *Math. Slovaca* **39** (1989), 7–11.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 377–401
https://doi.org/10.26493/1855-3974.1765.d73
(Also available at http://amc-journal.eu)

The Möbius function of $\mathrm{PSU}(3, 2^{2^n})$

Giovanni Zini [*]

*Department of Mathematics and Applications, University of Milano-Bicocca,
via Cozzi 55, 50125 Milano, Italy*

Received 25 July 2018, accepted 10 November 2018, published online 29 January 2019

Abstract

Let G be the simple group $\mathrm{PSU}(3, 2^{2^n})$, $n > 0$. For any subgroup H of G, we compute the Möbius function $\mu_L(H, G)$ of H in the subgroup lattice L of G, and the Möbius function $\mu_{\bar{L}}([H], [G])$ of $[H]$ in the poset $\bar{L}$ of conjugacy classes of subgroups of G. For any prime p, we provide the Euler characteristic of the order complex of the poset of non-trivial p-subgroups of G.

Keywords: Unitary groups, Möbius function, subgroup lattice.

Math. Subj. Class.: 20G40, 20D30, 05E15, 06A07

1 Introduction

The Möbius function $\mu(H, G)$ on the subgroups of a finite group G is defined recursively by $\mu(G, G) = 1$ and $\sum_{K \geq H} \mu(K, G) = 0$ if $H < G$. This function was used in 1936 by Hall [12] to enumerate k-tuples of elements of G which generate G, for a given k.

The combinatorial and group-theoretic properties of the Möbius function were investigated by many authors; see the paper [14] by Hawkes, Isaacs, and Özaydin. The Möbius function is defined more generally on a locally finite poset $(\mathcal{P}, \leq)$ by the recursive definition $\mu(x, x) = 1$, $\mu(x, y) = 0$ if $x \not\leq y$, and $\sum_{x \leq z \leq y} \mu(z, y) = 0$ if $x \leq y$; for instance, the poset taken into consideration may be the subgroup lattice L of a finite group G ordered by inclusion. Mann [19, 20] studied $\mu(H, G)$ in the broader context of profinite groups G and defined a probabilistic zeta function $P(G, s)$ associated to G, related to the probability of generating G with s elements when G is positively finitely generated.

The Möbius function on a poset $\mathcal{P}$ also appears in the context of topological invariants of the order simplicial complex $\Delta(\mathcal{P})$ associated to $\mathcal{P}$, see the works of Brown [2] and

[*] This research was partially supported by Ministry for Education, University and Research of Italy (MIUR) and by the Italian National Group for Algebraic and Geometric Structures and their Applications (GNSAGA - INdAM). The author would like to thank Francesca Dalla Volta and Martino Borello for their useful comments and suggestions, and Emilio Pierro for pointing out a mistake in a previous version of this paper.

E-mail address: giovanni.zini@unimib.it (Giovanni Zini)

Quillen [25]; if $\mathcal{P}$ is the subgroup lattice of a finite group G, then the reduced Euler characteristic of $\Delta(\mathcal{P})$ is equal to $\mu(\{1\}, G)$. This motivates the search for $\mu(\{1\}, G)$ independently of the knowledge of $\mu(H, G)$ for other subgroups H of G, see for instance [26, 27] and the references therein; $\mu(\{1\}, G)$ is often called the *Möbius number* of G. Shareshian provided a formula in [26] for $\mu(\{1\}, \mathrm{Sym}(n))$, and computed $\mu(\{1\}, G)$ in [27] when $G \in \{\mathrm{PGL}(2, q), \mathrm{PSL}(2, q), \mathrm{PGL}(3, q), \mathrm{PSL}(3, q), \mathrm{PGU}(3, q), \mathrm{PSU}(3, q)\}$ with q odd or G is a Suzuki group $\mathrm{Sz}(2^{2h+1})$.

Consider the poset $\bar{L}$ of conjugacy classes $[H]$ of subgroups H of a finite group G, ordered as follows: $[H] \leq [K]$ if and only if H is contained in some conjugate of K in G. After Hawkes, Isaacs, and Özaydin [14], we denote by $\lambda(H, G)$ the Möbius function $\mu([H], [G])$ in $\bar{L}$, while $\mu(H, G)$ is the Möbius function in L. Some attempt was done to search relations between the Möbius functions $\mu(H, G)$ and $\lambda(H, G)$; Hawkes, Isaacs, and Özaydin [14] proved that, if G is solvable, then

$$\mu(\{1\}, G) = |G'| \cdot \lambda(\{1\}, G). \tag{1.1}$$

The property (1.1), which we call (μ, λ)-property, does not hold in general for non-solvable groups; see [1]. Pahlings [23] proved that, if G is solvable, then

$$\mu(H, G) = [N_{G'}(H) : H \cap G'] \cdot \lambda(H, G) \tag{1.2}$$

for any subgroup H of G. The analysis of the generalized (μ, λ)-property (1.2), although false in general for non-solvable groups, is of interest since it relates the Möbius functions $\mu(H, G)$ and $\lambda(H, G)$.

A lot of work was done by several authors about probabilistic functions for groups; see for instance [6, 10, 19, 20]. In particular, Mann posed in [19] a conjecture, the validity of which would imply that the sum

$$\sum_{H} \frac{\mu(H, G)}{[G : H]^s}$$

over all subgroups $H < G$ of finite index of a positively finitely generated profinite group G is absolutely convergent for s in some right complex half-plane and, for $s \in \mathbb{N}$ large enough, represents the probability of generating G with s elements. Lucchini [18] showed that this problem can be reduced so that Mann's conjecture is reformulated as follows: there exist two constants $c_1, c_2 \in \mathbb{N}$ such that, for any finite monolithic group G with non-abelian socle,

1. $|\mu(H, G)| \leq [G : H]^{c_1}$ for any $H < G$ such that $G = H \operatorname{soc}(G)$, and
2. the number of subgroups $H < G$ of index n in G such that $H \operatorname{soc}(G) = G$ and $\mu(H, G) \neq 0$ is upper bounded by n^{c_2}, for any $n \in \mathbb{N}$.

It seems natural to investigate this conjecture on finite monolithic groups starting by almost simple groups. Mann's conjecture has been shown to be satisfied by the alternating and symmetric groups [3], as well as by those families of groups G for which $\mu(H, G)$ has been computed for any subgroup H; namely, $\mathrm{PSL}(2, q)$ [8, 12], $\mathrm{PGL}(2, q)$ [8], the Suzuki groups $\mathrm{Sz}(2^{2h+1})$ [9], and the Ree groups $R(3^{2h+1})$ [24].

In this paper, we take into consideration the three dimensional projective special unitary group $G = \mathrm{PSU}(3, q)$ over the field with $q = 2^{2^n}$ elements, for any positive n (note that $\mathrm{PSU}(3, q) = \mathrm{PGU}(3, q)$ as $3 \nmid (q + 1)$). In particular, the following results are obtained.

(i) We compute $\mu(H,G)$ for any subgroup H of G, as summarized in Table 1. This shows that the groups $\mathrm{PSU}(3,2^{2^n})$ satisfy Mann's conjecture.

(ii) We compute $\lambda(H,G)$ for any subgroup H of G, as summarized in Table 1. This shows that the groups $\mathrm{PSU}(3,2^{2^n})$ satisfy the (μ,λ)-property, but do not satisfy the generalized (μ,λ)-property.

(iii) We compute the Euler characteristic $\chi(\Delta(L_p \setminus \{1\}))$ of the order complex of the poset $L_p \setminus \{1\}$ of non-trivial p-subgroups of G, for any prime p, as summarized in Table 2.

For the subgroups listed in Table 1, the isomorphism type determines a unique conjugacy class in G.

Table 1: Subgroups H of $G = \mathrm{PSU}(3,q)$, $q = 2^{2^n}$, with $\mu(H) \neq 0$ or $\lambda(H) \neq 0$.

Isomorphism type of H	$\lvert H\rvert$	$N_G(H)$	$\mu(H,G)$	$\lambda(H,G)$
G	$q^3(q^3+1)(q^2-1)$	H	1	1
$(E_q \,.\, E_{q^2}) \rtimes C_{q^2-1}$	$q^3(q^2-1)$	H	-1	-1
$\mathrm{PSL}(2,q) \times C_{q+1}$	$q(q^2-1)(q+1)$	H	-1	-1
$(C_{q+1} \times C_{q+1}) \rtimes \mathrm{Sym}(3)$	$6(q+1)^2$	H	-1	-1
$C_{q^2-q+1} \rtimes C_3$	$3(q^2-q+1)$	H	-1	-1
$E_q \rtimes C_{q^2-1}$	$q(q^2-1)$	H	1	1
$(C_{q+1} \times C_{q+1}) \rtimes C_2$	$2(q+1)^2$	H	1	1
$\mathrm{Sym}(3)$	6	$\mathrm{Sym}(3) \times C_{q+1}$	$q+1$	1
C_3	3	$C_{q^2-1} \rtimes C_2$	$\frac{2(q^2-1)}{3}$	1
C_2	2	$(E_q \,.\, E_{q^2}) \rtimes C_{q+1}$	$-\frac{q^3(q+1)}{2}$	-1

Table 2: Euler characteristic of the order complex of the poset of proper p-subgroups of G.

Prime p	$p \nmid \lvert G\rvert$	$p=2$	$p \mid (q+1)$	$p \mid (q-1)$	$p \mid (q^2-q+1)$
$\chi(\Delta(L_p \setminus \{1\}))$	0	q^3+1	$-\frac{q^6-2q^5-q^4+2q^3-3q^2}{3}$	$\frac{q^6+q^3}{2}$	$-\frac{q^6+q^5-q^4-q^3}{3}$

The paper is organized as follows. Section 2 contains preliminary results on the Möbius functions $\mu(H,G)$ and $\lambda(H,G)$ and the relation between the Möbius function and the Euler characteristic of the order complex; this section contains also preliminary results on the groups $G = \mathrm{PSU}(3,2^{2^n})$, whose elements are described geometrically in their action on the Hermitian curve associated to G. Sections 3 and 4 are devoted to the determination of $\mu(H,G)$ and $\lambda(H,G)$, respectively, for any subgroup H of G. Section 5 provides the Euler characteristic of the order complex of the poset of proper p-subgroups of G, for any prime p.

2 Preliminary results

Let $(\mathcal{P}, \leq)$ be a finite poset. The Möbius function $\mu_{\mathcal{P}} \colon \mathcal{P} \times \mathcal{P} \to \mathbb{Z}$ is defined recursively as follows:

$$\mu_{\mathcal{P}}(x,y) = 0 \quad \text{if} \quad x \not\leq y; \qquad \mu_{\mathcal{P}}(x,x) = 1; \qquad \sum_{x \leq z \leq y} \mu_{\mathcal{P}}(z,y) = 0 \quad \text{if} \quad x < y.$$

If $x < y$, then $\mu_{\mathcal{P}}(x,y)$ can be equivalently defined by

$$\sum_{x \leq z \leq y} \mu_{\mathcal{P}}(x,z) = 0.$$

To the poset $\mathcal{P}$ we can associate a simplicial complex $\Delta(\mathcal{P})$ whose vertices are the elements of $\mathcal{P}$ and whose i-dimensional faces are the chains $a_0 < \cdots < a_i$ of length i in $\mathcal{P}$; $\Delta(\mathcal{P})$ is called the *order complex* of $\mathcal{P}$. Provided that $\mathcal{P}$ has a least element 0, the Euler characteristic of the order complex of $\mathcal{P} \setminus \{0\}$ is computed as follows (see [28, Proposition 3.8.6]):

$$\chi(\Delta(\mathcal{P} \setminus \{0\})) = - \sum_{x \in \mathcal{P} \setminus \{0\}} \mu_{\mathcal{P}}(0,x).$$

Given a finite group G, we will consider the following two Möbius functions associated to G.

(i) The Möbius function on the subgroup lattice L of G, ordered by inclusion. We will denote $\mu_L(H,G)$ simply by $\mu(H)$.

(ii) The Möbius function on the poset $\bar{L}$ of conjugacy classes $[H]$ of subgroups H of G, ordered as follows: $[H] \leq [K]$ if and only if H is contained in the conjugate gKg^{-1} for some $g \in G$. We will denote $\mu_{\bar{L}}([H],[G])$ simply by $\lambda(H)$.

Two facts will be used to compute $\mu(H)$. The first easy fact is that, if H and K are conjugate in G, then $\mu(H) = \mu(K)$. The second fact is due to Hall [12, Theorem 2.3], and is stated in the following lemma.

Lemma 2.1. *If $H < G$ satisfies $\mu(H) \neq 0$, then H is the intersection of maximal subgroups of G.*

For any prime p, let L_p be the subposet of L given by all p-subgroups of G, so that

$$\chi(\Delta(L_p \setminus \{1\})) = - \sum_{H \in L_p \setminus \{1\}} \mu_{L_p}(\{1\}, H). \tag{2.1}$$

By a result of Brown [2], $\chi(\Delta(L_p \setminus \{1\}))$ is congruent to 1 modulo the order $|G|_p$ of a Sylow p-subgroup of G. In order to compute explicitly $\chi(\Delta(L_p \setminus \{1\}))$ we will use the following result of Hall [12, Equation (2.7)]:

Lemma 2.2. *Let H be a p-group of order p^r. If H is not elementary abelian, then $\mu_{L_p}(\{1\}, H) = 0$. If H is elementary abelian, then $\mu_{L_p}(\{1\}, H) = (-1)^r p^{\binom{r}{2}}$.*

We describe now the group G which will be considered in the following sections. Let n be a positive integer, $q = 2^{2^n}$, $\mathbb{F}_q$ be the finite field with q element, and $\bar{\mathbb{F}}_q$ be the algebraic closure of $\mathbb{F}_q$. Let $\mathcal{U}$ be a non-degenerate unitary polarity of the plane $\mathrm{PG}(2,q^2)$ over $\mathbb{F}_{q^2}$, and $\mathcal{H}_q \subset \mathrm{PG}(2,\bar{\mathbb{F}}_q)$ be the Hermitian curve defined by $\mathcal{U}$. The following homogeneous equations define models for $\mathcal{H}_q$ which are projectively equivalent over $\mathbb{F}_{q^2}$:

$$X^{q+1}+Y^{q+1}+Z^{q+1}=0; \tag{2.2}$$

$$X^qZ+XZ^q-Y^{q+1}=0. \tag{2.3}$$

The models (2.2) and (2.3) are called the Fermat and the Norm-Trace model of $\mathcal{H}_q$, respectively. The set of $\mathbb{F}_{q^2}$-rational points of $\mathcal{H}_q$ is denoted by $\mathcal{H}_q(\mathbb{F}_{q^2})$, and consists of the q^3+1 isotropic points of $\mathcal{U}$. The full automorphism group $\mathrm{Aut}(\mathcal{H}_q)$ of $\mathcal{H}_q$ is defined over $\mathbb{F}_{q^2}$, and coincides with the unitary subgroup $\mathrm{PGU}(3,q)$ of $\mathrm{PGL}(3,q^2)$ stabilizing $\mathcal{H}_q(\mathbb{F}_{q^2})$, of order $|\mathrm{PGU}(3,q)|=q^3(q^3+1)(q^2-1)$.

The combinatorial properties of $\mathcal{H}_q(\mathbb{F}_{q^2})$ can be found in [16]. In particular, any line ℓ of $\mathrm{PG}(2,q^2)$ has either 1 or $q+1$ common points with $\mathcal{H}_q(\mathbb{F}_{q^2})$, that is, ℓ is either a tangent line or a chord of $\mathcal{H}_q(\mathbb{F}_{q^2})$; in the former case ℓ contains its pole with respect to $\mathcal{U}$, in the latter case ℓ doesn't. Also, $\mathrm{PGU}(3,q)$ acts 2-transitively on $\mathcal{H}_q(\mathbb{F}_{q^2})$ and transitively on $\mathrm{PG}(2,q^2)\setminus\mathcal{H}_q$; $\mathrm{PGU}(3,q)$ acts transitively also on the non-degenerate self-polar triangles $T=\{P_1,P_2,P_3\}\subset\mathrm{PG}(2,q^2)\setminus\mathcal{H}_q$ with respect to $\mathcal{U}$. Recall that, if $\sigma\in\mathrm{PGU}(3,q)$ stabilizes a point $P\in\mathrm{PG}(2,q^2)$, then σ stabilizes also the polar line of P with respect to $\mathcal{U}$, and vice versa.

The curve $\mathcal{H}_q$ is non-singular and $\mathbb{F}_{q^2}$-maximal of genus $g=\frac{q(q-1)}{2}$, that is, the size of $\mathcal{H}_q(\mathbb{F}_{q^2})$ attains the Hasse-Weil upper bound $q^2+1+2gq$. This implies that $\mathcal{H}_q$ is $\mathbb{F}_{q^4}$-minimal and $\mathbb{F}_{q^6}$-maximal, so that $\mathcal{H}_q(\mathbb{F}_{q^4})\setminus\mathcal{H}_q(\mathbb{F}_{q^2})=\emptyset$ and $|\mathcal{H}_q(\mathbb{F}_{q^6})\setminus\mathcal{H}_q(\mathbb{F}_{q^2})|=q^6+q^5-q^4-q^3$. Let Φ_{q^2} be the Frobenius map $(X,Y,Z)\mapsto(X^{q^2},Y^{q^2},Z^{q^2})$ over $\mathrm{PG}(2,\bar{\mathbb{F}}_{q^2})$; then the $\mathbb{F}_{q^6}\setminus\mathbb{F}_{q^2}$-rational points of $\mathcal{H}_q$ split into $\frac{q^6+q^5-q^4-q^3}{3}$ non-degenerate triangles $\{P,\Phi_{q^2}(P),\Phi_{q^2}^2(P)\}$. The group $\mathrm{PGU}(3,q)$ is transitive on such triangles.

Since $3\nmid(q+1)$, we have $\mathrm{PGU}(3,q)=\mathrm{PSU}(3,q)$; henceforth, we denote by G the simple group $\mathrm{PSU}(3,q)$. The following classification of subgroups of G goes back to Hartley [13]; here we use that $\log_2(q)$ has no odd divisors different from 1. The notation S_2 stands for a Sylow 2-subgroup of G, which is a non-split extension $E_q\,.\,E_{q^2}$ of its elementary abelian center of order q by an elementary abelian group of order q^2.

Theorem 2.3. *Let $n>0$, $q=2^{2^n}$, and $G=\mathrm{PSU}(3,q)$. Up the conjugation, the maximal subgroups of G are the following.*

(i) The stabilizer $M_1(P)\cong S_2\rtimes C_{q^2-1}$ of a point $P\in\mathcal{H}_q(\mathbb{F}_{q^2})$, of order $q^3(q^2-1)$.

(ii) The stabilizer $M_2(P)\cong\mathrm{PSL}(2,q)\times C_{q+1}$ of a point $P\in\mathrm{PG}(2,q^2)\setminus\mathcal{H}_q(\mathbb{F}_{q^2})$, of order $q(q^2-1)(q+1)$.

(iii) The stabilizer $M_3(T)\cong(C_{q+1}\times C_{q+1})\rtimes\mathrm{Sym}(3)$ of a non-degenerate self-polar triangle $T=\{P,Q,R\}\subset\mathrm{PG}(2,q^2)\setminus\mathcal{H}_q$ with respect to $\mathcal{U}$, of order $6(q+1)^2$.

(iv) The stabilizer $M_4(T)\cong C_{q^2-q+1}\rtimes C_3$ of a triangle $T=\{P,\Phi_{q^2}(P),\Phi_{q^2}^2(P)\}\subset\mathcal{H}_q(\mathbb{F}_{q^6})\setminus\mathcal{H}_q(\mathbb{F}_{q^2})$, of order $3(q^2-q+1)$.

For a detailed description of the maximal subgroups of G, both from an algebraic and a geometric point of view, we refer to [11, 21, 22].

In our investigation it is useful to know the geometry of the elements of $\mathrm{PGU}(3,q)$ on $\mathrm{PG}(2,\bar{\mathbb{F}}_q)$, and in particular on $\mathcal{H}_q(\mathbb{F}_{q^2})$. This can be obtained as a corollary of Theorem 2.3, and is stated in Lemma 2.2 with the usual terminology of collineations of projective planes; see [16]. In particular, a linear collineation σ of $\mathrm{PG}(2,\bar{\mathbb{F}}_q)$ is a (P,ℓ)-*perspectivity*, if σ preserves each line through the point P (the *center* of σ), and fixes each point on the line ℓ (the *axis* of σ). A (P,ℓ)-perspectivity is either an *elation* or a *homology* according to $P \in \ell$ or $P \notin \ell$. Lemma 2.4 was obtained in [21] in a more general form (i.e., for any prime power q).

Lemma 2.4. *For a nontrivial element $\sigma \in G = \mathrm{PSU}(3,q)$, $q = 2^{2^n}$, one of the following cases holds.*

- *(A)* $\mathrm{ord}(\sigma) \mid (q+1)$ *and σ is a homology, with center $P \in \mathrm{PG}(2,q^2) \setminus \mathcal{H}_q$ and axis ℓ_P which is a chord of $\mathcal{H}_q(\mathbb{F}_{q^2})$; (P,ℓ_P) is a pole-polar pair with respect to $\mathcal{U}$.*
- *(B)* $2 \nmid \mathrm{ord}(\sigma)$ *and σ fixes the vertices P_1, P_2, P_3 of a non-degenerate triangle $T \subset \mathrm{PG}(2,q^6)$.*
  - *(B1)* $\mathrm{ord}(\sigma) \mid (q+1)$*, $P_1, P_2, P_3 \in \mathrm{PG}(2,q^2) \setminus \mathcal{H}_q$, and the triangle T is self-polar with respect to $\mathcal{U}$.*
  - *(B2)* $\mathrm{ord}(\sigma) \mid (q^2-1)$ *and* $\mathrm{ord}(\sigma) \nmid (q+1)$*; $P_1 \in \mathrm{PG}(2,q^2) \setminus \mathcal{H}_q$ and $P_2, P_3 \in \mathcal{H}_q(\mathbb{F}_{q^2})$.*
  - *(B3)* $\mathrm{ord}(\sigma) \mid (q^2-q+1)$ *and $P_1, P_2, P_3 \in \mathcal{H}_q(\mathbb{F}_{q^6}) \setminus \mathcal{H}_q(\mathbb{F}_{q^2})$.*
- *(C)* $\mathrm{ord}(\sigma) = 2$*; σ is an elation with center $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and axis ℓ_P which is tangent to $\mathcal{H}_q$ at P, such that (P,ℓ_P) is a pole-polar pair with respect to $\mathcal{U}$.*
- *(D)* $\mathrm{ord}(\sigma) = 4$*; σ fixes a point $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and a line ℓ_P which is tangent to $\mathcal{H}_q$ at P, such that (P,ℓ_P) is a pole-polar pair with respect to $\mathcal{U}$.*
- *(E)* $\mathrm{ord}(\sigma) = 2d$ *where d is a nontrivial divisor of $q+1$; σ fixes two points $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and $Q \in \mathrm{PG}(2,q^2) \setminus \mathcal{H}_q$, the polar line PQ of P, and the polar line of Q which passes through P.*

For a detailed description of the elements and subgroups of G, both from an algebraic and a geometric point of view, we refer to [11, 21, 22], on which our geometric arguments are based.

Throughout the paper, a nontrivial element of G is said to be of type (A), (B), (B1), (B2), (B3), (C), (D), or (E), as given in Lemma 2.4. Also, the polar line to $\mathcal{H}_q$ at $P \in \mathrm{PG}(2,q^2)$ is denoted by ℓ_P. Note that, under our assumptions, any element of order 3 in G is of type (B2). We will denote a cyclic group of order d by C_d and an elementary abelian group of order d by E_d. The center $Z(S_2)$ of S_2 is elementary abelian of order q, and any element in $S_2 \setminus Z(S_2)$ has order 4; see [11, Section 3].

3 Determination of $\mu(H)$ for any subgroup H of G

Let $n > 0$, $q = 2^{2^n}$, $G = \mathrm{PSU}(3,q)$. This section is devoted to the proof of the following theorem.

Theorem 3.1. *Let H be a proper subgroup of G. Then H is the intersection of maximal subgroups of G if and only if H is one of the following groups:*

$$\begin{array}{lll} S_2 \rtimes C_{q^2-1}, & \mathrm{PSL}(2,q) \times C_{q+1}, & C_{q^2-q+1} \rtimes C_3, \\ (C_{q+1} \times C_{q+1}) \rtimes \mathrm{Sym}(3), & E_q \rtimes C_{q^2-1}, & (C_{q+1} \times C_{q+1}) \rtimes C_2, \\ C_{q+1} \times C_{q+1}, & C_{q^2-1}, & C_{2(q+1)}, \\ C_{q+1} = Z(M_2(P)) \textit{ for some } P, & E_q, & \mathrm{Sym}(3), \\ C_3, & C_2, & \{1\}. \end{array} \tag{3.1}$$

Given a type of groups in Equation (3.1)*, there is just one conjugacy class of subgroups of G of that isomorphism type.*

The normalizer $N_G(H)$ of H in G for the groups H in Equation (3.1) *are, respectively:*

$$\begin{array}{lll} H, & H, & H, \\ H, & H, & H, \\ H \rtimes \mathrm{Sym}(3), & H \rtimes C_2, & E_q \times C_{q+1}, \\ \mathrm{PSL}(2,q) \times H, & S_2 \rtimes C_{q^2-1}, & H \times C_{q+1}, \\ C_{q^2-1} \rtimes C_2, & S_2 \rtimes C_{q+1}, & G. \end{array} \tag{3.2}$$

The values $\mu(H)$ for the groups H in Equation (3.1) *are, respectively:*

$$\begin{array}{ccc} -1, & -1, & -1, \\ -1, & 1, & 1, \\ 0, & 0, & 0, \\ 0, & 0, & q+1, \\ \dfrac{2(q^2-1)}{3}, & -\dfrac{q^3(q+1)}{2}, & 0. \end{array} \tag{3.3}$$

The proof of Theorem 3.1 is divided into several propositions.

Proposition 3.2. *The group G contains exactly one conjugacy class for any group in Equation* (3.1).

Proof. **Case 1:** The first four groups in Equation (3.1), i.e.,

$$S_2 \rtimes C_{q^2-1},\ \mathrm{PSL}(2,q) \times C_{q+1},\ C_{q^2-q+1} \rtimes C_3, \text{ and } (C_{q+1} \times C_{q+1}) \rtimes \mathrm{Sym}(3),$$

are the maximal subgroups of G, for which there is just one conjugacy class by Theorem 2.3.

Case 2: Let $\alpha_1, \alpha_2 \in G$ have order 3, so that they are of type (B2) and α_i fixes two distinct points $P_i, Q_i \in \mathcal{H}_q(\mathbb{F}_{q^2})$. The group G is 2-transitive on $\mathcal{H}_q(\mathbb{F}_{q^2})$, and the pointwise stabilizer of $\{P_i, Q_i\}$ is cyclic of order $q^2 - 1$. Hence, $\langle\alpha_1\rangle$ and $\langle\alpha_2\rangle$ are conjugated in G.

Case 3: Let $\alpha_1, \alpha_2 \in G$ have order 2, so that they are of type (C) and α_i fixes exactly one point P_i on $\mathcal{H}_q(\mathbb{F}_{q^2})$. Up to conjugation $P_1 = P_2$, as G is transitive on $\mathcal{H}_q(\mathbb{F}_{q^2})$. The involutions fixing P_1 in G, together with the identity, form an elementary abelian group E_q, which is normalized by a cyclic group C_{q-1}; no nontrivial element of C_{q-1} commutes with any nontrivial element of E_q (see [11, Section 4]). Hence, α_1 and α_2 are conjugated under an element of C_{q-1}.

Case 4: Let $\alpha_1, \alpha_2, \beta_1, \beta_2 \in G$ satisfy $o(\alpha_i) = 3$, $o(\beta_i) = 2$, and $H_i := \langle \alpha_i, \beta_i \rangle \cong \mathrm{Sym}(3)$. As shown in the previous point, we can assume $\alpha_1 = \alpha_2$ up to conjugation. Let $P, Q \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and $R \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$ be the fixed points of α_1. Since $\beta_i \alpha_1 \beta_i^{-1} = \alpha_1^{-1}$, we have that β_i fixes R and interchanges P and Q; β is then uniquely determined from the $\mathbb{F}_{q^2}$-rational point of PQ fixed by β (namely, the intersection between PQ and the axis of β). Since the pointwise stabilizer C_{q^2-1} of $\{P, Q\}$ acts transitively on $PQ(\mathbb{F}_{q^2}) \setminus \mathcal{H}_q$, β_1 and β_2 are conjugated, and the same holds for H_1 and H_2.

Case 5: Any two groups isomorphic to C_{q^2-1} are conjugated in G, because they are generated by elements of type (B2) and G is 2-transitive on $\mathcal{H}_q(\mathbb{F}_{q^2})$.

Case 6: Any two groups isomorphic to E_q are conjugated in G, because any such group fixes exactly one point $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$, G is transitive on $\mathcal{H}_q(\mathbb{F}_{q^2})$, and the stabilizer $G_P = M_1(P)$ contains just one subgroup E_q.

Case 7: Any two groups $H_1, H_2 \cong E_q \rtimes C_{q^2-1}$ are conjugated in G. In fact, their Sylow 2-subgroups E_q coincide up to conjugation, as shown in the previous point. The normalizer $N_G(E_q)$ fixes the fixed point $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$ of E_q, and hence $N_G(E_q) = M_1(P) = S_2 \rtimes C_{q^2-1}$. The complements C_{q^2-1} are conjugated by Schur-Zassenhaus Theorem; hence, H_1 and H_2 are conjugated.

Case 8: Any two groups isomorphic to $C_{2(q+1)}$ are conjugated in G, because they are generated by elements of type (E) and two elements α_1, α_2 of type (E) of the same order are conjugated in G. In fact, α_i is uniquely determined by its fixed points $P_i \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and $Q_i \in \ell_{P_i}(\mathbb{F}_{q^2}) \setminus \mathcal{H}_q$; here, ℓ_{P_i} is the polar line of P_i. Up to conjugation $P_1 = P_2$, from the transitivity of G on $\mathcal{H}_q(\mathbb{F}_{q^2})$. Also, S_2 has order q^3 and acts on the q^2 points of $\ell_{P_i}(\mathbb{F}_{q^2}) \setminus \mathcal{H}_q$ with kernel E_q, hence transitively. We can then assume $Q_1 = Q_2$.

Case 9: Let Z_{P_i} be the center of $M_2(P_i)$, $i = 1, 2$. As shown in [5, Section 4], $Z_{P_i} \cong C_{q+1}$ and Z_{P_i} is made by the homologies with center P_i, together with the identity. Since G is transitive on $\mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$, we have up to conjugation that $M_2(P_1) = M_2(P_2)$ and $Z_{P_1} = Z_{P_2}$.

Case 10: Any two groups $H_1, H_2 \cong C_{q+1} \times C_{q+1}$ are conjugated in G. In fact, H_i is the pointwise stabilizer of a self-polar triangle $T_i = \{P_i, Q_i, R_i\} \subset \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$ (see [5, Section 3]), and the stabilizers of T_1 and T_2 are conjugated by Theorem 2.3.

Case 11: Any two groups $H_1, H_2 \cong (C_{q+1} \times C_{q+1}) \rtimes C_2$ are conjugated in G. In fact, their subgroups $C_{q+1} \times C_{q+1}$ coincide up to conjugation as shown above, and fix pointwise a self-polar triangle $T = \{P, Q, R\} \subset \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$. Let $\beta_i \in H_i$ have order 2, $i = 1, 2$. Then β_i fixes one vertex of T and interchanges the other two vertexes. Up to conjugation in $M_3(T)$ we have $\beta_1(P) = \beta_2(P) = P$. Then $H_1 = H_2$, as they coincide with the stabilizer of P in $M_3(T)$. □

Proposition 3.3. *The normalizers $N_G(H)$ of the groups H in Equation* (3.1) *are given in Equation* (3.2).

Proof. **Case 1:** Clearly $N_G(H) = H$ for any H from the first four groups of Equation (3.1) as H is maximal in G.

Case 2: Let $H = E_q \rtimes C_{q^2-1}$. Then $H \leq M_2(P)$, where P is the unique fixed point of C_{q^2-1} in $\mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$. The group H has a unique cyclic subgroup C_{q+1} of order $q+1$; namely, C_{q+1} is the center of $M_2(P)$ and is made by the homologies with center P; since q is even, H is a split extension $C_{q+1} \times (E_q \rtimes C_{q-1})$. Hence, $N_G(H) \leq N_G(C_{q+1}) = M_2(P)$. The group $H/C_{q+1} \cong E_q \rtimes C_{q-1}$ is maximal and hence self-normalizing in $M_2(P)/C_{q+1} = \mathrm{PSL}(2, q)$; thus, $N_G(E_q \rtimes C_{q-1}) = H$ and $N_G(H) = H$.

Case 3: Let $H = C_{q+1} \times C_{q+1}$. Then $N_G(H) \leq M_3(T)$, where T is the self-polar triangle fixed pointwise by H. Since H is the kernel of $M_3(T)$ in its action on T, we have $N_G(H) = M_3(T)$ and $|N_G(H)| = 6|H|$.

Case 4: Let $H = (C_{q+1} \times C_{q+1}) \rtimes C_2$. Then $C_{q+1} \times C_{q+1}$ is normal in $N_G(H)$, being the unique subgroup of index 2 in H. Hence $N_G(H) \leq M_3(T)$, where T is the self-polar triangle fixed pointwise by H. Also, $N_G(H)$ fixes the vertex P of T fixed by H, so that $N_G(H) \neq M_3(T)$. This implies $N_G(H) = H$.

Case 5: Let $H = C_{q^2-1}$. Then H is generated by an element α of type (B2) with fixed points $P, Q \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and $R \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$. Let β be an involution satisfying $\beta(R) = R$, $\beta(P) = Q$, and $\beta(Q) = P$; then $\beta \in N_G(H)$, because H coincides with the pointwise stabilizer of $\{P, Q\}$ in G. An explicit description is the following: given $\mathcal{H}_q$ with equation (2.3), we can assume up to conjugation that $\alpha = \mathrm{diag}(a^{q+1}, a, 1)$ where a is a generator if $\mathbb{F}_{q^2}^*$ (see [11]); then take

$$\beta = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \tag{3.4}$$

Since $N_G(H)$ acts on $\{P, Q\}$ and $\beta \in N_G(H)$, the pointwise stabilizer H of $\{P, Q\}$ has index 2 in $N_G(H)$. This implies $N_G(H) = C_{q^2-1} \rtimes C_2$ and $|N_G(H)| = 2|H|$.

Case 6: Let $H = C_{2(q+1)}$, so that H is generated by an element α of type (E) fixing exactly two points $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and $Q \in \ell_P(\mathbb{F}_{q^2}) \setminus \mathcal{H}_q$. Then $N_G(H)$ fixes P and Q. The subgroup E_q of $M_1(P)$ commutes with H elementwise, while any 2-element in $M_1(P) \setminus E_q$ has order 4 and does not fix Q; hence, the Sylow 2-subgroup of $N_G(H)$ is E_q. Also, $N_G(H) = E_q \rtimes C_d$, where C_d is a subgroup of C_{q^2-1} containing the subgroup C_{q+1} of H. Let C_2 be the subgroup of H of order 2; the quotient group $(C_2 \rtimes C_d)/C_{q+1} \cong C_2 \rtimes C_{\frac{d}{q+1}}$ acts faithfully as a subgroup of $\mathrm{PGL}(2, q)$ on the $q+1$ points of $\ell_Q \cap \mathcal{H}_q$. By the classification of subgroups of $\mathrm{PGL}(2, q)$ ([7]; see [17, Hauptsatz 8.27]), this implies $d = 1$; that is, $N_G(H) = E_q \rtimes C_{q+1}$ and $|N_G(H)| = \frac{q}{2}|H|$.

Case 7: Let $H = C_{q+1} = Z(M_2(P))$. Since H is the center of $M_2(P)$, $M_2(P) \leq N_G(H)$. Conversely, H is made by homologies with center P, and hence $N_G(H)$ fixes P. Thus, $N_G(H) = M_2(P)$ and $|N_G(H)| = q(q^2-1)|H|$.

Case 8: Let $H = E_q$. Since E_q has a unique fixed point P on $\mathcal{H}_q(\mathbb{F}_{q^2})$ and $E_q = Z(M_1(P))$, we have $N_G(H) \leq M_1(P)$ and $M_1(P) \leq N_G(H)$, so that $N_G(H) = M_1(P)$ and $|N_G(H)| = q^2(q^2-1)|H|$.

Case 9: Let $H = \mathrm{Sym}(3) = \langle \alpha, \beta \rangle$, with $o(\alpha) = 3$ and $o(\beta) = 2$. Let $P, Q \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and $R \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$ be the fixed points of α; β fixes R, interchanges P and Q, and fixes another point A_β on $\ell_R \cap \mathcal{H}_q$. The group $N_G(H)$ acts on $\{P, Q\}$ and on $\{A_\beta, A_{\alpha\beta}, A_{\alpha^2\beta}\}$.

The pointwise stabilizer C_{q^2-1} has a subgroup C_{q+1} which is the center of $M_2(P)$ and fixes PQ pointwise, while any element in $C_{q^2-1} \setminus C_{q+1}$ acts semiregularly on $PQ \setminus \{P, Q\}$; hence, $C_{q^2-1} \cap N_G(H) = C_{3(q+1)}$. If an element $\gamma \in N_G(H)$ fixes $\{P, Q\}$ pointwise, then γ fixes a point in $\{A_\beta, A_{\alpha\beta}, A_{\alpha^2\beta}\}$, and hence $\gamma \in \{\beta, \alpha\beta, \alpha^2\beta\}$. Therefore, $N_G(H) = C_{3(q+1)} \rtimes C_2 = H \times C_{q+1}$ and $|N_G(H)| = (q + 1)|H|$.

Case 10: Let $H = C_3$ and α be a generator of H, with fixed points $P, Q \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and $R \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$. The normalizer $N_G(H)$ fixes R and acts on $\{P, Q\}$. There exists an involution $\beta \in G$ normalizing H and interchanging P and Q (see Equation (3.4)). Then the pointwise stabilizer of $\{P, Q\}$ has index 2 in $N_G(H)$. Also, the pointwise stabilizer of $\{P, Q\}$ in G is cyclic of order $q^2 - 1$. Then $N_G(H) = C_{q^2-1} \rtimes C_2$ and $|N_G(H)| = \frac{2(q^2-1)}{3}|H|$.

Case 11: Let $H = C_2$ and α be a generator of H, with fixed point $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$. Then $N_G(H)$ fixes P, i.e. $N_G(H) \leq M_1(P) = S_2 \rtimes C_{q^2-1}$. Since any involution of $M_1(P)$ is in the center of S_2, the Sylow 2-subgroup of $N_G(H)$ has order q^3. Let $\beta \in C_{q^2-1}$. If $o(\beta) \mid (q + 1)$, then β commutes with any involution of S_2. If $o(\beta) \nmid (q + 1)$, then β does not commute with any element of S_2. This implies that $N_G(H) = S_2 \rtimes C_{q+1}$, and $|N_G(H)| = \frac{q^3(q+1)}{2}|H|$. □

Lemma 3.4. *Let $\alpha \in G$ be an involution, and hence an elation, with center P and axis ℓ_P. Then there exist exactly $q^3/2$ self-polar triangles $T_{i,j} = \{P_i, Q_{i,j}, R_{i,j}\}$, $i = 1, \ldots, q^2$, $j = 1, \ldots, \frac{q}{2}$, such that α stabilizes $T_{i,j}$. Also, $P_i \in \ell_P$ and $P \in Q_{i,j}R_{i,j}$ for any i and j.*

Proof. The number of involutions in G is $(q^3 + 1)(q - 1)$, since for any of the $q^3 + 1$ $\mathbb{F}_{q^2}$-rational points P of $\mathcal{H}_q$ the involutions fixing P form a group E_q. The number of self-polar triangles $T \subset \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$ is $[G : M_3(T)] = \frac{(q^3+1)q^3(q^2-1)}{6(q+1)^2}$. For any self-polar triangle $T = \{A_1, A_2, A_3\} \subset \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$, the number of involutions in G stabilizing T is $3(q + 1)$. In fact, for any of the 3 vertexes of T there are exactly $q + 1$ involutions $\alpha_1, \ldots, \alpha_{q+1}$ fixing that vertex, say A_1, and interchanging A_2 and A_3; α_i is uniquely determined by its center $A_2A_3 \cap \mathcal{H}_q$. Then, by double counting the size of

$$\{(\beta, T) \mid \beta \in G,\ o(\beta) = 2,\ T \subset \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q \text{ is a self-polar triangle, } \beta \text{ stabilizes } T\},$$

α stabilizes exactly $\frac{q^3}{2}$ self-polar triangles T. For any such T, one vertex P_i of T lies on the axis of α, because α is an elation, and the other two vertexes $\{Q_{i,j}, R_{i,j}\}$ of T lie on the polar line ℓ_{P_i} of P_i. Since $M_1(P)$ is transitive on the q^2 points $P_1, \ldots, P_{q^2}$ of $\ell_P(\mathbb{F}_{q^2}) \setminus \{P\}$, any point P_i is contained in the same number $\frac{q}{2}$ of self-polar triangles $T_{i,j}$ stabilized by α. □

Lemma 3.5. *Let $\alpha \in G$ have order 3. Then there are exactly $\frac{q^2-1}{3}$ self-polar triangles*

$$T_i \subset \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q, \quad i = 1, \ldots, \tfrac{q^2-1}{3},$$

which are stabilized by α. Also, there are exactly $\frac{2(q^2-1)}{3}$ triangles

$$\tilde{T}_j = \{P_j, \Phi_{q^2}(P_j), \Phi_{q^2}^2(P_j)\} \subset \mathcal{H}_q(\mathbb{F}_{q^6}) \setminus \mathcal{H}_q(\mathbb{F}_{q^2}), \quad j = 1, \ldots, \tfrac{2(q^2-1)}{3},$$

which are stabilized by α.

Proof. By Proposition 3.2, any two subgroups of G of order 3 are conjugated in G. Also, any element of order 3 is conjugated to its inverse by an involution of G. Hence, any two element of order 3 are conjugated in G.

Now the claim follows by double counting the size of

$$\{(\beta, T) \mid \beta \in G,\ o(\beta) = 3,\ T \subset \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q \text{ is a self-polar triangle,} \\ \beta \text{ stabilizes } T\},$$

and

$$\{(\beta, \tilde{T}) \mid \beta \in G,\ o(\beta) = 3,\ \tilde{T} = \{P, \Phi_{q^2}(P), \Phi_{q^2}^2(P)\} \text{ with} \\ P \in \mathcal{H}_q(\mathbb{F}_{q^6}) \setminus \mathcal{H}_q(\mathbb{F}_{q^2}),\ \beta \text{ stabilizes } \tilde{T}\},$$

using the following facts. The number of elements of order 3 in G is $\binom{q^3+1}{2} \cdot 2$. The number of self-polar triangles $T \subset \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$ is $[G : M_3(T)]$. The number of elements of order 3 stabilizing a fixed self-polar triangle T is $2(q+1)^2$, because any element acting as a 3-cycle on the vertexes of T has order 3 (see [5, Section 3]). The number of triangles $\tilde{T} = \{P, \Phi_{q^2}(P), \Phi_{q^2}^2(P)\} \subset \mathcal{H}_q(\mathbb{F}_{q^6}) \setminus \mathcal{H}_q(\mathbb{F}_{q^2})$ is $[G : M_4(\tilde{T})]$. The number of elements of order 3 stabilizing a fixed triangle $\tilde{T}$ is $2(q^2 - q + 1)$, because any element in $M_4(\tilde{T}) \setminus C_{q^2-q+1}$ has order 3 (see [4, Section 4]). □

Lemma 3.6. *Let $H < G$ be isomorphic to* $\mathrm{Sym}(3)$*, $H = \langle\alpha\rangle \rtimes \langle\beta\rangle$. Then there are exactly $q + 1$ self-polar triangles*

$$T_i = \{P_i, Q_i, R_i\} \subset \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q, \quad i = 1, \ldots, q + 1,$$

which are stabilized by H. Up to relabeling the vertexes, we have that $P_1, \ldots, P_{q+1}$ lie on the axis of the elation β, $Q_1, \ldots, Q_{q+1}$ lie on the axis of the elation $\alpha\beta$, and $R_1, \ldots, R_{q+1}$ lie on the axis of the elation $\alpha^2\beta$.

Proof. By Proposition 3.2, any two subgroups $K_1, K_2 < G$ with $K_i \cong \mathrm{Sym}(3)$ are conjugated, and $|N_G(K_i)| = 6(q + 1)$; hence, the number of subgroups of G isomorphic to $\mathrm{Sym}(3)$ is $[G : N_G(K_i)] = \frac{(q^3+1)q^3(q-1)}{6}$. The number of self-polar triangles T is $[G : M_3(T)] = \frac{(q^2-q+1)q^3(q-1)}{6}$. Then the claim on the number of self-polar triangles follows by double counting the size of

$$\{(K, T) \mid K < G,\ K \cong \mathrm{Sym}(3),\ T \subset \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q \text{ is a self-polar triangle,} \\ K \text{ stabilizes } T\},$$

once we show that, for any self-polar triangle $T = \{A, B, C\}$, there are in G exactly $(q + 1)^2$ subgroups isomorphic to $\mathrm{Sym}(3)$ which stabilize T.

Let $K < M_3(T)$, $K \cong \mathrm{Sym}(3)$, $K = \langle\alpha, \beta\rangle$ with $o(\alpha) = 3$, $o(\beta) = 2$. Let P, Q, R be the fixed points of α, with $P \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$, $Q, R \in \mathcal{H}_q(\mathbb{F}_{q^2})$. By Proposition 3.3, $N_G(K) = K \times C_{q+1}$ where C_{q+1} is made by homologies with center P; this implies $N_G(K) \cap M_3(T) = K$. Hence, there are at least $[M_3(T) : \mathrm{Sym}(3)] = (q + 1)^2$ distinct groups $\mathrm{Sym}(3)$ stabilizing T, namely the conjugates of K through elements of $M_3(T)$. On the other side, $M_3(T)$ contains exactly $(q + 1)^2$ subgroups K of order 3, with fixed points $P \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$, $Q, R \in \mathcal{H}_q(\mathbb{F}_{q^2})$. Any involution β of $M_3(T)$ normalizing

K is uniquely determined by the vertex of T that β fixes, because $\beta(P) = P$, $\beta(Q) = R$, and $\beta(R) = Q$. Thus, K is contained in exactly one subgroup of $M_3(T)$ isomorphic to $\mathrm{Sym}(3)$. Therefore the number of subgroups isomorphic to $\mathrm{Sym}(3)$ which stabilize T is $(q+1)^2$.

Finally, the configuration of the vertexes of $T_1, \dots, T_{q+1}$ on the axes of the involutions of H follows from Lemma 2.4 and the fact that every involution fixes a different vertex of T_i. □

Proposition 3.7. *Any group H in Equation* (3.1) *is the intersection of maximal subgroups of G.*

Proof. **Case 1:** The first four groups of Equation (3.1) are exactly the maximal subgroups of G.

Case 2: Let $H = E_q \rtimes C_{q^2-1}$. Let $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$ be the unique point of $\mathcal{H}_q$ fixed by E_q; E_q fixes ℓ_P pointwise. Also, the fixed points of C_{q^2-1} are $P, Q \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and $R \in \mathrm{PG}(2,q^2) \setminus \mathcal{H}_q$, where $R \in \ell_P$ and $PQ = \ell_R$. Then $H \leq M_1(P) \cap M_2(R)$. Conversely, from $M_1(P) \cap M_2(R) \leq M_1(P)$ follows $M_1(P) \cap M_2(R) = K \rtimes C_d$ with $K \leq S_2$ and $C_d \leq C_{q^2-1}$. From $M_1(P) \cap M_2(R) \leq M_2(R)$ follows that K does not contain any element of type (D), so that $K \leq E_q$. Thus, $M_1(P) \cap M_2(R) \leq H$, and $H = M_1(P) \cap M_2(R)$.

Case 3: Let $H = (C_{q+1} \times C_{q+1}) \rtimes C_2$. Let $T = \{P, Q, R\} \subset \mathrm{PG}(2,q^2) \setminus \mathcal{H}_q$ be the self-polar triangle fixed pointwise by $C_{q+1} \times C_{q+1}$, and let P be the vertex of T fixed by C_2. Then $H \leq M_3(T) \cap M_2(P)$. Conversely, since $M_3(T) \cap M_2(P)$ fixes P and acts on $\{Q, R\}$, the pointwise stabilizer $C_{q+1} \times C_{q+1}$ of T has index at most 2 in $M_3(T) \cap M_2(P)$, so that $M_3(T) \cap M_2(P) \leq H$. Thus, $H = M_3(T) \cap M_2(P)$.

Case 4: Let $H = C_{q+1} \times C_{q+1}$. Let $T = \{P, Q, R\} \subset \mathrm{PG}(2,q^2) \setminus \mathcal{H}_q$ be the self-polar triangle fixed pointwise by $C_{q+1} \times C_{q+1}$. Since H is the whole pointwise stabilizer of T in G, we have $H = M_2(P) \cap M_2(Q) \cap M_2(R)$.

Case 5: Let $H = C_{q^2-1}$ and let α be a generator of H, with fixed points $P, Q \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and $R \in \mathrm{PG}(2,q^2) \setminus \mathcal{H}_q$. The pointwise stabilizer of $\{P, Q\}$ in G is exactly H; thus, $H = M_1(P) \cap M_2(Q)$.

Case 6: Let $H = C_{2(q+1)}$ and let α be a generator of H, of type (E), with fixed points $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and $Q \in \ell_P(\mathbb{F}_{q^2}) \setminus \mathcal{H}_q$. By Lemma 3.4 there are $\frac{q}{2}$ self-polar triangles stabilized by the involution α^{q+1} having one vertex in Q and two vertexes on ℓ_Q; let $T = \{Q, R_1, R_2\}$ be one of these triangles. Then $H \leq M_1(P) \cap M_2(Q) \cap M_3(T)$.

Conversely, let $\sigma \in (M_1(P) \cap M_2(Q) \cap M_3(T)) \setminus \{1\}$. If σ fixes $\{R_1, R_2\}$ pointwise, then from $\sigma \in M_1(P)$ follows that σ is in the kernel $C_{q+1} \leq H$ of the action of $M_2(Q)$ on ℓ_Q. The quotient $(M_1(P) \cap M_2(Q) \cap M_3(T))/C_{q+1}$ acts on ℓ_Q as a subgroup of $\mathrm{PSL}(2,q)$ fixing P and interchanging R_1 and R_2. From [17, Hauptsatz 8.27] follows $(M_1(P) \cap M_2(Q) \cap M_3(T))/C_{q+1} \cong C_2$, and hence $H = M_1(P) \cap M_2(Q) \cap M_3(T)$.

Case 7: Let $H = C_{q+1} = Z(M_2(P))$. Then H is made by the homologies of G with center P, together with the identity. Thus, $H = M_1(P_1) \cap M_1(P_2) \cap M_1(P_3)$, where P_1, P_2, P_3 are distinct point in $\ell_P \cap \mathcal{H}_q$.

Case 8: Let $H = E_q$ and let P be the unique point of $\mathcal{H}_q(\mathbb{F}_{q^2})$ fixed by any element in H. Then $H = M_2(P_1) \cap M_2(P_2) \cap M_2(P_3)$, where P_1, P_2, P_3 are distinct points in $\ell_P(\mathbb{F}_{q^2}) \setminus \{P\}$.

Case 9: Let $H = C_2$, α be a generator of H with fixed point $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$, and $P_1, P_2, P_3 \in \ell_P(\mathbb{F}_{q^2}) \setminus \{P\}$. Let $T = \{P_1, Q_{1,1}, R_{1,1}\}$ be a self-polar triangle stabilized by α. Then $H \leq M_2(P_1) \cap M_2(P_2) \cap M_2(P_3) \cap M_3(T)$. Since the elation α is uniquely determined by the image of one point not on its axis ℓ_P, $H \leq M_3(T)$ implies $H = M_2(P_1) \cap M_2(P_2) \cap M_2(P_3) \cap M_3(T)$.

Case 10: Let $H = C_3$. By Lemma 3.5, H stabilizes $\frac{2(q^2-1)}{3}$ triangles $\tilde{T} \subset \mathcal{H}_q(\mathbb{F}_{q^6}) \setminus \mathcal{H}_q(\mathbb{F}_{q^2})$; let $\tilde{T}_1$ and $\tilde{T}_2$ be two of them. Then $H \leq M_4(\tilde{T}_1) \cap M_4(\tilde{T}_2)$. If $H < M_4(\tilde{T}_1) \cap M_4(\tilde{T}_2)$, then there exist a nontrivial $\sigma \in G$ stabilizing pointwise both $\tilde{T}_1$ and $\tilde{T}_2$, a contradiction to Lemma 2.4. Thus, $H = M_4(\tilde{T}_1) \cap M_4(\tilde{T}_2)$.

Case 11: Let $H = \mathrm{Sym}(3)$. By Lemma 3.6, H stabilizes $q+1$ self-polar triangles $T_1, \ldots, T_{q+1}$, so that $H \leq M_3(T_1) \cap \cdots \cap M_3(T_{q+1})$. Suppose by contradiction that $H \neq M_3(T_1) \cap \cdots \cap M_3(T_{q+1})$. Then $M_3(T_1) \cap \cdots \cap M_3(T_{q+1})$ contains a nontrivial element σ fixing every triangle T_i pointwise. Since the triangles T_i's do not have vertexes in common, this is a contradiction to Lemma 2.4. Thus, $H = M_3(T_1) \cap \cdots \cap M_3(T_{q+1})$.

Case 12: Let $H = \{1\}$. Since G is simple, H is the Frattini subgroup of G. □

Proposition 3.8. *If $H < G$ is the intersection of maximal subgroups, then H is one of the groups in Equation* (3.1).

Proof. We proceed as follows: we take every subgroup $K < G$ in Equation (3.1), starting from the maximal subgroups M_i of G; we consider the intersections $H = K \cap M_i$ of K with the maximal subgroups of G; here, we assume that $K \not\leq M_i$. We show that H is again one of the groups in Equation (3.1).

Case 1: Let $K = S_2 \rtimes C_{q^2-1} = M_1(P)$ for some $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$.

Let $H = K \cap M_1(Q)$, $Q \neq P$. Then H is the pointwise stabilizer of $\{P, Q\} \subset \mathcal{H}_q(\mathbb{F}_{q^2})$, which is cyclic of order $q^2 - 1$, i.e. $H = C_{q^2-1}$.

Let $H = K \cap M_2(Q)$. Suppose $Q \in \ell_P$. Then $H = E_{q^2} \rtimes C_{q^2-1}$, where E_{q^2} is made by the elations with axis PQ and C_{q^2-1} is generated by an element of type (B2) with fixed points Q, P, and another point $R \in \ell_Q$. Now suppose $Q \notin \ell_P$. Then H stabilizes ℓ_Q and hence also the point $R = \ell_P \cap \ell_Q$. Then H stabilizes QR and hence also the pole A of QR; by reciprocity, $A \in PQ$. Thus, H fixes three collinear point A, P, Q, and hence every point on AP. Then $H = C_{q+1} = Z(M_2(R))$.

Let $H = K \cap M_3(T)$, $T = \{A, B, C\}$, with P on a side of T, say $P \in AB$. Then H fixes C and acts on $\{A, B\}$. Thus, H is generated by an element of type (E) with fixed points P, C and fixed lines PC, AB; hence, $H = C_{2(q+1)}$.

Let $H = K \cap M_3(T)$, $T = \{A, B, C\}$, with P out of the sides of T. By reciprocity, no vertex of T lies on ℓ_P. This implies that no elation acts on T, so that $2 \nmid |H|$; this also implies that no homology in $M_3(T)$ fixes P, so that H has no nontrivial elements fixing T pointwise. Thus $H \leq C_3$.

Let $H = K \cap M_4(T)$. By Lagrange's theorem, $H \leq C_3$.

Case 2: Let $K = \mathrm{PSL}(2, q) \times C_{q+1} = M_2(P)$ for some $P \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$.

Let $H = K \cap M_2(Q)$, $Q \neq P$, and R be the pole of PQ. If $R \in PQ$, then H is the pointwise stabilizer of PQ and is made by the elations with center R; thus, $H = E_q$. If $R \notin PQ$, then H is the pointwise stabilizer of $T = \{P, Q, R\}$; thus, $H = C_{q+1} \times C_{q+1}$.

Let $H = K \cap M_3(T)$ with $T = \{A, B, C\}$. If P is a vertex of T, then $H = (C_{q+1} \times C_{q+1}) \rtimes C_2$. If P is on a side of T but is not a vertex, say $P \in AB$, then H fixes the pole $D \in AB$ of C. Then H fixes pointwise $T' = \{P, C, D\}$ and acts on $\{A, B\}$. This implies that H fixes AB pointwise and $H = C_{q+1} = Z(M_2(C))$. If P is out of the sides of T, then no nontrivial element of H fixes T pointwise; thus, $H \leq \mathrm{Sym}(3)$.

Let $H = K \cap M_4(T)$. By Lagrange's theorem, $H \leq C_3$.

Case 3: Let $K = (C_{q+1} \times C_{q+1}) \rtimes \mathrm{Sym}(3) = M_3(T)$ for some self-polar triangle $T = \{A, B, C\}$.

Let $H = K \cap M_3(T')$ with $T' = \{A', B', C'\} \neq T$. If T and T' have one vertex $A = A'$ in common, then $H = C_{2(q+1)}$ is generated by an element of type (E) fixing A and a point $D \in BC = B'C'$. If $A' \in AC \setminus \{A, C\}$, then H stabilizes $B'C'$, because $B'C'$ is the only line containing 4 points of $\{A, B, C, A', B', C'\}$. Then H fixes A', A, and C; hence also B. Since H acts on $\{B', C'\}$, H cannot be made by nontrivial homologies of center B; thus, $H = \{1\}$.

Let $H = K \cap M_4(T')$. By Lagrange's theorem, $H \leq C_3$.

Case 4: Let $K = C_{q^2-q+1} \rtimes C_3 = M_4(T)$ for some $T \subset \mathcal{H}_q(\mathbb{F}_{q^6})$. Let $H = K \cap M_4(T')$ with $T' \neq T$. Since 3 does not divide the order of the pointwise stabilized C_{q^2-q+1} of T, H contains no nontrivial elements fixing T or T' pointwise. Thus, $H \leq C_3$.

Case 5: Let $K = E_q \rtimes C_{q^2-1}$ and $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$, $Q \in \ell_P \setminus \{P\}$ be the fixed points of K.

Let $H = K \cap M_1(R)$ with $R \neq P$. If $R \in \ell_Q$, then $H = C_{q^2-1}$. If $R \notin \ell_Q$, then H fixes the pole S of PR; by reciprocity $S \in PQ$, so that H fixes PQ pointwise and also $R \notin PQ$. Thus, $H = \{1\}$.

Let $H = K \cap M_2(R)$ with $R \neq Q$. If $R \in \ell_P$, then H is the pointwise stabilizer E_q of PQ. If $R \notin \ell_P$, then H fixes pointwise the self-polar triangle $\{Q, R, S\}$ where S is the pole of QR. Hence, either $H = C_{q+1} = Z(M_2(Q))$ or $H = \{1\}$ according to $P \in RS$ or $P \notin RS$, respectively.

Let $H = K \cap M_3(T)$ with $T = \{A, B, C\}$. If P is on a side of T, say $P \in BC$, then either $H = \{1\}$ or $H = C_{q+1} = Z(M_2(A))$. If P is out of the sides of T, then no nontrivial element of H can fix T pointwise; thus, $H \leq \mathrm{Sym}(3)$.

Let $H = K \cap M_4(T)$. By Lagrange's theorem, $H \leq C_3$.

Case 6: Let $K = (C_{q+1} \times C_{q+1}) \rtimes C_2 = M_3(T) \cap M_2(A)$, where $T = \{A, B, C\}$.

Let $H = K \cap M_1(P)$. If $P \in BC$, then $H = C_{2(q+1)}$ is generated by an element of type (E). If $P \notin BC$, then $H = \{1\}$.

Let $H = K \cap M_2(P)$, $P \neq A$. If $P \in \{B, C\}$, then H is the pointwise stabilizer $C_{q+1} \times C_{q+1}$ of T. If $P \in AB \setminus \{A, B\}$ or $P \in AC \setminus \{A, C\}$, then $H = C_{q+1} = Z(M_2(C))$ or $H = C_{q+1} = Z(M_2(B))$, respectively. If $P \in BC \setminus \{B, C\}$, then H fixes A, P, the pole of AP, and acts on $\{B, C\}$; thus, $H = C_{q+1} = Z(M_2(A))$. If P is not on the sides of T, then no nontrivial element of H can fix T pointwise; thus, $H \leq C_2$.

Let $H = K \cap M_3(T')$ with $T' = \{A', B', C'\} \neq T$. Since $3 \nmid |H|$, H fixes a vertex of T', say A'. If $A' = A$, then $H = C_{2(q+1)}$. If $A' \in \{B, C\}$, then H fixes T pointwise and acts on $\{B', C'\}$; thus, $H = C_{q+1} = Z(M_2(A'))$. If $A' \in (AB \cup AC) \setminus \{A, B, C\}$, then H fixes AB or AC pointwise and acts on $\{B', C'\}$; thus, $H = \{1\}$. If $A' \in BC$, then H

fixes A, A', and the pole D of AA'; as H acts on $\{B, C\}$, this implies $H = \{1\}$. If A' is not on the sides of T, then no nontrivial element of H fixes T pointwise and $H \leq C_2$.

Let $H = K \cap M_4(T')$. By Lagrange's theorem, $H \leq C_3$.

Case 7: Let $K = C_{q+1} \times C_{q+1} = M_3(T) \cap M_2(A) \cap M_2(B) \cap M_2(C)$ with $T = \{A, B, C\}$.

Let $H = K \cap M_1(P)$ or $H = K \cap M_2(P)$. If P is not on the sides of T, then $H = \{1\}$; if P is on a side of T, say $P \in BC$, then $H = C_{q+1} = Z(M_2(A))$.

Let $H = K \cap M_3(T')$ with $T' = \{A', B', C'\}$. Since K is not divisible by 2 or 3, $H \neq \{1\}$ only if H fixes T' pointwise. Up to relabeling, this implies $A' = A$, $B', C' \in BC$, and $H = C_{q+1} = Z(M_2(A))$.

Let $H = K \cap M_4(T')$. By Lagrange's theorem, $H = \{1\}$.

Case 8: Let $K = C_{q^2-1} = \langle \alpha \rangle$, with α of type (B2) fixing the points $P \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$ and $Q, R \in \mathcal{H}_q(\mathbb{F}_{q^2})$.

Let $H = K \cap M_1(A)$ or $H = K \cap M_2(A)$. Since the nontrivial elements of H are either of type (B2) or of type (A) with axis QR, we have $H = \{1\}$ unless $A \in QR$; in this case, $H = C_{q+1} = Z(M_2(P))$.

Let $H = K \cap M_3(T)$ or $H = K \cap M_4(T)$. By Lagrange's theorem, $H \leq C_3$.

Case 9: Let $K = C_{2(q+1)} = \langle \alpha \rangle$ with α of type (E) fixing the points $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and $Q \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$.

Let $H = K \cap M_1(R)$ or $H = K \cap M_2(R)$. If $R \in \ell_Q$, then $H = C_{q+1} = Z(M_2(Q))$. If $R \notin \ell_Q$, then $H = \{1\}$.

Let $H = K \cap M_3(T)$; recall that $H < K$. If Q is a vertex of T, then $H = C_{q+1} = Z(M_2(Q))$. If Q is not a vertex of T, then no homology in K acts on T; hence, $H \leq C_2$.

Let $H = K \cap M_4(T)$. By Lagrange's theorem, $H = \{1\}$.

Case 10: Let $K = C_{q+1} = Z(M_2(P))$ for some $P \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$ and $\sigma \in K \setminus \{1\}$. Then σ fixes no points out of $\{P\} \cup \ell_P$; also, the triangles fixed by σ have one vertex in P and two vertexes on ℓ_P. Thus, $K \cap M_i = \{1\}$ for any maximal subgroup M_i of G not containing K.

Case 11: Let $K = E_q$ and $\sigma \in E_q \setminus \{1\}$. Recall that K fixes one point $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and the line ℓ_P pointwise. Also, σ fixes no points out of ℓ_P. If σ fixes a triangle $T = \{A, B, C\}$, then one vertex of T lies on $\ell_P(\mathbb{F}_{q^2})$, say A, and σ is uniquely determined by $\sigma(B) = C$. Thus, $K \cap M_1(Q) = K \cap M_2(Q) = K \cap M_4(T') = \{1\}$ and $K \cap M_3(T) \leq C_2$.

Case 12: Let $K \in \{\mathrm{Sym}(3), C_3, C_2, \{1\}\}$. Then every subgroup of K is in Equation (3.1). □

Proposition 3.9. *The values $\mu(H)$ for the groups in Equation* (3.1) *are given in Equation* (3.3).

Proof. Let H be one of the groups in Equation (3.1). By Lemma 2.1 and Proposition 3.8, $\mu(H)$ only depends on the subgroups K of G such that $H < K$ and K is in Equation (3.1).

Case 1: If H is one of the first four groups in Equation (3.1), then H is maximal in G, and hence $\mu(H) = -1$.

Case 2: Let $H = E_q \rtimes C_{q^2-1}$. Let $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and $Q \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$ be the fixed points of H. Then $H = M_1(P) \cap M_2(Q)$ and H is not contained in any other maximal subgroup of G. Thus, $\mu(H) = -\{\mu(G) + \mu(M_1(P)) + \mu(M_2(Q))\} = 1$.

Case 3: Let $H = (C_{q+1} \times C_{q+1}) \rtimes C_2$. Let $T = \{P, Q, R\}$ be the self-polar triangle stabilized by H, with $H(P) = P$. No point different from P is fixed by H. Also, if a triangle $T' = \{P', Q'\} \neq T$ is fixed by H, then P is a vertex of T', say $P = P'$, and $\{Q', R'\} \subset QR$; but $C_{q+1} \times C_{q+1}$ has orbits of length $q + 1 > |\{Q', R'\}|$, so that H cannot fix T'. Then $H = M_2(P) \cap M_3(T)$ and H is not contained in any other maximal subgroup of G. Thus, $\mu(H) = 1$.

Case 4: Let $H = C_{q+1} \times C_{q+1}$ and $T = \{P, Q, R\}$ be the self-polar triangle fixed pointwise by H. The vertexes of T are the unique fixed points of the elements of type (B1) in H. Also, any triangle $T' \neq T$ fixed by an element of type (A) in H has two vertexes on a side ℓ of T; but H has orbits of length $q + 1 > 2$ on ℓ, so that H does not fix T'. Then $H = M_3(T) \cap M_2(P) \cap M_2(Q) \cap M_2(R)$ and H is not contained in any other maximal subgroup of G.

If K is one of the groups $M_3(T) \cap M_2(P)$, $M_3(T) \cap M_2(P)$, $M_3(T) \cap M_2(P)$, then K contains H properly, and $\mu(K) = 1$ as shown in the previous point. The intersection of three groups between $M_3(T)$, $M_2(P)$, $M_2(Q)$, and $M_2(R)$ is equal to H. Thus, by direct computation, $\mu(H) = 0$.

Case 5: Let $H = C_{q^2-1}$ with fixed points $P \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$ and $Q, R \in \mathcal{H}_q(\mathbb{F}_{q^2})$. Then $H = M_1(Q) \cap M_1(R) = M_1(Q) \cap M_1(R) \cap M_2(P)$. We already know $\mu(M_1(Q) \cap M_2(P)) = \mu(M_1(R) \cap M_2(P)) = 1$. Moreover, C_{q^2-1} has no fixed triangles, by Lagrange's theorem, and no other fixed points. Thus, by direct computation, $\mu(H) = 0$.

Case 6: Let $H = C_{2(q+1)} = \langle\alpha\rangle$; α is of type (E), fixes the points $P \in \mathcal{H}_q(\mathbb{F}_{q^2})$ and $Q \in PG(2, q^2) \setminus \mathcal{H}_q$, and fixes the lines ℓ_P and ℓ_Q. Since α^2 is a homology with center Q, the orbits on ℓ_Q of H coincide with the orbits on ℓ_Q of the elation α^{q+1}. By Lemma 3.4, the self-polar triangles T_i stabilized by H have a vertex in Q and two vertexes on ℓ_Q; there are exactly $\frac{q}{2}$ such triangles $T_1, \ldots, T_{\frac{q}{2}}$. No other triangle and no other point different from P and Q is fixed by H, so that $H = M_1(P) \cap M_2(Q) \cap M_3(T_1) \cap \cdots \cap M_3(T_{\frac{q}{2}})$ and H is not contained in any other maximal subgroup of G.

If K is the intersection of $M_2(Q)$ with one of the groups $M_1(P), M_3(T_1), \ldots, M_3(T_{\frac{q}{2}})$, then $K = E_q \rtimes C_{q^2-1}$ or $K = (C_{q+1} \times C_{q+1}) \rtimes C_2$; hence, K contains H properly and $\mu(K) = 1$ as shown above. The intersection of K with a third maximal subgroup of G containing H coincides with H. Finally, the intersection of any two groups in $\{M_1(P), M_3(T_1), \ldots, M_3(T_{\frac{q}{2}})\}$ coincides with H. Thus, by direct computation, $\mu(H) = 0$.

Case 7: Let $H = C_{q+1} = Z(M_2(P))$. Denote $\ell_P \cap \mathcal{H}_q = \{P_1, \ldots, P_{q+1}\}$ and $\ell(\mathbb{F}_{q^2}) \setminus \mathcal{H}_q = \{Q_1, \ldots, Q_{q^2-q}\}$ such that, for $i = 1, \ldots, \frac{q^2-q}{2}$, $T_i = \{P, Q_i, Q_{i+\frac{q^2-q}{2}}\}$ are the self-polar triangles with a vertex in P. Then

$$H = \bigcap_{i=1}^{q+1} M_1(P_i) \cap M_2(P) \cap \bigcap_{i=1}^{q^2-q} M_2(Q_i) \cap \bigcap_{i=1}^{(q^2-q)/2} M_3(T_i)$$

and H is not contained in any other maximal subgroup of G. By direct inspection, the intersections K of some (at least two) maximal subgroups of G such that $H < K < G$ are exactly the following.

(i) $K = M_1(P_i) \cap M_1(P_j)$ for some $i \neq j$; in this case, $K = C_{q^2-1}$ and $\mu(K) = 0$.

(ii) $K = M_1(P_i) \cap M_2(P)$ with $i \in \{1, \dots, q+1\}$; in this case, $K = E_q \rtimes C_{q^2-1}$ and $\mu(K) = 1$. These $q+1$ groups are pairwise distinct.

(iii) $K = M_1(P_i) \cap M_3(T_j)$ for some i, j; in this case, $K = C_{2(q+1)}$ and $\mu(K) = 0$.

(iv) $K = M_2(P) \cap M_2(Q_i)$ for some i; in this case, $K = C_{q+1} \times C_{q+1}$ and $\mu(K) = 0$.

(v) $K = M_2(P) \cap M_3(T_i)$ with $i \in \{1, \dots, \frac{q^2-q}{2}\}$; in this case, $K = (C_{q+1} \times C_{q+1}) \rtimes C_2$ and $\mu(K) = 1$. These $\frac{q^2-q}{2}$ groups are pairwise distinct.

(vi) $K = M_2(Q_i) \cap M_3(T_i)$ or $K = M_2(Q_{i+\frac{q^2-q}{2}}) \cap M_3(T_i)$, with $i \in \{1, \dots, \frac{q^2-q}{2}\}$; in this case, $K = (C_{q+1} \times C_{q+1}) \rtimes C_2$ and $\mu(K) = 0$. These $q^2 - q$ groups are pairwise distinct.

To sum up, the only subgroups K with $H < K < G$ and $\mu(K) \neq 0$ are the maximal subgroups, $q+1$ distinct groups of type $E_q \rtimes C_{q^2-1}$, and $\frac{3(q^2-q)}{2}$ distinct groups of type $(C_{q+1} \times C_{q+1}) \rtimes C_2$. Thus, $\mu(H) = 0$.

Case 8: Let $H = E_q$. Let P be the point of $\mathcal{H}_q(\mathbb{F}_{q^2})$ fixed by H; H fixes ℓ_P pointwise. We have $H = M_1(P) \cap M_2(Q_1) \cap \cdots \cap M_2(Q_{q^2})$, where $Q_1, \dots, Q_{q^2}$ are the $\mathbb{F}_{q^2}$-rational points of $\ell_P \setminus \{P\}$; H is not contained in any other maximal subgroup of G. The intersections K of at least two maximal subgroups of G such that $H < K < G$ are exactly the q^2 groups $M_1(P) \cap M_2(Q_i) = E_q \rtimes C_{q^2-1}$, with $\mu(K) = 1$. Thus, by direct computation, $\mu(H) = 0$.

Case 9: Let $H = \mathrm{Sym}(3) = \langle \alpha, \beta \rangle$ with $o(\alpha) = 3$ and $o(\beta) = 2$. Let $P \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$ and $Q, R \in \mathcal{H}_q$ be the fixed points of α, and $A \in QR$ be the fixed point of β on $\mathcal{H}_q$, so that β fixes $\ell_A = AP$. By Lemma 3.6 and its proof, $H = M_2(P) \cap M_3(T_1) \cap \cdots \cap M_3(T_{q+1})$, where T_i has one vertex on $\ell_A \setminus \{P, A\}$ and the other two vertexes are collinear with A; H is not contained in any other maximal subgroup of G.

For any $i, j \in \{1, \dots, q+1\}$ with $i \neq j$, no vertex of T_j is on a side of T_i; hence, no nontrivial element of $M_3(T_i) \cap M_3(T_j)$ fixes T_i pointwise. This implies $M_3(T_i) \cap M_3(T_j) = H$. Analogously, no nontrivial element in $M_3(T_i) \cap M_2(P)$ fixes T_i pointwise, and this implies $M_3(T_i) \cap M_2(P) = H$. Thus, by direct computation, $\mu(H) = q+1$.

Case 10: Let $H = C_3 = \langle \alpha \rangle$ with fixed points $P \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$ and $Q, R \in \mathcal{H}_q$. By Lemma 3.5,

$$H = M_1(Q) \cap M_1(R) \cap M_2(P) \cap \bigcap_{i=1}^{(q^2-1)/3} M_3(T_i) \cap \bigcap_{i=1}^{2(q^2-1)/3} M_4(\tilde{T_i})$$

and H is not contained in any other maximal subgroup of G. By direct inspection, the intersections K of at least two maximal subgroups of G such that $H < K < G$ are exactly the following.

(i) $K = M_1(Q) \cap M_2(P)$ or $K = M_1(R) \cap M_2(P)$; in this case, $K = E_q \rtimes C_{q^2-1}$ and $\mu(K) = 1$.

(ii) $K = M_1(Q) \cap M_1(R)$; in this case, $K = C_{q^2-1}$ and $\mu(K) = 0$.

(iii) There are exactly $\frac{q-1}{3}$ groups K containing H with $K \cong \mathrm{Sym}(3)$, and hence $\mu(K) = q+1$. In fact, any involution $\beta \in G$ satisfying $\langle H, \beta\rangle \cong \mathrm{Sym}(3)$ interchanges Q and R and fixes a point of $(QR \cap \mathcal{H}_q) \setminus \{P, Q\}$; conversely, any of the $q-1$ points $A_1, \dots, A_{q-1}$ of $(QR \cap \mathcal{H}_q) \setminus \{P, Q\}$ determines uniquely the involution $\beta_i \in G$ such that $\beta(A_i)$, $\beta_i(Q) = R$, $\beta_i(R) = Q$, and hence $\langle H, \beta_i\rangle \cong \mathrm{Sym}(3)$. The involutions β_i, $\alpha\beta_i$, and $\alpha^2\beta_i$, together with H, generate the same group; thus, there are exactly $\frac{q-1}{3}$ groups $\mathrm{Sym}(3)$ containing H.

Thus, by direct computation, $\mu(H) = \frac{2(q^2-1)}{3}$.

Case 11: Let $H = C_2 = \langle\alpha\rangle$, where α has center P. Let $\ell_P(\mathbb{F}_{q^2}) \setminus \{P\} = \{P_1, \dots, P_{q^2}\}$. By Lemma 3.4,

$$H = M_1(P) \cap \bigcap_{i=1}^{q^2} M_2(P_i) \cap \bigcap_{i=1}^{q^2} \bigcap_{j=1}^{q/2} M_3(T_{i,j}),$$

where the triangles $T_{i,j}$ are described in Lemma 3.4; H is not contained in any other maximal subgroup of G. By direct inspection, the intersections K of at least two maximal subgroups of G such that $H < K < G$ are exactly the following.

(i) $K = M_1(P) \cap M_2(P_i)$ for $i = 1, \dots, q^2$; in this case, $K = E_q \rtimes C_{q^2-1}$ and $\mu(K) = 0$.

(ii) $K = M_2(P_i) \cap M_2(P_j)$ with $i \neq j$; in this case, $K = E_q$ and $\mu(K) = 0$.

(iii) $K = M_1(P) \cap M_3(T_{i,j})$; in this case, $K = E_q \rtimes C_{2(q+1)}$ and $\mu(K) = 0$.

(iv) $K = M_2(Q_i) \cap M_3(T_{i,j})$ with $i \in \{1, \dots, q^2\}$ and $j \in \{1, \dots, \frac{q}{2}\}$; these $\frac{q^3}{2}$ distinct groups are of type $(C_{q+1} \times C_{q+1}) \rtimes C_2$, so that $\mu(K) = 1$.

(v) There are exactly $N = \frac{q^3}{2}$ groups K containing H such that $K \cong \mathrm{Sym}(3)$, and hence $\mu(K) = q+1$. This follows by double counting the size of

$$I = \{(H, K) \mid H, K < G,\ H \cong C_2,\ K \cong \mathrm{Sym}(3),\ H < K\}.$$

Arguing as in the proof of Lemma 3.4, $|I| = (q^3+1)(q-1)N$; arguing as in the proof of Lemma 3.6, $|I| = \frac{q^3(q^3+1)(q-1)}{6} \cdot 3$. Hence, $N = \frac{q^3}{2}$.

Thus, by direct computation, $\mu(H) = -\frac{q^3(q+1)}{2}$.

Case 12: Let $H = \{1\}$. Then $\mu(H) = -\sum_{\{1\} < K \leq G} \mu(K, G)$. By the values $\mu(K)$ computed in the previous cases, Propositions 3.2, and Proposition 3.3, only the following groups K have to be considered:

(i) 1 group G;

(ii) q^3+1 groups $S_2 \rtimes C_{q^2-1}$;

(iii) $q^2(q^2-q+1)$ groups $\mathrm{PSL}(2, q) \times C_{q+1}$;

(iv) $\frac{q^3(q-1)(q^2-q+1)}{6}$ groups $(C_{q+1} \times C_{q+1}) \rtimes \mathrm{Sym}(3)$;

(v) $\frac{q^3(q+1)^2(q-1)}{3}$ groups $C_{q^2-q+1} \rtimes C_3$;

(vi) $(q^3+1)q^2$ groups $E_q \rtimes C_{q^2-1}$;

(vii) $\frac{q^3(q-1)(q^2-q+1)}{2}$ groups $(C_{q+1} \times C_{q+1}) \rtimes C_2$;

(viii) $\frac{q^3(q^3+1)(q-1)}{6}$ groups $\mathrm{Sym}(3)$;

(ix) $\frac{q^3(q^3+1)}{2}$ groups C_3;

(x) $(q^3+1)(q-1)$ groups C_2.

Thus, by direct computation, $\mu(H) = 0$. □

4 Determination of $\lambda(H)$ for any subgroup H of G

Let $n > 0$, $q = 2^{2^n}$, $G = \mathrm{PSU}(3, q)$. This section is devoted to the proof of the following theorem.

Theorem 4.1. *Let H be a proper subgroup of G. Then $\lambda(H) \neq 0$ if and only H is one of the following groups:*

$$\begin{array}{lll} E_q \rtimes C_{q^2-1}, & (C_{q+1} \times C_{q+1}) \rtimes C_2, & \mathrm{Sym}(3), \\ C_3, & S_2 \rtimes C_{q^2-1}, & \mathrm{PSL}(2,q) \times C_{q+1}, \\ (C_{q+1} \times C_{q+1}) \rtimes \mathrm{Sym}(3), & C_{q^2-q+1} \rtimes C_3, & C_2. \end{array} \tag{4.1}$$

For any isomorphism type in Equation (4.1) *there is just one conjugacy class of subgroups of G.*

If H is in the first row of Equation (4.1)*, then $\lambda(H) = -1$; if H is in the second row of Equation* (4.1)*, then $\lambda(H) = 1$.*

Proof. By Proposition 3.2, for any isomorphism type in Equation (4.1) there is just one conjugacy class of subgroups of G of that type. Hence, we can use the notation $[M_1]$, $[M_2]$, $[M_3]$ and $[M_4]$ for the conjugacy classes of $M_1(P)$, $M_2(P)$, $M_3(T)$ and $M_4(T)$, respectively. If $H = G$, then $\lambda(H) = 1$; if H is one of the groups in the second row of Equation (4.1) and $H \neq C_2$, then $\lambda(H) = -1$ as H is maximal in G.

Case 1: Firstly, we assume that H is not a subgroup of $\mathrm{Sym}(3)$, and that H is not a group of homologies, i.e. $H \not\leq C_{q+1} = Z(M_2(Q))$ for any point Q.

(i) Let $H < M_4(T)$ for some T. From $H \neq C_3$ follows that some nontrivial element in H fixes T pointwise; hence, H is not contained in any maximal subgroup of G other than $M_4(T)$. Thus, inductively, $\lambda(H) = -\{\lambda(G) + \lambda(M_4(T))\} = 0$.

(ii) Let $H < M_1(P)$ for some P; we assume in addition that $\gcd(|H|, q-1) > 1$. Here, the assumption $H \not\leq \mathrm{Sym}(3)$ reads $H \notin \{\{1\}, C_2, C_3\}$. If H contains an element of order 4, then H is not contained in any maximal subgroup of G other than $M_1(P)$. Thus, inductively, $\lambda(H) = 0$.

We can then assume that the 2-elements of H are involutions, so that $H = E_{2^r} \rtimes C_d$ with $0 \leq r \leq 2^n$ and $d \mid (q^2-1)$ (see [15, Theorem 11.49]). This implies that $H \leq M_1(P) \cap M_2(Q)$ for some $Q \in \ell_P$; the eventual nontrivial elements in H whose order divides $q+1$ are homologies with center Q. Then we have $[H] \leq [M_1]$, $[H] \leq [M_2]$; by Lagrange's theorem, $[H] \not\leq [M_4]$. From the assumptions $\gcd(|H|, q-1) > 1$ and $H \not\leq \mathrm{Sym}(3)$ follows $[H] \not\leq [M_3]$.

If $H = E_q \rtimes C_{q^2-1}$, then no proper subgroup of $M_1(P)$ or $M_2(Q)$ contains H properly; thus, $\lambda(H) = 1$. If $H \neq E_q \rtimes C_{q^2-1}$, then $H < E_q \rtimes C_{q^2-1} = M_1(P) \cap$

$M_2(Q)$ up to conjugation. Thus, inductively, the only classes $[K]$ with $[H] \le [K]$ and $\lambda(K) \neq 0$ are $[K] \in \{[G], [M_1], [M_2], [E_q \rtimes C_{q^2-1}]\}$. This implies $\lambda(H) = 0$.

(iii) Let $H < M_2(Q)$ for some Q, and assume also $H \not\le M_1(P)$ for any P. As $H \not\le C_3$, we have $[H] \not\le [M_4]$. The group $\bar{H} := H/(H \cap Z(M_2(Q)))$ acts as a subgroup of $\mathrm{PSL}(2, q)$ on $\ell_Q \cap \mathcal{H}_q$; we assume in this point that H is one of the following groups (see [17, Hauptsatz 8.27]): $\mathrm{PSL}(2, 2^{2^h})$ with $0 < h \le n$; a dihedral group of order $2d$ where d is a divisor of $q-1$ greater than 3; $\mathrm{Alt}(5)$. Then, by Lagrange's theorem, $[H] \not\le [M_3]$. Thus, inductively, G and $M_2(Q)$ are the only groups K with $H < K$ and $\lambda(K) \neq 0$, so that $\lambda(H) = 0$.

Note that, since we are under the assumptions $H \not\le M_1(P)$ for any P, $H \not\le \mathrm{Sym}(3)$, and $H \not\le C_{q+1} = Z(M_2(Q))$, we have that the only subgroups $\bar{H}$ of $\mathrm{PSL}(2, q)$ for which $\lambda(H)$ still has not been computed are the cyclic or dihedral groups of order d or $2d$ (respectively), where d is a nontrivial divisor of $q + 1$.

(iv) Let $H < M_3(T)$ for some T, and assume also $H \not\le M_1(P)$ for any P. As $H \not\le C_3$, we have $[H] \not\le [M_4]$. Here, the assumption $H \not\le \mathrm{Sym}(3)$ means that some nontrivial element of H fixes T pointwise. Hence, the assumption $H \not\le C_{q+1} = Z(M_2(Q))$ for any vertex Q of T, together with $H \not\le M_1(P)$, implies that H contains some element of type (B1). Write $H = L \rtimes K$, with $K \le \mathrm{Sym}(3)$ and $L < C_{q+1} \times C_{q+1}$.

If $K = C_3$ or $K = \mathrm{Sym}(3)$, then $[H] \not\le [M_2]$; thus, inductively, G and $M_3(T)$ are the only groups K with $H < K$ and $\lambda(K) \neq 0$, so that $\lambda(H) = 0$.

If $K = C_2$ and $L = C_{q+1} \times C_{q+1}$, then $H \le M_2(Q)$ for some vertex Q of T. Since $\bar{H} := H/(H \cap Z(M_2(Q)))$ is dihedral of order $2(q + 1)$, [17, Haptsatz 8.27] implies the non-existence of groups K with $H < K < M_2(Q)$ (except for $q = 4$ and $\bar{K} = \mathrm{Alt}(5)$; in this case, $\lambda(K) = 0$ by the previous point). Thus, $\lambda(H) = -\{\lambda(G) + \lambda(M_2(Q)) + \lambda(M_3(T))\} = 1$.

If $K = C_2$ and $L < C_{q+1} \times C_{q+1}$, then again $H \le M_2(Q)$ with Q vertex of T. The group $\bar{H}$ is dihedral of order $2d$, where $d \mid (q + 1)$; $d > 1$ because L contains elements of type (B1). By the previous point and [17, Hauptsatz 8.27], the only groups K with $H < K < M_2(Q)$ are such that $\bar{K}$ is dihedral of order dividing $q + 1$. Thus, inductively, $\lambda(H) = 0$.

If $K = \{1\}$, then $H \in M_2(Q)$ for any vertex Q of T. The group $\bar{H} < \mathrm{PSL}(2, q)$ on the line $\ell_Q \cap \mathcal{H}_q$ is cyclic of order $d \mid (q+1)$; $d > 1$ because H has elements of type (B1). By [17, Hauptsatz 8.27], the groups K with $H < K < M_2(Q)$ are such that either $\bar{K}$ is cyclic of order dividing $q + 1$, or we have already proved that $\lambda(K) = 0$. Thus, inductively, $\lambda(K) = 0$.

(v) Let $H < M_2(Q)$ for some Q. Let $\bar{H} \neq \{1\}$ be the induced subgroup of $\mathrm{PSL}(2, q)$ acting on $\ell_Q \cap \mathcal{H}_q$. If $\bar{H}$ is cyclic or dihedral of order d or $2d$ (respectively) with $d \mid (q + 1)$, then $H \le M_3(T)$ for some T. Hence, $\lambda(H) = 0$, as already computed in the previous point in the case $K = \{1\}$ if $\bar{H}$ is cyclic, or in the case $K = C_2$ if H is dihedral.

(vi) Under the assumptions that $H \not\le \mathrm{Sym}(3)$ and H is not a group of homologies, the only remaining case is $H < M_1(P)$ for some P with $\gcd(|H|, q - 1) = 1$. In this case $H = E_{2^r} \times C_d$, where C_d is cyclic of order $d \mid (q+1)$ and made by homologies, whose axis passes through P and whose center Q lies on ℓ_P. We have $r > 0$, because $H \not\le Z(M_2(Q))$.

If $r = 1$, then H is cyclic of order $2d$ generated by an element of type (E). By Lemma 3.4, $H \leq M_3(T)$, where T has a vertex in Q and two vertexes on ℓ_Q. Hence, $[H] \leq [M_1]$, $[H] \leq [M_2]$, $[H] \leq [M_3]$, and $[H] \not\leq [M_4]$. Let K be such that $H < K \leq G$ and K is not of the same type of H, i.e. K is not cyclic of order $2d'$ with $d' \mid (q+1)$. As shown in the previous points, $\lambda(K) \neq 0$ if and only if $[K] \in \{[G], [M_1], [M_2], [M_3], [E_q \rtimes C_{q^2-1}], [(C_{q+1} \times C_{q+1}) \rtimes C_2]\}$. Thus, inductively, $\lambda(H) = 0$.

Case 2: Let $H \leq C_{q+1} = Z(M_2(Q))$ for some Q and K be a subgroup of G properly containing H. As shown above, $\lambda(K) \neq 0$ if and only if

$$[K] \in \{[G], [M_1], [M_2], [M_3], [E_q \rtimes C_{q^2-1}], [(C_{q+1} \times C_{q+1}) \rtimes C_2]\}.$$

Thus $\lambda(Z(M_2(Q))) = 0$ and, inductively, $\lambda(H) = 0$.

Case 3: Let $H = \mathrm{Sym}(3) = \langle\alpha\rangle \rtimes \langle\beta\rangle$ with $o(\alpha) = 3$ and $o(\beta) = 2$. Let $P \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$ and $Q, R \in \mathcal{H}_q(\mathbb{F}_{q^2})$ be the fixed point of α, so that β fixes P and interchanges Q and R. This implies $[H] \leq [M_2]$. By Lemma 3.6, $[H] \leq [M_3]$. From the computations above and Lagrange's theorem, no class $[K]$ with $K \leq G$ other than $[G]$, $[M_2]$ and $[M_3]$ satisfies $[H] \leq [K]$ and $\lambda(H) \neq 0$. Thus, $\lambda(H) = 1$.

Case 4: Let $H = C_3$. By Lagrange's theorem and Proposition 3.2, $H < K \leq G$ and $\lambda(K) \neq 0$ if and only if

$$[K] \in \{[G], [M_1], [M_2], [M_3], [M_4], [E_q \rtimes C_{q^2-1}], [\mathrm{Sym}(3)]\}.$$

Thus, $\lambda(H) = 1$.

Case 5: Let $H = C_2$. By Lagrange's theorem and Proposition 3.2, $H < K \leq G$ and $\lambda(K) \neq 0$ if and only if

$$[K] \in \{[G], [M_1], [M_2], [M_3], [E_q \rtimes C_{q^2-1}], [(C_{q+1} \times C_{q+1}) \rtimes C_2], [\mathrm{Sym}(3)]\}.$$

Thus, $\lambda(H) = -1$.

Case 6: Let $H = \{1\}$. Collecting all the classes $[K]$ with $\lambda(K) \neq 0$, we have by direct computation $\lambda(H) = 0$. □

5 Determination of $\chi(\Delta(L_p \setminus \{1\}))$ for any prime p

Let $n > 0$, $q = 2^{2^n}$, $G = \mathrm{PSU}(3, q)$. If p is a prime number, we denote by L_p the poset of p-subgroups of G ordered by inclusion, by $L_p \setminus \{1\}$ its subposet of proper p-subgroups of G, and by $\Delta(L_p \setminus \{1\})$ the order complex of $L_p \setminus \{1\}$. In this section we determine the Euler characteristic $\chi(\Delta(L_p \setminus \{1\}))$ of $\Delta(L_p \setminus \{1\})$ for any prime p, using Equation (2.1) and Lemma 2.2. The results are stated in Theorem 5.1 and in Table 2.

Theorem 5.1. *For any prime number p one of the following cases holds:*

(i) $p \nmid |G|$ *and* $\chi(\Delta(L_p \setminus \{1\})) = 0$*;*

(ii) $p = 2$ *and* $\chi(\Delta(L_2 \setminus \{1\})) = q^3 + 1$*;*

(iii) $p \mid (q+1)$ *and* $\chi(\Delta(L_p \setminus \{1\})) = -\frac{q^6 - 2q^5 - q^4 + 2q^3 - 3q^2}{3}$*;*

(iv) $p \mid (q-1)$ *and* $\chi(\Delta(L_p \setminus \{1\})) = -\frac{q^6+q^3}{2}$*;*

(v) $p \mid (q^2-q+1)$ *and* $\chi(\Delta(L_p \setminus \{1\})) = -\frac{q^6+q^5-q^4-q^3}{3}$.

Proof. Since $|G| = q^3(q+1)^2(q-1)(q^2-q+1)$, q is even, and $3 \mid (q-1)$, the cases $p \nmid |G|$, $p = 2$, $p \mid (q+1)$, $p \mid (q-1)$, and $p \mid (q^2-q+1)$ are exhaustive and pairwise incompatible. We denote by S_p a Sylow p-subgroup of G.

Case 1: Let $p \nmid |G|$. Then $\Delta(L_p \setminus \{1\}) = \emptyset$, and hence $\chi(\Delta(L_p \setminus \{1\})) = \chi(\emptyset) = 0$.

Case 2: Let $p = 2$. The group G has q^3+1 Sylow 2-subgroups, and any two of them intersect trivially; see [15, Theorem 11.133]. Any nontrivial element σ of S_2 fixes exactly one point P on $\mathcal{H}_q(\mathbb{F}_{q^2})$ which is the same for any $\sigma \in S_2$; S_2 is uniquely determined among the Sylow 2-subgroups of G by P. Hence, Equation (2.1) reads

$$\chi(\Delta(L_2 \setminus \{1\})) = -(q^3+1) \sum_{H \in L_2 \setminus \{1\},\ H(P)=P} \mu_{L_2}(\{1\}, H),$$

where P is a given point of $\mathcal{H}_q(\mathbb{F}_{q^2})$. By Lemma 2.2, we only consider those 2-groups in $M_1(P)$ which are elementary abelian. Then we consider all nontrivial subgroups H of an elementary abelian 2-group E_q of order q. For any such group $H = E_{2^r}$ of order 2^r, with $1 \leq r \leq 2^n$, we have $\mu_{L_2}(\{1\}, H) = (-1)^r \cdot 2^{\binom{r}{2}}$ by Lemma 2.2. Thus,

$$\chi(\Delta(L_2 \setminus \{1\})) = -(q^3+1) \sum_{r=1}^{2^n} (-1)^r \, 2^{\binom{r}{2}} \binom{2^n}{r}_2$$

where the Gaussian coefficient $\binom{2^n}{r}_2$ counts the subgroups of E_q of order 2^r. Using the property

$$\binom{2^n}{r}_2 = \binom{2^n-1}{r-1}_2 + 2^r \binom{2^n-1}{r}_2$$

we obtain

$$\begin{aligned}
&\sum_{r=1}^{2^n} (-1)^r \, 2^{\binom{r}{2}} \binom{2^n}{r}_2 \\
&\qquad = \sum_{r=1}^{2^n} (-1)^r \, 2^{\binom{r}{2}} \binom{2^n-1}{r-1}_2 + \sum_{r=1}^{2^n} (-1)^r \, 2^{\binom{r}{2}+r} \binom{2^n-1}{r}_2 \\
&\qquad = \sum_{r=0}^{2^n-1} (-1)^{r+1} \, 2^{\binom{r+1}{2}} \binom{2^n-1}{r}_2 + \sum_{r=1}^{2^n} (-1)^r \, 2^{\binom{r+1}{2}} \binom{2^n-1}{r}_2 \\
&\qquad = (-1)^0 \, 2^{\binom{1}{2}} \binom{2^n-1}{0}_2 + (-1)^{2^n} \, 2^{\binom{2^n+1}{2}} \binom{2^n-1}{2^n}_2 = -1.
\end{aligned}$$

Thus, $\chi(\Delta(L_2 \setminus \{1\})) = q^3+1$.

Case 3: Let $p \mid (q+1)$. Then $S_p \leq C_{q+1} \times C_{q+1}$, and hence $S_p \cong C_{p^s} \times C_{p^s}$, where $p^s \mid (q+1)$ and $p^{s+1} \nmid (q+1)$. Let H be a subgroup of S_p. By Lemma 2.2, $\mu_{L_p}(\{1\}, H) \neq 0$ only if H is elementary abelian of order p or p^2; in this cases, $\mu_{L_p}(\{1\}, C_p) = -1$ and $\mu_{L_p}(\{1\}, C_p \times C_p) = r$. Now we count the number of elementary abelian subgroups of order p or p^2 in G.

(i) A subgroup E_{p^2} of G of type $C_p \times C_p$ is uniquely determined by the maximal subgroup $M_3(T)$ such that E_{p^2} is the Sylow p-subgroup of $M_3(T)$. Hence, G contains exactly $[G : N_G(M_3(T))] = \frac{q^3(q^2-q+1)(q-1)}{6}$ elementary abelian subgroups of order p^2.

(ii) A subgroup C_p made by homologies is uniquely determined by its center $P \in \mathrm{PG}(2, q^2) \setminus \mathcal{H}_q$ of homology, because the group of homologies with center P is cyclic. Hence, G contains exactly $|\,\mathrm{PG}(2, q^2) \setminus \mathcal{H}_q| = q^2(q^2 - q + 1)$ cyclic subgroups of order p made by homologies.

(iii) A subgroup C_p which is not made by homologies is made by elements of type (B1), and fixes pointwise a unique self-polar triangle T. The Sylow p-subgroup $C_p \times C_p$ of $M_3(T)$ contains exactly 3 subgroups C_p made by homologies, namely the groups of homologies with center one of the vertexes of T. Since $C_p \times C_p$ contains $p + 1$ subgroups C_p altogether, $C_p \times C_p$ contains exactly $p - 2$ subgroups C_p not made by homologies. Thus, the number of subgroups C_p of G not made by homologies is $(p - 2) \cdot [G : N_G(M_3(T))] = \frac{q^3(q^2-q+1)(q-1)(p-2)}{6}$.

Thus, by direct computation,

$$\begin{aligned}
&\chi(\Delta(L_p \setminus \{1\})) \\
&\qquad = -\left\{\frac{q^3(q^2 - q + 1)(q - 1)(p - 2)}{6} \cdot r \right. \\
&\qquad\qquad \left. + \left[q^2(q^2 - q + 1) + \frac{q^3(q^2 - q + 1)(q - 1)(p - 2)}{6}\right] \cdot (-1)\right\} \\
&\qquad = -\frac{q^6 - 2q^5 - q^4 + 2q^3 - 3q^2}{3}.
\end{aligned}$$

Case 4: Let $p \mid (q - 1)$. By Lemma 2.4, S_p is a subgroup of the cyclic group C_{q^2-1} fixing two points P, Q on $\mathcal{H}_q(\mathbb{F}_{q^2})$; then a proper p-subgroup H of G satisfies $\mu_{L_p}(\{1\}) \neq 0$ if and only if H has order p; in this case, $\mu_{L_p}(\{1\}, H) = -1$. Also, by Lemma 2.4, any two Sylow p-subgroups of G have trivial intersection. Then the number of subgroups C_p of G is equal to the number $\binom{q^3+1}{1}$ of couples of points in $\mathcal{H}_q(\mathbb{F}_{q^2})$; equivalently, this number is equal to $[G : N_G(C_{q^2})]$, where $|N_G(C_{q^2-1})| = 2(q^2 - 1)$ by Proposition 3.3. Thus, $\chi(\Delta(L_p \setminus \{1\})) = -\frac{q^6+q^3}{2}$.

Case 5: Let $p \mid (q^2 - q + 1)$. Then $S_p \leq C_{q^2-q+1}$, and hence a proper p-subgroup H of G satisfies $\mu_{L_p}(\{1\}, H) \neq 0$ if and only if H has order p; in this case, $\mu_{L_p}(\{1\}, H) = -1$. The number of subgroups C_p of G is equal to the number of subgroups C_{q^2-q+1}, and hence to the number $[G : N_G(M_4(\tilde{T}))] = \frac{q^3(q+1)^2(q-1)}{3}$ of maximal subgroups of type $M_4(\tilde{T})$ in G. Thus, $\chi(\Delta(L_p \setminus \{1\})) = -\frac{q^3(q+1)^2(q-1)}{3} = -\frac{q^6+q^5-q^4-q^3}{3}$. □

References

[1] M. Bianchi, A. Gillio Berta Mauri and L. Verardi, On Hawkes-Isaacs-Özaydin's conjecture, *Istit. Lombardo Accad. Sci. Lett. Rend. A* **124** (1990), 99–117.

[2] K. S. Brown, Euler characteristics of groups: the p-fractional part, *Invent. Math.* **29** (1975), 1–5, doi:10.1007/bf01405170.

[3] V. Colombo and A. Lucchini, On subgroups with non-zero Möbius numbers in the alternating and symmetric groups, *J. Algebra* **324** (2010), 2464–2474, doi:10.1016/j.jalgebra.2010.07.040.

[4] A. Cossidente, G. Korchmáros and F. Torres, Curves of large genus covered by the Hermitian curve, *Comm. Algebra* **28** (2000), 4707–4728, doi:10.1080/00927870008827115.

[5] F. Dalla Volta, M. Montanucci and G. Zini, On the classification problem for the genera of quotients of the Hermitian curve, 2018, `arXiv:1805.09118 [math.AG]`.

[6] E. Damian and A. Lucchini, The probabilistic zeta function of finite simple groups, *J. Algebra* **313** (2007), 957–971, doi:10.1016/j.jalgebra.2007.02.055.

[7] L. E. Dickson, *Linear Groups with an Exposition of the Galois Field Theory*, B. G. Teubner, Leipzig, 1901.

[8] M. Downs, The Möbius function of $\mathrm{PSL}_2(q)$, with application to the maximal normal subgroups of the modular group, *J. London Math. Soc.* **43** (1991), 61–75, doi:10.1112/jlms/s2-43.1.61.

[9] M. Downs and G. A. Jones, Möbius inversion in Suzuki groups and enumeration of regular objects, in: J. Širáň and R. Jajcay (eds.), *Symmetries in Graphs, Maps, and Polytopes*, Springer, Cham, volume 159 of *Springer Proceedings in Mathematics & Statistics*, pp. 97–127, 2016, doi:10.1007/978-3-319-30451-9_5, papers from the 5th SIGMAP Workshop held in West Malvern, July 7–11, 2014.

[10] D. H. Dung and A. Lucchini, Rationality of the probabilistic zeta functions of finitely generated profinite groups, *J. Group Theory* **17** (2014), 317–335, doi:10.1515/jgt-2013-0037.

[11] A. Garcia, H. Stichtenoth and C.-P. Xing, On subfields of the Hermitian function field, *Compositio Math.* **120** (2000), 137–170, doi:10.1023/a:1001736016924.

[12] P. Hall, The Eulerian functions of a group, *Q. J. Math. (Oxford Series)* **7** (1936), 134–151, doi:10.1093/qmath/os-7.1.134.

[13] R. W. Hartley, Determination of the ternary collineation groups whose coefficients lie in the $\mathrm{GF}(2^n)$, *Ann. Math.* **27** (1925), 140–158, doi:10.2307/1967970.

[14] T. Hawkes, I. M. Isaacs and M. Özaydin, On the Möbius function of a finite group, *Rocky Mountain J. Math.* **19** (1989), 1003–1034, doi:10.1216/rmj-1989-19-4-1003.

[15] J. W. P. Hirschfeld, G. Korchmáros and F. Torres, *Algebraic Curves over a Finite Field*, Princeton Series in Applied Mathematics, Princeton University Press, Princeton, NJ, 2008.

[16] D. R. Hughes and F. C. Piper, *Projective Planes*, volume 6 of *Graduate Texts in Mathematics*, Springer-Verlag, Berlin, 1973.

[17] B. Huppert, *Endliche Gruppen I*, volume 134 of *Die Grundlehren der Mathematischen Wissenschaften*, Springer-Verlag, Berlin, 1967, doi:10.1007/978-3-642-64981-3.

[18] A. Lucchini, On the subgroups with non-trivial Möbius number, *J. Group Theory* **13** (2010), 589–600, doi:10.1515/jgt.2010.009.

[19] A. Mann, Positively finitely generated groups, *Forum Math.* **8** (1996), 429–459, doi:10.1515/form.1996.8.429.

[20] A. Mann, A probabilistic zeta function for arithmetic groups, *Internat. J. Algebra Comput.* **15** (2005), 1053–1059, doi:10.1142/s0218196705002633.

[21] M. Montanucci and G. Zini, Some Ree and Suzuki curves are not Galois covered by the Hermitian curve, *Finite Fields Appl.* **48** (2017), 175–195, doi:10.1016/j.ffa.2017.07.007.

[22] M. Montanucci and G. Zini, Quotients of the Hermitian curve from subgroups of $\mathrm{PGU}(3,q)$ without fixed points or triangles, 2018, `arXiv:1804.03398 [math.AG]`.

[23] H. Pahlings, On the Möbius function of a finite group, *Arch. Math. (Basel)* **60** (1993), 7–14, doi:10.1007/bf01194232.

[24] E. Pierro, The Möbius function of the small Ree groups, *Australas. J. Combin.* **66** (2016), 142–176, `https://ajc.maths.uq.edu.au/pdf/66/ajc_v66_p142.pdf`.

[25] D. Quillen, Homotopy properties of the poset of nontrivial p-subgroups of a group, *Adv. Math.* **28** (1978), 101–128, doi:10.1016/0001-8708(78)90058-0.

[26] J. Shareshian, On the Möbius number of the subgroup lattice of the symmetric group, *J. Combin. Theory Ser. A* **78** (1997), 236–267, doi:10.1006/jcta.1997.2762.

[27] J. W. Shareshian, *Combinatorial Properties of Subgroup Lattices of Finite Groups*, Ph.D. thesis, Rutgers, The State University of New Jersey, New Brunswick, 1996, `https://search.proquest.com/docview/304280477`.

[28] R. P. Stanley, *Enumerative Combinatorics, Volume 1*, volume 49 of *Cambridge Studies in Advanced Mathematics*, Cambridge University Press, Cambridge, 2nd edition, 2012.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 403–410
https://doi.org/10.26493/1855-3974.1749.84e
(Also available at http://amc-journal.eu)

Regular self-dual and self-Petrie-dual maps of arbitrary valency*

Jay Fraser , Olivia Jeans

The Open University, Milton Keynes, U.K.

Jozef Širáň

The Open University, Milton Keynes, U.K. and
Slovak University of Technology, Bratislava, Slovakia

Received 13 July 2018, accepted 18 December 2018, published online 30 January 2019

Abstract

The existence of a regular, self-dual and self-Petrie-dual map of any given even valency has been proved by D. Archdeacon, M. Conder and J. Širáň (2014). In this paper we extend this result to any odd valency ≥ 5. This is done using algebraic number theory and maps defined on the groups $\mathrm{PSL}(2,p)$ in the case of odd prime valency ≥ 5 and valency 9, and using coverings for the remaining odd valencies.

Keywords: Regular map, automorphism group, self-dual map, self-Petrie-dual map.

Math. Subj. Class.: 05C25, 05C10

1 Introduction

In this paper we consider regular maps (that is, cellular embeddings of graphs on closed surfaces) with the highest ‘level of symmetry’, which are, in addition, invariant under the operators of duality and Petrie duality. Regular maps have been addressed in a number of papers and we refer here to the latest survey [11] for a large number of details; here we just sum up the essentials needed for our purposes.

*The third author thanks the Banff International Research Station (BIRS) for supporting his participation in the stimulating workshop ‘Symmetries of Surfaces, Maps and Dessins’ (September 2017) at which the statement of the problem addressed in this paper crystallised and attracted the attention of a number of colleagues, notably M. Conder who produced important initial examples.

The same author also acknowledges support from the APVV Research Grants 15-0220 and 17-0428, and from the VEGA Research Grants 1/0026/16 and 1/0142/17.

E-mail addresses: jay.fraser@open.ac.uk (Jay Fraser), olivia.jeans@open.ac.uk (Olivia Jeans), , jozef.siran@open.ac.uk (Jozef Širáň)

From an algebraic point of view, a regular map M can be identified with a finite group G with three distinguished involutory generators x, y, z and relators $(yz)^k$, $(zx)^\ell$ and $(xy)^2$ so that x and y commute; we will formally write $M = (G; x, y, z)$ to encapsulate the situation. The pair (k, ℓ) is the *type* of M, and we will assume throughout that $k, \ell \geq 3$; the type is *hyperbolic* if $1/k + 1/\ell < 1/2$. Geometrically and topologically, elements of G may be identified with flags (which correspond to mutually incident vertex-edge-face triples) and left cosets of the subgroups $\langle x, y\rangle$, $\langle y, z\rangle$ and $\langle z, x\rangle$ represent edges, vertices and faces of the embedded graph, with incidence given by non-empty intersection of cosets. Moreover, left multiplication by elements of G on the cosets induce map automorphisms of M and, in fact, G is isomorphic to the (full) automorphism group $\mathrm{Aut}(M)$ of M. Conjugates of x, y and z, respectively, induce automorphisms that locally act on M as reflections along some edge, in some edge, and in an axis of some corner of M. Similarly, conjugates of $r = yz$ and $s = zx$ represent rotations about vertices and face centres of the map; in particular, every vertex has valency k and every face is bounded by a closed walk of length ℓ. The map M is *orientable* (meaning that its underlying surface is orientable) if and only if $G^+ = \langle r, s\rangle$ is a subgroup of G of index two, and *non-orientable* otherwise. Thus, in the non-orientable case, the entire group G can be generated by the two rotations r and s only, and the involutions x, y, z are then expressible in terms of r and s; in such a situation we also write $M = (G; r, s)$.

Every automorphism of a map, regarded as a permutation of flags that preserves incidence along and across edges and within corners, is completely determined by its action on a single flag. If the automorphism group is transitive (and hence regular) on flags, one may identify the group with the flag set and arrive at the description outlined above. But even then a map may still exhibit ‘external symmetries’ induced by invariance under the operators of duality and Petrie-duality. The two operators are well known; informally, duality interchanges the roles of vertices and faces, and the Petrie dual of a map is formed by re-embedding its underlying graph so that the new faces are the left-right (‘zig-zag’) closed walks in the original map. A map is *self-dual* or *self-Petrie-dual* if it is isomorphic to its dual or Petrie dual, respectively. In the case of a *regular* map $M = (G; x, y, z)$ as above, it is also well known (cf. [11]) that M is self-dual if and only if the group G admits an automorphism interchanging x with y and fixing z, and M is self-Petrie-dual if G has an automorphism interchanging x with xy and fixing y and z. In [1], regular maps that are both self-dual and self-Petrie-dual have been said to have *trinity symmetry*.

The natural question regarding the existence of regular maps with trinity symmetry for any valency was raised more than four decades ago. In [15] it was suggested that the map $M = (G; x, y, z)$ for the group $G = \langle x, y, z; x^2, y^2, z^2, (xy)^2, (yz)^{2n}, (zx)^{2n}, (xyz)^{2n}, (xzyzxyz)^2\rangle$ is a regular map with trinity symmetry, of valency $2n$ for every $n \geq 1$. This was eventually proved in [1] in a much more general form, including also invariance under the so-called hole operators that represent additional levels of ‘external symmetries’ not discussed here. However, the question remained almost completely open for odd-valent regular maps with trinity symmetry, as pointed out by the third author at the 2017 BIRS Workshop ‘Symmetries of Surfaces, Maps and Dessins’ [4, Part 4.7]. Note that such a map must necessarily be non-orientable because of self-Petrie-duality with Petrie walks of odd length. There is no such map of valency 3 since the only regular map of type $(3, 3)$ is the 2-skeleton of a tetrahedron. At the time of publication of the report [4] the only two sets of known examples of regular maps with trinity symmetry of odd valency $k \geq 5$ were those discovered computationally by M. Conder for $5 \leq k \leq 19$ and the ones resulting from the work of G. Jones [7]. The method of Jones actually has potential to produce examples for infinitely many odd values of k but in [7] explicit examples have been given only for $k = 15$ (found also in [1] by a different method) and $k = 455$.

Here we completely settle the problem by showing that for every odd $k \geq 5$ there exists a regular, self-dual and self-Petrie-dual map of valency k. Our strategy is to establish this result first for every *prime* $k \geq 5$ and for $k = 9$ by algebraic methods motivated by those used in [8], and applied to more detailed results of [6] on regular maps defined on linear fractional groups. We then extend this to non-prime odd values of $k \geq 5$ by an analogue of a covering tool from [1]. The paper is organised accordingly: in Sections 2 and 3 we review results on regular self-dual and self-Petrie-dual maps on linear fractional groups and develop the algebraic methods needed for our purposes, and in Section 4 we prove our general result and make a few concluding remarks.

2 Regular maps on linear fractional groups

Classification of all orientably-regular maps with orientation-preserving automorphism group isomorphic to $\mathrm{PSL}(2,q)$ or $\mathrm{PGL}(2,q)$ follows from [9] and can be found in a somewhat more explicit form in [10]; the latter was re-interpreted and extended to regular maps (orientable or not) in [5]. Since we will be interested only in the special case of odd valency and face length, we just reproduce the corresponding part of the classification result here (the cases when one of the entries in the type of the map is even are more involved and we refer to [8] for details).

Proposition 2.1. *Let $(k,\ell) \neq (5,5)$ be a hyperbolic pair with both entries odd and let p be an odd prime dividing neither k nor ℓ. Let $e = e(k,\ell)$ be the smallest positive integer j such that $2n \mid (p^j - \varepsilon_n)$ for each $n \in \{k,\ell\}$ and some $\varepsilon_n \in \{+1,-1\}$, and let ξ_n be a primitive $2n$-th root of unity in $\mathrm{GF}(p^e)$ if $\varepsilon_n = 1$ or in $\mathrm{GF}(p^{2e})$ if $\varepsilon_n = -1$. Further, let $D = \xi_k^2 + \xi_k^{-2} + \xi_\ell^2 + \xi_\ell^{-2} \neq 0$ and let*

$$R = \pm \begin{bmatrix} \xi_k & 0 \\ 0 & \xi_k^{-1} \end{bmatrix} \quad \text{and} \quad S = \pm(\xi_k - \xi_k^{-1})^{-1} \begin{bmatrix} -(\xi_\ell + \xi_\ell^{-1})\xi_k^{-1} & -D \\ 1 & (\xi_\ell + \xi_\ell^{-1})\xi_k \end{bmatrix}$$

be elements of $\mathrm{PSL}(2,p^e)$ if $\varepsilon_k = 1$ and of $\mathrm{PSL}(2,p^{2e})$ otherwise. Then,

(a) *the group $G_{k,\ell} = \langle R,S\rangle$ is isomorphic to $\mathrm{PSL}(2,p^e)$, with R of order k and S of order ℓ;*

(c) *$M = (G_{k,\ell}; R,S)$ is a regular map of type (k,ℓ), which is non-orientable if and only if $-D$ is a square in $\mathrm{GF}(p^e)$.*

We note that if $p^e \equiv \pm 1 \pmod{10}$, the group $\mathrm{PSL}(2,p^e)$ contains (up to conjugacy) two exceptional pairs R,S as above for $(k,\ell) = (5,5)$ with the property that $\langle R,S\rangle \cong A_5$; this case (omitted from [8, Theorem 2.2]) is addressed in [6]. However, this situation does not apply in what follows.

Necessary and sufficient conditions for self-duality and self-Petrie-duality of the maps $M = (G_{k,\ell}; R,S)$ from Proposition 2.1 were established in [6]. As they are also quite complex we present here only a simple sufficient condition appearing as Corollary 4.3 in [6] which (in terms and notation of Proposition 2.1) can be re-stated as follows.

Proposition 2.2. *Let $k \geq 5$ be odd, and let $p \geq 5$ be a prime not dividing k. Further, let $\ell = k$ and let $\xi = \xi_k = \xi_\ell$ be a primitive $2k$-th root of unity in $\mathrm{GF}(p^e)$ or in $\mathrm{GF}(p^{2e})$*

for $e = e(k,\ell)$ *such that* $3(\xi^2 + \xi^{-2}) + 2 = 0$*. Then,* $M = (G_{k,\ell}; R, S)$ *is a (non-orientable) self-dual and self-Petrie-dual regular map of valency* k*, with automorphism group isomorphic to* $\mathrm{PSL}(2, p^e)$*.*

The condition $3(\xi^2+\xi^{-2})+2 = 0$ is equivalent to $3(\xi+\xi^{-1})^2 = 4$ and for its fulfilment it is necessary that 3 be a square in $\mathrm{GF}(p^e)$, $p \geq 5$. For $e = 1$, this holds if and only if $p \equiv \pm 1 \pmod{12}$, and it is always the case if $e \geq 2$. But we can say more. Namely, the element $\zeta = \xi^2$ in Proposition 2.2 is a primitive k-th root of unity in $F = \mathrm{GF}(p^e)$ or $F = \mathrm{GF}(p^{2e})$, and the condition $3(\zeta + \zeta^{-1}) + 2 = 0$ represents a quadratic equation in the prime field F_p of F; it also says that $\zeta + \zeta^{-1} \in F_p$. The last fact is equivalent to $(\zeta + \zeta^{-1})^p = \zeta + \zeta^{-1}$, which reduces to $(\zeta^{p-1} - 1)(\zeta^{p+1} - 1) = 0$ in F. It follows that either $\zeta \in F_p$ and $p \equiv 1 \pmod{2k}$, or ζ lies in a quadratic extension of F_p and $p \equiv -1 \pmod{2k}$, and in both cases we have $e = 1$ (recall that k is assumed to be odd). The bulk of Proposition 2.2 may now be restated in a form more suitable for our future use.

Corollary 2.3. *Let* $k \geq 5$ *be odd. Assume that there exists a prime* $p \geq 5$ *such that* $p \equiv \pm 1 \pmod{2k}$ *and* $p \equiv \pm 1 \pmod{12}$*, and a primitive* k*-th root of unity* ζ *in a finite field of order* p *or* p^2 *with the property that* $3(\zeta + \zeta^{-1}) + 2 = 0$*. Then, there exists a non-orientable self-dual and self-Petrie-dual regular map of valency* k *with automorphism group* $\mathrm{PSL}(2, p)$*.*

3 Algebraic preliminaries

For any $k \geq 3$, let α be a primitive complex k-th root of unity; its minimal polynomial is the k-th cyclotomic polynomial. Let $h = \alpha + \alpha^{-1}$ and let $K = \mathbb{Q}(h)$ be the field obtained by adjoining h to the rationals. It is known [13, Proposition 2.16] that the ring $\mathcal{O}$ of algebraic integers of K is $\mathbb{Z}(h)$. We will focus on the algebraic integer $g = 3h+2 \in \mathcal{O}$. Observe that $g \neq 0$, for otherwise α would be a root of a quadratic polynomial over $\mathbb{Z}$, contrary to $k \geq 3$.

Recall that the norm $N(y)$ of an element $y \in \mathcal{O}$ is defined as the product $\prod_t \sigma_t(y)$, where σ_t denotes the injective homomorphism $\mathcal{O} \to \mathbb{C}$ into the field of complex numbers, uniquely determined by $\sigma_t(\alpha) = \alpha^t$, and t ranges over all integers between 1 and $(k-1)/2$ that are relatively prime to k. It is well known that $N(y)$ is an integer for any $y \in \mathcal{O}$, which is a consequence of the invariance of $N(y)$ under the endomorphisms σ_t.

For the norm of our element $g \in \mathcal{O}$ we thus have $N(g) = \prod_t (3\sigma_t(h) + 2)$, the product being taken over all t between 1 and $(k-1)/2$, coprime to k. The $\varphi(k)/2$ images $\sigma_t(h)$ appearing in this product are precisely the roots of the minimal polynomial $\Psi(x)$ of degree $\varphi(k)/2$ for $h = \alpha + \alpha^{-1}$, see e.g. [8]. So, if $\Psi(x) = \prod_t (x - \sigma_t(h)) = \sum_j a_j x^j$ where j ranges from 0 to $\varphi(k)/2$, then the integral coefficients a_j will also appear in the expansion of the above product. More precisely, letting $r = \varphi(k)/2$ and $u = -2/3$, one has

$$\begin{aligned} N(g) &= \prod_t (3\sigma_t(h) + 2) = (-3)^r \prod_t (u - \sigma_t(h)) \\ &= (-3)^r \sum_{j=0}^{r} a_j u^j = \sum_{j=0}^{r} (-3)^{r-j} 2^j a_j. \end{aligned} \tag{3.1}$$

Let us consider what happens when we look at (3.1) modulo 9. Up to the last two terms all the remaining ones are a multiple of 9 and so, noting that $a_r = 1$, we have

$$N(g) \equiv 2^r - 3 \cdot 2^{r-1} a_{r-1} \pmod{9}. \tag{3.2}$$

We will show that if $k \geq 5$ and k is a *prime*, then the norm $N(g)$ is not equal to ± 1, which means that g is then *not* a unit of the ring $\mathcal{O}$. Indeed, let $k \geq 5$ be a prime, so that $r = \varphi(k)/2 = (k-1)/2$. By [12] we then also have $a_{r-1} = 1$, and the congruence (3.2) becomes

$$N(g) \equiv 2^{(k-1)/2} - 3 \cdot 2^{(k-3)/2} \equiv -2^{(k-3)/2} \pmod 9.$$

It is easy to check that $2^j \equiv \pm 1 \pmod 9$ for a positive integer j if and only if j is a multiple of 3. This means that if k is prime, the norm $N(g)$ can be congruent to $\pm 1 \pmod 9$ only if $(k-3)/2$ is a multiple of 3, giving a contradiction if $k \geq 5$. Further, from (3.1) with the help of $a_r = 1$ and $a_0 = \pm 1$ [12] it follows that if $k \geq 5$ is a prime, then $N(g)$ is divisible neither by 2 nor by 3. We thus have:

Lemma 3.1. *If $k \geq 5$ is a prime, then $N(g) \neq \pm 1$; in particular, the non-zero element $g \in \mathcal{O}$ is not a unit of the ring $\mathcal{O}$. Moreover, for every prime factor p of $N(g)$ one has $p \geq 5$.* □

Consider now the field $K' = \mathbb{Q}(\alpha)$, an extension of K of degree two. Let $\mathcal{O}'$ be the ring of algebraic integers of K'; it is well known [13, Theorem 2.6] that $\mathcal{O}' = \mathbb{Z}(\alpha)$, and, of course, $[\mathcal{O}' : \mathcal{O}] = 2$. The (integral) norm $N'(z)$ of any $z \in \mathcal{O}'$ is now the product $\prod_t \sigma_t(z)$ taken over all injective homomorphism $\sigma_t \colon \mathcal{O}' \to \mathbb{C}$ given by $\sigma_t(\alpha) = \alpha^t$ for t between 1 and $k-1$ coprime to k, and again one has $N'(z) \in \mathbb{Z}$. The two norms, N on $\mathcal{O}$ and N' on $\mathcal{O}'$, are related by $N'(y) = (N(y))^2$ for each $y \in \mathcal{O}$.

We will keep assuming that $k \geq 5$ is an odd prime, and we let $p \geq 5$ be an arbitrary prime divisor of $N(g)$, which exists by Lemma 3.1. We continue by considering the ideal $\langle g, p \rangle$ of $\mathcal{O}' = \mathbb{Z}(\alpha)$ generated by the elements g and p.

Lemma 3.2. *If $k \geq 5$ is a prime and if $p \geq 5$ is a prime divisor of $N(g)$, the ideal $\langle g, p \rangle$ is proper in the ring $\mathcal{O}'$.*

Proof. Suppose that $\langle g, p \rangle = \mathcal{O}'$, which means that $1 = Ag + Bp$ for some $A, B \in \mathcal{O}'$. Clearly $A \neq 0$, for otherwise $1 = N'(B)N'(p) = N'(B)p^{k-1}$ and so $N'(B)$ would not be an integer. Now, $1 = N'(1) = N'(Ag + Bp) = \prod_\sigma \sigma(Ag + Bp)$, where the product is being taken over all the $\varphi(k) = k-1$ embeddings $\sigma \colon \mathcal{O}' \to \mathbb{C}$. Expansion of this product gives $N'(Ag + Bp) = N'(A)N'(g) + cp$ for some $c \in \mathcal{O}'$. Thus, $cp \in \mathbb{Z}$ and so either $c \in \mathbb{Z}$ or $c = \pm 1/p$. As p is a divisor of $N'(g) = (N(g))^2$ and $N'(A)$ is a non-zero integer, in either case it follows that $N'(Ag + Bp) \neq 1$, a contradiction. □

By Lemma 3.2, the ideal $\langle g, p \rangle$ is contained in some maximal ideal $J = J_p$ of the ring $\mathcal{O}'$. Since $\mathcal{O}'$ is a Dedekind domain, the ideal J has finite index in $\mathcal{O}'$ and so $\mathcal{O}'/J$ is a finite field F of characteristic p, that is, $F \cong \mathrm{GF}(p^m)$ for some $m \geq 1$. Recalling our assumption of primality of k we show that the (multiplicative) order of the element $\overline{\alpha} = \alpha + J$ in the field $F = \mathcal{O}'/J$ is equal to k. Indeed, suppose this is not the case. Then, because of primality of k, the order of $\overline{\alpha}$ in F would have to be one, meaning that $\overline{\alpha} = 1$ in F. But then, since the element $\overline{g} = g + J$ is equal to zero in F, we would have $0 = \overline{g} = 3(\overline{\alpha} + \overline{\alpha}^{\,-1}) + 2 = 8$ in F, a contradiction as p is odd. Observe also that $k \neq p$ since no element in F has multiplicative order p.

This way we have constructed a finite field F of characteristic p containing a primitive k-th root $\overline{\alpha}$ of unity such that $3(\overline{\alpha} + \overline{\alpha}^{\,-1}) + 2 = 0$. We now invoke the analysis immediately preceding Corollary 2.3 in Section 2, which fully applies to our situation. As the result we conclude that F is the prime field F_p if and only if $\overline{\alpha} \in F_p$ for $p \equiv 1 \pmod{2k}$; otherwise

F is a quadratic extension of F_p for $p \equiv -1 \pmod{2k}$. In both cases we have $p \equiv \pm 1 \pmod{12}$ because 3 has to be a square in F_p. Summing up, we have proved:

Proposition 3.3. *Let $k \geq 5$ be an odd prime and let α be a primitive complex k-th root of unity. Further, let $g = 3(\alpha + \alpha^{-1}) + 2$ and let $N(g)$ be the norm of g in the ring $\mathbb{Z}(\alpha)$. Then, $N(g) \notin \{0, \pm 1\}$, every prime divisor p of $N(g)$ satisfies $p \geq 5$, and $p \equiv \pm 1 \pmod{2k}$ and $p \equiv \pm 1 \pmod{12}$, and for every such p there is a finite field F of order p or p^2 containing a primitive k-th root $\overline{\alpha}$ of 1 such that $\overline{g} = 3(\overline{\alpha} + \overline{\alpha}^{\,-1}) + 2 = 0$ in F.* □

4 The main result

To obtain a restricted version of our main result for prime valencies at least five we just need to put the pieces together. Indeed, taking $\zeta = \overline{\alpha}$ in Proposition 3.3 and combining it with Proposition 2.2 and Corollary 2.3 immediately gives:

Theorem 4.1. *For every odd prime $k \geq 5$ there exists a prime $p \equiv \pm 1 \pmod{2k}$ and $p \equiv \pm 1 \pmod{12}$ such that $\mathrm{PSL}(2, p)$ is the automorphism group of a (non-orientable) regular, self-dual and self-Petrie-dual map of valency k.* □

We know that there is no 3-valent regular map with trinity symmetry, but there is one of valency 3^2 that can be constructed by the machinery of Section 2 as follows. The element 2 is a primitive 9-th root of unity mod 73, and so is $\zeta = 2^4$ and its multiplicative inverse $\zeta^{-1} = 2^5$, with ζ and ζ^{-1} satisfying the condition $3(\zeta + \zeta^{-1}) + 2 = 0 \pmod{73}$. By Proposition 2.2 the group $\mathrm{PSL}(2, 73)$ carries a self-dual and self-Petrie-dual regular map of valency 9.

Based on Theorem 4.1 and the above remark we are now in position to prove a full version of our main result. As alluded to in the Introduction (Section 1), this will be done with the help of coverings, and more specifically using a non-orientable analogue of Theorem 2.1 of [1]. We state it here in a restricted version sufficient for our purpose.

Theorem 4.2. *If there is a non-orientable regular map of odd valency $d \geq 5$ with trinity symmetry and with automorphism group G, then for any odd integer $n \geq 3$ there is a non-orientable regular map of degree nd with trinity symmetry and automorphism group isomorphic to $(\mathbb{Z}_n)^{1+|G|/4} \rtimes G$.*

Sketch of a proof. As indicated, this result was proved in [1, Theorem 2.1] for orientable maps (and, in this category, in a much more general setting that included also external symmetries induced by hole operators). The parts of the proof in [1] that refer to regularity, self-duality and self-Petrie-duality apply almost word-by-word to the non-orientable case and we thus give only a sketch of the arguments here. We will assume familiarity with the theory of lifts of maps by corner voltage assignments as explained e.g. in [1, 2, 3]; a *corner* of a regular map $M = (G; x, y, z)$ is any 2-subset of the form $\{g, gz\}$ for $g \in G$.

Now let $M = (G; x, y, z)$ be a regular map as in the statement. For odd $n \geq 3$ let $H = \mathbb{Z}_n^{|G|/2}$ be the space of all $|G|/2$-tuples with entries from $\mathbb{Z}_n$ and let $\mathcal{E}$ be the set of unit vectors (those with exactly one non-zero coordinate, equal to 1) in H. Define a corner voltage assignment σ on flags of M – that is, on the elements of G – in the group H by assigning the $|G|/2$ two-element subsets $\{\varepsilon, -\varepsilon\}$ for $\varepsilon \in \mathcal{E}$ to the $|G|/2$ corners $\{g, gz\}$ for $g \in G$ in an arbitrary one-to-one fashion. By arguments in the proof of Theorem 2.1 in [1] that do not depend on orientability, the lift of the map M of type

(d, d) by the voltage assignment σ has $n^{-1+|G|/4}$ components, each isomorphic to a regular map $M^\sigma = (G^\sigma; x^\sigma, y^\sigma, z^\sigma)$ of type (nd, nd) for the group $G^\sigma = (\mathbb{Z}_n)^{1+|G|/4} \rtimes G$ and suitable involutory generators $x^\sigma, y^\sigma, z^\sigma$ of G^σ. Moreover, by the reasoning in the same proof (again applying also to non-orientable maps), trinity symmetry of M implies trinity symmetry of M^σ. Note that both M and M^σ are non-orientable as their Petrie walks (of length d and nd) have odd length. □

Collecting our findings we arrive at the main result of this paper as a consequence of Theorem 4.1 and the remark following it, both in combination with Theorem 4.2.

Theorem 4.3. *For every odd $d \geq 5$ there exists a regular, self-dual and self-Petrie-dual map of valency d.* □

A few remarks are in order. The reader may have observed that if the conclusion of Proposition 3.3 in Section 3 was valid for all odd $k \geq 5$ (and not just for *prime* $k \geq 5$), we would have a proof of our main result that would be independent on coverings and the resulting regular maps with trinity symmetry would have automorphism group isomorphic to $\mathrm{PSL}(2, p)$ for suitable primes depending on k. Research in this direction is currently being undertaken by the first two authors of this paper. Here we include a table of the first few values of $N(g)$ for odd k between 5 and 29, with $\Phi(n)$ standing for the prime factorisation of n; observe that all the primes p in the prime factorization of $|N(g)|$ satisfy $p \equiv \pm 1 \pmod{2k}$ and $p \equiv \pm 1 \pmod{12}$:

k	$N(g)$	$\Phi(\|N(g)\|)$
5	-11	prime
7	-13	prime
9	-73	prime
11	$+263$	prime
13	-131	prime
15	-239	prime
17	-4079	prime
19	$+15503$	37×419
21	$+5209$	prime
23	-4093	prime
25	$+56149$	prime
27	-16417	prime
29	$+3161869$	59×53591

As noted earlier, existence of the regular maps for the first eight entries in this table was discovered by M. Conder, who also found such maps of valency 7 and 17 for the Janko simple groups J_2 and J_3.

We conclude by noting that a strategy for proving Theorem 4.3 was also outlined by S. Wilson [14] by reducing the problem to a number-theoretic question related to Chebyshev polynomials over finite fields.

References

[1] D. Archdeacon, M. Conder and J. Širáň, Trinity symmetry and kaleidoscopic regular maps, *Trans. Amer. Math. Soc.* **366** (2014), 4491–4512, doi:10.1090/s0002-9947-2013-06079-5.

[2] D. Archdeacon, P. Gvozdjak and J. Širáň, Constructing and forbidding automorphisms in lifted maps, *Math. Slovaca* **47** (1997), 113–129.

[3] D. Archdeacon, R. B. Richter, J. Širáň and M. Škoviera, Branched coverings of maps and lifts of map homomorphisms, *Australas. J. Combin.* **9** (1994), 109–121, `http://ajc.maths.uq.edu.au/pdf/9/ocr-ajc-v9-p109.pdf`.

[4] M. Conder, N. Boston, G. González-Diez, G. Jones and T. Tucker, Report on the BIRS Workshop 'Symmetries of Surfaces, Maps and Dessins', September 2017, `http://www.birs.ca/workshops/2017/17w5162/report17w5162.pdf`.

[5] M. Conder, P. Potočnik and J. Širáň, Regular hypermaps over projective linear groups, *J. Aust. Math. Soc.* **85** (2008), 155–175, doi:10.1017/s1446788708000827.

[6] G. Erskine, K. Hriňáková and O. Jeans, Self-dual, self-Petrie-dual and Moebius regular maps on linear fractional groups, 2018, `arXiv:1807.11307 [math.CO]`.

[7] G. A. Jones, Combinatorial categories and permutation groups, *Ars Math. Contemp.* **10** (2016), 237–254, doi:10.26493/1855-3974.545.fd5.

[8] G. A. Jones, M. Mačaj and J. Širáň, Nonorientable regular maps over linear fractional groups, *Ars Math. Contemp.* **6** (2013), 25–35, doi:10.26493/1855-3974.251.044.

[9] A. M. Macbeath, Generators of the linear fractional groups, in: W. J. LeVeque and E. G. Straus (eds.), *Number Theory*, American Mathematical Society, Providence, Rhode Island, volume 12 of *Proceedings of Symposia in Pure Mathematics*, 1969 pp. 14–32, proceedings of the Seventy-Third Annual Meeting of the American Mathematical Society held at Houston, Texas, January 24 – 28, 1967.

[10] C.-H. Sah, Groups related to compact Riemann surfaces, *Acta Math.* **123** (1969), 13–42, doi:10.1007/bf02392383.

[11] J. Širáň, How symmetric can maps on surfaces be?, in: S. R. Blackburn, S. Gerke and M. Wildon (eds.), *Surveys in Combinatorics 2013*, Cambridge University Press, Cambridge, volume 409 of *London Mathematical Society Lecture Note Series*, 2013 pp. 161–238, doi:10.1017/cbo9781139506748.006, papers from the 24th British Combinatorial Conference held in Egham, July 2013.

[12] D. Surowski and P. McCombs, Homogeneous polynomials and the minimal polynomial of $\cos(2\pi/n)$, *Missouri J. Math. Sci.* **15** (2003), 4–14, `http://cs.ucmo.edu/~mjms/2003.1/Surow.pdf`.

[13] L. C. Washington, *Introduction to Cyclotomic Fields*, volume 83 of *Graduate Texts in Mathematics*, Springer-Verlag, New York, 1982, doi:10.1007/978-1-4684-0133-2.

[14] S. E. Wilson, personal communication, September 2017.

[15] S. E. Wilson, *New Techniques for the Construction of Regular Maps*, Ph.D. thesis, University of Washington, Seattle, Washington, 1976, `https://search.proquest.com/docview/302844349`.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 411–417
https://doi.org/10.26493/1855-3974.1704.34c
(Also available at http://amc-journal.eu)

A generalization of the parallelogram law to higher dimensions

Alessandro Fonda [*]

Dipartimento di Matematica e Geoscienze, Università di Trieste,
P.le Europa 1, I-34127 Trieste, Italy

Received 24 May 2018, accepted 9 July 2018, published online 8 February 2019

Abstract

We propose a generalization of the parallelogram identity in any dimension $N \geq 2$, establishing the ratio of the quadratic mean of the diagonals to the quadratic mean of the faces of a parallelotope. The proof makes use of simple properties of the exterior product of vectors.

Keywords: Parallelogram law, parallelotope.

Math. Subj. Class.: 51M04

1 Introduction and statement of the result

The well known parallelogram law states:

> *For any parallelogram, the sum of the squares of the lengths of its two diagonals is equal to the sum of the squares of the lengths of its four sides.*

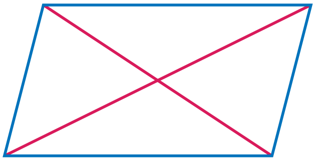


Figure 1: The two diagonals of a parallelogram.

[*] I would probably never have written this paper without the support of my son Marcello. I warmly thank him, in particular, for helping me finding the formula for the three-dimensional case.

E-mail address: a.fonda@units.it (Alessandro Fonda)

Equivalently: given two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, one has

$$\|\boldsymbol{a}+\boldsymbol{b}\|^2+\|\boldsymbol{a}-\boldsymbol{b}\|^2=2(\|\boldsymbol{a}\|^2+\|\boldsymbol{b}\|^2)\,.$$

This identity holds in any inner product space, but, since the two vectors belong to the same plane, we can see it as being of a two-dimensional nature. The aim of this paper is to provide a generalization to higher dimensions.

The parallelogram law has a natural geometric interpretation, involving the areas of the squares constructed on the sides and on the diagonals of the parallelogram. In particular, when $\|\boldsymbol{a}+\boldsymbol{b}\|=\|\boldsymbol{a}-\boldsymbol{b}\|$, it reduces to the Pythagorean theorem. In this paper, however, we will look at the parallelogram law from a rather unusual point of view: writing it as

$$\frac{\|\boldsymbol{a}+\boldsymbol{b}\|^2+\|\boldsymbol{a}-\boldsymbol{b}\|^2}{2}=2\,\frac{\|\boldsymbol{a}\|^2+\|\boldsymbol{b}\|^2+\|\boldsymbol{a}\|^2+\|\boldsymbol{b}\|^2}{4}\,,$$

and taking the square roots, we can state it in the following equivalent form.

> *For any parallelogram, the ratio of the quadratic mean of the lengths of its diagonals to the quadratic mean of the lengths of its sides is equal to $\sqrt{2}$.*

Now, instead of a parallelogram, we will consider an N-dimensional parallelotope, and our goal will be to prove that the same type of proposition holds in this general case. Indeed, our result can be stated as follows.

Theorem 1.1. *For any N-dimensional parallelotope, the ratio of the quadratic mean of the $(N-1)$-dimensional measures of its diagonals to the quadratic mean of the $(N-1)$-dimensional measures of its faces is equal to $\sqrt{2}$.*

For $N=2$, the 1-dimensional measure is the length, and we recover the parallelogram law. In the general case, we first need to specify what a *diagonal* should be, and indeed this will be clarified in the following sections. For example, if $N=3$, the diagonals of a parallelepiped are precisely the parallelograms obtained joining the opposite edges of the parallelepiped (see Figure 2 below), so that the 2-dimensional measures of the diagonals are the areas of these parallelograms.

Notice that our definition of a diagonal is not the same as the one given in [1, 2], where a different generalization of the parallelogram law has been proposed; in the three-dimensional case, e.g., their diagonals are triangles. We believe that our definition is somewhat more natural, since here the diagonals share the same geometrical shape of the faces.

We provide the proof of our main theorem in Section 3. However, for the reader's convenience, we thought it useful to first explain its proof in detail in the more familiar three-dimensional case. This is what we are going to do next.

2 The three-dimensional case

To start with, let us consider a three-dimensional parallelepiped $\mathcal{P}$, and see how to extend the parallelogram law to this case. Instead of the lengths of the four sides of the parallelogram, we would like to take the areas of the *six faces* of the parallelepiped. On the other hand, the lengths of the two diagonals of the parallelogram should naturally be replaced by the areas of the *six diagonals* of the parallelepiped, i.e., the six parallelograms obtained joining the opposite edges of the parallelepiped. In this case, Theorem 1.1 can be rephrased as follows.

For any three-dimensional parallelepiped, the sum of the squares of the areas of its six diagonals is equal to twice the sum of the squares of the areas of its six faces.

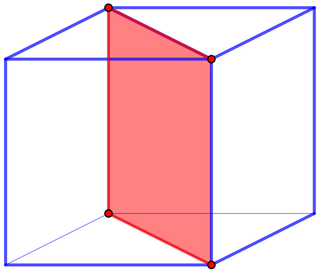


Figure 2: One of the six diagonals of a parallelepiped.

In order to prove this statement, assume the parallelepiped to be *generated* by the following three vectors:

$$\boldsymbol{a} = (a_1, a_2, a_3), \quad \boldsymbol{b} = (b_1, b_2, b_3), \quad \boldsymbol{c} = (c_1, c_2, c_3).$$

By this we mean that $\mathcal{P}$ is the set of points obtained as linear combinations of these three vectors, with coefficients in the interval $[0, 1]$:

$$\mathcal{P} = \{\alpha\boldsymbol{a} + \beta\boldsymbol{b} + \gamma\boldsymbol{c} : \alpha, \beta, \gamma \in [0, 1]\}.$$

The six faces of $\mathcal{P}$ are defined as

$$\begin{aligned}
\mathcal{F}_1^- &= \{\alpha\boldsymbol{a} + \beta\boldsymbol{b} + \gamma\boldsymbol{c} \in \mathcal{P} : \alpha = 0\},\\
\mathcal{F}_1^+ &= \{\alpha\boldsymbol{a} + \beta\boldsymbol{b} + \gamma\boldsymbol{c} \in \mathcal{P} : \alpha = 1\},\\
\mathcal{F}_2^- &= \{\alpha\boldsymbol{a} + \beta\boldsymbol{b} + \gamma\boldsymbol{c} \in \mathcal{P} : \beta = 0\},\\
\mathcal{F}_2^+ &= \{\alpha\boldsymbol{a} + \beta\boldsymbol{b} + \gamma\boldsymbol{c} \in \mathcal{P} : \beta = 1\},\\
\mathcal{F}_3^- &= \{\alpha\boldsymbol{a} + \beta\boldsymbol{b} + \gamma\boldsymbol{c} \in \mathcal{P} : \gamma = 0\},\\
\mathcal{F}_3^+ &= \{\alpha\boldsymbol{a} + \beta\boldsymbol{b} + \gamma\boldsymbol{c} \in \mathcal{P} : \gamma = 1\}.
\end{aligned}$$

So,

$$\begin{aligned}
&\mathcal{F}_1^- \text{ is generated by } \boldsymbol{b} \text{ and } \boldsymbol{c},\\
&\mathcal{F}_2^- \text{ is generated by } \boldsymbol{a} \text{ and } \boldsymbol{c},\\
&\mathcal{F}_3^- \text{ is generated by } \boldsymbol{a} \text{ and } \boldsymbol{b},
\end{aligned}$$

while $\mathcal{F}_k^+$ is congruent to $\mathcal{F}_k^-$, for each $k = 1, 2, 3$.

The six diagonals of $\mathcal{P}$ are defined as

$$\begin{aligned}
\mathcal{D}^1_{1,2} &= \{\alpha\boldsymbol{a} + \beta\boldsymbol{b} + \gamma\boldsymbol{c} \in \mathcal{P} : \alpha = \beta\},\\
\mathcal{D}^2_{1,2} &= \{\alpha\boldsymbol{a} + \beta\boldsymbol{b} + \gamma\boldsymbol{c} \in \mathcal{P} : \alpha + \beta = 1\},\\
\mathcal{D}^1_{1,3} &= \{\alpha\boldsymbol{a} + \beta\boldsymbol{b} + \gamma\boldsymbol{c} \in \mathcal{P} : \alpha = \gamma\},\\
\mathcal{D}^2_{1,3} &= \{\alpha\boldsymbol{a} + \beta\boldsymbol{b} + \gamma\boldsymbol{c} \in \mathcal{P} : \alpha + \gamma = 1\},\\
\mathcal{D}^1_{2,3} &= \{\alpha\boldsymbol{a} + \beta\boldsymbol{b} + \gamma\boldsymbol{c} \in \mathcal{P} : \beta = \gamma\},\\
\mathcal{D}^2_{2,3} &= \{\alpha\boldsymbol{a} + \beta\boldsymbol{b} + \gamma\boldsymbol{c} \in \mathcal{P} : \beta + \gamma = 1\}.
\end{aligned}$$

So,

$$\begin{aligned}
&\mathcal{D}^1_{1,2} \text{ is generated by } \boldsymbol{a} + \boldsymbol{b} \text{ and } \boldsymbol{c},\\
&\mathcal{D}^1_{1,3} \text{ is generated by } \boldsymbol{a} + \boldsymbol{c} \text{ and } \boldsymbol{b},\\
&\mathcal{D}^1_{2,3} \text{ is generated by } \boldsymbol{b} + \boldsymbol{c} \text{ and } \boldsymbol{a},
\end{aligned}$$

while

$$\begin{aligned}
&\mathcal{D}^2_{1,2} \text{ is congruent to the set generated by } \boldsymbol{a} - \boldsymbol{b} \text{ and } \boldsymbol{c},\\
&\mathcal{D}^2_{1,3} \text{ is congruent to the set generated by } \boldsymbol{a} - \boldsymbol{c} \text{ and } \boldsymbol{b},\\
&\mathcal{D}^2_{2,3} \text{ is congruent to the set generated by } \boldsymbol{b} - \boldsymbol{c} \text{ and } \boldsymbol{a}.
\end{aligned}$$

Our proposition is thus translated into the following identity:

$$\begin{aligned}
&\|(\boldsymbol{a} + \boldsymbol{b}) \times \boldsymbol{c}\|^2 + \|(\boldsymbol{a} - \boldsymbol{b}) \times \boldsymbol{c}\|^2\\
&+ \|(\boldsymbol{a} + \boldsymbol{c}) \times \boldsymbol{b}\|^2 + \|(\boldsymbol{a} - \boldsymbol{c}) \times \boldsymbol{b}\|^2\\
&+ \|(\boldsymbol{b} + \boldsymbol{c}) \times \boldsymbol{a}\|^2 + \|(\boldsymbol{b} - \boldsymbol{c}) \times \boldsymbol{a}\|^2 = 4(\|\boldsymbol{b} \times \boldsymbol{c}\|^2 + \|\boldsymbol{a} \times \boldsymbol{c}\|^2 + \|\boldsymbol{a} \times \boldsymbol{b}\|^2).
\end{aligned}$$

Here, we have used the vector product, so that, e.g.,

$$\begin{aligned}
\|\boldsymbol{a} \times \boldsymbol{b}\|^2 &= \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix}^2 + \begin{vmatrix} a_3 & a_1 \\ b_3 & b_1 \end{vmatrix}^2 + \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix}^2\\
&= (a_2b_3 - b_2a_3)^2 + (a_3b_1 - b_3a_1)^2 + (a_1b_2 - b_1a_2)^2.
\end{aligned}$$

In order to prove the above identity, we just notice that, by the parallelogram law,

$$\begin{aligned}
\|(\boldsymbol{a} + \boldsymbol{b}) \times \boldsymbol{c}\|^2 + \|(\boldsymbol{a} - \boldsymbol{b}) \times \boldsymbol{c}\|^2 &=\\
&= \|(\boldsymbol{a} \times \boldsymbol{c}) + (\boldsymbol{b} \times \boldsymbol{c})\|^2 + \|(\boldsymbol{a} \times \boldsymbol{c}) - (\boldsymbol{b} \times \boldsymbol{c})\|^2\\
&= 2(\|\boldsymbol{a} \times \boldsymbol{c}\|^2 + \|\boldsymbol{b} \times \boldsymbol{c}\|^2),
\end{aligned}$$

and similarly

$$\begin{aligned}
\|(\boldsymbol{a} + \boldsymbol{c}) \times \boldsymbol{b}\|^2 + \|(\boldsymbol{a} - \boldsymbol{c}) \times \boldsymbol{b}\|^2 &= 2(\|\boldsymbol{a} \times \boldsymbol{b}\|^2 + \|\boldsymbol{c} \times \boldsymbol{b}\|^2),\\
\|(\boldsymbol{b} + \boldsymbol{c}) \times \boldsymbol{a}\|^2 + \|(\boldsymbol{b} - \boldsymbol{c}) \times \boldsymbol{a}\|^2 &= 2(\|\boldsymbol{b} \times \boldsymbol{a}\|^2 + \|\boldsymbol{c} \times \boldsymbol{a}\|^2).
\end{aligned}$$

Summing up the three formulas, our identity is proved.

Remark 2.1. There surely are several ways to extend the parallelogram law to higher dimensions. Just to mention one of these, in the three-dimensional case we have

$$\begin{aligned}&\|\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c}\|^2+\|\boldsymbol{a}+\boldsymbol{b}-\boldsymbol{c}\|^2\\ &+\|\boldsymbol{a}-\boldsymbol{b}+\boldsymbol{c}\|^2+\|\boldsymbol{a}-\boldsymbol{b}-\boldsymbol{c}\|^2=4(\|\boldsymbol{a}\|^2+\|\boldsymbol{b}\|^2+\|\boldsymbol{c}\|^2).\end{aligned}$$

We acknowledge the referee for pointing out this identity. It is proved directly (by the use of the classical parallelogram law) and can be easily extended to any dimension.

3 Proof of the main theorem

We now provide a proof for the general N-dimensional case. Let $\mathcal{P}$ be the parallelotope generated by the vectors $\boldsymbol{a}_1,\dots,\boldsymbol{a}_N$, i.e.,

$$\mathcal{P}=\left\{\sum_{k=1}^N c_k\boldsymbol{a}_k : c_k\in[0,1],\ \text{for } k=1,\dots,N\right\}.$$

Its $2N$ faces are defined by

$$\mathcal{F}_n^-=\left\{\sum_{k=1}^N c_k\boldsymbol{a}_k\in\mathcal{P} : c_n=0\right\},\quad \mathcal{F}_n^+=\left\{\sum_{k=1}^N c_k\boldsymbol{a}_k\in\mathcal{P} : c_n=1\right\},$$

with $n=1,\dots,N$. Each $\mathcal{F}_n^-$ is generated by the vectors $\boldsymbol{a}_1,\dots,\widehat{\boldsymbol{a}_n},\dots,\boldsymbol{a}_N$, where, as usual, $\widehat{\boldsymbol{a}_n}$ means that $\boldsymbol{a}_n$ is missing, while $\mathcal{F}_n^+$ is a translation of $\mathcal{F}_n^-$, for every $n=1,\dots,N$.

Concerning the diagonals, they are defined as

$$\mathcal{D}_{i,j}^1=\left\{\sum_{k=1}^N c_k\boldsymbol{a}_k\in\mathcal{P} : c_i=c_j\right\},\quad \mathcal{D}_{i,j}^2=\left\{\sum_{k=1}^N c_k\boldsymbol{a}_k\in\mathcal{P} : c_i+c_j=1\right\},$$

with indices $i<j$ varying from 1 to N. There are $N(N-1)$ of them. Hence, we have that

$$\mathcal{D}_{i,j}^1 \text{ is generated by } \boldsymbol{a}_i+\boldsymbol{a}_j \text{ and } \boldsymbol{a}_1,\dots,\widehat{\boldsymbol{a}_i},\dots,\widehat{\boldsymbol{a}_j},\dots,\boldsymbol{a}_N,$$

while

$$\mathcal{D}_{i,j}^2 \text{ is a translation of the set generated by } \boldsymbol{a}_i-\boldsymbol{a}_j \text{ and } \boldsymbol{a}_1,\dots,\widehat{\boldsymbol{a}_i},\dots,\widehat{\boldsymbol{a}_j},\dots,\boldsymbol{a}_N.$$

In order to compute the $(N-1)$-dimensional measures of the faces and the diagonals of our parallelotope, we make use of the following proposition involving the exterior product of vectors in $\mathbb{R}^N$. (See, e.g., [3] for the definition and the main properties of the exterior product.)

Proposition 3.1. *The M-dimensional measure of a parallelotope generated by M vectors $\boldsymbol{v}_1,\dots,\boldsymbol{v}_M$ in $\mathbb{R}^N$, with $1\le M\le N$, is given by $\|\boldsymbol{v}_1\wedge\cdots\wedge\boldsymbol{v}_M\|$.*

Proof. If $\boldsymbol{v}_1,\dots,\boldsymbol{v}_M$ are linearly dependent, the M-dimensional measure of the parallelotope generated by $\boldsymbol{v}_1,\dots,\boldsymbol{v}_M$ is equal to zero, hence coincides with $\|\boldsymbol{v}_1\wedge\cdots\wedge\boldsymbol{v}_M\|$.

Assume now that the vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_M$ are linearly independent, and let V be the subspace generated by them. Choose an orthonormal basis $\boldsymbol{e}_1, \dots, \boldsymbol{e}_M$ of V, and write

$$\begin{aligned} \boldsymbol{v}_1 &= v_{11}\boldsymbol{e}_1 + \cdots + v_{1M}\boldsymbol{e}_M, \\ &\vdots \\ \boldsymbol{v}_M &= v_{M1}\boldsymbol{e}_1 + \cdots + v_{MM}\boldsymbol{e}_M. \end{aligned}$$

Then,

$$\boldsymbol{v}_1 \wedge \cdots \wedge \boldsymbol{v}_M = \det \begin{pmatrix} v_{11} & \cdots & v_{1M} \\ \vdots & & \vdots \\ v_{M1} & \cdots & v_{MM} \end{pmatrix} \boldsymbol{e}_1 \wedge \cdots \wedge \boldsymbol{e}_M,$$

so that

$$\|\boldsymbol{v}_1 \wedge \cdots \wedge \boldsymbol{v}_M\| = \left| \det \begin{pmatrix} v_{11} & \cdots & v_{1M} \\ \vdots & & \vdots \\ v_{M1} & \cdots & v_{MM} \end{pmatrix} \right|,$$

which is indeed the M-dimensional measure of the parallelotope generated by the vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_M$. □

Hence, the $(N-1)$-dimensional measures of the faces $\mathcal{F}_n^{\pm}$ are given by

$$\|\boldsymbol{a}_1 \wedge \cdots \wedge \widehat{\boldsymbol{a}_n} \wedge \cdots \wedge \boldsymbol{a}_N\|,$$

while the $(N-1)$-dimensional measures of the diagonals $\mathcal{D}^1_{i,j}$ are equal to

$$\|(\boldsymbol{a}_i + \boldsymbol{a}_j) \wedge \textstyle\bigwedge_{k\neq i,j} \boldsymbol{a}_k\|,$$

and those of the diagonals $\mathcal{D}^2_{i,j}$ are equal to

$$\|(\boldsymbol{a}_i - \boldsymbol{a}_j) \wedge \textstyle\bigwedge_{k\neq i,j} \boldsymbol{a}_k\|.$$

Choosing any couple $i < j$, by the parallelogram law we have that

$$\begin{aligned} &\|(\boldsymbol{a}_i + \boldsymbol{a}_j) \wedge \textstyle\bigwedge_{k\neq i,j} \boldsymbol{a}_k\|^2 + \|(\boldsymbol{a}_i - \boldsymbol{a}_j) \wedge \textstyle\bigwedge_{k\neq i,j} \boldsymbol{a}_k\|^2 = \\ &\qquad = 2\Big(\|\boldsymbol{a}_1 \wedge \cdots \wedge \widehat{\boldsymbol{a}_j} \wedge \cdots \wedge \boldsymbol{a}_N\|^2 + \|\boldsymbol{a}_1 \wedge \cdots \wedge \widehat{\boldsymbol{a}_i} \wedge \cdots \wedge \boldsymbol{a}_N\|^2\Big). \end{aligned}$$

We now want to take the sum of all these equalities, with $i < j$ varying form 1 to N. We claim that, for any $n = 1, \dots, N$, when performing such a sum, in the right hand side,

the term $2\|\boldsymbol{a}_1 \wedge \cdots \wedge \widehat{\boldsymbol{a}_n} \wedge \cdots \wedge \boldsymbol{a}_N\|^2$ will appear $N-1$ times.

Indeed, this term may appear with $j = n$, while i varies from 1 to $n-1$, or with $i = n$, while j varies from $n+1$ to N, and there are exactly $N-1$ of such possibilities. Hence, summing all the equalities, we have that

$$\begin{aligned} &\sum_{i<j} \Big(\|(\boldsymbol{a}_i + \boldsymbol{a}_j) \wedge \textstyle\bigwedge_{k\neq i,j} \boldsymbol{a}_k\|^2 + \|(\boldsymbol{a}_i - \boldsymbol{a}_j) \wedge \textstyle\bigwedge_{k\neq i,j} \boldsymbol{a}_k\|^2\Big) = \\ &\qquad\qquad = (N-1) \sum_{n=1}^{N} 2\|\boldsymbol{a}_1 \wedge \cdots \wedge \widehat{\boldsymbol{a}_n} \wedge \cdots \wedge \boldsymbol{a}_N\|^2. \end{aligned}$$

So, we have proved the following.

> *For any N-dimensional parallelotope, the sum of the squares of the $(N-1)$-dimensional measures of its $N(N-1)$ diagonals is equal to $N-1$ times the sum of the squares of the $(N-1)$-dimensional measures of its $2N$ faces.*

The proof of the theorem is now easily completed, dividing each of the two sums by the number of their addends and taking the square roots.

Remark 3.2. Since our result is valid in any dimension N, it would be interesting to investigate whether it could be extended also to some infinite-dimensional vector spaces. This seems to be a remarkable problem which could lead to further insight on the nature of these identities.

References

[1] M. Khosravi and M. D. Taylor, The wedge product and analytic geometry, *Amer. Math. Monthly* **115** (2008), 623–644, doi:10.1080/00029890.2008.11920573.

[2] A. Nash, A generalized parallelogram law, *Amer. Math. Monthly* **110** (2003), 52–57, doi:10.2307/3072345.

[3] I. R. Shafarevich and A. O. Remizov, *Linear Algebra and Geometry*, Springer, Heidelberg, 2013, doi:10.1007/978-3-642-30994-6, translated from the 2009 Russian original by D. Kramer and L. Nekludova.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 419–443
https://doi.org/10.26493/1855-3974.1401.7d9
(Also available at http://amc-journal.eu)

S^2 coverings by isosceles and scalene triangles – adjacency case I

Catarina P. Avelino [*]

*Centre of Mathematics of the University of Minho – UTAD Pole (CMAT-UTAD),
University of Trás-os-Montes e Alto Douro, Vila Real, Portugal*
affiliated also with: *Center for Computational and Stochastic Mathematics (CEMAT),
University of Lisboa (IST-UL), Portugal*

Altino F. Santos

*Centre of Mathematics of the University of Minho – UTAD Pole (CMAT-UTAD),
University of Trás-os-Montes e Alto Douro, Vila Real, Portugal*

Received 9 May 2017, accepted 16 October 2017, published online 9 February 2019

Abstract

The aim of this paper is the study and classification of spherical f-tilings by scalene triangles T and isosceles triangles T'. Due to the complexity of this wide class of tilings, we consider a subclass performed by the adjacency of the shortest side of T and the longest side of T'. It consists of seven families of f-tilings (four families with one discrete parameter and one continuous parameter, two families with one discrete parameter and one sporadic f-tiling). We also analyze the combinatorial structure of all these families of f-tilings, as well as the group of symmetries of each tiling and the transitivity classes of isohedrality and isogonality.

Keywords: Dihedral f-tilings, combinatorial properties, spherical trigonometry, symmetry groups.

Math. Subj. Class.: 52C20, 52B05, 20B35

1 Introduction

A *folding tessellation* or *folding tiling* (f-tiling, for short) of the sphere S^2 is an edge-to-edge finite polygonal tiling τ of S^2 such that all vertices of τ satisfy the angle-folding relation, i.e., each vertex is of even valency and the sums of alternate angles around each vertex are equal to π.

[*] This research was partially supported by Fundação para a Ciência e a Tecnologia (FCT) through projects UID/MAT/00013/2013 and UID/Multi/04621/2013.

E-mail addresses: cavelino@utad.pt (Catarina P. Avelino), afolgado@utad.pt (Altino F. Santos)

F-tilings are intrinsically related to the theory of isometric foldings of Riemannian manifolds, introduced by Robertson [8] in 1977. In some situations (beyond the scope of this paper), the edge-complex associated to a spherical f-tiling is the set of singularities of some spherical isometric folding.

The classification of f-tilings was initiated by Breda [1], with a complete classification of all spherical monohedral (triangular) f-tilings. Afterwards, in 2002, Ueno and Agaoka [9] have established the complete classification of all triangular monohedral tilings of the sphere (without any restrictions on angles). Curiously, the triangular tilings of even valency at any vertex are necessarily f-tilings. Dawson has also been interested in special classes of spherical tilings, see [3, 4, 5], for instance. Spherical f-filings by two noncongruent classes of isosceles triangles have recently obtained [2, 7].

From now on,

(i) T denotes a spherical scalene triangle with internal angles $\alpha > \beta > \gamma$ and side lengths $a > b > c$;

(ii) T' denotes a spherical isosceles triangle with internal angles $(\delta, \delta, \varepsilon)$, $\delta \neq \varepsilon$, and side lengths (d, d, e),

as illustrated in Figure 1.

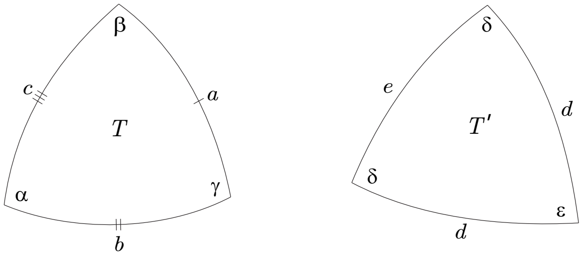


Figure 1: A spherical scalene triangle, T, and a spherical isosceles triangle, T'.

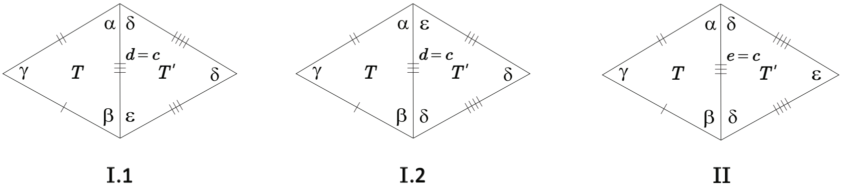
Taking into account the area of the prototiles T and T', we have

$$\alpha + \beta + \gamma > \pi \quad \text{and} \quad 2\delta + \varepsilon > \pi.$$

As $\alpha > \beta > \gamma$, we also have $\alpha > \frac{\pi}{3}$. In [6] it was established that any f-tiling by T and T' has necessarily vertices of valency four.

We begin by pointing out that any f-tiling by T and T', in which the shortest side of T is equal to the longest side of T', has at least two cells congruent to T and T', respectively, such that they are in adjacent positions and in one and only one of the situations illustrated in Figure 2. Our aim in this paper is to classify f-tilings in the first case of adjacency (Figure 2-Case I).

Next section contains the main results of this paper. In Subsection 2.1 we describe six families of spherical f-tilings and one single f-tiling that we may obtain in this case of adjacency. The combinatorial structure of these f-tilings and the classification of the group of symmetries and also the transitivity classes of isogonality and isohedrality are presented in Subsection 2.2. The proof of the main result consists in a long and exhaustive methodology and it is presented in Section 3.



Case I: $\delta > \varepsilon$ Case II: $\delta < \varepsilon$

Figure 2: Distinct cases of adjacency.

2 Main result

2.1 f-tilings in the adjacency case I

Theorem 2.1. *Let T and T' be a spherical scalene triangle and a spherical isosceles triangle, respectively, such that they are in one of the adjacent positions illustrated in Figure 2-Case I. Then, from this we obtain six families of spherical f-tilings and one isolated f-tiling,*

$$\begin{array}{llll} \mathcal{D}_\delta^k \ (k \geq 3), & \mathcal{G}^k \ (k \geq 4), & \bar{\mathcal{G}}^k \ (k \geq 4), & \mathcal{H}, \\ \mathcal{F}_\beta^k \ (k \geq 4), & \mathcal{I}_\beta^k \ (k \geq 3), & \mathcal{J}_\beta^k \ (k \geq 4), & \end{array}$$

that satisfy, respectively:

(i) $\alpha + \delta = \pi,\ \delta + \beta + \varepsilon = \pi,\ k\gamma = \pi,\ \varepsilon = \varepsilon_k(\delta),\ \delta \in \left(\delta_{\min}^k, \frac{\pi}{2}\right),\ k \geq 3$, *where*

$$\varepsilon_k(\delta) = 2 \operatorname{arccot}\left(2 \cos\frac{\pi}{k} \csc 2\delta - \cot\delta\right) \quad \textit{and}$$

$$\delta_{\min}^k = \arccos\frac{\sqrt{1 + 8\cos\frac{\pi}{k}} - 1}{4};$$

(ii) $\alpha + \delta = \pi,\ \alpha + \beta + \varepsilon = \pi,\ \delta + \beta + \gamma = \pi,\ k\gamma = \pi,\ \delta = \delta_k,\ k \geq 4$, *where*

$$\delta_k = \operatorname{arccot}\left(\frac{1}{2}\tan\frac{\pi}{2k}\left(2 - \sec^2\frac{\pi}{2k}\right)\right);$$

(iii) $\alpha + \delta = \pi,\ \alpha + \beta + \varepsilon = \pi,\ \delta + \beta + \gamma = \pi,\ 2\beta + \gamma + \varepsilon = \pi,\ k\gamma = \pi,\ \delta = \delta_k,\ k \geq 4$*;*

(iv) $\alpha + \delta = \pi,\ \alpha + \gamma + \gamma = \pi,\ 3\beta + \varepsilon = \pi,\ 5\gamma = \pi$, *where*

$$\beta = \beta^0 = 4\arctan\sqrt{3 + 4\sqrt{5} - 2\sqrt{22 + 6\sqrt{5}}};$$

(v) $\alpha + \delta = \pi,\ 2\beta + \gamma + \varepsilon = \pi,\ k\gamma = \pi,\ \alpha = \alpha_k^1(\beta),\ \beta \in \left(\beta_{\min}^{1k}, \beta_{\max}^{1k}\right),\ k \geq 4$, *where*

$$\alpha_k^1(\beta) = \arccos\left(-\cos\frac{\pi}{k}\sec\frac{\pi}{2k}\cos\left(\beta + \frac{\pi}{2k}\right)\right),$$

$$\beta_{\min}^{1k} = \max\left\{\frac{\pi}{k}, \arccos\left(\frac{1}{2}\sec\frac{\pi}{2k}\right) - \frac{\pi}{2k}\right\} \quad \textit{and}$$

$$\beta_{\max}^{1k} = \frac{(k-1)\pi}{2k};$$

(vi) $\alpha+\delta=\pi,\ 2\beta+\varepsilon=\pi,\ k\gamma=\pi,\ \alpha=\alpha_k^2(\beta),\ \beta\in\left(\beta_{\min}^{2k},\frac{\pi}{2}\right),\ k\geq 3$, *where*

$$\alpha_k^2(\beta)=\arccos\left(-\cos\frac{\pi}{k}\cos\beta\right)\quad \textit{and}$$
$$\beta_{\min}^{2k}=\max\left\{\frac{\pi}{k},\arccos\frac{\sqrt{\cos^2\frac{\pi}{k}+8}-\cos\frac{\pi}{k}}{4}\right\};$$

(vii) $\alpha+\varepsilon=\pi,\ \beta+2\delta=\pi,\ k\gamma=\pi,\ \alpha=\alpha_k^3(\beta),\ \beta\in\left(\frac{\pi}{k},\beta_{\max}^{3k}\right),\ k\geq 4$, *where*

$$\alpha_k^3(\beta)=\arccos\left(2\sin^2\frac{\beta}{2}-\cos\frac{\pi}{k}\right)\quad \textit{and}$$
$$\beta_{\max}^{3k}=2\arcsin\frac{\sqrt{1+8\cos\frac{\pi}{k}}-1}{4}.$$

For each family of f-tilings we present the distinct classes of congruent vertices in Figure 3 (including the respective number of vertices in each tiling).

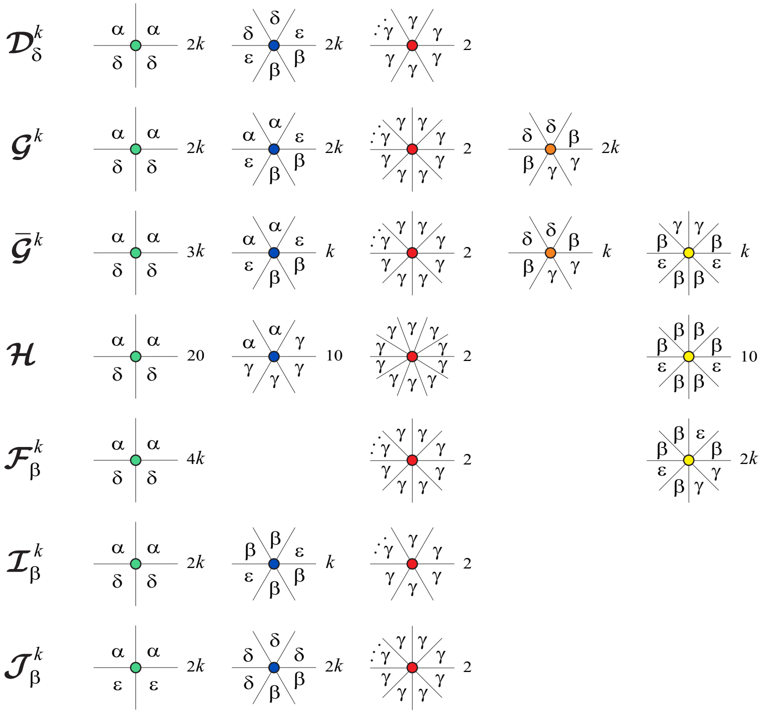
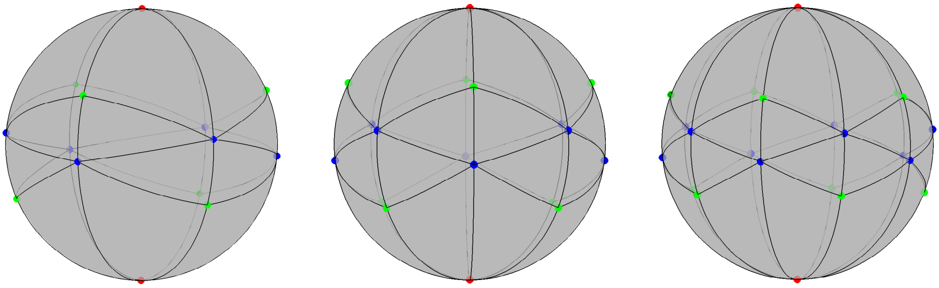


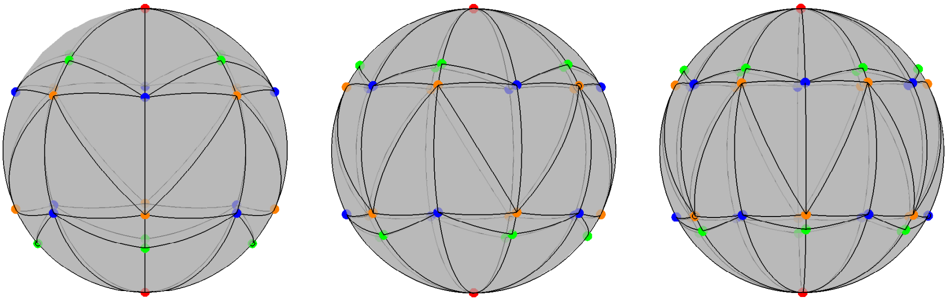
Figure 3: Distinct classes of congruent vertices.

Particularizing suitable values for the parameters involved in each case, the corresponding 3D representations of these families of f-tilings are given in Figures 4 – 10.



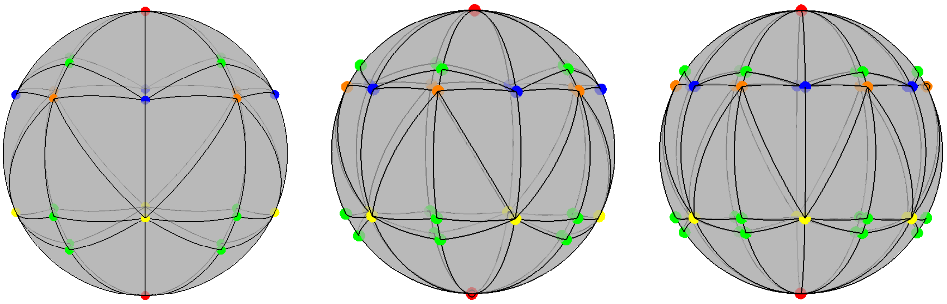
(a) $\mathcal{D}_\delta^3$ (b) $\mathcal{D}_\delta^4$ (c) $\mathcal{D}_\delta^5$

Figure 4: f-tilings in the adjacency case I; the $\mathcal{D}_\delta^k$ family.



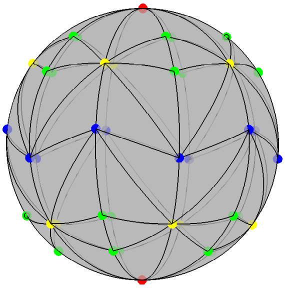
(d) $\mathcal{G}^4$ (e) $\mathcal{G}^5$ (f) $\mathcal{G}^6$

Figure 5: f-tilings in the adjacency case I; the $\mathcal{G}^k$ family.



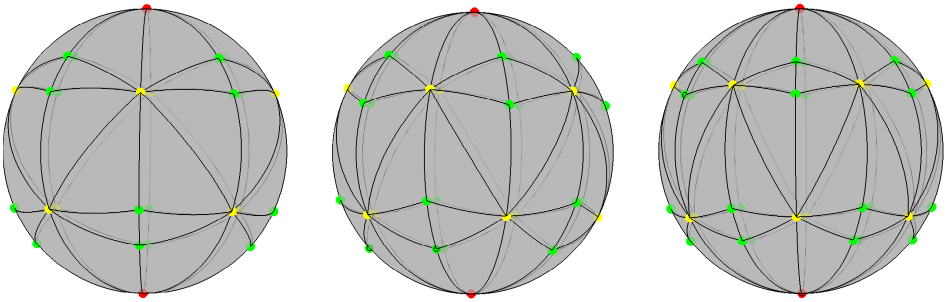
(g) $\bar{\mathcal{G}}^4$ (h) $\bar{\mathcal{G}}^5$ (i) $\bar{\mathcal{G}}^6$

Figure 6: f-tilings in the adjacency case I; the $\bar{\mathcal{G}}^k$ family.



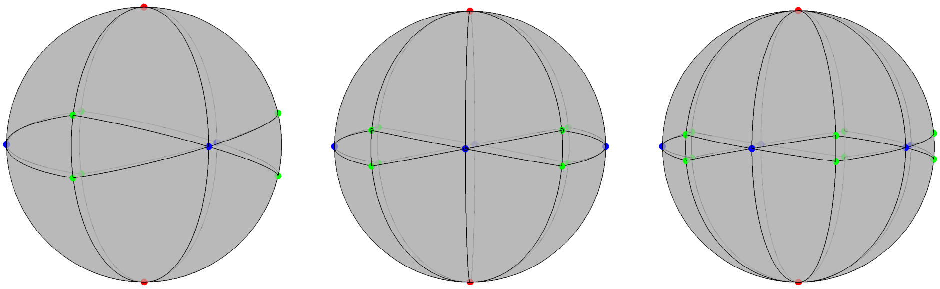
(j) $\mathcal{H}$

Figure 7: f-tilings in the adjacency case I; the isolated f-tiling.



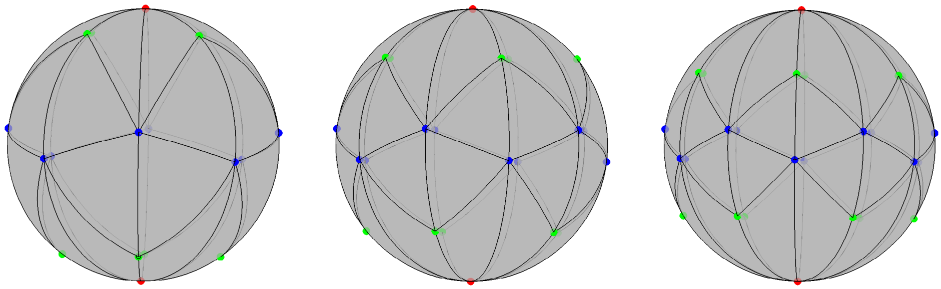
(k) $\mathcal{F}_{\beta}^{4}$ (l) $\mathcal{F}_{\beta}^{5}$ (m) $\mathcal{F}_{\beta}^{6}$

Figure 8: f-tilings in the adjacency case I; the $\mathcal{F}_{\beta}^{k}$ family.



(n) $\mathcal{I}_{\beta}^{3}$ (o) $\mathcal{I}_{\beta}^{4}$ (p) $\mathcal{I}_{\beta}^{5}$

Figure 9: f-tilings in the adjacency case I; the $\mathcal{I}_{\beta}^{k}$ family.



(q) $\mathcal{J}_\beta^4$ (r) $\mathcal{J}_\beta^5$ (s) $\mathcal{J}_\beta^6$

Figure 10: f-tilings in the adjacency case I; the $\mathcal{J}_\beta^k$ family.

2.2 Symmetry groups and combinatorial structure

In this subsection we present the group of symmetries of each spherical f-tiling mentioned in Theorem 2.1. The number of transitivity classes of tiles and vertices of each tiling is indicated in Table 1.

Any symmetry of $\mathcal{D}_\delta^k$, $k \geq 3$, fixes the north pole $N = (0, 0, 1)$ (and consequently the south pole $S = -N$) or maps N into S (and consequently S into N). The symmetries that fix N are generated, for instance, by the rotation $R^z_{\frac{2\pi}{k}}$ (of an angle $\frac{2\pi}{k}$ around the z axis) and the reflection ρ^{yz} (on the coordinate plane $y \circ z$) giving rise to a subgroup of $G(\mathcal{D}_\delta^k)$ isomorphic to D_k, the dihedral group of order $2k$. Now, the map

$$\phi = R^z_{\frac{\pi}{k}} \circ \rho^{xy} = \rho^{xy} \circ R^z_{\frac{\pi}{k}}$$

is a symmetry of $\mathcal{D}_\delta^k$ that changes N and S. One has $\phi^{2k-1} \circ \rho^{yz} = \rho^{yz} \circ \phi$ and ϕ has order $2k$. It follows that ϕ and ρ^{yz} generate $G(\mathcal{D}_\delta^k)$, and so it is isomorphic to D_{2k}. Moreover, $\mathcal{D}_\delta^k$ is 2-tile-transitive and 3-vertex-transitive with respect to this group.

The analysis considered to the combinatorial structure of $\mathcal{D}_\delta^k$ also applies to the family of f-tilings $\mathcal{G}^k$, $k \geq 4$. And so $G(\mathcal{G}^k) = D_{2k}$. $\mathcal{G}^k$ is 3-isohedral and 4-isogonal.

Concerning the family of f-tilings $\bar{\mathcal{G}}^k$, $k \geq 4$, we have that $G(\bar{\mathcal{G}}^k) = D_k$, since in this case there is no symmetry sending the north pole into the south pole. Moreover, $\bar{\mathcal{G}}^k$ has 6 transitivity classes of tiles, and so it is 6-isohedral. The vertices of $\bar{\mathcal{G}}^k$ form 8 transitivity classes.

Regarding the symmetry group of $\mathcal{H}$, the symmetries that fix N are generated by the rotation $R^z_{\frac{2\pi}{5}}$ and the reflection ρ^{yz} on the plane $x = 0$. On the other hand,

$$\phi = R^z_{\frac{\pi}{5}} \circ \rho^{xy}$$

is also a symmetry of $\mathcal{H}$ that sends N into S. Thus, we conclude that $G(\mathcal{H})$ is isomorphic to D_{10}, the dihedral group of order 20. $\mathcal{H}$ is 4-tile-transitive and 5-vertex-transitive.

Any symmetry of $\mathcal{I}_\beta^k$, $k \geq 3$, fixes N or maps N into S. The symmetries that fix N are generated, for instance, by the rotation $R^z_{\frac{2\pi}{k}}$ of order k and the reflection ρ^{yz}, giving rise to a subgroup $\mathcal{S}$ of $G(\mathcal{I}_\beta^k)$ isomorphic to D_k. To obtain the symmetries that send N into

S it is enough to compose each element of $\mathcal{S}$ with ρ^{xy}. Since ρ^{xy} commutes with $R^{z}_{\frac{\pi}{k}}$ and ρ^{yz}, we may conclude that $G(\mathcal{I}^k_\beta)$ is isomorphic to $C_2 \times D_k$. $\mathcal{I}^k_\beta$ has 2 transitivity classes of tiles with respect to the group of symmetries and 3 transitivity classes of vertices.

Similarly to previous cases, we have $G(\mathcal{F}^k_\beta) = G(\mathcal{J}^k_\beta) = D_{2k}$. $\mathcal{F}^k_\beta$ is 3-isohedral and 4-isogonal and $\mathcal{J}^k_\beta$ is 2-isohedral and 3-isogonal.

The combinatorial structure of the class of spherical f-tilings described in the previous subsection, including the symmetry groups, is summarized in Table 1. Our notation is as follows:

- $|V|$ is the number of distinct classes of congruent vertices;
- N_1 and N_2 are, respectively, the number of triangles congruent to T and T', respectively;
- $G(\tau)$ is the symmetry group of each tiling τ and the indices of isohedrality and isogonality for the symmetry group are denoted, respectively, by #isoh. and #isog.

3 Proof of Theorem 2.1

In the case of adjacency I, any f-tiling by T and T' has at least two cells congruent to T and T', respectively, such that they are in adjacent positions and in one and only one of the situations illustrated in Figure 2. After certain initial assumptions are made, it is usually possible to deduce sequentially the nature and orientation of most of the other tiles. Eventually, either a complete tiling or an impossible configuration proving that the hypothetical tiling fails to exist is reached. In the diagrams that follow, the order in which these deductions can be made is indicated by the numbering of the tiles. For $j \geq 2$, the location of tiling j can be deduced directly from the configurations of tiles $(1, 2, \dots, j-1)$ and from the hypothesis that the configuration is part of a complete f-tiling, except where otherwise indicated.

Observe that we have $\delta > \frac{\pi}{3}$. Also, as $d = c$ and using spherical trigonometric formulas, we get

$$\frac{\cos\gamma + \cos\alpha\cos\beta}{\sin\alpha\sin\beta} = \cot\delta\cot\frac{\varepsilon}{2}. \tag{3.1}$$

Proof of Theorem 2.1. We consider separately the subcases illustrated in Figure 2-Case I.

Case I.1: With the labeling of Figure 11(a), at vertex v_1 we must have

$$\alpha + \delta < \pi \quad \text{or} \quad \alpha + \delta = \pi.$$

Case I.1.1: Suppose firstly that $\alpha + \delta < \pi$. If $\alpha < \delta$, we must have $\alpha + \delta + k\varepsilon = \pi$, with $k \geq 1$. Due to the existence of vertices of valency four, it follows that $\delta = \frac{\pi}{2}$, and consequently, by Equation (3.1), $\cos\gamma + \cos\alpha\cos\beta = 0$. Nevertheless, this is not possible, since $\cos\gamma > \cos\beta > \cos\alpha > 0$. Therefore, $\alpha \geq \delta$. It follows that $\alpha > \beta > \delta > \varepsilon > \gamma$ and $\alpha + \delta + k\gamma = \pi$, with $k \geq 1$; see Figure 11(b). Note that $\theta_1 = \gamma$, otherwise at vertex v_2 we get $\alpha + \beta = \pi = \gamma + \varepsilon$, which is an impossibility. Now, we have

$$\theta_2 = \gamma, \quad \theta_2 = \delta \quad \text{or} \quad \theta_2 = \varepsilon.$$

Case I.1.1.1: If $\theta_2 = \gamma$, we obtain the configuration illustrated in Figure 12(a). Due to the edge lengths, at vertex v_3 we must have $\theta_3 + \beta + \rho \leq \pi$, with $\rho \geq \varepsilon$, which implies $\theta_3 = \varepsilon$. At vertex v_4 we reach a contradiction, as $\alpha + \delta + \rho > \pi$, for all $\rho \in \{\alpha, \beta, \delta, \varepsilon\}$.

Table 1: Combinatorial structure of the dihedral f-tilings of S^2 by scalene triangles T and isosceles triangles T' performed by the shortest side of T and the longest side of T' in the case of adjacency I.

f-tiling	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	$\boldsymbol{\delta}$	$\boldsymbol{\varepsilon}$	$\|V\|$	N_1	N_2	$G(\tau)$	**#isoh.**	**#isog.**
$\boldsymbol{\mathcal{D}_{\delta}^{k},\ k\geq 3}$	$\pi-\delta$	$\pi-\delta-\varepsilon$	$\frac{\pi}{k}$	$\left(\delta_{\min}^{k},\frac{\pi}{2}\right)$	$\varepsilon_k(\delta)$	3	$4k$	$4k$	D_{2k}	2	3
$\boldsymbol{\mathcal{G}^{k},\ k\geq 4}$	$\pi-\delta$	$\frac{(k-1)\pi}{k}-\delta$	$\frac{\pi}{k}$	δ_k	$2\delta-\frac{(k-1)\pi}{k}$	4	$8k$	$4k$	D_{2k}	3	4
$\boldsymbol{\bar{\mathcal{G}}^{k},\ k\geq 4}$	$\pi-\delta$	$\frac{(k-1)\pi}{k}-\delta$	$\frac{\pi}{k}$	δ_k	$2\delta-\frac{(k-1)\pi}{k}$	5	$8k$	$4k$	D_k	6	8
$\boldsymbol{\mathcal{H}}$	$\frac{3\pi}{5}$	β^0	$\frac{\pi}{5}$	$\frac{2\pi}{5}$	$\pi-3\beta^0$	4	60	20	D_{10}	4	5
$\boldsymbol{\mathcal{F}_{\beta}^{k},\ k\geq 4}$	$\alpha_k^1(\beta)$	$\left(\beta_{\min}^{1k},\beta_{\max}^{1k}\right)$	$\frac{\pi}{k}$	$\pi-\alpha$	$\frac{(k-1)\pi}{k}-2\beta$	3	$8k$	$4k$	D_{2k}	3	4
$\boldsymbol{\mathcal{I}_{\beta}^{k},\ k\geq 3}$	$\alpha_k^2(\beta)$	$\left(\beta_{\min}^{2k},\frac{\pi}{2}\right)$	$\frac{\pi}{k}$	$\pi-\alpha$	$\pi-2\beta$	3	$4k$	$2k$	$C_2\times D_k$	2	3
$\boldsymbol{\mathcal{J}_{\beta}^{k},\ k\geq 4}$	$\alpha_k^3(\beta)$	$\left(\frac{\pi}{k},\beta_{\max}^{3k}\right)$	$\frac{\pi}{k}$	$\frac{\pi-\beta}{2}$	$\pi-\alpha$	3	$4k$	$4k$	D_{2k}	2	3

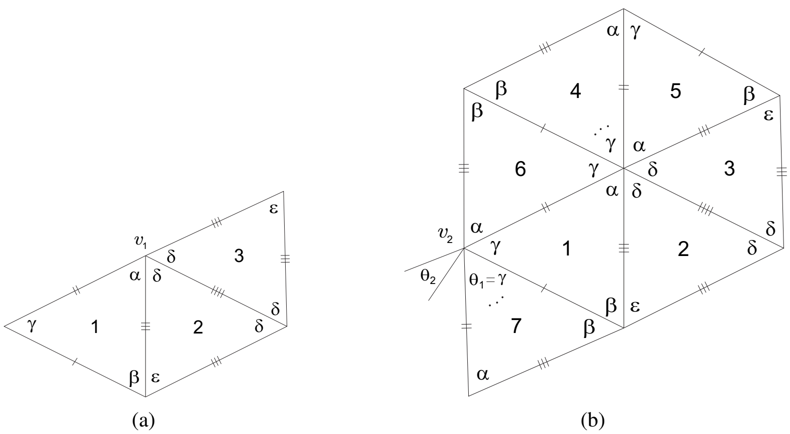


Figure 11: Local configurations.

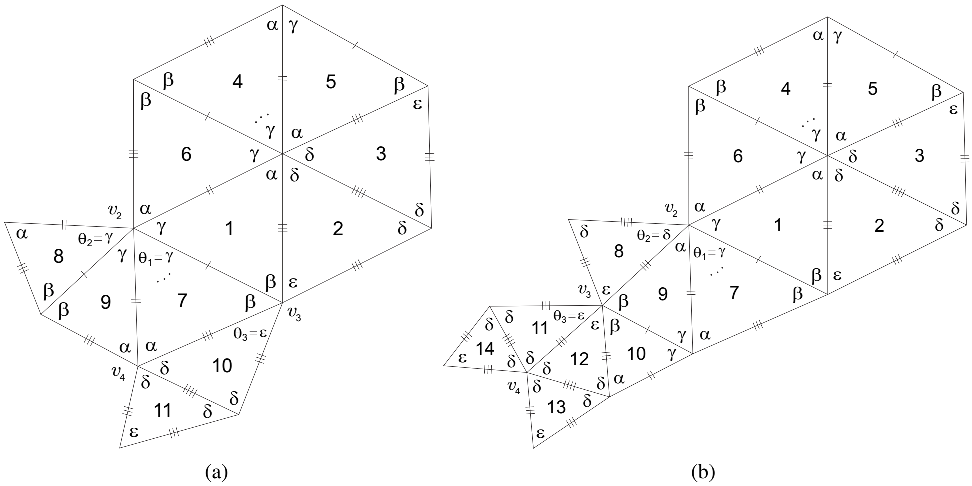
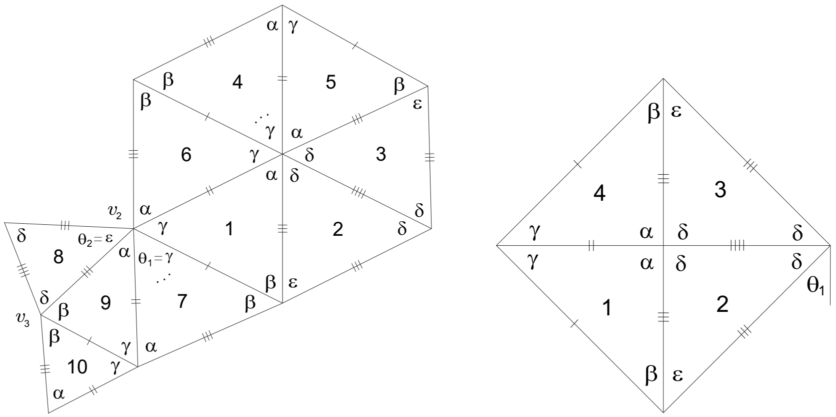


Figure 12: Local configurations.

Case I.1.1.2: If $\theta_2 = \delta$ (Figure 12(b)), we reach an impossibility at vertex v_4, since $\delta + \delta + \rho > \pi$, for all $\rho \in \{\alpha, \beta, \delta, \varepsilon\}$. Note that θ_3 cannot be γ (tile 11), as it implies a sum of alternate angles at vertex v_3 including the angles β, ρ_1 and ρ_2, with $\rho_1 \in \{\alpha, \beta\}$ and $\rho_2 \in \{\alpha, \beta, \delta, \varepsilon\}$, which is not possible due to the dimensions of the involved angles.

Case I.1.1.3: Finally we consider $\theta_2 = \varepsilon$ (Figure 13(a)). At vertex v_3 we must have $\delta + \beta + \bar{k}\gamma = \pi$, $\bar{k} > k$. Nevertheless, an incompatibility between sides at this vertex cannot be avoided.

Case I.1.2: Suppose now that $\alpha + \delta = \pi$ (consequently $\beta + \gamma > \delta > \frac{\pi}{3}$). If $\alpha = \delta = \frac{\pi}{2}$, we also get $\gamma = \frac{\pi}{2}$, which is not possible. On the other hand, if $\delta > \frac{\pi}{2} > \alpha\,(> \beta > \gamma)$, we obtain $\cot \delta < 0$, thereby making Equation (3.1) infeasible. Thus, $\alpha > \frac{\pi}{2} > \delta$. With the



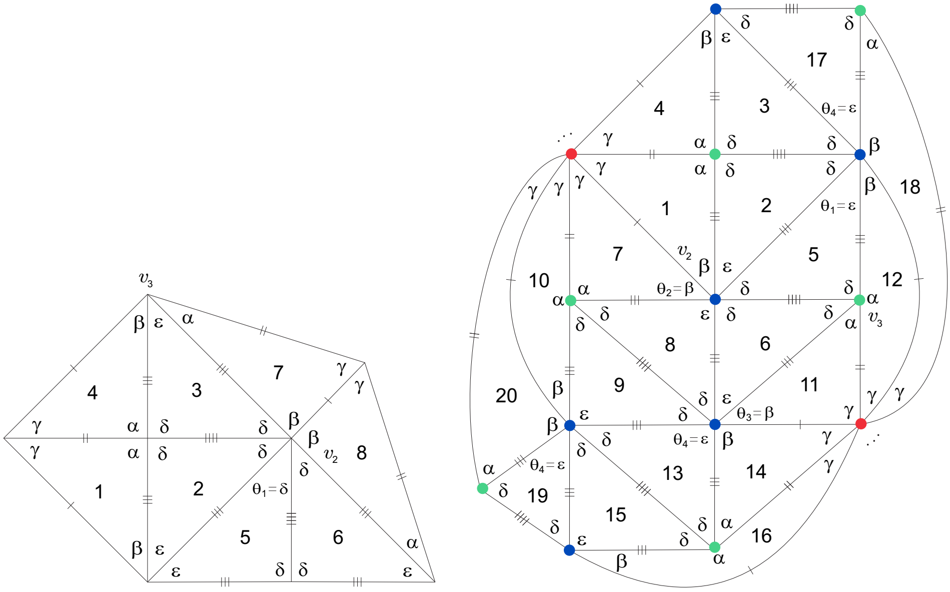
(a) (b)

Figure 13: Local configurations.

labeling of Figure 13(b), we have

$$\theta_1 = \delta, \quad \theta_1 = \varepsilon, \quad \theta_1 = \beta \quad \text{or} \quad \theta_1 = \alpha.$$

Case I.1.2.1: If $\theta_1 = \delta$, we get the configuration illustrated in Figure 14(a). Note that, at vertex v_2, it is not possible to have $\delta+\delta+k\gamma=\pi$, with $k\geq 1$, and $\delta+\delta+\beta+\gamma>\pi$. At vertex v_3 we must have $\alpha+\beta+k\varepsilon=\pi$, with $k\geq 1$. Nevertheless, at this vertex we



(a) (b) Planar representation of $\mathcal{D}_{\delta}^{k}$.

Figure 14: Local configurations.

reach a contradiction, since $(\delta+\delta+\beta)+(\alpha+\beta+\varepsilon) > (\delta+\delta+\varepsilon)+(\alpha+\beta+\gamma) > 2\pi$.

Case I.1.2.2: If $\theta_1 = \varepsilon$, we obtain the configuration of Figure 14(b). Note that if $\theta_2 = \gamma$, we would get the angles $(\delta, \varepsilon, \gamma, \beta, \ldots)$ in one of the sum of alternate angles at vertex v_2; but $(\delta+\varepsilon+\gamma+\beta)+(\alpha+\delta) = (\delta+\delta+\varepsilon)+(\alpha+\beta+\gamma) > 2\pi$, which is not possible; at tile 11, it is easy to observe that $\theta_3 \neq \alpha, \gamma, \delta$; on the other hand, θ_3 cannot be ε, otherwise, at vertex v_3, we get $\delta+\delta+\beta = \pi$, but $(\alpha+\delta)+(\delta+\delta+\beta)+(\varepsilon+\delta+\beta+\varepsilon+\cdots) > 2(\delta+\delta+\varepsilon)+(\alpha+\beta+\gamma) > 3\pi$, which is a contradiction; a similar reasoning applies to the choice of θ_4 and the fact that $k = 1$ in the sum $\delta+\beta+k\varepsilon = \pi$, at vertex v_2. We denote the continuous family of f-tilings illustrated in Figure 14(b) by $\mathcal{D}_\delta^k$, where

$$\alpha+\delta = \pi, \quad \delta+\beta+\varepsilon = \pi \quad \text{and} \quad k\gamma = \pi, \text{ with } k \geq 3.$$

As $0 < \varepsilon < \delta < \frac{\pi}{2}$, using Equation (3.1) we get

$$\begin{aligned} \frac{\cos\frac{\pi}{k}+\cos\delta\cos(\delta+\varepsilon)}{\sin(\delta+\varepsilon)} = \frac{\cos\delta\cos\frac{\varepsilon}{2}}{\sin\frac{\varepsilon}{2}} &\iff \cos\frac{\pi}{k}\sin\frac{\varepsilon}{2} = \cos\delta\sin\left(\delta+\frac{\varepsilon}{2}\right) \\ &\iff \cos\frac{\pi}{k} = \cos\delta\sin\delta\cot\frac{\varepsilon}{2}+\cos^2\delta \\ &\iff \cot\frac{\varepsilon}{2} = 2\cos\frac{\pi}{k}\csc 2\delta - \cot\delta. \end{aligned}$$

Therefore,

$$\varepsilon = \varepsilon_k(\delta) = 2\operatorname{arccot}\left(2\cos\frac{\pi}{k}\csc 2\delta - \cot\delta\right), \ k \geq 3,$$

with $\delta \in \left(\delta_{\min}^k, \frac{\pi}{2}\right)$, where

$$\delta_{\min}^k = \arccos\frac{\sqrt{1+8\cos\frac{\pi}{k}}-1}{4} > \frac{\pi}{3}$$

is obtained when $\varepsilon = \delta$. The graph of this function for $\delta_{\min}^k < \delta < \frac{\pi}{2}$ is outlined in Figure 15, for different values of k. 3D representations of $\mathcal{D}_\delta^3$, $\mathcal{D}_\delta^4$ and $\mathcal{D}_\delta^5$ are given in Figures 4(a) – 4(c).

Case I.1.2.3: Consider $\theta_1 = \beta$ (Figure 16(a)). At vertex v_1 we cannot have $\alpha+\beta = \pi = \varepsilon+\gamma$, as $\alpha > \delta > \varepsilon$ and $\beta > \gamma$. Thus, $\alpha > \frac{\pi}{2} > \delta > \beta > \gamma > \varepsilon$ and $\alpha+\beta+k\varepsilon = \pi$, $k \geq 1$. It is easy to observe that $k = 1$, as $k > 1$ lead to a vertex with a sum of alternate angles including the angles δ, δ and ρ, with $\rho \in \{\alpha, \beta, \delta, \varepsilon\}$, which is not possible due to the dimensions of the involved angles. The last configuration extends to the one illustrated in Figure 16(b). At vertex v_2 we have necessarily one of the following situations:

(i) $\delta+\beta+\beta = \pi$;

(ii) $\delta+\beta+\gamma = \pi$.

Note that $\delta+\beta+k\varepsilon = \pi$, $k > 1$ lead to a vertex with a sum of alternate angles including the angles δ, δ and ρ, with $\rho \in \{\alpha, \beta, \delta, \varepsilon\}$.

(i) If $\delta+\beta+\beta = \pi$, we obtain the configuration illustrated in Figure 17(a). Note that, at vertex v_3, we cannot have $\alpha+\gamma+\gamma+k\rho = \pi$, with $\rho \in \{\gamma, \varepsilon\}$ and $k \geq 1$, otherwise we get $(\alpha+\gamma+\gamma+k\rho)+(\alpha+\delta)+(\delta+\beta+\beta) \geq (\alpha+\beta+\gamma)+(\alpha+\beta+\gamma)+(\delta+\delta+\varepsilon) > 3\pi$, which is not possible.

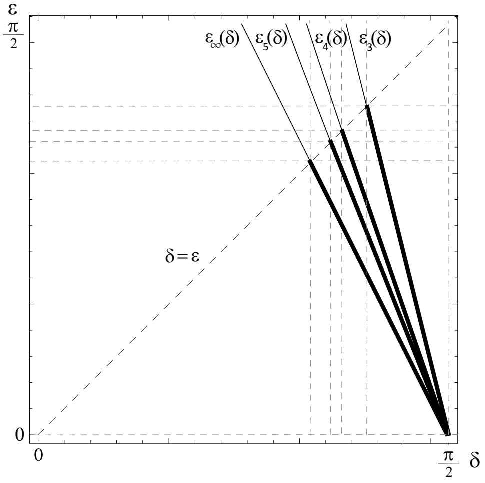


Figure 15: $\varepsilon = \varepsilon_k(\delta)$, with $\delta^k_{\min} < \delta < \frac{\pi}{2}$, and for $k = 3, 4, 5, \ldots, \infty$.

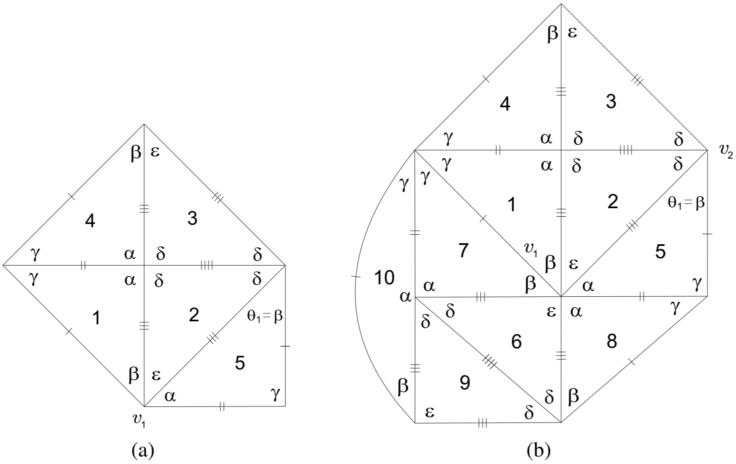


Figure 16: Local configurations.

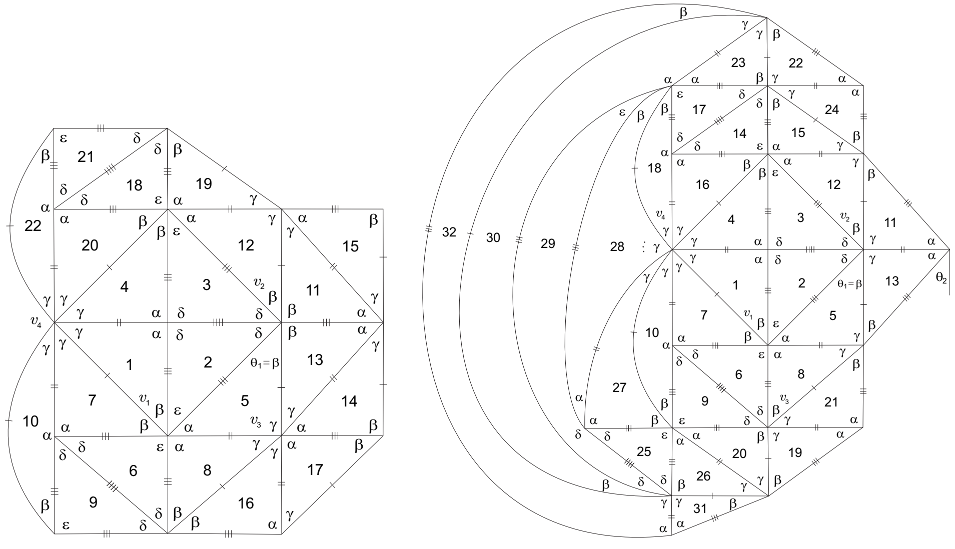
At vertex v_4 we must have $k\gamma = \pi$, with $k \geq 4$. As $\delta = 2\gamma$ and $\pi < \delta + \delta + \varepsilon = 4\gamma + \varepsilon$, we conclude that $k = 4$, which is not possible as $\delta < \frac{\pi}{2}$.

(ii) If $\delta + \beta + \gamma = \pi$, the last configuration gives rise to the one illustrated in Figure 17(b), where θ_2 can be ε or δ. According to the selection for θ_2, we obtain the planar representations illustrated in Figures 18(a) and 18(b), respectively. In the first case we have

$$\alpha + \delta = \pi, \quad \alpha + \beta + \varepsilon = \pi, \quad \delta + \beta + \gamma = \pi, \quad k\gamma = \pi, \text{ with } k \geq 4,$$

and

$$\delta = \delta_k = \operatorname{arccot}\left(\frac{1}{2}\tan\frac{\pi}{2k}\left(2 - \sec^2\frac{\pi}{2k}\right)\right).$$



(a) (b)

Figure 17: Local configurations.

Note that by Equation (3.1) we have

$$\frac{\cos\frac{\pi}{k}+\cos\delta\cos(\delta+\frac{\pi}{k})}{\sin(\delta+\frac{\pi}{k})} = -\frac{\cos\delta\sin\left(\delta+\frac{\pi}{2k}\right)}{\cos\left(\delta+\frac{\pi}{2k}\right)}$$

$$\iff \cos\frac{\pi}{k}\cos\left(\delta+\frac{\pi}{2k}\right)+\cos\delta\cos\frac{\pi}{2k}=0$$

$$\iff 2\cos\delta\cos^3\frac{\pi}{2k}-\sin\delta\cos\frac{\pi}{k}\sin\frac{\pi}{2k}=0$$

$$\iff \cot\delta=\tan\frac{\pi}{2k}-\frac{1}{2}\tan\frac{\pi}{2k}\sec^2\frac{\pi}{2k}.$$

We denote this family of f-tilings by $\mathcal{G}^k$, $k\geq 4$. 3D representations of $\mathcal{G}^k$, $k=4,5,6$, are presented in Figures 5(d) – 5(f).

In the second case we have

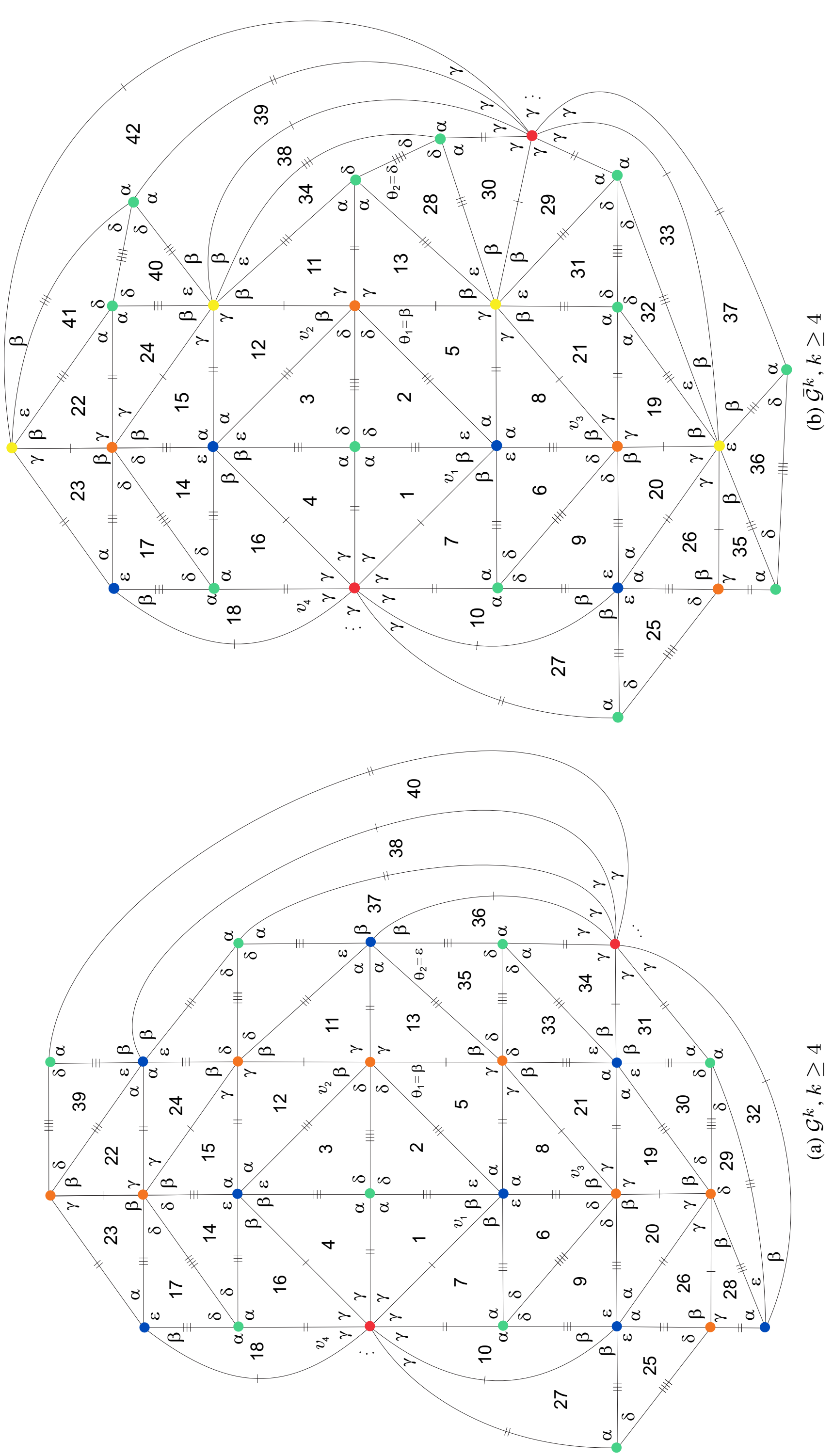
$$\alpha+\delta=\pi,\quad \alpha+\beta+\varepsilon=\pi,\quad \delta+\beta+\gamma=\pi,\quad 2\beta+\gamma+\varepsilon=\pi,$$
$$k\gamma=\pi,\ \text{with } k\geq 4,\quad \text{and } \delta=\delta^k;$$

we denote this family of f-tilings by $\bar{\mathcal{G}}^k$. 3D representations, for $k=4,5,6$, are presented in Figures 6(g) – 6(i).

Case I.1.2.4: If $\theta_1=\alpha$ (Figure 19(a)), we must have $\beta<\delta$, otherwise there is no way to satisfy the angle-folding relation around vertex v_1. Then, $\alpha>\frac{\pi}{2}>\delta>\beta>\gamma$ and $\delta>\varepsilon$. Now, we have

$$\theta_2=\beta,\quad \theta_2=\gamma\quad\text{or}\quad \theta_2=\varepsilon.$$

Note that θ_2 cannot be δ, as $\delta+\beta+\varepsilon+\rho>\pi$, for all $\rho\in\{\beta,\gamma\}$.



(a) $\mathcal{G}^k$, $k \geq 4$

(b) $\bar{\mathcal{G}}^k$, $k \geq 4$

Figure 18: Planar representations.

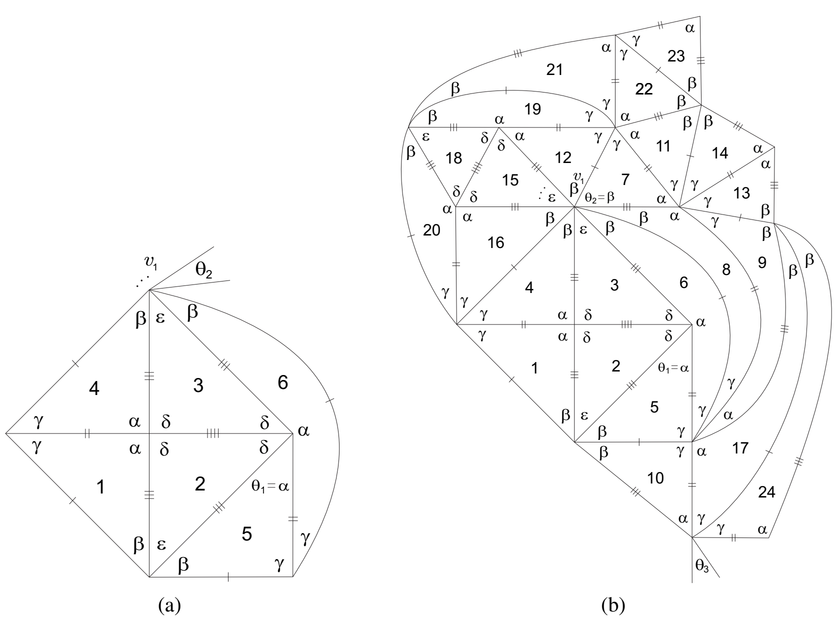


Figure 19: Local configurations.

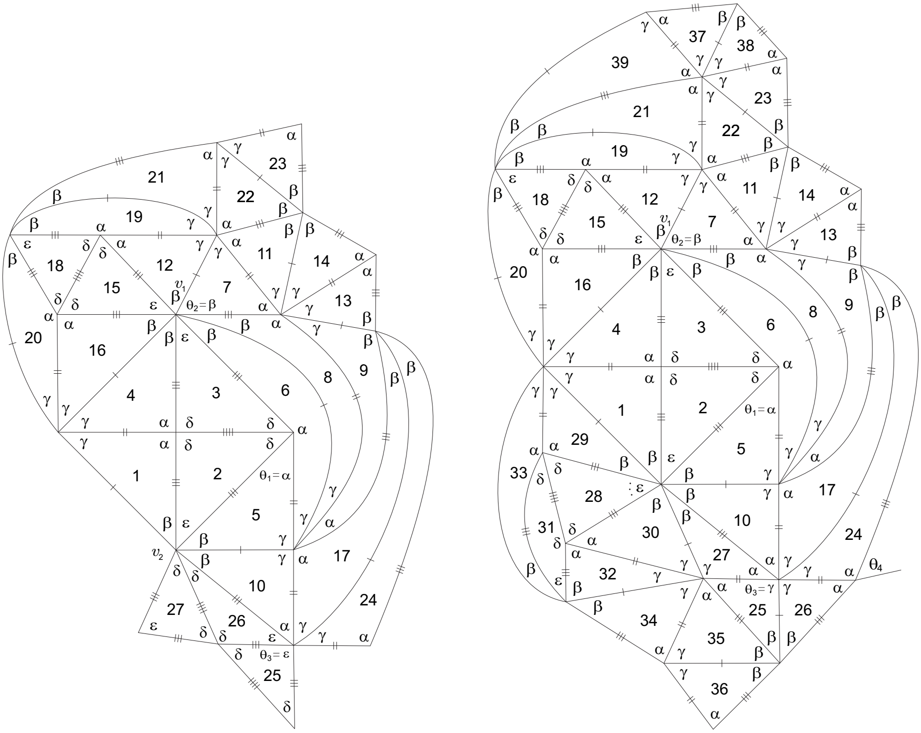
Case I.1.2.4.1: If $\theta_2 = \beta$, we get the configuration illustrated in Figure 19(b), where $\alpha + 2\gamma = \pi$ and, at vertex v_1, $3\beta + k\varepsilon = \pi$, $k \geq 1$. As $k > 1$ implies the existence of a vertex with a sum of alternate angles containing $\delta + \delta + \beta$, and $(3\beta + k\varepsilon) + (2\delta + \beta) + (\alpha + \delta) \geq (\alpha + \beta + \gamma) + 2(2\delta + \varepsilon) > 3\pi$, we conclude that $k = 1$. Now, $\theta_3 \in \{\varepsilon, \gamma\}$. If $\theta_3 = \varepsilon$ (Figure 20(a)), at vertex v_2 we reach a contradiction, as for $\rho \in \{\beta, \gamma\}$, we get $\delta + \beta + \varepsilon + \rho \geq \delta + \beta + \varepsilon + \gamma > 2\delta + \varepsilon > \pi$. On the other hand, if $\theta_3 = \gamma$, the last configuration extends to the one illustrated in Figure 20(b).

If $\theta_4 = \varepsilon$ (Figure 21(a)), at vertex v_3 we must have $\delta + 2\beta = \pi$, as $\delta + 2\beta + \rho > \pi$, for all $\rho \in \{\alpha, \beta, \gamma, \delta, \varepsilon\}$ (note that $\alpha + \beta + \varepsilon = 3\beta + \varepsilon = \pi$, implying $\gamma > \varepsilon$; consequently $\alpha > \frac{\pi}{2} > \delta > \beta > \gamma > \varepsilon$). As $k\gamma = \pi$, $4\gamma = \delta + 2\gamma < \alpha + 2\gamma = \pi$ and $6\gamma = 3\delta > \pi$, we conclude that $k = 5$. Jointly with the remaining conditions, we obtain $\alpha = \frac{3\pi}{5}$, $\beta = \frac{3\pi}{10}$, $\gamma = \frac{\pi}{5}$, $\delta = \frac{2\pi}{5}$ and $\varepsilon = \frac{\pi}{10}$. Nevertheless, under these conditions, Equation (3.1) is impossible. On the other hand, if $\theta_4 = \delta$, we obtain the planar representation illustrated in Figure 21(b). We have

$$\alpha = \frac{3\pi}{5}, \quad \beta = 4 \arctan \sqrt{3 + 4\sqrt{5} - 2\sqrt{22 + 6\sqrt{5}}},$$
$$\gamma = \frac{\pi}{5}, \quad \delta = \frac{2\pi}{5}, \quad \text{and} \quad \varepsilon = \pi - 3\beta.$$

We denote this f-tiling by $\mathcal{H}$, whose 3D representation is presented in Figure 7(j).

Case I.1.2.4.2: If $\theta_2 = \gamma$, we obtain the configuration illustrated in Figure 22(a). Note that, at vertex v_1, all the alternate angle sums containing $\beta + \beta + \gamma + \rho$, with $\rho \in \{\alpha, \beta, \gamma, \delta\}$, exceed π, and so $\beta + \beta + \gamma + k\varepsilon = \pi$, with $k = 1$ ($k \geq 1$ implies the existence of a vertex with alternate sum $\delta + \delta + \beta = \pi$ and $\varepsilon > \beta$).



(a) (b)

Figure 20: Local configurations.

Now, θ_3 must be β or γ.

In the first case (Figure 22(b)), we observe that at vertex v_3 we must have $\delta+\delta+\beta=\pi$, implying at vertex v_4 the existence of an alternate angle sum containing $\alpha+\beta+\gamma>\pi$, which is an impossibility.

On the other hand, if $\theta_3=\gamma$, the last configuration extends to the one illustrated in Figure 23. We denote this family of f-tilings by $\mathcal{F}_{\beta}^{k}$, where

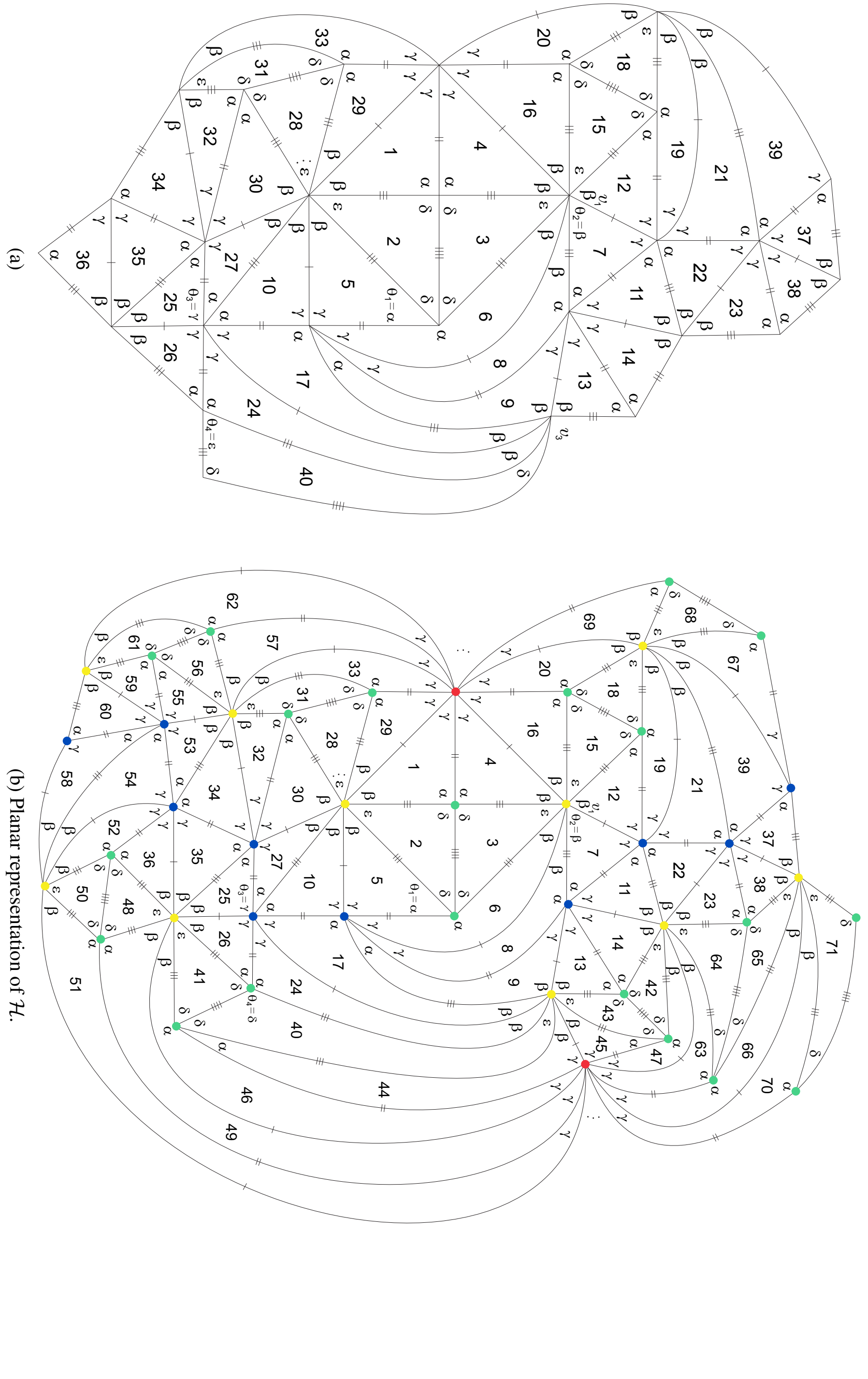
$$\alpha+\delta=\pi, \quad 2\beta+\gamma+\varepsilon=\pi \quad \text{and} \quad k\gamma=\pi, \text{ with } k\geq 4.$$

As $\gamma=\frac{\pi}{k}<\beta<\delta$, $\beta+\gamma>\delta$, using Equation (3.1) we get

$$\begin{aligned}\frac{\cos\frac{\pi}{k}+\cos\alpha\cos\beta}{\sin\beta} &= \frac{-\cos\alpha\sin\left(\beta+\frac{\pi}{2k}\right)}{\cos\left(\beta+\frac{\pi}{2k}\right)}\\ &\iff \cos\frac{\pi}{k}\cos\left(\beta+\frac{\pi}{2k}\right)+\cos\alpha\cos\frac{\pi}{2k}=0\\ &\iff \cos\alpha=-\cos\frac{\pi}{k}+\sec\frac{\pi}{2k}\cos\left(\beta+\frac{\pi}{2k}\right).\end{aligned}$$

Therefore,

$$\alpha=\alpha_k^1(\beta)=\arccos\left(-\cos\frac{\pi}{k}\sec\frac{\pi}{2k}\cos\left(\beta+\frac{\pi}{2k}\right)\right),\ k\geq 4,$$



(a)

(b) Planar representation of $\mathcal{H}$.

Figure 21: Local configurations.

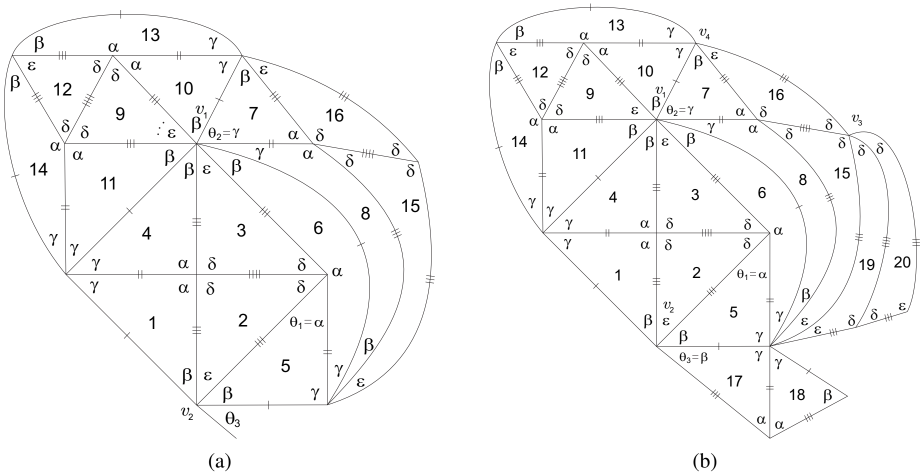


Figure 22: Local configurations.

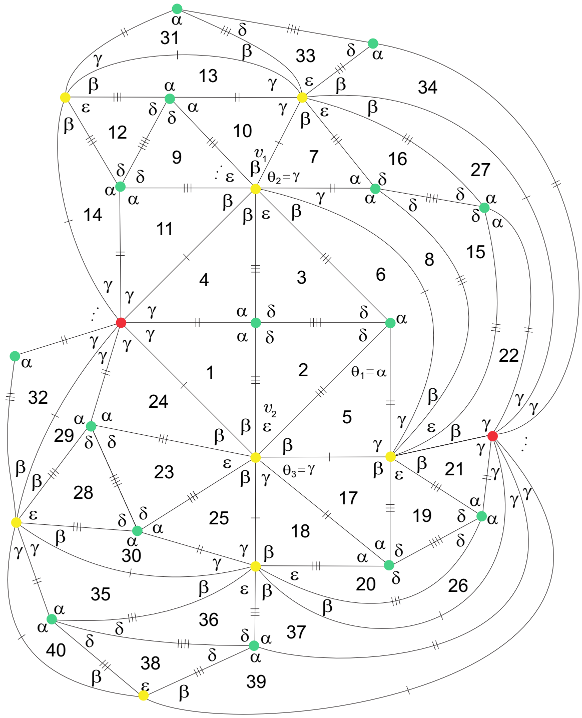


Figure 23: Planar representation of $\mathcal{F}_{\beta}^{k}$.

with $\beta \in \left(\beta^{1k}_{\min}, \beta^{1k}_{\max}\right)$, where

$$\beta^{1k}_{\min} = \max\left\{\frac{\pi}{k}, \arccos\left(\frac{1}{2}\sec\frac{\pi}{2k}\right) - \frac{\pi}{2k}\right\} \quad \text{and} \quad \beta^{1k}_{\max} = \frac{(k-1)\pi}{2k}$$

are obtained, respectively, when $\varepsilon = \gamma$ or $\varepsilon = \delta$ and $\alpha = \delta$. Note that if $\varepsilon = \delta$, we get

$$\begin{aligned}
\cos\left(2\beta + \frac{\pi}{k}\right) &= -\cos\frac{\pi}{k}\sec\frac{\pi}{2k}\cos\left(\beta + \frac{\pi}{2k}\right) \\
\iff \cos\frac{\pi}{2k}\left(2\cos^2\left(\beta + \frac{\pi}{2k}\right) - 1\right) &= -\cos\frac{\pi}{k}\cos\left(\beta + \frac{\pi}{2k}\right) \\
\iff \cos\left(\beta + \frac{\pi}{2k}\right) &= \frac{-\cos\frac{\pi}{k} + \sqrt{\cos^2\frac{\pi}{k} + 8\cos^2\frac{\pi}{2k}}}{4\cos\frac{\pi}{2k}} \\
\iff \cos\left(\beta + \frac{\pi}{2k}\right) &= \frac{-\cos\frac{\pi}{k} + \left(2\cos^2\frac{\pi}{2k} + 1\right)}{4\cos\frac{\pi}{2k}} \\
\iff \cos\left(\beta + \frac{\pi}{2k}\right) &= \frac{1}{2}\sec\frac{\pi}{2k}.
\end{aligned}$$

The graph of $\alpha = \alpha^1_k(\beta)$, for $\beta^{1k}_{\min} < \beta < \beta^{1k}_{\max}$, is outlined in Figure 24, for different values of k. Note that the condition $\varepsilon < \delta$ is equivalent to $\alpha < 2\beta + \frac{\pi}{k}$.

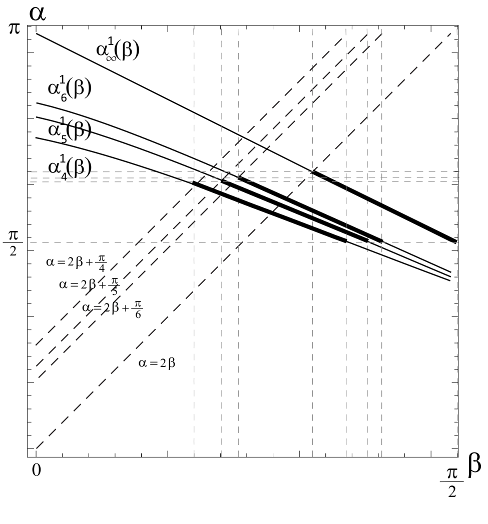


Figure 24: $\alpha = \alpha^1_k(\beta)$, with $\beta^{1k}_{\min} < \beta < \beta^{1k}_{\max}$, and for and for $k = 4, 5, 6, \ldots, \infty$.

3D representations of $\mathcal{F}^4_\beta$, $\mathcal{F}^5_\beta$ and $\mathcal{F}^6_\beta$ are given in Figures 8(k) – 8(m).

Case I.1.2.4.3: If $\theta_2 = \varepsilon$ (Figure 19(a)), at vertex v_1 we must have

(i) $\beta + \beta + k\varepsilon = \pi$, $k \geq 1$,

(ii) $\beta + \beta + \beta + k\varepsilon = \pi$, $k \geq 1$ or

(iii) $\beta + \beta + \gamma + k\varepsilon = \pi$, $k \geq 1$.

Note that in all these cases k must be one, otherwise we reach a vertex with alternate sum $\delta+\delta+\beta=\pi$ and other vertex surrounded in cyclic order by $(\alpha,\varepsilon,\beta,\ldots)$, which is not possible.

In case (i), $\beta+\beta+\varepsilon=\pi$, we obtain the planar representation of Figure 25. We denote

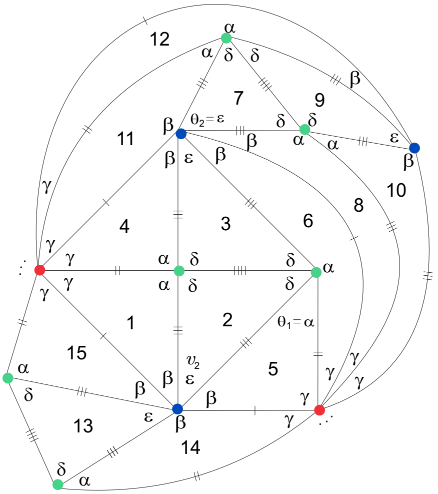


Figure 25: Planar representation of $\mathcal{I}_\beta^k$.

this family of f-tilings by $\mathcal{I}_\beta^k$, where $\alpha+\delta=\pi$, $2\beta+\varepsilon=\pi$ and $k\gamma=\pi$, with $k\geq 3$. Using Equation (3.1), we get

$$\alpha=\alpha_k^2(\beta)=\arccos\left(-\cos\frac{\pi}{k}\cos\beta\right),\ k\geq 3,$$

with

$$\max\left\{\frac{\pi}{k},\arccos\frac{\sqrt{\cos^2\frac{\pi}{k}+8}-\cos\frac{\pi}{k}}{4}\right\}<\beta<\frac{\pi}{2},$$

where the lower and upper bounds are obtained, respectively, when $\varepsilon=\gamma$ or $\varepsilon=\delta$ and $\alpha=\delta$. The graph of this function is outlined in Figure 26, for different values of k. Note that the condition $\varepsilon<\delta$ is equivalent to $\alpha<2\beta$.

3D representations of $\mathcal{I}_\beta^k$, for $k=3,4,5$, are illustrated in Figures 9(n) – 9(p).

In case (ii), $\beta+\beta+\beta+\varepsilon=\pi$, using similar arguments applied before, the local configuration extends to the f-tiling $\mathcal{H}$, obtained in Case I.1.2.4.1.

In the last case, by symmetry we obtain the families of f-tilings $\mathcal{F}_\beta^k$ and $\bar{\mathcal{G}}^k$, $k\geq 4$, of Cases I.1.2.4.2 and I.1.2.3(ii), respectively.

Case I.2: With the labeling of Figure 27(a), at vertex v_1 we must have

$$\beta+\delta=\pi \quad \text{or} \quad \beta+\delta<\pi.$$

Case I.2.1: Suppose firstly that $\beta+\delta=\pi$. As $\delta=\beta=\frac{\pi}{2}$ implies $\gamma=\frac{\pi}{2}$, we have $\delta\neq\beta$. If $\delta>\beta$, by Equation (3.1), we conclude that $\alpha>\frac{\pi}{2}$, preventing a feasible assignment for θ_1 and θ_2. In turn, if $\delta<\beta$, we obtain a vertex (v_2) surrounded by four angles δ. As $2\delta<\beta+\delta=\pi$, we would have $2\delta+\rho\leq\pi$, with $\rho\in\{\alpha,\beta,\gamma,\delta,\varepsilon\}$, which is not possible.

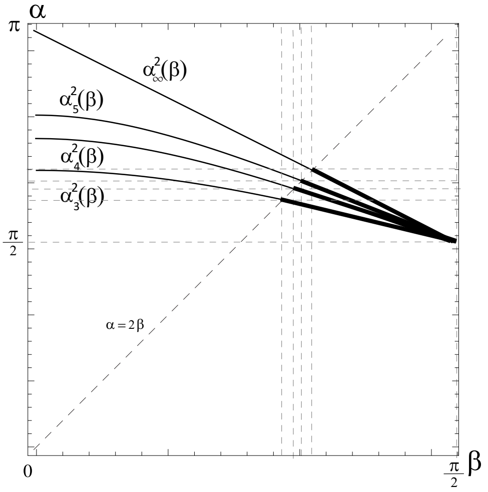


Figure 26: $\alpha = \alpha_k^2(\beta)$, with $\beta_{\min}^{2k} < \beta < \frac{\pi}{2}$, and for $k = 3, 4, 5, \ldots, \infty$.

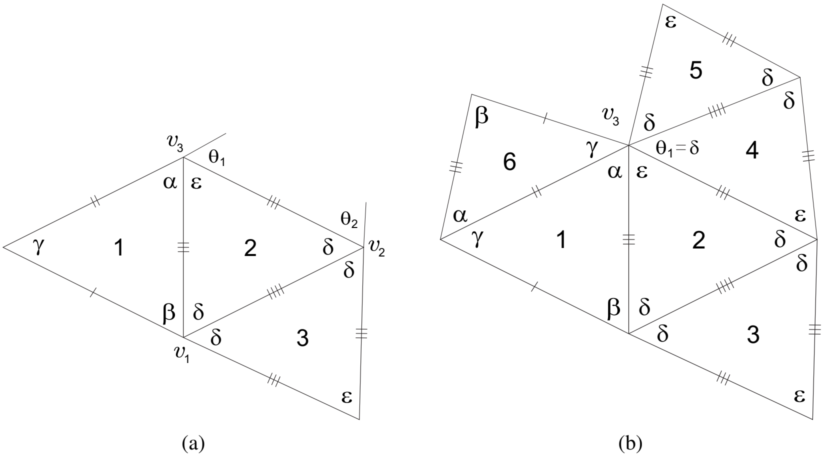


Figure 27: Local configurations.

Case I.2.2: Suppose now that $\beta + \delta < \pi$. As in Case I.2.1, if $\delta \geq \frac{\pi}{2}$, we obtain $\alpha > \frac{\pi}{2}$ and no assignment for θ_1 and θ_2 is possible. Thus, $\delta < \frac{\pi}{2}$ and, as any tiling has necessarily vertices of valency four, we have $\alpha \geq \frac{\pi}{2}$. Now, observing Figure 27(a), we have $\theta_1 \in \{\delta, \varepsilon, \beta\}$.

Case I.2.2.1: If $\theta_1 = \delta$, we obtain the configuration illustrated in Figure 27(b). Vertex v_3 must have valency three, but in this case we get $\alpha + \beta + \delta = \pi = \delta + \varepsilon + \gamma$, implying $\varepsilon > \alpha$, which is not possible.

Case I.2.2.2: If $\theta_1 = \varepsilon$, the last configuration extends uniquely to the one illustrated in Figure 28. Note that at vertex v_4, θ_2 must be β and the vertex must have valency three

($2\beta > \beta+\gamma > \delta$). We denote this family of f-tilings by $\mathcal{J}_{\beta}^{k}$, where $\alpha+\varepsilon=\pi$, $2\delta+\beta=\pi$

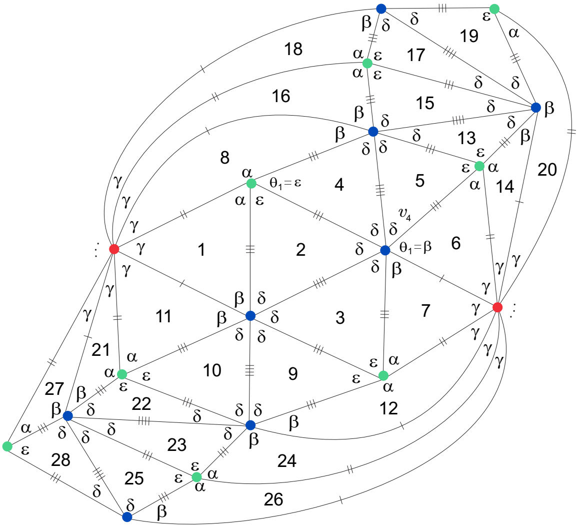


Figure 28: Planar representation of $\mathcal{J}_{\beta}^{k}$.

and $k\gamma=\pi$, with $k\geq 4$. Using Equation (3.1) we get

$$\frac{\cos\frac{\pi}{k}+\cos\alpha\cos\beta}{2\sin\frac{\beta}{2}}=\sin\frac{\beta}{2}(1-\cos\alpha)$$

$$\iff \cos\frac{\pi}{k}+\cos\alpha\left(2\cos^2\frac{\beta}{2}-1\right)=2\left(1-\cos^2\frac{\beta}{2}\right)(1-\cos\alpha)$$

$$\iff \cos\alpha=2\sin^2\frac{\beta}{2}-\cos\frac{\pi}{k}.$$

Therefore,

$$\alpha=\alpha_k^3(\beta)=\arccos\left(2\sin^2\frac{\beta}{2}-\cos\frac{\pi}{k}\right),\ k\geq 4,$$

with

$$\frac{\pi}{k}<\beta<2\arcsin\frac{\sqrt{1+8\cos\frac{\pi}{k}}-1}{4},$$

where the lower and upper bounds are obtained, respectively, when $\beta=\gamma$ and $\varepsilon=\delta$. The graph of this function is outlined in Figure 29, for different values of k.

3D representations of $\mathcal{J}_{\beta}^{k}$, for $k=4,5,6$, are illustrated in Figures 10(q) – 10(s).

Case I.2.2.3: Finally, if $\theta_1=\beta$, at vertex v_3 (see Figure 27(a)) we have $\alpha+\beta\leq\pi$. $\alpha+\beta=\pi=\varepsilon+\gamma$ implies $\varepsilon>\alpha>\frac{\pi}{2}>\delta$, which is a contradiction. As any tiling has necessarily vertices of valency four, we conclude that $\alpha+\beta+k\varepsilon=\pi$, $k\geq 1$, and $\alpha+\delta=\pi$ at vertex v_2, as illustrated in Figure 30, configuration coincident with the one presented in Figure 16(b), which leads to the families of f-tilings $\mathcal{G}^k$ and $\bar{\mathcal{G}}^k$ (Case I.1). □

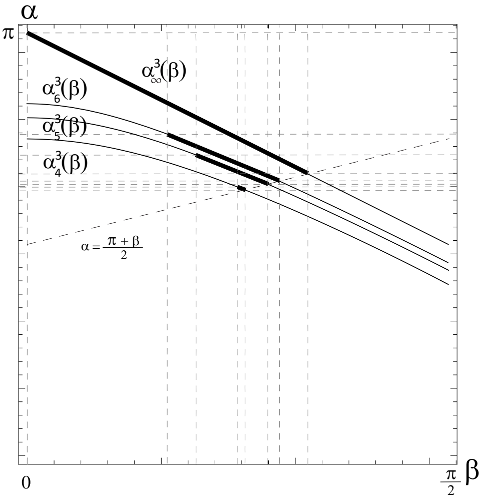


Figure 29: $\alpha = \alpha^3_k(\beta)$, with $\frac{\pi}{k} < \beta < \beta^{3k}_{\max}$, and for $k = 4, 5, 6, \dots, \infty$.

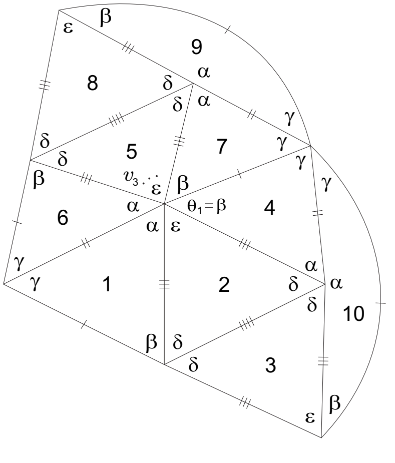


Figure 30: Local configuration.

References

[1] A. M. R. Azevédo Breda, A class of tilings of S^2, *Geom. Dedicata* **44** (1992), 241–253, doi:10.1007/bf00181393.

[2] A. Breda, R. Dawson and P. Ribeiro, Spherical f-tilings by two noncongruent classes of isosceles triangles – II, *Acta Math. Sin. (English Series)* **30** (2014), 1435–1464, doi:10.1007/s10114-014-3302-5.

[3] R. J. M. Dawson, Tilings of the sphere with isosceles triangles, *Discrete Comput. Geom.* **30** (2003), 467–487, doi:10.1007/s00454-003-2846-4.

[4] R. J. M. Dawson and B. Doyle, Tilings of the sphere with right triangles I: The asymptotically right families, *Electron. J. Combin.* **13** (2006), #R48, `http://www.combinatorics.org/ojs/index.php/eljc/article/view/v13i1r48`.

[5] R. J. M. Dawson and B. Doyle, Tilings of the sphere with right triangles II: The $(1, 3, 2)$, $(0, 2, n)$ subfamily, *Electron. J. Combin.* **13** (2006), #R49, `http://www.combinatorics.org/ojs/index.php/eljc/article/view/v13i1r49`.

[6] A. M. d'Azevedo Breda and A. F. Santos, Dihedral f-tilings of the sphere by spherical triangles and equiangular well-centered quadrangles, *Beiträge Algebra Geom.* **45** (2004), 447–461, `https://www.emis.de/journals/BAG/vol.45/no.2/8.html`.

[7] A. M. R. d'Azevedo Breda and P. dos Santos Ribeiro, Spherical f-tilings by two non congruent classes of isosceles triangles – I, *Math. Commun.* **17** (2012), 127–149, `https://hrcak.srce.hr/82991`.

[8] S. A. Robertson, Isometric folding of Riemannian manifolds, *Proc. Roy. Soc. Edinburgh Sect. A* **79** (1978), 275–284, doi:10.1017/s0308210500019788.

[9] Y. Ueno and Y. Agaoka, Classification of tilings of the 2-dimensional sphere by congruent triangles, *Hiroshima Math. J.* **32** (2002), 463–540, `http://projecteuclid.org/euclid.hmj/1151007492`.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 445–463
https://doi.org/10.26493/1855-3974.1679.ad3
(Also available at http://amc-journal.eu)

Ascending runs in permutations and valued Dyck paths*

Marilena Barnabei [†], Flavio Bonetti ,
Niccolò Castronuovo , Matteo Silimbani

Dipartimento di Matematica, Università di Bologna, Bologna, 40126, Italy

Received 18 April 2018, accepted 2 October 2018, published online 11 February 2019

Abstract

We define a bijection between permutations and valued Dyck paths, namely, Dyck paths whose odd vertices are labelled with an integer that does not exceed their height. This map allows us to characterize the set of permutations avoiding the pattern 132 as the preimage of the set of Dyck paths with minimal labeling. Moreover, exploiting this bijection we associate to the set of n-permutations a polynomial that generalizes at the same time Eulerian polynomials, Motzkin numbers, super-Catalan numbers, little Schröder numbers, and other combinatorial sequences. Lastly, we determine the Hankel transform of the sequence of such polynomials.

Keywords: Permutation, Dyck path, pattern avoidance, Hankel transform.

Math. Subj. Class.: 05A05, 05A15, 05A19

1 Introduction

Many bijections are present in the literature between the symmetric group $\mathbf{S}_n$ and the set of Dyck paths of semilength n with some kind of labeling on their steps (see e.g. [3, 8, 18]). In this paper, inspired by [3], we define a bijection Γ between permutations and valued Dyck paths, namely, Dyck paths whose odd vertices are labelled with an integer that does not exceed their height.

More precisely, we write a permutation π as the juxtaposition of ascending runs, and associate to every integer i from 1 to n a pair of consecutive steps in the path according

[*] This work was partially supported by the University of Bologna, funds for selected research topics.

[†] Corresponding author.

E-mail addresses: marilena.barnabei@unibo.it (Marilena Barnabei), flavio.bonetti@unibo.it (Flavio Bonetti), niccolo.castronuovo2@unibo.it (Niccolò Castronuovo), matteo.silimbani4@unibo.it (Matteo Silimbani)

to the fact that i is the unique element of an ascending run (a *head-tail*) in π or the initial (*head*), final (*tail*) or middle element (*boarder*) of an ascending run of length greater or equal to two. Every pair of consecutive steps is labelled with an integer that depends on the respective position of the ascending runs in π.

Observe that a similar construction was described in [11] in terms of peaks, valleys, double descents and double rises of the permutation. Given a permutation $\pi = \pi_1\pi_2\dots\pi_n$, it turns out that for $i \neq 1, n$ the entry π_i is

(i) a head if and only if it is a valley;

(ii) a tail if an only if it is a peak;

(iii) a head-tail if and only if it is double descent;

(iv) a boarder if and only if it is a double rise.

However, π_1 and π_n may play different roles in the two environments, and this fact leads to different results. The present construction seems to shed new light on combinatorial properties of permutations. In particular the results of Section 5 seem to be difficult to obtain with the construction in [11].

The map Γ allows us to characterize the set of permutations avoiding the pattern 132 (213, resp.) as the preimage of the set of Dyck paths with minimal (maximal, respectively) labeling. In these particular cases it is possible to translate the ascending runs of the permutation directly in terms of tunnels of the Dyck path. As a consequence we get a bijection between the permutations avoiding 132 and those avoiding 213 that is new, up to our knowledge.

If a permutation avoids 132 its ascending runs are the blocks of a non-crossing partition. Hence our map provides also a bijection between Dyck paths and non-crossing partitions, that turns out to be the same as the bijection introduced in [24].

In Section 5 we consider monomials in the variables H, S, B associated with each permutation according to the number of heads, head-tails and boarders. In this way we construct a polynomial $F_n(H, S, B)$ as the sum of such monomials over all the permutations of length n. These polynomials generalize at the same time Eulerian numbers, factorials and many other sequences. We exploit the results of the previous sections to deduce a recurrence relation for these polynomials and a functional equation for their generating function. We also determine the Hankel transform of the sequence $(F_n)_{n\geq 0}$, hence obtaining both new and known results about the Hankel transform of various specializations of these polynomials. Finally, we consider the sequence of polynomials $\widehat{F}_n(H, S, B)$, defined as the sum of the monomials that correspond to permutations avoiding 132. These polynomials specialize in many well-known sequences related to Catalan and Motzkin numbers.

2 The bijection

A *Dyck path* of semilength n is a lattice path contained in $\mathbb{N} \times \mathbb{N}$, starting in $(0, 0)$, ending in $(2n, 0)$, consisting of unitary north-east steps of the form $(1, 1)$ and of unitary south-east steps of the form $(1, -1)$ and lying above the x-axis. The north-east steps are called *up steps* (denoted by U) and the south-east steps are called *down steps* (denoted by D).

As usual, a Dyck path can be identified with a word $w = S_1S_2\dots S_{2n}$ of length $2n$ in the alphabet $\{U, D\}$ with the constraint that the number of occurrences of the letter U is equal to the number of occurrences of the letter D and, for every i, the number of occurrences of U in the subword $S_1 S_2 \dots S_i$ is not smaller than the number of occurrences of D. The word w is called a *Dyck word.* In the following we will not distinguish between a Dyck path and the corresponding word.

We denote by $\mathcal{D}_n$ the set of Dyck path of semilength n.

Given a Dyck path $d \in \mathcal{D}_n$, decompose it into 2-step subpaths $d = d_1 d_2 \dots d_n$. The subpaths d_i will be called *dimers* and the decomposition $d = d_1 \dots d_n$ will be called the *dimer decomposition* of d. For every $i = 1, \dots, n$, let k_i be the y-coordinate of the middle point of the dimer d_i. We associate to d the n-tuple $m(d) = (m_1, m_2, \dots, m_n)$, where $m_i = \frac{k_i - 1}{2}$. We will call the integer m_i the *height* of the dimer d_i, and the n-tuple $m(d)$ the *height list* of d.

Example 2.1. Consider the Dyck path $d = UU|UU|DD|UD|DD$ in Figure 1. Then

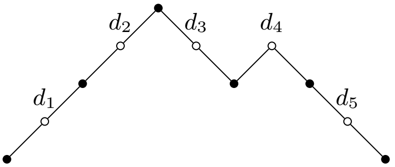


Figure 1: The Dyck path $d = UU|UU|DD|UD|DD$.

$m(d) = (0, 1, 1, 1, 0)$.

Let $\pi = \pi_1 \pi_2 \dots \pi_n$ be a permutation in $\mathbf{S}_n$ written in one-line notation.

An *ascending run* in π is a maximal increasing subsequence of π. For example, the ascending runs of 346512 are $w_1 = 346$, $w_2 = 5$ and $w_3 = 12$. Write π as

$$\pi = w_1 w_2 \dots w_k,$$

where the w_i's are the ascending runs in π. Let h_i and t_i be the first and the last element of w_i. Note that h_i and t_i can coincide. We call h_i and t_i the *head* and the *tail* of w_i. Clearly $t_i > h_{i+1}$ for $1 \leq i \leq k - 1$.

Now we associate to every permutation of length n a Dyck path d of semilength n defined as follows. For $i = 1, \dots, n$,

(i) if i is both a head and a tail, set $d_i = UD$;

(ii) if i is a head but not a tail, set $d_i = UU$;

(iii) if i is a tail but not a head, set $d_i = DD$;

(iv) if i is neither a head nor a tail, set $d_i = DU$.

Then $d = d_1 d_2 \dots d_n$.

Obviously the correspondence $\gamma \colon \pi \to d$ is far from being injective. For example, both the permutations 3124 and 1243 in $\mathbf{S}_4$ correspond to the Dyck path $UUDUUDDD$.

In order to get a bijection, we associate to the permutation π a *valued Dyck path*, namely a pair (d, l), where d is the Dyck path defined above and $l = (l_1, l_2, \dots, l_n)$ is the sequence of non-negative integers given by

$$l_i = |\{j \mid h_j < i < t_j,\ t_j \text{ precedes } i \text{ in } \pi\}|.$$

Example 2.2. Consider the permutation $\pi = 1254367$. The ascending runs of π are $w_1 = 125$, $w_2 = 4$ and $w_3 = 367$. The heads and the tails of π are $h_1 = 1$, $h_2 = 4$, $h_3 = 3$, $t_1 = 5$, $t_2 = 4$, and $t_3 = 7$. The Dyck path associated with π is $d = UU|DU|UU|UD|DD|DU|DD$ (in Figure 2). and the list l associated with the permutation π is $(0, 0, 1, 1, 0, 0, 0)$.

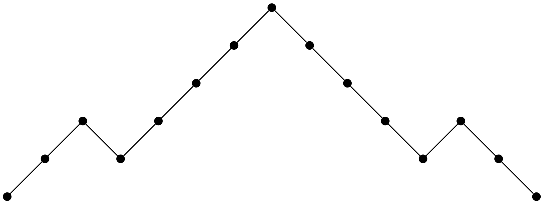


Figure 2: The Dyck path $d = UU|DU|UU|UD|DD|DU|DD$.

We denote by $\Gamma(\pi)$ the valued Dyck path associated with the permutation π. The next proposition describes the connection between the list l associated with π and the height list $m(d)$.

Proposition 2.3. *Let π be a permutation in $\mathbf{S}_n$. Set $\Gamma(\pi) = (d, l)$, with $l = (l_1, \dots, l_n)$. Let $m(d) = (m_1, \dots, m_n)$ be the height list of d. Then, for all $1 \leq i \leq n$,*

$$l_i \leq m_i.$$

Proof. Let i be an integer such that $1 \leq i \leq n$. If $d = d_1 d_2 \dots d_n$ is the dimer decomposition of d, the integer m_i can be written as

$$m_i = |\{j \mid d_j = UU,\, 0 < j < i\}| - |\{j \mid d_j = DD,\, 0 < j < i\}| - \epsilon$$

where

$$\epsilon = \begin{cases} 0 & \text{if the first step of } d_i \text{ is an up step;} \\ 1 & \text{if the first step of } d_i \text{ is a down step.} \end{cases}$$

We notice that, denoting by h_j and t_j the j-th head and tail of π, respectively, the integer $|\{j \mid d_j = UU,\, 0 < j < i\}|$ is the number of heads h_j such that $h_j < i$, while $|\{j \mid d_j = DD,\, 0 < j < i\}|$ is the number of tails t_j such that $t_j < i$. Hence

$$\begin{aligned} m_i &= |\{j \mid h_j < i\}| - |\{j \mid t_j < i\}| - \epsilon \\ &= |\{j \mid h_j < i \leq t_j\}| - \epsilon \\ &= |\{j \mid h_j \leq i \leq t_j\}| - 1. \end{aligned}$$

The assertion now follows immediately from the definition of l_i. □

The above proposition shows that the image of the map Γ is contained in the set of all pairs (d, l), where d is a Dyck path of semilength n and $l = (l_1, \dots, l_n)$ is a sequence of positive integers such that, for all $1 \leq i \leq n$, $l_i \leq m_i$, the i-th element of the height list of d. We denote by $\mathcal{DL}_n$ the set of such pairs. Our next goal is to prove that the map Γ is actually a bijection between $\mathbf{S}_n$ and $\mathcal{DL}_n$.

To this aim we describe a procedure whose iteration will be proved to provide the inverse of Γ. In order to describe such a procedure we need the notion of tagged Dyck path.

A *tagged Dyck path* is a pair $[d, \lambda]$ where d is a Dyck path of semilength n and $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$ is an increasing sequence of positive integers. Intuitively, we think of the tag λ_i, $1 \leq i \leq n$, as attached to the dimer d_i.

Example 2.4. See Figure 3.

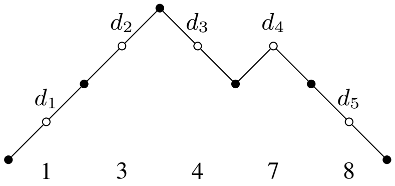


Figure 3: A tagged Dyck path with tags $\lambda = (1, 3, 4, 7, 8)$.

Now we describe a procedure P whose input is a triple (d, λ, l) where $[d, \lambda]$ is a tagged Dyck path of semilength r and $(d, l) \in \mathcal{DL}_r$. P produces another triple (d', λ', l') with the same properties, with d' a Dyck path of smaller semilength.

Let $d = d_1 d_2 \dots d_r$ be the dimer decomposition of d. Set $l = (l_1, l_2, \dots, l_r)$ and $\lambda = (\lambda_1, \dots, \lambda_r)$.

(I) If d is the empty path, end.

(II) Else, find the smallest index i such that $l_i = m_i$ and the second step of d_i is a down step (such an i exists since the integer r satisfies the conditions above).

(II.a) If the first step of d_i is an up step, the output of P is the triple (d', λ', l'), where

- d' is obtained from d by removing d_i,
- l' is obtained from l by removing l_i,
- λ' is obtained from λ by removing λ_i.

Note that $m(d')$ is obtained from $m(d)$ by removing m_i, so, for all i, the i-th element of l' is not greater than the i-th element of m', hence $(d', l') \in \mathcal{DL}_{r-1}$.

(II.b) Otherwise, follow the path d backwards, starting from d_i, until you find a dimer d_j, $j < i$, such that $d_j = UU$ and $l_j = m_j$. Let $\{j_1, j_2, \dots, j_k\}$ be the set of indices j_q such that $j < j_q < i$ and $l_{j_q} = m_{j_q}$. Then, the output of P is the triple (d', λ', l'), where

- d' is obtained from d by removing the dimers $d_j, d_{j_1}, \dots, d_{j_k}, d_i$,
- l' is obtained from l by removing the entries $l_j, l_{j_1}, \dots, l_{j_k}, l_i$,
- λ' is obtained from λ by removing the entries $\lambda_j, \lambda_{j_1}, \dots, \lambda_{j_k}, \lambda_i$.

It is easily seen that for all i, the i-th element of l' is not greater than the i-th element of $m(d')$, hence $(d', l') \in \mathcal{DL}_{r-k-2}$.

Let now $(d, l) \in \mathcal{DL}_n$. We apply t times, say, the procedure P starting from the triple $(d, (1, \dots, n), l)$, until we get the empty path. At the i-th step, denote by w_i the list of symbols which have been removed from the set of tags, written in increasing order.

Let $\Lambda(d,l)$ be the permutation π obtained by juxtaposing the lists $w_t w_{t-1} \dots w_1$. It remains to show that $w_t, \dots, w_1$ are precisely the ascending runs of the permutation π, namely, that the last element of w_{i+1} is greater than the first element of w_i. For the sake of simplicity, we prove this fact for w_1 and w_2. Denote by k_1 the first element of w_1, k_2 the last element of w_1 (i.e., the first and the last tag removed at step 1) and h the last element of w_2 (i.e., the last tag removed in step 2). Note that $l_h = m_h$. Since the index h has not been chosen at step 1, we must have $h > k_2 > k_1$. A similar argument can be used for the general case.

Example 2.5. Consider the path $d = UU|UD|UU|DD|DU|DD|UU|DD$ (see Figure 4). Then, $m(d) = (0,1,1,1,0,0,0,0)$. In order to construct $\Lambda(d,(0,0,0,1,0,0,0,0))$, we apply the procedure P iteratively, starting with the triple

$$(d,(1,2,3,4,5,6,7,8),(0,0,0,1,0,0,0,0)).$$

We represent the triple (d,λ,l) as a Dyck path with the tags written below the path, while each integer l_i is placed close to the respective dimer d_i. A dashed line is drawn for every height m_i. The white dots represent the vertices involved in the application of P.

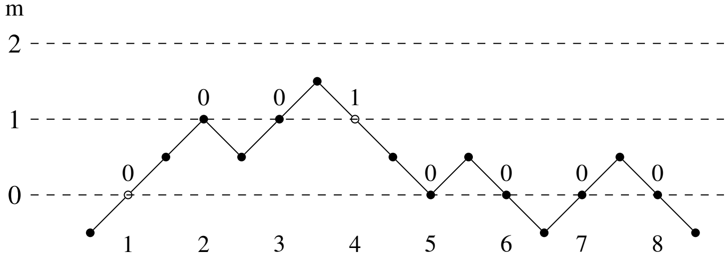


Figure 4: The Dyck path $d = UU|UD|UU|DD|DU|DD|UU|DD$.

At the first application of P we have $i = 4$, $d_4 = DD$, $j = 1$, $k = 0$, $\lambda_1 = 1$ and $\lambda_4 = 4$. The last ascending run of the permutation π is $w_1 = 14$ and we get the new triple (see Figure 5):

$$(UDUUDUDDUUDD,(2,3,5,6,7,8),(0,0,0,0,0,0)).$$

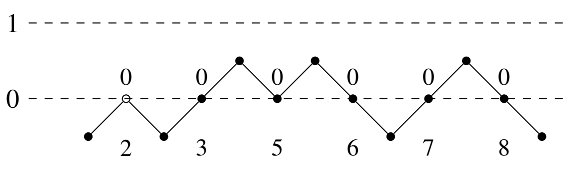


Figure 5: Output of the first application of P.

At the second application of P, $i = 1$, $d_1 = UD$, and $\lambda_1 = 2$, hence the permutation π ends now by $w_2w_1 = 214$. The output is now the triple (see Figure 6):

$$(UUDUDDUUDD, (3, 5, 6, 7, 8), (0, 0, 0, 0, 0)).$$

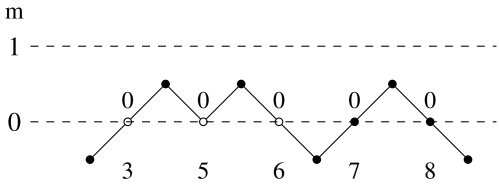


Figure 6: Output of the second application of P.

At the third step, $i = 3$, $d_3 = DD$, $j = 1$, $k = 1$, $j_1 = 2$, $\lambda_1 = 3$, $\lambda_2 = 5$, $\lambda_3 = 6$, so the permutation π ends now by $w_3w_2w_1 = 356214$ and we get the new triple (see Figure 7):

$$(UUDD, (7, 8), (0, 0)).$$

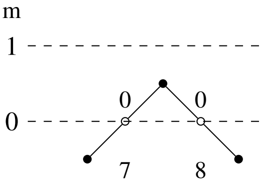


Figure 7: Output of the third application of P.

At this step we have $i = 2$, $d_2 = DD$, $j = 1$, $k = 0$, $\lambda_1 = 7$, $\lambda_2 = 8$. Since the application of P produces now the empty triple, the permutation $\Lambda(d, l)$ is

$$\pi = 78356214.$$

The following result assures that the maps Γ and Λ are inverse of each other. The proof is based on an alternative description of the map Γ. Although this description is more complicated than the previous one, it is the most suitable for that purpose.

Theorem 2.6. *Let π be a permutation of length n. Then*

$$\Lambda(\Gamma(\pi)) = \pi.$$

Moreover, let (d, l) be an element of $\mathcal{DL}_n$. Then

$$\Gamma(\Lambda(d, l)) = (d, l).$$

Proof. The map Γ can be described as the result of the iteration of the following procedure Q.

The procedure Q takes as input a triple (d, λ, l) and an increasing sequence of numbers w where, as usual, $[d, \lambda]$ is a tagged Dyck path of semilength r, $(d, l) \in \mathcal{DL}_r$ and the elements of w are different from the elements of λ. Q produces a triple (d', λ', l') with the same properties, with d' a Dyck path of greater semilength.

Set $\lambda = (\lambda_1, \dots, \lambda_r)$, $l = (l_1, \dots, l_r)$, $w = x_1 \dots x_k$, and let $d = d_1 \dots d_r$, be the dimer decomposition of d. Hence, λ_i is the tag of d_i. Then

(I) The new list of tags λ' is the union of λ with the elements of w.

(II) For $i = 1, \dots, k$ define a new dimer $\widehat{d_i}$ as follows.

$$\widehat{d_i} = \begin{cases} UD & \text{if } i = 1 = k \\ UU & \text{if } i = 1 < k \\ DU & \text{if } 1 < i < k \\ DD & \text{if } 1 < i = k. \end{cases}$$

Tag the dimer $\widehat{d_i}$ with the symbol x_i. Set

$$\{e_1, e_2, \dots, e_{k+r}\} = \{d_1, \dots, d_r\} \cup \{\widehat{d_1}, \dots, \widehat{d_k}\}.$$

Then the path d' is $e_{i_1} e_{i_2} \dots e_{i_{k+r}}$, written in increasing order of the corresponding tags.

Roughly speaking, the new path is obtained by interlacing the new dimers with the dimers of d, following the increasing order of the tags in λ'.

(III) Set $m(d') = (m'_1, m'_2, \dots, m'_{r+k})$. The sequence l' is $(l'_1, \dots, l'_{r+k})$, where $l'_i = l_{t_i}$ if i is an index corresponding to a symbol in λ, and $l'_i = m'_i$ if i is an index corresponding to a symbol in w.

For every permutation $\pi \in \mathbf{S}_n$, let $\pi = w_1 \dots w_k$ be the decomposition of π into ascending runs. The pair $\Gamma(\pi)$ is obtained by applying k times the procedure Q, starting from the empty triple, and using the increasing sequence w_i at the i-th application of Q, $1 \le i \le k$.

It is easily seen that the procedures P and Q are inverse of each other. □

As a consequence of the preceding theorem we have that the map Γ is a bijection between the set of permutations of length n and the set $\mathcal{DL}_n$. Hence

$$|\mathcal{DL}_n| = n!.$$

3 Properties of the map Γ

Given a permutation $\sigma \in \mathbf{S}_n$, one can partition the set $\{1, 2, \dots, n\}$ into intervals $A_1, \dots, A_t$ so that $\sigma(A_i) = A_i$ for every i. The restrictions of σ to the intervals in the finest of these decompositions are called *connected components* of σ. A permutation σ with a single connected component is called *connected.* The *right connected components* of $\sigma = \sigma_1 \sigma_2 \dots \sigma_n$ are the connected components of the reverse $R(\sigma) = \sigma_n \sigma_{n-1} \dots \sigma_1$

of σ. A permutation is said to be *right connected* if $R(\sigma)$ is connected. As an example, the right connected components of the permutation 45132 are 45 and 132 while the permutation 2314 is right connected.

The function Γ maps right connected permutations to irreducible Dyck paths.

We recall that a *return* of a Dyck path d is a down step whose ending point lies on the x-axis. An *irreducible Dyck path* is a Dyck path whose only return is its last step. Every Dyck path d can be uniquely written as $d = p_1p_2\dots p_k$, where each p_i is an irreducible Dyck path.

Proposition 3.1. *Let π be a permutation in $\mathbf{S}_n$ and let $\pi = u_1u_2\dots u_k$ be the decomposition of π into right connected components. Let $\Gamma(\pi) = (d, l)$. Then d decomposes into irreducible components $d_k \dots d_1$, where d_i is the path corresponding to the normalization of u_i.*

Proof. The assertion follows immediately from the definition of the map Γ. □

Example 3.2. Consider the same pair $(d, l) \in \mathcal{DL}_8$ as in Example 2.5 and the corresponding permutation $\pi = 78356214$. We have that the right-connected components of π are 78 and 356214. The Dyck path corresponding to the permutation 356214 is $UUUDUUDDDUDD$, the Dyck path corresponding to the normalization of 78, i.e., to the permutation 12, is $UUDD$, and these are precisely the irreducible components of the Dyck path d.

Denote by $RC(\pi)$ the *reverse-complement* of the permutation π, namely, if $\pi = \pi_1\pi_2\dots\pi_n$, then $RC(\pi) = (n+1-\pi_n)(n+1-\pi_{n-1})\dots(n+1-\pi_1)$.

The following assertion relates the images of the permutations π and $RC(\pi)$ under the map Γ.

Proposition 3.3. *Let π a permutation and let $\Gamma(\pi) = (d, l)$. Then $\Gamma(RC(\pi)) = (d', l')$ where d' is the path obtained from d by reflecting the path d along a vertical line, and $l'_i = m_{n+1-i} - l_{n+1-i}$, $1 \le i \le n$.*

Proof. Let $w_1 \dots w_k$ be the decomposition of $\pi \in \mathbf{S}_n$ into ascending runs with $w_i = x_{i,1}\dots x_{i,l_i}$. For all i, set $y_{i,j} = n+1-x_{i,l_i+1-j}$. Then the decomposition into ascending runs of $RC(\pi)$ is $\widehat{w_k}\dots\widehat{w_1}$ with $\widehat{w_i} = y_{i,1}\dots y_{i,l_i}$.

The assertion follows now immediately from the definition of the map Γ. □

Example 3.4. Consider the permutation $\pi = 78356214$ of Example 2.5. Hence $RC(\pi) = 58734612$. We have $\Gamma(\pi) = (d, l)$, where

$$\begin{aligned} d &= UUUDUUDDDUDDUUDD \quad \text{and} \\ l &= (0,0,0,1,0,0,0,0). \end{aligned}$$

The height list of d is $m(d) = (0,1,1,1,0,0,0,0)$. Then $\Gamma(RC(\pi)) = (d', l')$ with

$$\begin{aligned} d' &= UUDDUUDUUUDDUDDD \quad \text{and} \\ l' &= (0,0,0,0,0,1,1,0). \end{aligned}$$

Now we define a map $\Phi\colon \mathbf{S}_n \to \mathbf{S}_n$. Let $\pi \in \mathbf{S}_n$ and $w_1 \dots w_k$ its decomposition into ascending runs. Given two consecutive ascending runs w_i and w_{i+1}, we say that they are *contigue* if $w_{i+1}w_i$ is an increasing sequence of integers or, equivalently, if the tail of w_{i+1} is smaller than the head of w_i. Then we consider the decomposition of $\pi = B_1 \dots B_h$ where the B_i's are maximal sequences of contigue ascending runs. We define the image of π under the map Φ as follows:

$$\Phi(\pi) = B_h \dots B_1.$$

For example, if $\pi = 5623147$, then $B_1 = 5623$, $B_2 = 147$ and $\Phi(\pi) = 1475623$.

Note that Φ is an involution, namely, Φ^2 is the identity over $\mathbf{S}_n$.

Theorem 3.5. *Let $\pi \in \mathbf{S}_n$. Then*

$$\Gamma(\pi) = (d, l) \quad \textit{if and only if} \quad \Gamma(\Phi(\pi)) = (d, m(d) - l).$$

Proof. Let $\Gamma(\pi) = (d, l)$ and $\Gamma(\Phi(\pi)) = (d', l')$. Since permutation $\Phi(\pi)$ has the same heads and tails as π, $d = d'$. Moreover the definition of the map Φ implies immediately that $l' = m(d) - l$. □

It follows from Proposition 3.3 and Theorem 3.5 that

$$\Phi \circ RC = RC \circ \Phi.$$

4 Pattern avoiding permutations

Let $\pi \in \mathbf{S}_n$ and $\tau \in \mathbf{S}_m$. We say that $\pi = \pi_1 \dots \pi_n$ *contains the pattern* $\tau = \tau_1 \dots \tau_m$ if there exists an index subsequence $1 \le i_1 < i_2 < \dots < i_m \le n$ such that $\pi_{i_j} < \pi_{i_k}$ iff $\tau_j < \tau_k$ for $1 \le j, k \le m$. Otherwise, π *avoids the pattern* τ. The set of permutations of length n that avoid the pattern τ is denoted by $\mathbf{S}_n(\tau)$.

In this section we study the behavior of the map Γ when restricted to some subsets of pattern-avoiding permutations.

4.1 Permutations avoiding 132

The following proposition shows that the set $\mathbf{S}_n(132)$ corresponds to a particular subset of $\mathcal{DL}_n$.

Proposition 4.1. *Let $\pi \in \mathbf{S}_n$ and let $\Gamma(\pi) = (d, l)$. Then*

$$\pi \in \mathbf{S}_n(132) \quad \textit{if and only if} \quad l = (0, \dots, 0).$$

Proof. Set $l = (l_1, \dots, l_n)$. We recall that

$$l_i = |\{j \mid h_j < i < t_j,\ t_j \text{ precedes } i \text{ in } \pi\}|.$$

Suppose that there exists an index i such that $l_i > 0$. Then there are at least a head h_j and a tail t_j with $h_j < i < t_j$ such that t_j precedes i in π. As a consequence, $h_j t_j i$ is an occurrence of 132 in π.

This suffices to conclude the proof, since both the cardinalities of $\mathbf{S}_n(132)$ and $\mathcal{D}_n$ are given by the n-th Catalan number (see [21] and [23], respectively). □

Many bijections between 132-avoiding permutations and Dyck paths are present in the literature (see [12, Chapter 4] for an exhaustive description). The preceding proposition implies that the map Γ, when restricted to the set $\mathbf{S}_n(132)$, provides yet another one bijection between this set and $\mathcal{D}_n$, whose inverse has an easy description in terms of multitunnels.

Given a Dyck path d, a *multitunnel* in d is a maximal horizontal segment between two lattice points of d lying always below d (see [9]). A multitunnel can consist of a single point. For our purpose we will be interested in *odd multitunnels*, namely, multitunnels whose points have odd y-coordinate.

For example, the three odd multitunnels of the path in Figure 8 are the dashed segments (one of them reduces to a point).

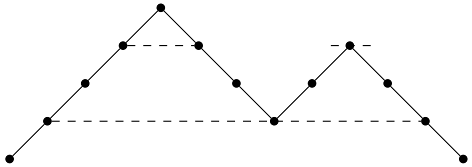


Figure 8: A Dyck path with three odd multitunnels (denoted with dashed segments).

Proposition 4.2. *Let d be a Dyck path and π be the corresponding permutation in $\mathbf{S}_n(132)$, namely $\pi = \Lambda(d, (0, \dots, 0))$. Consider the tagged Dyck path $[d, (1, \dots, n)]$. Every ascending run of π is given by the tags of the points of d lying on the same odd multitunnel. Moreover, the sequence of the heads of π is a decreasing sequence.*

Proof. At the first application of the procedure P the first chosen dimer is the dimer containing the first return of d, and the removed tags correspond to the dimers at height 0 in the leftmost irreducible component of d, whose middle points are precisely the points of the leftmost and lowest odd multitunnel. The same argument can be used for the following steps. □

Example 4.3. Consider the Dyck path d in Figure 9. The three multitunnels of d contain the

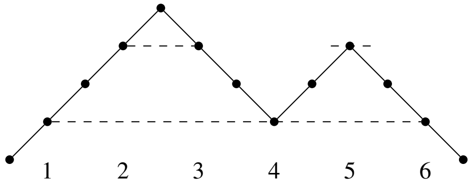


Figure 9: A Dyck path with three multitunnels.

points whose tags are $\{1, 4, 6\}$, $\{2, 3\}$, $\{5\}$, and the corresponding permutation in $\mathbf{S}_6(132)$ is 523146.

We recall that a *descent* of a permutation π is an index i, $1 \leq i \leq n-1$, such that $\pi_i > \pi_{i+1}$. If $\pi = w_1 \dots w_k$ is the decomposition of π into ascending runs, a descent of π occurs at the end of each ascending run, except the last one: Hence, the descents of π are given by the positions of the first $k-1$ tails.

Since it is well known (see e.g. [4]) that the number of permutations in $\mathbf{S}_n(132)$ with h descents is the Narayana number

$$N(n, h+1) = \frac{1}{n}\binom{n}{h+1}\binom{n}{h},$$

we get that the Dyck paths of semilength n with h odd multitunnels are counted by the Narayana number $N(n, h)$.

4.2 Non-crossing partitions

The set of 132-avoiding permutations of length n corresponds bijectively to the set of non-crossing partitions of $\{1, 2, \dots, n\}$. Such partitions were introduced by Kreweras in [15] and extensively studied by many authors in recent years (see [1, 17, 19, 22], to name but a few).

A partition of the set $\{1, 2, \dots, n\}$ is said to be *non-crossing* if, whenever four elements $a, b, c, d \in \{1, \dots, n\}$ with $a < b < c < d$ are such that a, c are in the same block and b, d are in the same block, then the two blocks coincide. We denote by $NC(n)$ the set of non-crossing partitions of $\{1, \dots, n\}$.

As usual (see e.g. [19]) we represent non-crossing partitions graphically by plotting n points on the real line labelled with $1, 2, \dots, n$ and joining points corresponding to successive elements of the same block by arcs. Since we consider a non-crossing partition, the arcs of this diagram do not intersect in points different from $1, 2, \dots, n$.

For example, the non-crossing partition whose blocks are $\{1, 5\}$, $\{2, 3\}$, $\{4\}$, $\{6, 7, 8\}$ is graphically represented in Figure 10.

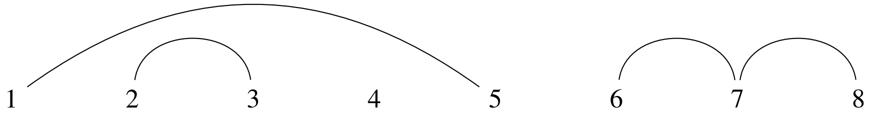


Figure 10: The non-crossing partition with blocks $\{1, 5\}$, $\{2, 3\}$, $\{4\}$, $\{6, 7, 8\}$.

Many bijections between $\mathbf{S}_n(132)$ and $NC(n)$ are available in the literature (see e.g. [19, 24]).

The following proposition provides another bijection between these two sets.

Proposition 4.4. *The permutation* $\pi \in \mathbf{S}_n$ *avoids* 132 *if and only if*

(a) the heads of π *are in decreasing order and*

(b) the partition of n *given by the ascending runs of* π *is a non-crossing partition.*

Proof. Proposition 4.2 implies immediately that if the permutation π avoids 132 conditions (a) and (b) are fulfilled.

Conversely, let π be a permutation containing an occurrence of 132. Let $\pi_i \pi_j \pi_k$ be this occurrence. Without loss of generality, suppose that π_i is the head of an ascending run. If π_j is in the same ascending run, π_k is surely in a different ascending run whose head we denote by h. If $h > \pi_i$, condition (a) is not satisfied. If h is smaller than π, the quadruple h, π_i, π_k, π_j does not satisfy condition (b). If π_j and π_k are in different ascending runs, denote by $\widehat{h}$ the head of the ascending run containing π_k. If $\widehat{h} > \pi_i$, condition (a) is not satisfied. If $\widehat{h} < \pi_i$, the triple $\widehat{h}, \pi_j, \pi_k$ is an occurrence of 132 with $\widehat{h}$ and π_j in the same ascending run. This completes the proof. □

The result of the previous proposition induces a bijection between non-crossing partitions and Dyck paths. The same bijection can be found in [24, Proposition 2.1].

4.3 Permutations avoiding 213

We now turn our attention to the pattern 213. We recall that, since $RC(132) = 213$, we have $\pi \in \mathbf{S}_n(132)$ whenever $RC(\pi) \in \mathbf{S}_n(213)$. Hence, Proposition 3.3 implies immediately the following results:

Proposition 4.5. *Let $\pi \in \mathbf{S}_n$ and let $\Gamma(\pi) = (d, l)$. Then*

$$\pi \in \mathbf{S}_n(213) \quad \textit{if and only if} \quad l = m(d).$$

Proposition 4.6. *Let d be a Dyck path and π be the corresponding permutation in $\mathbf{S}_n(213)$, namely, $\pi = \Lambda(d, m(d))$. Consider the tagged Dyck path $[d, (1, \ldots, n)]$. Every ascending run of π is given by the tags of the points of d which lie on the same odd multitunnel. The sequence of the tails of π is a decreasing sequence.*

The preceding results imply that the map Φ, when restricted to $\mathbf{S}_n(132)$, becomes a bijection onto $\mathbf{S}_n(213)$ that can be described as follows.

Proposition 4.7. *Consider a permutation $\pi \in \mathbf{S}_n(132)$ whose decomposition into ascending runs is $\pi = w_1 \ldots w_k$. The corresponding permutation $\Phi(\pi)$ in $\mathbf{S}_n(213)$ is the permutation with the same ascending runs rearranged so that the tails are in decreasing order.*

For example, if $\pi = 56\,23\,147$ then $\Phi(\pi) = 147\,56\,23$.

To the best of our knowledge, this map is new.

Let $\pi \in \mathbf{S}_n$. A *left-to-right* (LR) *minimum* of π is an element π_i of π such that $\pi_i < \pi_j$ for all $j < i$. Similarly, a *right-to-left* (RL) *maximum* is an element π_i such that $\pi_j < \pi_i$ for all $j > i$. The next proposition describes the behavior of the map Φ with respect to the statistics “number of LR minima”, “number of RL maxima”, and “number of descents”.

Proposition 4.8. *Let π be a permutation in $\mathbf{S}_n(132)$. Then the permutations π and $\Phi(\pi)$ have the same number of descents. Moreover, the number of LR minima of π equals the number of RL maxima of $\Phi(\pi)$.*

Proof. Since $\pi \in \mathbf{S}_n(132)$, the heads of its ascending runs are in decreasing order, hence they are precisely the LR minima of π. Similarly, the tails of $\Phi(\pi)$ are its RL maxima. Finally, we recall that a permutation with k ascending runs has $k - 1$ descents. The proof now follows immediately by Proposition 4.7. □

5 Polynomials associated with permutations and their Hankel transform

Now we associate with a permutation $\pi \in S_n$ the monomial $\theta(\pi) = H^{|H_\pi|} S^{|HT_\pi|} B^{|B_\pi|}$ (tails are not considered because they are always in bijection with heads). Note that if $\gamma(\pi) = \gamma(\sigma)$, namely, π and σ correspond to the same Dyck path, then $\theta(\pi) = \theta(\sigma)$.

Set

$$F_n(H, S, B) = \sum_{\pi \in \mathbf{S}_n} \theta(\pi) = \sum_{h,s,b \geq 0} a_{h,s,b,n} H^h S^s B^b,$$

where $a_{h,s,b,n}$ is the number of permutations of $\mathbf{S}_n$ with h proper heads, s head-tails, and b boarders. Note that $a_{h,s,b,n} = 0$ when $2h + s + b \neq n$.

We want to study the generating function

$$F(H,S,B,X)=\sum_{n\geq 0}F_nX^n=\sum_{h,s,b,n\geq 0}a_{h,s,b,n}H^hS^sB^bX^n.$$

We recall that every permutation in $\mathbf{S}_n$ can be obtained from a permutation in $\mathbf{S}_{n-1}$ by adding the symbol n in any position. Table 1 shows how each insertion of the element n between the entries a and b into a permutation $\pi\in\mathbf{S}_{n-1}$ modifies the number of proper heads, proper tails, head-tails and boarders (in this table, ϵ denotes the empty word).

Table 1: The insertion of n between two symbols a and b of $\sigma\in S_{n-1}$.

a	b	becomes	a	n	b
h	t	→	h	t	ht
h	b	→	h	t	h
ht	h	→	h	t	h
ht	ht	→	h	t	ht
b	b	→	b	t	h
b	t	→	b	t	ht
t	h	→	b	t	h
t	ht	→	b	t	ht
ϵ	h	→	ϵ	ht	h
ϵ	ht	→	ϵ	ht	ht
t	ϵ	→	b	t	ϵ
ht	ϵ	→	h	t	ϵ

We have the following mutually exclusive options:

(i) n is placed immediately after a tail. In this case, the number of boarders increases by one.

(ii) n is placed immediately before a tail. In this case, the number of head-tail increases by one.

(iii) n is placed immediately before a boarder. In this case, the number of boarders decreases by one while the number of heads and tails increases by one.

(iv) n is placed immediately after a head-tail. In this case, the number of head-tails decreases by one while the number of heads and tails increases by one.

(v) n is placed at the first position. In this case, the number of head-tail increases by one.

As a consequence we have

$$\begin{aligned}a_{h,s,b,n}=&\,h\,a_{h,s,b-1,n-1}+h\,a_{h,s-1,b,n-1}\\&+b\,a_{h-1,s,b+1,n-1}+s\,a_{h-1,s+1,b,n-1}+a_{h,s-1,b,n-1},\end{aligned}\tag{5.1}$$

with the obvious boundary conditions.

This recurrence relation shows that the polynomials $F_n(H,S,B)$ satisfy

$$F_n=\left[(HB+HS)\frac{\partial F_{n-1}}{\partial H}+H\left(\frac{\partial F_{n-1}}{\partial B}+\frac{\partial F_{n-1}}{\partial S}\right)+S\cdot F_{n-1}\right]\qquad\forall n\geq 1,$$

with $F_0 = 1$.

As a consequence we get the following functional equation for the generating function $F(H,S,B,X)$:

$$F = 1 + X\left[(HB + HS)\frac{\partial F}{\partial H} + H\left(\frac{\partial F}{\partial B} + \frac{\partial F}{\partial S}\right) + S \cdot F\right].$$

We recall that the Hankel matrix H_n of order $n+1$ of a sequence $(a_n)_{n\in\mathbb{N}}$ is the $(n+1)\times(n+1)$ matrix whose (i,j)-th entry is a_{i+j} where the indices range between 0 and n. The *Hankel transform* of the sequence (a_n) is the sequence $(b_n)_{n\in\mathbb{N}}$ where

$$b_n = \det H_n = \det\begin{bmatrix} a_0 & a_1 & \dots & a_n \\ a_1 & a_2 & \dots & a_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ a_n & a_{n+1} & \dots & a_{2n}\end{bmatrix}.$$

The problem of determining an explicit expression for the Hankel transform of combinatorial sequences is an active area in combinatorics. An exhaustive review of different methods for determinant evaluations, including Hankel determinants, is given by Krattenthaler in [13, 14].

The main result of this section is the following theorem which gives an explicit formula for the Hankel transform of the polynomial sequence (F_n).

Theorem 5.1. *The Hankel transform of the sequence* $(F_n(H,S,B))_{n\geq 0}$ *is given by*

$$\left(H^{m(m+1)/2} \cdot \prod_{i=0}^{m} (i!)^2\right)_{m\geq 0}.$$

Proof. To prove the result we use the Gessel-Viennot lemma (see e.g. [2, p. 217]).

Consider in the lattice plane the two sets of points $\{A_0, A_1, \dots, A_m\}$ and $\{B_0, B_1, \dots, B_m\}$ with $A_i = (-2i, 0)$ and $B_j = (2j, 0)$. For every permutation ρ of the set $\{0, 1, \dots, m\}$ consider a set $p_0, p_1, \dots, p_m$ of $m+1$ valued Dyck paths such that p_i starts at A_i and ends at $B_{\rho(i)}$.

Each valued Dyck path with initial point A_i and ending point B_j corresponds to a permutation in S_{i+j}, by the results of Section 2. As we noticed above, for any Dyck path d, all the permutations σ such that $\gamma(\sigma) = d$ share the same monomial $\theta(\sigma)$. For this reason, this monomial will be denoted by $\theta(d)$.

Associate with the $(m+1)$-tuple $p_0, p_1, \dots, p_m$, where $p_i = (d_i, l^{(i)})$, the monomial $\prod_{i=0}^{m} \theta(d_i)$. This monomial will be called the weight of the $(m+1)$-tuple.

Consider the set of points $C_k = (0, 2k)$, with $0 \leq k \leq m$. Notice that every path $d_0, d_1, \dots, d_m$ contains exactly a point C_k. There are only two possibilities:

(i) there are at least two paths among $d_0, \dots, d_m$ that intersect in one of the points C_k, $0 \leq k \leq m-1$;

(ii) ρ is the identity permutation and $d_i = U^{2i}D^{2i}$, $i = 0, \dots, m$. We will call this configuration the *trivial* $(m+1)$-tuple.

We define a weight-preserving involution over the set of non trivial $(m+1)$-tuples of valued Dyck paths. This involution changes the sign of the corresponding permutation of $\{0, 1, \ldots, m\}$. Given a $(m+1)$-tuple $p_0, p_1, \ldots, p_m$, with $p_i = (d_i, l^{(i)})$, find the greatest value of k such that there exist at least two paths intersecting at C_k, $0 \leq k \leq m-1$. Take the two paths d_i and d_j that intersect in C_k with minimal indices. Then associate with the $(m+1)$-tuple $p_0, p_1, \ldots, p_m$ a new $(m+1)$-tuple $q_0, q_1, \ldots, q_m$ as follows:

(i) if $s \neq i, j$, $q_s = p_s$;

(ii) q_i goes from A_i to C_k along p_i and along p_j from C_k to $B_{\rho(j)}$. Similarly, q_j goes from A_j to C_k along p_j and along p_i from C_k to $B_{\rho(i)}$. The tags of the new paths are defined accordingly.

By the construction of the map Γ and its inverse Λ, we have that

$$\sum_{(d,l)\in\mathcal{D}L_n} \theta(d) = F_n(H, S, B),$$

hence, the polynomial associated to the valued Dyck paths from A_i to B_j is precisely F_{i+j}.

As a consequence, by the Gessel-Viennot lemma, the determinant of the m-th Hankel matrix of the sequence $(F_n(H, S, B))_{n\geq 0}$ is equal to the product of the monomials corresponding to non-intersecting valued Dyck paths, precisely, valued Dyck paths starting at A_i and ending at B_i, of the form $U^{2i}D^{2i}$, for all $0 \leq i \leq m$. Note that there are $((i)!)^2$ such valued Dyck paths, each of which has monomial H^i. This completes the proof. □

Here we mention some specializations of the polynomials F_n. The preceding theorem allows us to compute the Hankel transform of all the following sequences, specializing the variable H accordingly. The first one of these Hankel transforms was previously obtained (see [6]). The other three are new, to the best of our knowledge.

1. Recalling that $2h + b + s = n$,

$$\widehat{a}_{m,n} := \sum_{\substack{h+s=m \\ h,s\geq 0 \\ s+2h\leq n}} a_{h,s,n-2h-s,n}$$

   is precisely the number of permutations of length n with m ascending runs, i.e., the (m, n)-th Eulerian number (see e.g. [7] for the definition and main properties of these numbers) and Equation (5.1) reduces to the well-known recurrence for the Eulerian numbers

$$\widehat{a}_{m,n} = m\,\widehat{a}_{m,n-1} + (n - (m-1))\,\widehat{a}_{m-1,n-1}.$$

   Hence, under the identification $B = 1$, $H = S = t$, the polynomial $F_n(H, S, B)$ reduces to the n-th Eulerian polynomial $F_n(t)$ (see [7]).

2. Under the identification $B = t$, $H = S = 1$, the coefficient of t^k in $F_n(t)$ turns out to be the cardinality of the set of n-permutations with k boarders, i.e., permutations with k occurrences of the consecutive pattern 123 (see [12]). An explicit expression for the exponential generating function of the polynomials $F_n(t)$ can be found in [10].

3. If we set $H = 1$ and $B = S = 0$, we get the number of n-permutations without boarders and head-tails, i.e., *down-up permutations*. As a consequence F_n is the n-th unsigned secant number (sequence A122045 in [20]).

4. If $H = 2$, $B = S = 1$, (F_n) is the sequence of Springer numbers (sequence A001586 in [20]). In fact, these numbers count *down-up signed permutations*, the equivalent of down-up permutations in the hyperoctaedral group. As an example $(3, -4, -2, -5, 1, -6)$ is a down-up signed permutation of the hyperoctaedral group $\mathbf{B}_6$.

   There is an obvious bijection between permutations of $\mathbf{S}_n$ with two-colored heads and down-up signed permutations. Let $\pi \in \mathbf{S}_n$ and let $\pi = w_1 w_2 \dots w_k$ be its decomposition into ascending runs. Consider the reverse of π, $R(\pi)$. If we want to construct a signed permutation whose corresponding unsigned permutation is $R(\pi)$ we have two possible choices for the sign of every head of π and only one choice for any other element. As an example consider $\pi = 81235467 \in \mathbf{S}_8$. The heads are 1 and 4. The four possible down-up signed permutations whose unsigned elements form $R(\pi) = 76453218$ are $(7, -6, 4, -5, 3, -2, 1, -8)$, $(7, -6, -4, -5, 3, -2, 1, -8)$, $(7, -6, 4, -5, 3, -2, -1, -8)$, and $(7, -6, -4, -5, 3, -2, -1, -8)$. This is clearly a bijective correspondence.

Now we consider the polynomials

$$\widehat{F}_n(H, S, B) = \sum_{\pi \in \mathbf{S}_n(132)} \theta(\pi).$$

Theorem 5.2. *The Hankel transform of the sequence* $(\widehat{F}_n(H, S, B))_{n \geq 0}$ *is given by*

$$\left(H^{m(m+1)/2}\right)_{m \geq 0}.$$

Proof. The proof is similar to the previous one, keeping in mind that permutations avoiding the pattern 132 correspond bijectively to valued Dyck paths where the sequence l is identically zero. □

Also in this case suitable specializations yield well-known results.

1. Specializing $B = 1$ and $H = S = t$ the polynomials $\widehat{F}_n$ turns out to be the *Narayana polynomials* (see e.g. [16]).
2. If $B = t$ and $H = S = 1$, $\widehat{F}_n(t)$ is the generating function of the set of n-permutations that avoid the pattern 132 where the variable t takes into account the occurrences of the consecutive pattern 123. An explicit formula for the generating function $\sum_{n \geq 0} \widehat{F}_n$ can be found in [5]. In particular, when $t = 0$, F_n is the n-th Motzkin number. Theorem 5.2 implies that this sequence is among the many sequences whose Hankel transform is the constant sequence $(1)_{n \geq 0}$.
3. When $H = B = 2$, $S = 1$, (F_n) is the sequence of *super-Catalan numbers* or *little Schroeder numbers* (sequence A001003 in [20]). These numbers are known to count Motzkin paths of length $n - 1$ where every up step has two colors and every horizontal step has three colors. It is easy to find a bijection between this set and the set of Dyck path of semilength n where the diods UU and DU have two colors.
4. If $H = t$ and $B = S = 1$, where t is a positive integer, F_n turns out to be the number of lattice paths from $(0, 0)$ to $(2n, 0)$ composed by steps of the form $U = (1, 1)$, $D = (1, -1)$ and $L = (3, 1)$, where the L steps have $t-1$ colors. In fact it is possible

to find a bijection between this last set and the set of Dyck paths of semilength n where diods UU have t colors. To define such a bijection consider a Dyck path of semilength n where each diod UU has t colors. Replace each diod UU labelled with the color k, $1 \leq k \leq t-1$ with an L step with the same color and replace the corresponding diod DD with a step D. If a diod UU is labelled with the color t, leave it and the corresponding DD unchanged. Leave also the diods UD and DU unchanged. This is the required bijection. Note that this last case reduces to sequence A052709 in [20] when $t = 2$ and to sequence A129147 when $t = 3$.

Theorem 5.2, applied to the particular cases above, shows that the Hankel transform of all the previous sequences is given by

$$\left(H^{n(n+1)/2}\right)_{n\geq 0}$$

specializing H accordingly.

References

[1] M. Aigner, Catalan and other numbers: a recurrent theme, in: H. Crapo and D. Senato (eds.), *Algebraic Combinatorics and Computer Science: A tribute to Gian-Carlo Rota*, Springer-Verlag Italia, Milan, pp. 347–390, 2001, doi:10.1007/978-88-470-2107-5_15.

[2] M. Aigner, *A Course in Enumeration*, volume 238 of *Graduate Texts in Mathematics*, Springer, Berlin, 2007, doi:10.1007/978-3-540-39035-0.

[3] J.-C. Aval, A. Boussicault and S. Dasse-Hartaut, Dyck tableaux, *Theoret. Comput. Sci.* **502** (2013), 195–209, doi:10.1016/j.tcs.2011.11.038.

[4] M. Barnabei, F. Bonetti and M. Silimbani, The descent statistic on 123-avoiding permutations, *Sém. Lothar. Combin.* **63** (2010), Article B63a (8 pages), https://www.mat.univie.ac.at/~slc/wpapers/s63barnbosi.html.

[5] M. Barnabei, F. Bonetti and M. Silimbani, The joint distribution of consecutive patterns and descents in permutations avoiding 3-1-2, *European J. Combin.* **31** (2010), 1360–1371, doi:10.1016/j.ejc.2009.11.011.

[6] P. Barry, Eulerian polynomials as moments, via exponential Riordan arrays, *J. Integer Seq.* **14** (2011), Article 11.9.5 (14 pages), https://cs.uwaterloo.ca/journals/JIS/VOL14/Barry7/barry172.html.

[7] M. Bóna, *Combinatorics of Permutations*, Discrete Mathematics and Its Applications, Chapman & Hall/CRC, Boca Raton, Florida, 2004, doi:10.1201/9780203494370, with a foreword by Richard Stanley.

[8] R. Cori, Indecomposable permutations, hypermaps and labeled Dyck paths, *J. Comb. Theory Ser. A* **116** (2009), 1326–1343, doi:10.1016/j.jcta.2009.02.008.

[9] S. Elizalde and E. Deutsch, A simple and unusual bijection for Dyck paths and its consequences, *Ann. Comb.* **7** (2003), 281–297, doi:10.1007/s00026-003-0186-y.

[10] S. Elizalde and M. Noy, Consecutive patterns in permutations, *Adv. Appl. Math.* **30** (2003), 110–125, doi:10.1016/s0196-8858(02)00527-4.

[11] J. Françon and G. Viennot, Permutations selon leurs pics, creux, doubles montées et double descentes, nombres d'Euler et nombres de Genocchi, *Discrete Math.* **28** (1979), 21–35, doi:10.1016/0012-365x(79)90182-1.

[12] S. Kitaev, *Patterns in Permutations and Words*, Monographs in Theoretical Computer Science, Springer, Heidelberg, 2011, doi:10.1007/978-3-642-17333-2, with a foreword by Jeffrey B. Remmel.

[13] C. Krattenthaler, Advanced determinant calculus, *Sém. Lothar. Combin.* **42** (1999), Article B42q (67 pages), `https://www.mat.univie.ac.at/~slc/wpapers/s42kratt.html`.

[14] C. Krattenthaler, Advanced determinant calculus: a complement, *Linear Algebra Appl.* **411** (2005), 68–166, doi:10.1016/j.laa.2005.06.042.

[15] G. Kreweras, Sur les partitions non croisées d'un cycle, *Discrete Math.* **1** (1972), 333–350, doi:10.1016/0012-365x(72)90041-6.

[16] M. D. Petković, P. Barry and P. Rajković, Closed-form expression for Hankel determinants of the Narayana polynomials, *Czechoslovak Math. J.* **62** (2012), 39–57, doi:10.1007/s10587-012-0015-8.

[17] H. Prodinger, A correspondence between ordered trees and noncrossing partitions, *Discrete Math.* **46** (1983), 205–206, doi:10.1016/0012-365x(83)90255-8.

[18] A. Randrianarivony, Correspondances entre les différents types de bijections entre le groupe symétrique et les chemins de Motzkin valués, *Sém. Lothar. Combin.* **35** (1995), Article B35h (12 pages), `https://www.mat.univie.ac.at/~slc/wpapers/s35arthur.html`.

[19] R. Simion, Combinatorial statistics on non-crossing partitions, *J. Comb. Theory Ser. A* **66** (1994), 270–301, doi:10.1016/0097-3165(94)90066-3.

[20] N. J. A. Sloane (ed.), The On-Line Encyclopedia of Integer Sequences, published electronically at `https://oeis.org`.

[21] R. P. Stanley, *Enumerative Combinatorics, Volume I*, The Wadsworth & Brooks/Cole Mathematics Series, Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, California, 1986, doi:10.1007/978-1-4615-9763-6.

[22] R. P. Stanley, Parking functions and noncrossing partitions, *Electron. J. Combin.* **4** (1997), #R20 (14 pages), `https://www.combinatorics.org/ojs/index.php/eljc/article/view/v4i2r20`.

[23] R. P. Stanley, *Enumerative Combinatorics, Volume 2*, volume 62 of *Cambridge Studies in Advanced Mathematics*, Cambridge University Press, Cambridge, 1999, doi:10.1017/cbo9780511609589.

[24] C. Stump, More bijective Catalan combinatorics on permutations and on signed permutations, *J. Comb.* **4** (2013), 419–447, doi:10.4310/joc.2013.v4.n4.a3.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 465–472
https://doi.org/10.26493/1855-3974.1699.74e
(Also available at http://amc-journal.eu)

Some extensions of optimal stopping with financial applications

Mihael Perman [*]
Faculty of Mathematics and Physics, University of Ljubljana,
Jadranska 19, SI-1000 Ljubljana, Slovenia and
University of Primorska, Faculty of Mathematics, Natural Sciences and Information
Technologies, Glagoljaška 8, SI-6000 Koper, Slovenia

Ana Zalokar [†]
University of Primorska, Faculty of Mathematics, Natural Sciences and Information
Technologies, Glagoljaška 8, SI-6000 Koper, Slovenia and
University of Primorska, Andrej Marušič Institute,
Muzejski trg 2, SI-6000 Koper, Slovenia

Received 15 May 2018, accepted 19 June 2018, published online 22 February 2019

Abstract

Finite horizon optimal stopping problems for Markov chains are a well researched topic. Frequently they are phrased in terms of cost or return because many financial models are based on Markov chains. In this paper we will apply optimal stopping to certain random walks on binary trees motivated by insurance considerations. The results are direct extensions of known results but the implications for insurance are of interest.

Keywords: Optimal stopping for Markov chains, equity-linked life insurance with guarantees.

Math. Subj. Class.: 60G40, 91B30

1 Introduction

Modern insurance regulation requires companies to apply market valuation to assets and liabilities. The value of assets can be determined directly from market prices, or through appropriate approximations using fair value methodology. For insurance liabilities, however, there is no regulated market to determine their value. The particular case that we

[*] This work is supported in part by the Slovenian Research Agency (research program P1-0285).

[†] This work is supported in part by the Slovenian Research Agency (research program P1-0285 and Young Researchers Grant).

E-mail addresses: mihael.perman@fmf.uni-lj.si (Mihael Perman), ana.zalokar@iam.upr.si (Ana Zalokar) will be considering here are equity-linked life policies with guarantees. The policyholder invests her premium in an underlying fund managed by the insurance company. A typical example are long term pension saving products. In recent years there has been a tendency to attach guarantees such as a minimum return or a minimum death benefit guarantee to these investments which gives rise to new liabilities. In many cases guarantees can be interpreted as contingent claims on the underlying fund, for example guarantees in equity-linked products or complementary health policies with equalization schemes, see [10].

Nonnenmacher was one of the first authors to interpret guarantees as put options on the value on the underlying fund, [8] and [7]. Once a stochastic model for the dynamics of the fund value is formulated, the liabilities arising from guarantees can be valued using the methods to value derivative securities. Paper [6] considers equity-linked products as contingent claims on the value of the underlying asset but introduces mortality as an independent additional source of randomness. The assumption of independence is often made, see [5] for some implications. With this addition the model is no longer complete and the paper considers optimal hedging strategies that minimize the expected cost for the insurer.

In this paper we will present some extensions of optimal stopping rules motivated by financial questions in insurance. The proofs follow the steps of classical proofs but the formulation of the problems is slightly different. These results will then be applied to investigate relative merits of different ways an insurance company can hedge its liabilities. The models are simplified versions of reality but can shed some light on what strategies may lead to best results.

2 Variations of optimal stopping rules

The classical finite horizon optimal stopping problem for a general finite length inhomogeneous Markov chain $X_0, X_1, \ldots, X_N$ and general state space is to minimize the expression

$$E\left[g(X_\nu) + \sum_{j=1}^{\nu-1} c(X_j)\right] \tag{2.1}$$

where ν runs over all stopping times with respect to the filtration of the Markov chain, and g and c are given functions. For the sake of simplicity it will be assumed that g and c are bounded. For Markov chains it is enough to solve the problem assuming that $X_0 = x$ for x in the state space. Denote the value function v_N by

$$v_N(x) = \inf_{\{\nu : P(\nu \leq N)=1\}} E\left[g(X_\nu) + \sum_{j=1}^{\nu-1} c(X_j) | X_0 = x\right] \tag{2.2}$$

The dynamic programming equations for this problem are defined recursively as

$$\begin{aligned} V_N(x) &:= g(x) \\ V_n(x) &:= \min\left\{g(x), c(x) + E\left[V_{n+1}\left(X_{n+1}\right)|X_n = x\right]\right\} \end{aligned} \tag{2.3}$$

for $n = 0, 1, \ldots, N-1$. The solution to the stopping problem is given by

Theorem 2.1. *The value function is given by $v_N(x) = V_0(x)$, and the optimal stopping time is given by*

$$\nu = \inf\left\{j \geq 0 : V_j(X_j) = g(X_j)\right\}. \tag{2.4}$$

See [9] for proofs.

The above stopping problem has many possible extensions and generalizations. For the financial application in this paper we will minimize the expression

$$E\left[g_\nu(X_1,\dots,X_\nu)+\sum_{j=1}^{\nu-1}c_j(X_1,\dots,X_j)\right] \tag{2.5}$$

for all stopping times $\nu \leq N$ for given functions $c_1,\dots,c_N$ and $g_1,\dots,g_N$. The dynamic programming equations in this more general setup are

$$\begin{aligned} V_N(x_0,\dots,x_N) &:= g_N(x_0,\dots,x_N) && (2.6)\\ V_n(x_0,\dots,x_n) &:= \min\big\{g_n(x_0,\dots,x_n), && (2.7)\\ &\qquad c_n(x_0,\dots,x_n)+E\left[V_{n+1}\left(x_0,\dots,x_n,X_{n+1}\right)|X_n=x_n\right]\big\} \end{aligned}$$

The optimal time is given by

$$\nu_N=\inf\left\{j\geq 0: V_j(X_0,\dots,X_j)=g_j(X_0,\dots,X_j)\right\}\,. \tag{2.8}$$

For the sake of completness we give the proof of this more general theorem. Define

$$Z_n=\sum_{j=0}^{n-1}c_j(X_0,\dots,X_j)+V_n(X_0,\dots,X_n) \tag{2.9}$$

for $j=0,1,\dots,N$. With this definition we have

Theorem 2.2. *The process $(Z_n)_{0\leq n\leq N}$ is a submartingale with respect to the filtration of the Markov chain.*

Proof. Denote $\mathcal{F}_n=\sigma(X_0,\dots,X_n)$ for $1\leq n\leq N$. We compute

$$\begin{aligned} E\left[Z_{n+1}|\mathcal{F}_n\right] &= \sum_{j=0}^{n}c_j(X_0,\dots,X_j)+E\left[V_{n+1}(X_0,\dots,X_{n+1})|\mathcal{F}_n\right]\\ &\geq \sum_{j=0}^{n-1}c_j(X_0,\dots,X_j)+V_n(X_0,\dots,X_n)\\ &= Z_n. \end{aligned}$$

□

Theorem 2.3. *For the time ν_N defined in (2.6) the expression (2.5) attains its minimum which equals $E(V_0(X_0))$.*

Proof. By Theorem 2.2

$$E\left[Z_\nu\right]\geq E\left[Z_0\right]=E\left[V_0(X_0)\right] \tag{2.10}$$

for all stopping times ν. By definition we have

$$E\left[V_0(X_0)\right]\leq E\left[Z_\nu\right]\leq E\left[\sum_{j=0}^{\nu-1}c_j(X_0,\dots,X_j)+g_\nu(X_0,\dots,X_\nu)\right]. \tag{2.11}$$

Replacing ν by ν_N we have

$$\begin{aligned}
&E\left[Z_{\nu_N\wedge(n+1)}|\mathcal{F}_n\right]\\
&\quad= 1(\nu_N \le n)Z_{\nu_N} + 1(\nu_N \ge n+1)E\left[Z_{n+1}|\mathcal{F}_n\right]\\
&\quad= 1(\nu_N \le n)Z_{\nu_N}\\
&\qquad + 1(\nu_N \ge n+1)\left[\sum_{j=1}^{n} c_j(X_0,\dots,X_j) + E\left[V_{n+1}(X_0,\dots,X_{n+1}|X_0,\dots,X_n\right]\right]\\
&\quad= 1(\nu_N \le n)Z_{\nu_N} + 1(\nu_N \ge n+1)\left[\sum_{j=1}^{n-1} c_j(X_0,\dots,X_j) + V_n(X_0,\dots,X_n)\right]\\
&\quad= 1(\nu_N \le n)Z_{\nu_N} + 1(\nu_N \ge n+1)Z_n\\
&\quad= Z_{\nu_N\wedge n}.
\end{aligned}$$

It follows that

$$E\left[V_0(X_0)\right] = E\left[Z_{\nu_N\wedge 1}\right] = \cdots = E\left[Z_{\nu_N\wedge N}\right]. \tag{2.12}$$

By (2.10) – (2.12) the minimum $E\left[V_0(X_0)\right]$ of (2.5) is attained at $\nu = \nu_N$. □

3 Application to insurance

Assume that the net premium of the m policyholders is invested in an equity-linked fund whose price follows the dynamics of the Cox-Ross-Rubinstein model, see [4]. Denote the prices by $S_0, S_1, \dots, S_N$. At time $j = 0$ the total investment of the policyholders is mS_0. In the next time instant the price of the fund is multiplied by u or d with probabilities p and $q = 1-p$ respectively with the usual assumptions $d < 1 < 1+r < u$. Many guarantees can be interpreted as contingent claims on the underlying fund. The minimum yield guarantee, to give the simplest example, stipulates that the payment to the policyholder at time N will be equal to at least

$$G = (1+r)^N S_0 \tag{3.1}$$

for some interest rate agreed to in the contract, which we assume to be constant throughout the lifetime of policies. Other types of guarantees can be included as well. If at the expiration the price of the fund reduced by possible fees exceeds G the policyholder gets the bigger of the two sums. If at time $j = 0$ the insurer buys m put options on the fund price with strike price $k = (1+r)^N S_0$ and expiration N that completely offsets the financial risk due to fund price fluctuations. But such a strategy does not take mortality into account. The strategy we will investigate will be a combination of charging fees towards the fund and at an optimal time buy options that at least partially offset financial risks. Paper [2] considers fund linked products with guarantees and an optimal fee structure which means that the insurance company charges a fee towards the underlying fund in an optimal way so that the expected discounted loss for the company is zero. In this paper we will consider a mixed approach. The company will set aside a portion of the fund value as a reserve possibly subject to some conditions. At any time the company can decide to switch to hedging future liabilities with derivatives based on the fund value and the number of surviving policyholders. The fees accumulated will partially offset the cost of the options. We will derive the optimal time to switch which will minimize the expected loss for the company.

Assume that the m policyholders are of the same age x. Denote the number of surviving policyholders at times $j = 0, 1, \ldots, N$ by $\alpha_0, \alpha_1, \ldots, \alpha_N$. We will assume that mortality is independent of the movement of the fund value. For the sake of simplicity we will consider contracts with no guaranteed minimum death benefit. Note that the sequence $\alpha_0, \ldots, \alpha_N$ is an inhomogeneous Markov chain due to ageing with $P(\alpha_{j+1} = i - k|\alpha_j = i) = \binom{i}{k} q_{x+j}^k p_{x+j}^{i-k}$ for $k = 0, 1, \ldots, i$ in the usual actuarial notation. For mortality simulation we use [1].

We will apply the theory developed in Section 2 to the Markov chain $(S_j, \alpha_j)_{0 \leq j \leq N}$ and the functions c_j and g_j that we now proceed to identify. We consider the following strategy: The company at each time j either sets aside the difference between the fund price S_j and the accumulated value $S_0(1 + r)^j$ if this difference is positive and 0 else, or the company buys a number of put options on the fund at strike price $k = S_0(1 + r)^N$. The number of options to be bought will be determined below in two cases. The options will offset some of the financial risk due to fund price fluctuation but the cost of buying the options will be incurred. In the notation of Section 2 what we set aside will reduce the loss and we define

$$c_j(S_0, S_1, \ldots, S_j) = -m \max\left(S_j - S_0(1 + r)^j, 0\right).$$

Let the price of the put option on the price of the fund at strike price $k = S_0(1 + r)^N$ at time j be denoted by $\pi_j(S_j, k, N)$ determined in the standard way for the binomial model, see [3]. In our scenario two possible numbers of put options can be considered: the first is to buy α_j put options at time j which means that the financial risk is eliminated because the options cover any possible shortfall of the fund price. The insurer will be able to cover liabilities towards surviving policyholders but will incur a cost that will contribute towards the loss. In our notation we put

$$g_j(S_0, \ldots, S_j, \alpha_0, \ldots, \alpha_j) = \alpha_j \pi_j(S_j, k, N). \tag{3.2}$$

The second possibility is to buy $E(\alpha_N|\alpha_j)$ options. This only partially offsets the risk of shortfall because there may be more surviving policyhoders than expected. In this case define for $j < N$

$$g_j(S_0, \ldots, S_j, \alpha_0, \ldots, \alpha_j) = E(\alpha_N|\alpha_j)\pi_j(S_j, k, N) \tag{3.3}$$

and

$$g_N(S_0, \ldots, S_N, \alpha_0, \ldots, \alpha_N) = \alpha_N \cdot \pi_N(S_N, k, N). \tag{3.4}$$

In both cases the expression

$$L_j = (1 + r)^{-j} g_j(S_0, \ldots, S_j, \alpha_0, \ldots, \alpha_j) + \sum_{i=1}^{j-1} (1 + r)^{-i} c_i(S_0, \ldots, S_i) \tag{3.5}$$

will be the discounted loss for the company if the option is bought at time j. We choose the stopping rule ν in such a way that the expected loss

$$E(L_\nu) \tag{3.6}$$

will be minimized. Note that the optimality depends on the probabilities p in the underlying model for the fund price. The solution is provided by Theorem 2.3. Note also that $\nu = N$ means that the insurer does not buy options but covers any shortfall from own funds.

Explicit calculations are not possible so we present two simulations to illustrate the results. The table below defines the parameters used in the simulations:

S_0	1
u	1.04
d	0.98
r	0.01
N	5
m	1000
x	30
z	5000

where z denotes the number of simulations. Table 1 summarizes some statistics for the final loss L_N for different p when the insurer buys α_ν options.

Table 1: A few selected descriptive statistics of the final loss distribution when α_ν options are bought.

p	$E(L_N)$	$\mathrm{SD}(L_N)$	90th percentile	$E(\nu)$	$\mathrm{SD}(\nu)$
0.51	−32.45	135.55	87.54	4.71	0.69
0.50	−21.08	122.52	60.58	3.70	1.57
0.49	−20.45	123.41	42.08	2.85	1.77
0.48	−17.62	117.49	41.72	2.63	1.74

Let us now look at results of simulations when the insurer at time j considers buying $E(\alpha_N|\alpha_j)$ options (Table 2).

Table 2: A few selected descriptive statistics of the final loss distribution when options cover the expected number of surviving policyholders.

p	$E(L_N)$	$\mathrm{SD}(L_N)$	90th percentile	$E(\nu)$	$\mathrm{SD}(\nu)$
0.51	0.69	165.27	195.57	3.45	1.60
0.50	9.42	160.78	197.46	3.22	1.64
0.49	11.45	158.66	196.50	3.05	1.63
0.48	18.36	152.96	197.24	2.75	1.64

In the first case the loss is negative and the financial risk is completely offset. It is true that such a strategy depends on assumptions about availability of derivatives but in more realistic settings it can still be used to reduce the cost of guarantees. In the second case note that the loss due to α_T exceeding the expected number of survivors needs to be taken into account because it contributes to the overall loss.

For the case $p = 0.48$ the distribution of final loss L_N and optimal stopping time ν are shown in Figure 1 for both numbers of options.

4 Conclusions

We propose a strategy of hedging liabilities arising from equity-linked products with minimum guarantees with financial derivatives. Simulations show that there is an optimal time

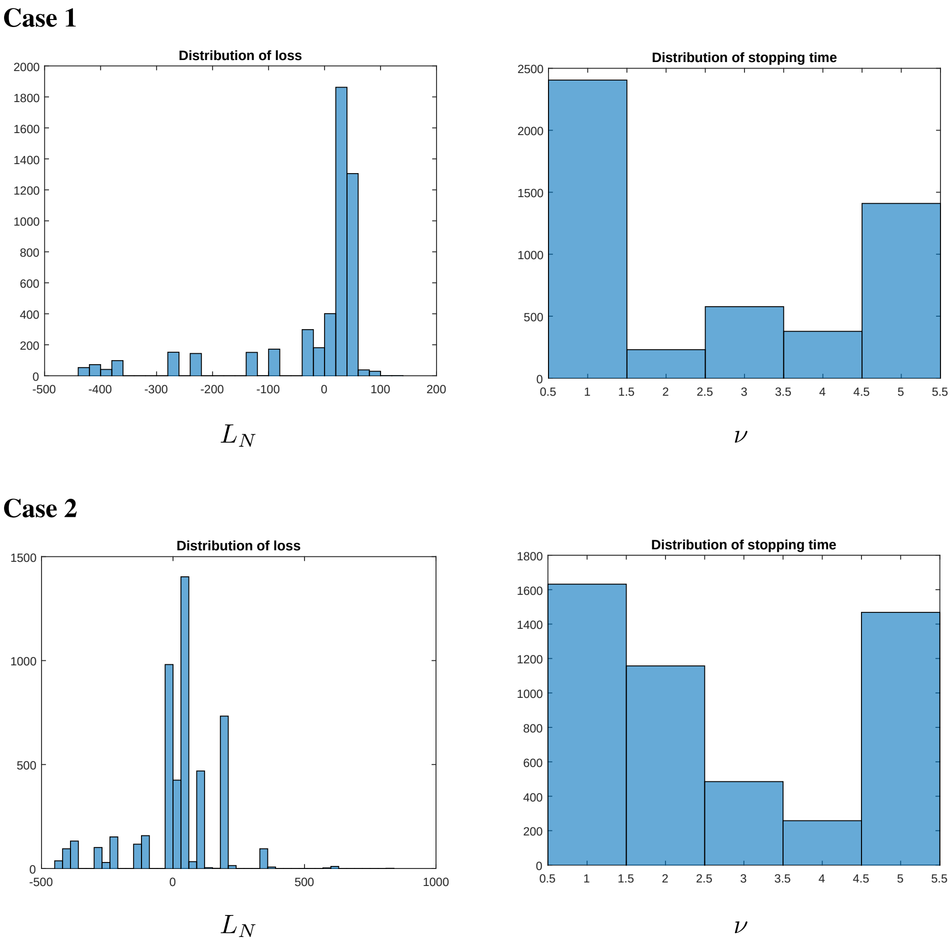


Figure 1: The distribution of final loss and optimal time ν.

to switch from charging a fee towards the fund to buying a put option that will offset the financial risk. The cases considered are simplified but may be an indication that strategies for more realistic settings are possible.

References

[1] Slovenian mortality tables SIA65, 2010 (Slovenske rentne tablice smrtnosti), *Official Gazette of the Republic of Slovenia* **26** (2016), 2148–2149, `https://www.uradni-list.si/_pdf/2016/Ur/u2016018.pdf`.

[2] A. R. Bacinello, P. Millossovich, A. Olivieri and E. Pitacco, Variable annuities: a unifying valuation approach, *Insurance Math. Econom.* **49** (2011), 285–297, doi:10.1016/j.insmatheco.2011.05.003.

[3] T. Björk, *Arbitrage Theory in Continuous Time*, Oxford Finance Series, Oxford University Press, 3rd edition, 2009.

[4] J. C. Cox, S. A. Ross and M. Rubinstein, Option pricing: a simplified approach, *J. Financial Econ.* **7** (1979), 229–263, doi:10.1016/0304-405x(79)90015-1.

[5] J. Dhaene, A. Kukush, E. Luciano, W. Schoutens and B. Stassen, On the (in-)dependence between financial and actuarial risks, *Insurance Math. Econom.* **52** (2013), 522–531, doi:10.1016/j.insmatheco.2013.03.003.

[6] T. Møller, Hedging equity-linked life insurance contracts, *N. Am. Actuar. J.* **5** (2001), 79–95, doi:10.1080/10920277.2001.10595986.

[7] D. J. F. Nonnenmacher, Guaranteed equity-linked products, in: *Proceedings of the 8th International AFIR Colloquium*, 1998 pp. 413–428, held in Robinson College at Cambridge University, UK, September 15 – 17, 1998, `http://www.actuaries.org/AFIR/colloquia/Cambridge/Nonnenmacher.pdf`.

[8] D. J. F. Nonnenmacher and J. Ruß, Equity-linked life insurance in Germany: quantifying the risk of additional policy reserves, in: *Proceedings of the 7th International AFIR Colloquium, Volume 2*, 1997 pp. 719–738, held in the Cairns International Hotel, North Queensland, Australia, August 11 – 15, 1997, `http://www.actuaries.org/AFIR/colloquia/Cairns/Nonnenmacher_Russ.pdf`.

[9] A. N. Shiryayev, *Optimal Stopping Rules*, volume 8 of *Stochastic Modelling and Applied Probability*, Springer-Verlag, New York, 2008, doi:10.1007/978-3-540-74011-7, translated from Russian by A. B. Aries (original Russian edition published by Nauka, Moscow, 1976).

[10] B. Zgrablić, The equalization scheme of the residual voluntary health insurance in Slovenia, *Ars Math. Contemp.* **8** (2015), 225–234, doi:10.26493/1855-3974.667.fec.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 473–486
https://doi.org/10.26493/1855-3974.1477.1c7
(Also available at http://amc-journal.eu)

Convertible subspaces that arise from different numberings of the vertices of a graph*

Henrique F. da Cruz [†], Ilda Inácio , Rogério Serôdio

Universidade da Beira Interior, Centro de Matemática e Aplicações (CMA-UBI),
Rua Marquês d'Ávila e Bolama, 6201-001, Covilhã, Portugal

Received 6 September 2017, accepted 3 January 2019, published online 28 February 2019

Abstract

In this paper, we describe subspaces of generalized Hessenberg matrices where the determinant is convertible into the permanent by affixing ± signs. These subspaces can arise from different numberings of the vertices of a graph. With this numbering process, we obtain some well-known sequences of integers. For instance, in the case of a path of length n, we prove that the number of these subspaces is the $(n+1)^{\text{th}}$ Fibonacci number.

Keywords: Determinant, permanent, Hessenberg matrix.

Math. Subj. Class.: 15A15, 05C05, 05C30

1 Introduction

Let $M_n(\mathbb{C})$ be the linear space of all n-square matrices over the complex field $\mathbb{C}$, and let S_n be the symmetric group of degree n. For $A = [a_{ij}] \in M_n(\mathbb{C})$, the permanent function is defined as

$$\operatorname{per}(A) = \sum_{\sigma \in S_n} \prod_{i=1}^{n} a_{i\sigma(i)}.$$

In a very similar way, the determinant function is defined as

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i\sigma(i)},$$

*Research partially supported by the project **UID/MAT/00212/2013**.

[†]Correspondent author.

E-mail addresses: hcruz@ubi.pt (Henrique F. da Cruz), ilda@ubi.pt (Ilda Inácio), rserodio@ubi.pt (Rogério Serôdio)

where sgn is the sign function.

The determinant is undoubtedly one of the most well-known functions in mathematics with applications in many areas. The permanent function is also a well-studied function, since it has many applications in combinatorics, but while the determinant can be easily computed, no efficient algorithm for computing the permanent is known. This difficulty leads to the idea of trying to compute it by using determinants. This idea dates back to 1913 in a work of Pólya [6], and it has been under intensive investigation since then. Pólya observed that the permanent of a 2 by 2 matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

is equal to the determinant of the related matrix

$$B = \begin{bmatrix} a_{11} & -a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

However, Szegö [8] proved that if $n \geq 3$, then there is no way to generalize this formula, i.e., there is no uniform way of changing the signs of the entries of a matrix $A \in M_n(\mathbb{C})$ in order to obtain a matrix B satisfying

$$\det(B) = \operatorname{per}(A). \tag{1.1}$$

Szegö's result didn't put an end to this question. In fact, the possibility that the permanent can be converted into the determinant by affixing $\pm$ signs is a research topic that remains active until today (see [1, 2] or [5] for recent works on this subject). In 1969, Gibson [3] proved that the linear space of lower Hessenberg matrices is a *convertible subspace*. That is, it is possible to change in a uniform way the signs of the entries of a matrix A in this subspace in order to obtain a matrix B satisfying (1.1).

More recently, C. Fonseca presented a new class of convertible subspaces. These subspaces are constructed using simple graphs as follows:

Definition 1.1 ([2])**.** Given a simple graph G with n vertices, numbered by the integers in $\{1, \dots, n\}$, a G-lower Hessenberg matrix $A = [a_{ij}]$ is an n-square matrix such that $a_{ij} = 0$ whenever $i < j$ and $\{i, j\}$ is not an edge of G.

If, in addition, $a_{ij} \neq 0$ whenever $i \geq j$ or $\{i, j\}$ is an edge of G, then we say that A is a *full G-lower Hessenberg matrix*. Obviously, two numberings of the vertices of a graph are the same if the sets of edges are equal.

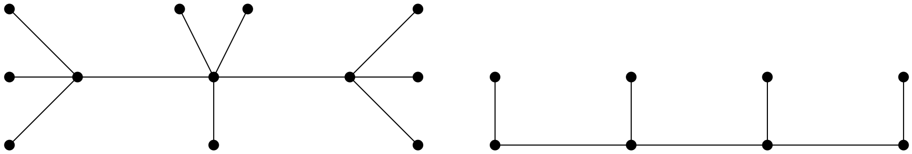
C. Fonseca [2] proved that if G is a generalized double star (the tree resulting from joining the central vertices of two stars by a path) whose vertices are numbered in a natural way (consecutive integers from left to right), the linear space of these G-lower Hessenberg matrices is a convertible subspace.

Let G be a simple graph with n vertices numbered with $1, \dots, n$. We say that G is *well-numbered* if the linear space of all G-lower Hessenberg matrices that arise from this numbering of the vertices of G is convertible. As we will see in the third section, not all graphs admit a well-numbering of its vertices. The characterization of the connected graphs for which such a numbering exist is our first main result.

Theorem 1.2. *A connected graph G admits a well numbering of its vertices if and only if G is a caterpillar.*

Recall that a caterpillar is a tree in which all vertices are within distance 1 of a central path. The interior vertices of the path are called *nodes*. Every caterpillar results from a sequence of stars $(R_1, R_2, \dots, R_t)$, such that the central vertex of R_i is linked to the central vertex of R_{i-1}, $i = 2, \dots, t$, by a single edge. If a caterpillar results from a sequence of stars with an equal number r of vertices, then we will denote this type of caterpillars by $C_t[r]$, where t is the number of the stars involved. So, $C_t[1]$ is a path with t vertices, and $C_1[r]$ is a star with r vertices.

Example 1.3. The caterpillar $C_3[4]$ is shown in Figure 1(a). The caterpillar $C_4[2]$ is shown in Figure 1(b).



(a) $C_3[4]$

(b) $C_4[2]$

Figure 1: Two examples of caterpillars.

Theorem 1.2 states that the vertices of a caterpillar G can be numbered in a way such that the subspace of all G-lower Hessenberg matrices arising from that numbering is convertible. A natural question is to know if all numberings of the vertices of a caterpillar produce a convertible subspace. The answer is negative, and a simple example can be provided with a path. Let G be a path with n vertices. If the vertices of G are numbered in a natural way (consecutive integers from left to right), then from Definition 1.1, we obtain the linear space of classical lower Hessenberg matrices, which Gibson [3] proved to be a convertible subspace. However, if we enumerate the vertices of G in a different way, the subspace of all G-lower Hessenberg matrices arising from that numbering of the vertices of G may no longer be convertible.

Example 1.4. Consider the path of length four numbered as in Figure 2.

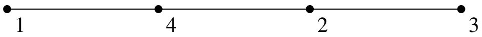


Figure 2: A numbered path on 4 vertices.

The subspace of the $C_4[1]$-lower Hessenberg matrices with maximal dimension constructed from this numbered path is the subspace

$$\left\{ \begin{pmatrix} a_{11} & 0 & 0 & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} : a_{ij} \in \mathbb{C},\ i, j = 1, \dots, 4 \right\},$$

which is not convertible as we will see in the next section.

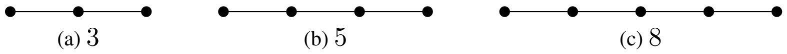
Thus, it is pertinent to ask how many different convertible subspaces, with maximal dimension, arise from different numberings of the vertices of a caterpillar. We restrict our study to caterpillars of the form $C_t[r]$. The answer is given in the next theorem which is our second main result:

Theorem 1.5. *Let r be a fixed integer, and $G = C_t[r]$. Assume that G has at least three vertices. The number of different convertible subspaces, with maximal dimension, that arise from the different numberings of the vertices of G is the $(t+1)^{th}$ term of the sequence*

$$a_n = \begin{cases} 0, & \textit{if } n = 0; \\ 1, & \textit{if } n = 1; \\ ra_{n-1} + a_{n-2}, & \textit{if } n \geq 2. \end{cases}$$

Example 1.6. In this example, we apply Theorem 1.5 to some caterpillars. Below each representation of the caterpillar (in Figures 3, 4 and 5), we give the number of convertible subspaces of G-lower Hessenberg matrices, with maximal dimension, that arise from different numberings of the vertices of that caterpillar. As we are going to see some of the sequences are well known.

1. For a path with three, four and five vertices see Figure 3.

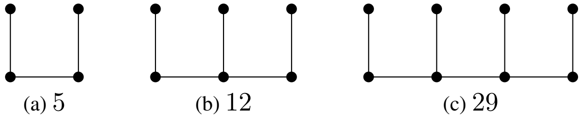


(a) 3 (b) 5 (c) 8

Figure 3: The number of convertible subspaces arising from $C_t[1]$ for $t = 3, 4, 5$.

   In general, the number of convertible subspaces arising from $C_t[1]$, $t \geq 3$ is the $(t+1)^{\text{th}}$ term of the OEIS [7] sequence `A000045`, the sequence of the Fibonacci numbers.

2. If $G = C_t[2]$, then for $t = 2, 3, 4$ see Figure 4.

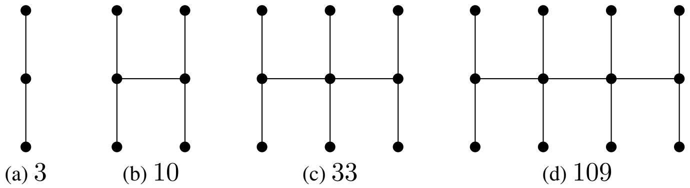


(a) 5 (b) 12 (c) 29

Figure 4: The number of convertible subspaces arising from $C_t[2]$ for $t = 2, 3, 4$.

   In general, the number of convertible subspaces arising from $C_t[2]$, $t \geq 2$ is the $(t+1)^{\text{th}}$ term of the OEIS sequence `A000129`, the sequence of the Pell numbers, also known as lambda numbers.

3. If $G = C_t[3]$, then for $t = 1, 2, 3, 4$ see Figure 5.



(a) 3 (b) 10 (c) 33 (d) 109

Figure 5: The number of convertible subspaces arising from $C_t[3]$ for $t = 1, 2, 3, 4$.

In general, the number of convertible subspaces arising from $C_t[3]$, $t \geq 1$, is the $(t+1)^{\text{th}}$ term of the OEIS sequence `A006190`.

This paper is organized as follows. In the next section, we present a simple criterium to decide when a subspace of G-lower Hessenberg matrices with maximum dimension is convertible. As we are going to see the convertibility of such subspace can be decided from the convertibility of a matrix of zeros and ones. In the third section, we prove our first main result, and describe how vertices numbering should be in order G be well-numbered. The characterization of such numberings allows proving Theorem 1.5.

2 Preliminary results

An n-square $(0,1)$-matrix is a square matrix whose entries are just zeros and ones. Similarly, for an n-square $(-1,1)$-matrix.

Let $S = [s_{i,j}]$ be an n-square $(0,1)$-matrix. For each $i \in \{1,\dots,n\}$ let

$$r_i = \sum_{k=1}^{n} s_{i,k}, \qquad \text{and} \qquad c_i = \sum_{k=1}^{n} s_{k,i}.$$

The sequence $\mathcal{R} = (r_1,\dots,r_n)$ is the *row-sum sequence* of S and the sequence $\mathcal{C} = (c_1,\dots,c_n)$ is the *column-sum sequence* of S.

Definition 2.1. Two matrices A and B are permutation equivalent, if there exist permutation matrices P and Q of suitable sizes such that $B = PAQ$.

An n-square $(0,1)$-matrix S defines a *coordinate subspace*

$$M_n(S) = \{S \star X : X \in M_n(\mathbb{C})\},$$

where $\star$ denotes the Hadamard product. We say that $M_n(S)$ is *convertible* if there exists an n-square $(-1,1)$-matrix C, such that

$$\det(C \star X) = \operatorname{per}(X),$$

for all $X \in M_n(S)$.

Let $T_n = [t_{i,j}]$ be an n-square $(0,1)$-matrix such that

$$t_{i,j} = 0 \quad \text{if and only if} \quad i < j+1.$$

The coordinate subspace $M_n(T_n)$ is the subspace of the lower Hessenberg matrices. Gibson [3] proved that if $C = [c_{i,j}]$ is the n-square $(-1,1)$-matrix such that

$$c_{i,j} = -1 \quad \text{if and only if} \quad j = i+1,$$

then

$$\det(C \star X) = \operatorname{per}(X),$$

for all $X \in M_n(T_n)$.

Another important result due to Gibson states the maximum number of nonzero entries in a convertible matrix.

Theorem 2.2 ([4]). *Let $A = [a_{ij}]$ be an n-square $(0,1)$-matrix such that $\operatorname{per}(A) > 0$, and the permanent of A can be converted into a determinant by affixing ± signs to the elements of A. Then A has at most $\Omega_n = \frac{1}{2}(n^2 + 3n - 2)$ positive entries, with equality if and only if A is permutation equivalent to T_n.*

Observe that the row- and column-sum sequences of T_n are, respectively,

$$\mathcal{R} = (2, 3, \dots, n-1, n, n) \quad \text{and} \quad \mathcal{C} = (n, n, n-1, , \dots, 3, 2).$$

So, Theorem 2.2 gives a simple criterium to decide when an n-square $(0,1)$-matrix S, with Ω_n nonzero entries satisfying $\operatorname{per}(S) > 0$ is convertible.

Proposition 2.3. *Let S be an n-square $(0,1)$-matrix with Ω_n nonzero entries satisfying $\operatorname{per}(S) > 0$. Then S is convertible if and only if S is permutation equivalent to a $(0,1)$-matrix whose row-sum sequence is $\mathcal{R} = (2, 3, \dots, n-1, n, n)$, and the column-sum sequence is $\mathcal{C} = (n, n, n-1, \dots, 3, 2)$.*

Proof. ($\Leftarrow$): The Hessenberg matrix T_n has row- and column-sum sequences $\mathcal{R} = (2, 3, \dots, n-1, n, n)$ and $\mathcal{C} = (n, n, n-1, \dots, 3, 2)$, respectively. By hypothesis, and by transitivity, S is permutation equivalent to T_n. Hence, by Theorem 2.2, S is convertible.

($\Rightarrow$): If S is convertible, then S is permutation equivalent to T_n, and T_n has row-sum sequence $\mathcal{R} = (2, 3, \dots, n-1, n, n)$, and column-sum sequence $\mathcal{C} = (n, n, n-1, \dots, 3, 2)$. □

Example 2.4. Consider the following matrices

$$S_1 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad S_2 = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Then, S_1 is not convertible because the row sum sequence of $(2, 3, 5, 5, 5)$, but S_2 is convertible because we can reorder the rows and the columns S_2 in order to obtain a matrix whose row-sum sequence and column-sum sequence are $(2, 3, 4, 5, 5)$ and $(5, 5, 4, 3, 2)$, respectively.

The next result states that the convertibility of a coordinate subspace $M_n(S)$ can be decided by the convertibility of S.

Proposition 2.5. *Let S be an n-square $(0,1)$-matrix with Ω_n nonzero entries, and*

$$\operatorname{per}(S) > 0.$$

Then, $M_n(S)$ is a convertible subspace if and only if S is convertible.

Proof. If $M_n(S)$ is a convertible subspace, then S is convertible.

Assume that S is convertible. Then, there exists an n-square $(-1,1)$-matrix C such that $\det(C \star S) = \operatorname{per}(S)$.

Let S'_n be the set of permutations $\sigma \in S_n$ such that $\prod_{i=1}^{n} s_{i\sigma(i)} \neq 0$. Since $\operatorname{per}(S) > 0$, $S'_n \neq \emptyset$. We have

$$\begin{aligned}\det(C \star S) &= \sum_{\sigma\in S'_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} c_{i\sigma(i)} s_{i\sigma(i)} \\ &= \sum_{\sigma\in S'_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} c_{i\sigma(i)} \prod_{i=1}^{n} s_{i\sigma(i)},\end{aligned}$$

where we conclude that $\operatorname{sgn}(\sigma)\prod_{i=1}^{n} c_{i\sigma(i)} = 1$ for all $\sigma \in S'_n$.

For any matrix $A \in M_n(S)$, we have

$$\begin{aligned}\det(C \star A) &= \sum_{\sigma\in S'_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} c_{i\sigma(i)} a_{i\sigma(i)} \\ &= \sum_{\sigma\in S'_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} c_{i\sigma(i)} \prod_{i=1}^{n} a_{i\sigma(i)} \\ &= \sum_{\sigma\in S'_n} \prod_{i=1}^{n} a_{i\sigma(i)} \\ &= \operatorname{per}(A),\end{aligned}$$

hence, A is convertible. Since A is arbitrary, we conclude that $M_n(S)$ is a convertible subspace. □

Let $A = [a_{ij}]$ be an n-square matrix. We denote by $A(i;j)$ the $(n-1)$-square submatrix obtained from A after removing the i^{th} row and the j^{th} column. Generalizing, $A(i_1,\ldots,i_k; j_1,\ldots,j_k)$ denotes the square submatrix of A after removing the rows $i_1,\ldots,i_k$ and the columns $j_1,\ldots,j_k$.

Gibson proved the following result.

Lemma 2.6 ([3]). *Let $S = [s_{ij}]$ be a convertible matrix. If $s_{rs} = 1$, then $S(r;s)$ is also convertible.*

A subspace version of this Lemma is as follows.

Proposition 2.7. *If $M_n(S)$ is a convertible subspace, and $s_{ij} = 1$, then $M_{n-1}(S(i;j))$ is also a convertible subspace.*

Proof. It follows easy from Lemma 2.6, and Proposition 2.5. □

Corollary 2.8. *If $M_n(S)$ is a convertible subspace and $s_{i_1,j_1},\ldots,s_{i_k,j_k}$ are k nonzero elements of S, then $M_{n-k}(S(i_1,\ldots,i_k;j_1,\ldots,j_k))$ is also a convertible subspace.*

Proof. Trivial by induction. □

3 Proofs of the main results

We start this section with a result that comes easily from Proposition 2.5.

Proposition 3.1. *A connected graph G is well-numbered if and only if the correspondent full G-lower Hessenberg matrix of* 0*'s and* 1*'s is convertible.*

Let L_n be the anti-identity matrix of order n. This matrix has in position (i, j) the element 1, if $i + j = n + 1$, and 0 otherwise. Let A be an n-square matrix. We denote by A^{at} the matrix $A^{at} := L_n A^t L_n$, where A^t is the transpose of A. The matrix A^{at} is the *anti-transpose* of A.

Remark 3.2. Let $A = [a_{ij}]$ be an n-square matrix, and let $A^{at} = [a^{at}_{ij}]$. Then

1. $a^{at}_{ij} = a_{n-j+1,n-i+1}$, for all $i, j = 1, \dots, n$;
2. $(A^{at})^{at} = A$.

The next result allows simplifying some of the proofs.

Lemma 3.3. *Let G be a graph with n vertices. If G is well-numbered, then G is also well-numbered replacing vertex i by vertex $n - i + 1$, for all $i = 1, \dots, n$.*

Proof. Let G be a well-numbered graph with n vertices, and let S be the correspondent full G-lower Hessenberg matrix of 0's and 1's. Since G is well-numbered, by Proposition 2.3, S is permutation equivalent to a matrix whose row sum sequence is $\mathcal{R} = (2, 3, \dots, n-1, n, n)$, and the column sum sequence is $\mathcal{C} = (n, n, n-1, \dots, 3, 2)$. By definition S^{at} is also permutation equivalent to a matrix whose row and column sum sequences are equal to $\mathcal{R}$ and $\mathcal{C}$ respectively, and for all $i, j \in \{1, \dots, n\}$ such that $i \geq j$, $s^{at}_{ij} = 1$. Consider the new enumeration of the vertices of G where the vertex i is replaced by $n - i + 1$, for all $i = 1, \dots, n$, and let $S' = [s'_{i,j}]$ be the correspondent full G-lower Hessenberg matrix of 0's and 1's. Let $i, j \in \{1, \dots, n\}$ such that $i < j$, and i is adjacent to j in the initial enumeration of the vertices of G. Since $n - j + 1 < n - i + 1$ we conclude that $s'_{n-j+1,n-i+1} = 1 = s_{i,j}$. Therefore $(S')^{at} = S$, that is $S' = S^{at}$, and the result follows. □

Let G be a well-numbered connected simple graph with n vertices, and let $S = [s_{ij}]$ be the correspondent full G-lower Hessenberg matrix of 0's and 1's. Since G is connected, and $s_{ij} = 1$ whenever $i \geq j$ we conclude, by Theorem 2.2, that S has exact Ω_n nonzero entries, $n - 1$ of them above the main diagonal. By Definition 1.1 we conclude that G is a connected graph with n vertices and $n - 1$ edges, that is G is a tree. However, not all trees can be well-numbered.

Proof of Theorem 1.2. By Proposition 2.5 we only have to consider $(0, 1)$-matrices.

($\Leftarrow$): Suppose G is a caterpillar with numbering as shown in Figure 6. We will prove by induction on the number of nodes that such numbering produces a convertible full G-lower Hessenberg matrix.

If we have only one node (see Figure 7) the correspondent full G-lower Hessenberg

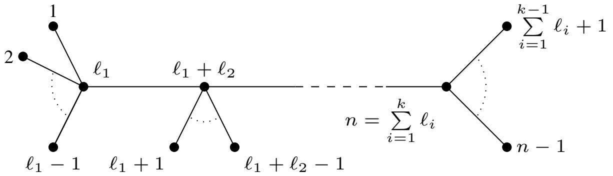


Figure 6: Numbering of the caterpillar G.

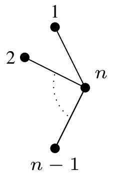


Figure 7: Special case with only one node.

matrix is

$$L_1=\begin{bmatrix}1&0&0&\dots&0&0&1\\1&1&0&\dots&0&0&1\\1&1&1&\dots&0&0&1\\\vdots&\vdots&\vdots&\ddots&\vdots&\vdots&\vdots\\1&1&1&\dots&1&0&1\\1&1&1&\dots&1&1&1\\1&1&1&\dots&1&1&1\end{bmatrix}.$$

This matrix is convertible by Proposition 2.3.

Let's suppose that the enumeration is valid for caterpillars with $k-1$ nodes, and that A is the correspondent full G-lower Hessenberg matrix. For a caterpillar with k nodes (see Figure 8) the correspondent full G-lower Hessenberg matrix is

$$S=\left[\begin{array}{c|c}\mathrm{A} & \begin{array}{c}\mathbf{0}\\\hline 0\ \dots\ 0\ 1\end{array}\\\hline \mathbb{1} & L_k\end{array}\right],$$

where $\mathbb{1}$ is a matrix whose entries are all 1's, the line above L_k corresponds to the edge between the last two nodes, and L_k is the full G-lower Hessenberg matrix arising from the last node and has structure as L_1.

By induction hypothesis A is convertible. Thus, by Proposition 2.3, A is permutation equivalent to a matrix with row- and column-sum sequence $\left(2,3,\dots,\sum_{i=1}^{k-1}\ell_i,\sum_{i=1}^{k-1}\ell_i\right)$. It is straightforward to see that S is permutation equivalent to a matrix with row- and column-sum sequence $\left(2,3,\dots,\sum_{i=1}^{k}\ell_i,\sum_{i=1}^{k}\ell_i\right)$. Hence, there exists at least one enumeration for the caterpillars that produces a convertible full G-lower Hessenberg matrix.

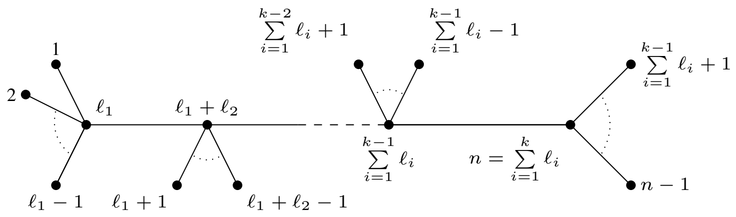

1
2
ℓ_1
$\ell_1 + \ell_2$
$\sum_{i=1}^{k-2} \ell_i + 1$
$\sum_{i=1}^{k-1} \ell_i - 1$
$\sum_{i=1}^{k-1} \ell_i + 1$
$\sum_{i=1}^{k-1} \ell_i$
$n = \sum_{i=1}^{k} \ell_i$
$n - 1$
$\ell_1 - 1$
$\ell_1 + 1$
$\ell_1 + \ell_2 - 1$

Figure 8: A caterpillar with k nodes.

($\Rightarrow$): We will prove by contradiction.

Every tree that is not a caterpillar has at least one node connected with three other nodes. Therefore every tree which is not a caterpillar contains the graph in Figure 9 as an induced subgraph. Let G be this graph and S be the full G-lower Hessenberg matrix of zeros and ones arising from an enumeration of the vertices of G. By Corollary 2.8, it is enough to prove that G cannot be well enumerated, that is, for every numbering of the vertices of G the rows or the columns of the correspondent full G-lower Hessenberg matrix of zeros and ones cannot be reordered to obtain $(2, 3, 4, 5, 6, 7, 7)$. By Proposition 3.3, we only have to consider the numberings where $v_3 \in \{1, 2, 3, 4\}$.

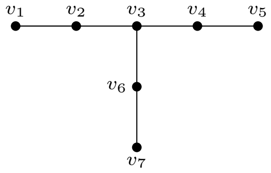


Figure 9: The subgraph of a tree which is not a caterpillar.

If $v_3 = 1$, then no row of S sums two.

If $v_3 = 2$, then the first row of S is the only one that the sum equals two. Thus, the vertex numbered with 1 must be a terminal and the second row of S sums 5. To be convertible, the third row of S is the only one whose sum equals three. So, we have the subgraph in Figure 10. Therefore, no row of S has four ones.

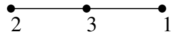


Figure 10: The subgraph in case of $v_3 = 2$.

If $v_3 = 3$, then to have row-sums equal to two and three, and two column-sums equal to seven, vertices numbered with 1 and 2 must be adjacent. Similarly, to have two row-sums equal to seven vertices numbered with 6 and 7 are adjacent. Thus, no row has four ones.

If $v_3 = 4$, then vertices numbered with 1 and 2 cannot be simultaneously adjacent to v_3 (otherwise, no row of S has two ones). The vertices numbered with 1 and 2 must be adjacent otherwise S doesn't have two column-sums equal to seven. By Lemma 3.3, vertices numbered with 6 and 7 must also be adjacent. Hence, if a row sums to four then no column can sum to four, and vice-versa. □

Definition 3.4. Let T be a tree, and let R be a star with central vertex x. We say that R is a pendant star of T if:

1. R is an induced subgraph of T;
2. If we remove all the vertices of R, then we obtain a tree denoted by T/R;
3. x is adjacent to a single vertex of T/R.

Remark 3.5. Note that every caterpillar with at least one edge has exactly two pendant stars.

For proving Theorem 1.5, some lemmas are needed:

Lemma 3.6. *Let $G = C_t[r]$ with $n = tr$ vertices. If G is well-numbered, then the vertices numbered with $1, 2, \dots, r-1$, must be in the same pendant star, and the vertices numbered with $n, n-1, \dots, n-r+1$ must be in the other pendant star.*

Proof. We only need to prove that the vertices of G numbered with $1, 2, \dots, r-1$, must be in the same pendant star, because by Lemma 3.3 we conclude that $n, n-1, \dots, n-r+1$ must be in the other pendant star. If G is well-numbered, then the correspondent full G-lower Hessenberg matrix S of 0's and 1's is convertible. Therefore S is permutation equivalent to a matrix whose row-sum sequence is $\mathcal{R} = (2, 3, \dots, n-1, n, n)$, and column-sum sequence is $\mathcal{C} = (n, n, n-1, \dots, 3, 2)$. An i^{th} column of S has n 1's if and only if the vertices numbered with $1, \dots, i-1$ are adjacent to the vertex numbered with i. Since $G = C_t[r]$, the maximum degree of a vertex in G is $r+1$, and then the two columns of S with n 1's must be in the first $r+2$ columns. Taking into account that caterpillars are sequences of stars connected by central vertices, we start showing that the vertices numbered with $1, \dots, r-1$ must be one of these stars, that is, the induced subgraph of $G = C_t[r]$ that is the one of the stars involved in the construction of G. After that, we prove that this star is a pendant one.

Assume that the vertices numbered with $1, \dots, r-1$ are not in the same star. Denote by R the star containing 1, and let j be the least integer less that r not in R. Denote by R' the star containing j. Several cases are needed to consider.

Case 1: If $1, \dots, j-1, j$ are pendant vertices, then having in mind the previous discussion, there are no two columns of S with n 1's, which is a contradiction.

Case 2: If $\ell \in \{1, \dots, j-1\}$ is the central vertex, and j is a pendant vertex, then the ℓ^{th} column of S has n 1's. If $1 \leq i \leq \ell-1$, then the i^{th} row sums $i+1$, the ℓ^{th} row sums at least $\ell + (r-1) - (\ell-1) + 1 = r+1$, and the i^{th} row, $\ell+1 \leq i \leq j-1$, sums i. Since j is a pendant vertex of R', we conclude that no row of S sums j, which is a contradiction.

Case 3: If the vertices $1, \dots, j-1$ are pendant vertices, and j is the central vertex of R', then i^{th} row sums to $i+1$, for all $1 \leq i \leq j-1$, and the j^{th} row sums $j + (r-1) + 1$ or $j + (r-1) + 2$. Since the only row that can sum $j+1$ is the row $j+1$, the vertex $j+1$ must be a pendant one of R'. Hence, S does not have two columns with n 1's, which is a contradiction.

Case 4: If $\ell \in \{1, \dots, j-1\}$ is the central vertex of R, and j is the central vertex of R', then the i^{th} row of S, $1 \leq i \leq \ell-1$, sums $i+1$, the ℓ^{th} row sums at least $\ell + (r-1) - (\ell-1) + 1 = r+1$, and the i^{th} row of S, $\ell+1 \leq i \leq j-1$, sums i. Since j is a central vertex, no row of S sums j, which is a contradiction.

We have proved that the vertices of G numbered with $1, \dots, r-1$ must be in the same star R. Now we will prove that R is pendant.

Suppose that this is not the case. Let ℓ, $1 \le \ell \le r-1$ be the central vertex of R. Then the i^{th} row of S, $1 \le i \le \ell - 1$, sums $i+1$, the ℓ^{th} row sums $\ell + (r-1) - (\ell-1) + 2 = r+2$, because R is not pendant, and the i^{th} row, $\ell+1 \le i \le j-1$, sums i. Since the ℓ^{th} row sums $r+2$, the $(r+1)^{\text{th}}$ row must sum $r+1$ and the r^{th} row must sum r. Then the vertices r and $r+1$ must be pendant vertices of R, which is a impossible.

If $1, \dots, r-1$ are pendant vertices of R, the central vertex of this star must be numbered with r, $r+1$ or $r+2$, otherwise there are no two columns of S with n 1's.

- For r, since R is not pendant, there are no row of S with $r+1$ 1's.
- For $r+1$, the vertex r must be adjacent to $r+1$, otherwise there are no two columns of S with n 1's. Then r is the central vertex of another star. Then, the r^{th} row of S sums at least $r + (r-1) + 1 = 2r$, the $(r+1)^{\text{th}}$ row sums at least $r+2$. Then S does not have a row with $r+1$ 1's.
- Finally, for $r+2$, by a similar procedure, we conclude that there is no row of S with $r+1$ 1's. □

Lemma 3.7. *Let $G = C_t[r]$. If G is well-numbered, then the vertices of one of the pendant stars R are numbered with 1 up to $r+1$. If $r+1$ is a vertex of R, then $r+1$ is central, and r is adjacent to $r+1$. If r is a vertex of R, then the central vertex can be any ℓ, $1 \le \ell \le r$.*

Proof. In the previous Lemma, we proved that $1, \dots, r-1$ must lay all in a pendant star. We only have to prove that the remaining vertex must be numbered with r or $r+1$. We have to consider two cases:

Case 1: If the pendant vertices of R are numbered with $1, \dots, r-1$, then the central vertex can only be numbered with r or $r+1$. Otherwise, S does not have two columns with n 1's. By the same reason, if $r+1$ is central, then r must be adjacent.

Case 2: If the central vertex of R is ℓ, $1 \le \ell \le r-1$, then there remains a pendant vertex. Suppose that this pendant vertex is $j \ge r+1$. In this case, the row r sums at least $r+1$, and no row sums r. Hence, R is numbered with $1, \dots, r$.

If r is a vertex of R, and ℓ, $1 \le \ell \le r$, is the central vertex, then the correspondent full G-lower Hessenberg matrix of 0' s and 1' s satisfies the condition of Proposition 2.3. □

Corollary 3.8. *Let $G = C_t[r]$. If G is well-numbered, then the vertices of one of the pendant stars R are numbered with $n = tr$ down to $n-1-r$. If $n-1-r$ is a vertex of R, then $n-r-1$ is central and $n-r$ is adjacent to $n-1-r$. If $n-r$ is a vertex of R, then the central vertex can be any ℓ, $n-r \le \ell \le n$.*

Proof. It follows directly from Lemmas 3.3 and 3.7. □

We are now in condition to prove our second main result:

Proof of Theorem 1.5. The proof is by induction on t, the number of stars in a caterpillar.

If $t = 1$, then $r \ge 3$. By Lemma 3.7, the central vertex of G can be numbered with any $\ell \in \{1, \dots, r\}$. So, the number of different convertible subspaces is $a_2 = r$.

If $t = 2$, then G has $n = 2r$, $r \ge 2$, vertices. By Lemma 3.7, each star has r different ways to be well numbered by consecutive numbers. Therefore, this gives r^2 different well numberings for G with each star having consecutive numbers. Besides these, by the same Lemma, it is possible to interchange vertices r and $r+1$, giving another convertible subspace. So the total number of different convertible subspaces is $a_3 = r^2 + 1$.

Let $t > 2$, and assume that the theorem holds for all $j \leq t$. Let $G = C_{t+1}[r]$. By hypothesis there are a_{t+1} convertible subspaces that arise from the different numberings of the caterpillar $C_t[r]$. The caterpillar G is obtained from $C_t[r]$ by augmenting with a star with r vertices. By induction hypothesis and Corollary 3.8, there are ra_{t+1} convertible subspaces that arise from a numbering of G, where the vertices of the new pendant star are numbered consecutively with $tr+1, \dots, (t+1)r$. There are also convertible subspaces that arise from a numbering of G, where the central vertex of new pendant star is numbered with tr. By induction hypothesis, this number is a_t. Then the total number of convertible subspaces that arise from the different numberings of the vertices of G is $ra_{t+1} + a_t = a_{t+2}$. □

As we already saw, the number of well numberings of the vertices of a path with t vertices is the $(t+1)^{\text{th}}$ Fibonacci number. In Table 1, we present all well-numbered paths of lengths 3, 4, and 5.

Table 1: All well-numbered paths with 3, 4, and 5 vertices.

$n = 3$	1 2 3	2 1 3	1 3 2
$n = 4$	1 2 3 4	2 1 3 4	1 3 2 4
	1 2 4 3	2 1 4 3	
$n = 5$	1 2 3 4 5	2 1 3 4 5	1 3 2 4 5
	1 2 4 3 5	2 1 4 3 5	1 3 2 5 4
	1 2 3 5 4	2 1 3 5 4	

References

[1] M. P. Coelho and M. A. Duffner, Subspaces where an immanant is convertible into its conjugate, *Linear Multilinear Algebra* **48** (2001), 383–408, doi:10.1080/03081080108818681.

[2] C. M. da Fonseca, An identity between the determinant and the permanent of Hessenberg-type matrices, *Czechoslovak Math. J.* **61** (2011), 917–921, doi:10.1007/s10587-011-0059-1.

[3] P. M. Gibson, An identity between permanents and determinants, *Amer. Math. Monthly* **76** (1969), 270–271, doi:10.2307/2316368.

[4] P. M. Gibson, Conversion of the permanent into the determinant, *Proc. Amer. Math. Soc.* **27** (1971), 471–476, doi:10.2307/2036477.

[5] A. Guterman, G. Dolinar and B. Kuzma, Pólya convertibility problem for symmetric matrices, *Math. Notes* **92** (2012), 624–635, doi:10.1134/s0001434612110053.

[6] G. Pólya, Aufgabe 424, *Arch. Math. Phys. Ser. 3* **20** (1913), 271.

[7] N. J. A. Sloane (ed.), The On-Line Encyclopedia of Integer Sequences, published electronically at `https://oeis.org`.

[8] G. Szegö, Lösung zu Aufgabe 424, *Arch. Math. Phys. Ser. 3* **21** (1913), 291–292.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 487–525
https://doi.org/10.26493/1855-3974.1330.916
(Also available at http://amc-journal.eu)

Linear separation of connected dominating sets in graphs[*]

Nina Chiarelli [†], Martin Milanič [‡]

University of Primorska, UP FAMNIT, Glagoljaška 8, SI-6000 Koper, Slovenia
University of Primorska, UP IAM, Muzejski trg 2, SI-6000 Koper, Slovenia

Received 21 February 2017, accepted 12 November 2018, published online 15 March 2019

Abstract

A connected dominating set in a graph is a dominating set of vertices that induces a connected subgraph. Following analogous studies in the literature related to independent sets, dominating sets, and total dominating sets, we study in this paper the class of graphs in which the connected dominating sets can be separated from the other vertex subsets by a linear weight function. More precisely, we say that a graph is connected-domishold if it admits non-negative real weights associated to its vertices such that a set of vertices is a connected dominating set if and only if the sum of the corresponding weights exceeds a certain threshold. We characterize the graphs in this non-hereditary class in terms of a property of the set of minimal cutsets of the graph. We give several characterizations for the hereditary case, that is, when each connected induced subgraph is required to be connected-domishold. The characterization by forbidden induced subgraphs implies that the class properly generalizes two well known classes of chordal graphs, the block graphs and the trivially perfect graphs. Finally, we study certain algorithmic aspects of connected-domishold graphs. Building on connections with minimal cutsets and properties of the derived hypergraphs and Boolean functions, we show that our approach leads to new polynomially solvable cases of the weighted connected dominating set problem.

Keywords: Connected dominating set, connected domination, connected-domishold graph, forbidden induced subgraph characterization, split graph, chordal graph, minimal cutset, minimal separator, 1*-Sperner hypergraph, threshold hypergraph, threshold Boolean function, polynomial-time algorithm.*

[*] The authors are grateful to the anonymous reviewers for the many insightful remarks, to Endre Boros, Andreas Brandstädt, Yves Crama, Vladimir Gurvich, Pinar Heggernes, Haiko Müller, and Vito Vitrih for helpful discussions, and to Dieter Kratsch and Haiko Müller for providing them with a copy of paper [43]. The work for this paper was partly done in the framework of a bilateral project between Argentina and Slovenia, financed by the Slovenian Research Agency (BI-AR/15–17–009) and MINCYT-MHEST (SLO/14/09).

Part of this work appeared as an extended abstract in [17].

[†] This work is supported in part by the Slovenian Research Agency (research project J1-7051).

[‡] This work is supported in part by the Slovenian Research Agency (research core funding No. I0-0035 and P1-0285, and projects N1-0032, J1-7051, and J1-9110).

Math. Subj. Class.: 05C69, 05C75, 05C65, 05C85

1 Introduction

1.1 Background

Threshold concepts have been a subject of investigation for various discrete structures, including graphs [18,20,48], Boolean functions [19,22,29,32,53,55], and hypergraphs [34, 58]. A common theme of these studies is a quest for necessary and sufficient conditions for the property that a given combinatorial structure defined over some finite ground set U admits non-negative real weights associated to elements of U such that a subset of U satisfies a certain property, say π, if and only if the sum of the corresponding weights exceeds a certain threshold. A more general framework has also been proposed, where the requirement is that a subset of U satisfies property π if and only if the sum of the corresponding weights belongs to a set T of thresholds given by a membership oracle [50].

Having the set U equipped with weights as above can have useful algorithmic implications. Consider for example the optimization problem of finding a subset of U with property π that has either maximum or minimum cost (according to a given linear cost function on the elements of the ground set). It was shown in [50] that if the weights as above are known and integer, then the problem can be solved by a dynamic programming approach in time $\mathcal{O}(|U|M)$ and with M calls of the membership oracle, where M is a given upper bound for T. The pseudo-polynomial running time should be expected, since the problem is very general and captures also the well-known knapsack problem [41]. Note, however, that the problem admits a much simpler, polynomial-time solution in the special case when the costs are unit and if we assume the monotone framework, where a set satisfies property π as soon as its total weight exceeds a certain threshold. Under these assumptions, a minimum-sized subset of U satisfying property π can be found by a simple greedy algorithm starting with the empty set and adding the elements in order of non-increasing weight until the threshold is exceeded.

Many interesting graph classes can be defined within the above framework, including threshold graphs [20, 42, 48], domishold graphs [1], total domishold graphs [16, 18], equistable graphs [54], and equidominating graphs [54]. In general, the properties of the resulting graph classes depend both on the choice of property π and on the constraints imposed on the structure of the set T of thresholds. For example, if U is the vertex set of a graph, property π denotes the property of being an independent (stable) set in a graph, and T is restricted to be an interval unbounded from below, we obtain the class of threshold graphs, which is a very well understood class of graphs, admitting many characterizations and linear-time algorithms for recognition and various optimization problems (see, e.g., [48]). If π denotes the property of being a dominating set and T is an interval unbounded from above, we obtain the class of domishold graphs, which enjoys similar properties as the class of threshold graphs. On the other hand, if π is the property of being a *maximal* stable set and T is restricted to consist of a single number, we obtain the class of equistable graphs, for which the recognition complexity is open (see, e.g., [47]), no structural characterization is known, and several NP-hard optimization problems remain intractable [50].

Notions and results from the theory of Boolean functions [22] and hypergraphs [2] can

E-mail addresses: nina.chiarelli@famnit.upr.si (Nina Chiarelli), martin.milanic@upr.si (Martin Milanič)

be useful for the study of graph classes defined within the above framework. For instance, the characterization of hereditarily total domishold graphs in terms of forbidden induced subgraphs given in [18] is based on the facts that every threshold Boolean function is 2-asummable [19] and that every dually Sperner hypergraph is threshold [16].[1] Moreover, the fact that threshold Boolean functions are closed under dualization and (when given by their complete DNF) can be recognized in polynomial time [55] leads to efficient algorithms for recognizing total domishold graphs and for finding a minimum total dominating set in a given total domishold graph [16]. The relationship also goes the other way around. For instance, total domishold graphs can be used to characterize threshold hypergraphs and threshold Boolean functions [18].

1.2 Aims and motivation

The aim of this paper is to further explore and exploit this fruitful interplay between threshold concepts in graphs, hypergraphs, and Boolean functions. We do this by studying the class of *connected-domishold* graphs, a new class of graphs that can be defined in the above framework, as follows. A *connected dominating set* (*CD set* for short) in a connected graph G is a set S of vertices of G that is *dominating*, that is, every vertex of G is either in S or has a neighbor in S, and *connected*, that is, the subgraph of G induced by S is connected. The ground set U is the vertex set of a connected graph $G = (V, E)$, property π is the property of being a connected dominating set in G, and T is any interval unbounded from above.

Our motivations for studying the notion of connected domination in the above threshold framework are twofold. First, connected domination is one of the most basic of the many variants of domination, with applications in modeling wireless networks, see, e.g., [6, 11, 12, 26, 27, 31, 35, 36, 60–62, 66]. The connected dominating set problem is the problem of finding a minimum connected dominating set in a given connected graph. This problem is NP-hard (and hard to approximate) for general graphs and remains intractable even under significant restrictions, for instance, for the class of split graphs (see Section 6.2). On the other hand, as outlined above, the problem is polynomially solvable in the class of connected-domishold graphs equipped with weights as in the definition. This motivates the study of connected-domishold graphs. In particular, identification of subclasses of connected-domishold graphs might lead to new classes of graphs where the connected dominating set problem (or its weighted version) is polynomially solvable.

Second, despite the increasingly large variety of graph domination concepts studied in the literature (see, e.g., [35, 36]), so far a relatively small number of "threshold-like" graph classes was studied with respect to notions of domination: the classes of domishold and equidominating graphs (corresponding to the usual domination), the class of equistable graphs (corresponding to independent domination), and the class of total domishold graphs (corresponding to total domination). These graph classes differ significantly with respect to their structural and algorithmic properties. For instance, while the class of domishold graphs is a highly structured hereditary subclass of cographs, the classes of equistable and of total domishold graphs are not contained in any nontrivial hereditary class of graphs and are not structurally understood.[2] In particular, the class of total domishold graphs is as rich in its combinatorial structure as the class of threshold hypergraphs [18], for which (despite

[1] In [16, 18], the hereditarily total domishold graphs were named hereditary total domishold graphs. We prefer to adopt the grammatically correct term "hereditarily total domishold".

[2] A class of graphs is said to be *hereditary* if it is closed under vertex deletion.

being recognizable in polynomial time via linear programming [22, 55]) the existence of a "purely combinatorial" polynomial-time recognition algorithm is an open problem [22]. These results, differences, and challenges provide further motivation for the study of structural and algorithmic properties of connected-domishold graphs.

1.3 The definition

Since a disconnected graph G does not have any connected dominating sets, we restrict our attention to connected graphs in the following definition.

Definition 1.1. A connected graph $G = (V, E)$ is said to be *connected-domishold* (*CD* for short) if there exists a pair (w, t) where $w\colon V \to \mathbb{R}_+$ is a weight function and $t \in \mathbb{R}_+$ is a threshold such that for every subset $S \subseteq V$, $w(S) := \sum_{x\in S} w(x) \geq t$ if and only if S is a connected dominating set in G. Such a pair (w, t) will be referred to as a *connected-domishold (CD) structure* of G.

We emphasize that the class of connected-domishold graphs is not the intersection of the classes of connected and domishold graphs. In fact, the two classes are incomparable: the 4-vertex cycle is connected and domishold [1] but not connected-domishold, see Example 1.3 below; the 4-vertex path is connected-domishold but not domishold. The hyphen in the name is meant to indicate this fact.

Example 1.2. The complete graph of order n is connected-domishold. Indeed, any nonempty subset $S \subseteq V(K_n)$ is a connected dominating set of K_n, and the pair $(w, 1)$ where $w(x) = 1$ for all $x \in V(K_n)$ is a CD structure of K_n.

Example 1.3. The 4-cycle C_4 is not connected-domishold. Denoting its vertices by v_1, v_2, v_3, v_4 in a cyclic order, we see that a subset $S \subseteq V(C_4)$ is CD if and only if it contains an edge. Therefore, if (w, t) is a CD structure of C_4, then $w(v_i) + w(v_{i+1}) \geq t$ for all $i \in \{1, 2, 3, 4\}$ (indices modulo 4), which implies $w(V(C_4)) \geq 2t$. On the other hand, $w(v_1) + w(v_3) < t$ and $w(v_2) + w(v_4) < t$, implying $w(V(C_4)) < 2t$.

1.4 Overview of results

Our results can be divided into four interrelated parts and can be summarized as follows:

1) Characterizations in terms of derived hypergraphs (resp., derived Boolean functions); a necessary and a sufficient condition.

In a previous work [18, Proposition 4.1 and Theorem 4.5], total domishold graphs were characterized in terms of thresholdness of a derived hypergraph and a derived Boolean function. We give similar characterizations of connected-domishold graphs (Proposition 3.4). The characterizations lead to a necessary and a sufficient condition for a graph to be connected-domishold, respectively, expressed in terms of properties of the derived hypergraph (equivalently: of the derived Boolean function; Corollary 3.5).

2) The case of split graphs. A characterization of threshold hypergraphs.

While the classes of connected-domishold and total domishold graphs are in general incomparable, we show that they coincide within the class of connected split graphs (Theorem 4.3). Building on this equivalence, we characterize threshold hypergraphs in terms of the connected-domisholdness property of a derived split graph (Theorem 4.4).

We also give examples of connected split graphs showing that neither of the two conditions for a graph to be connected-domishold mentioned above (one necessary and one sufficient) characterizes this property.

3) **The hereditary case.**

We observe that, contrary to the classes of threshold and domishold graphs, the class of connected-domishold graphs is not hereditary. This motivates the study of so-called *hereditarily connected-domishold graphs*, defined as graphs every connected induced subgraph of which is connected-domishold. As our main result (Theorem 5.4), we give several characterizations of the class of hereditarily connected-domishold graphs. The characterizations in terms of forbidden induced subgraphs implies that the class of hereditarily connected-domishold graphs is a subclass of the class of chordal graphs properly containing two well known classes of chordal graphs, the class of block graphs and the class of trivially perfect graphs.

4) **Algorithmic aspects via vertex separators.**

Finally, we build on all these results, together with some known results from the literature on connected dominating sets and minimal vertex separators in graphs, to study certain algorithmic aspects of the class of connected-domishold graphs and their hereditary variant. We identify a sufficient condition, capturing a large number of known graph classes, under which the CD property can be recognized efficiently (Theorem 6.1). We also show that the same condition, when applied to classes of connected-domishold graphs, results in classes of graphs for which the weighted connected dominating set problem (which is NP-hard even on split graphs) is polynomial-time solvable (Theorem 6.5). This includes the classes of hereditarily connected-domishold graphs and F_2-free split graphs (see Figure 1), leading to new polynomially solvable cases of the problem.

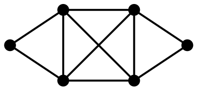


Figure 1: Graph F_2.

Structure of the paper. In Section 2, we state the necessary definitions and preliminary results on graphs, hypergraphs, and Boolean functions. In Section 3, we give characterizations of connected-domishold graphs in terms of thresholdness of derived hypergraphs and Boolean functions. Connected-domishold split graphs are studied in Section 4, where their relation to threshold hypergraphs is also observed. The main result of the paper, Theorem 5.4, is stated in Section 5, where some of its consequences are also derived. Section 6 discusses some algorithmic aspects of connected-domishold graphs. Our proof of Theorem 5.4 relies on a technical lemma, which is proved in Section 7.

2 Preliminaries

2.1 Graphs

All graphs in this paper will be finite, simple and undirected. The *(open) neighborhood* of a vertex v is the set of vertices in a graph G adjacent to v, denoted by $N_G(v)$ (or simply

$N(v)$ if the graph is clear from the context); the *closed neighborhood* of v is denoted by $N_G[v]$ and defined as $N_G(v) \cup \{v\}$. The *degree* of a vertex v in a graph G is the cardinality of its neighborhood. The complete graph, the path and the cycle of order n are denoted by K_n, P_n and C_n, respectively. A *clique* in a graph is a subset of pairwise adjacent vertices, and an *independent* (or *stable*) set is a subset of pairwise non-adjacent vertices. A *universal vertex* in a graph G is a vertex adjacent to all other vertices. For a set S of vertices in a graph G, we denote by $G[S]$ the subgraph of G induced by S. For a set $\mathcal{F}$ of graphs, we say that a graph is $\mathcal{F}$-free if it does not contain any induced subgraph isomorphic to a member of $\mathcal{F}$. Given a graph G, a vertex $v \in V(G)$, and a set $U \subseteq V(G) \setminus \{v\}$, we say that v *dominates* U if v is adjacent to every vertex in U.

The main notion that will provide the link between threshold Boolean functions and hypergraphs is that of cutsets in graphs. A *cutset* in a graph G is a set $S \subseteq V(G)$ such that $G - S$ is disconnected. A cutset is *minimal* if it does not contain any other cutset. For a pair of disjoint vertex sets A and B in a graph G such that no vertex in A has a neighbor in B, an *A, B-separator* is a set of vertices $S \subseteq V(G) \setminus (A \cup B)$ such that A and B are in different components of $G - S$. An A, B-separator is said to be *minimal* if it does not contain any other A, B-separator. When sets A and B are singletons, say $A = \{u\}$ and $B = \{v\}$, we will refer to a (minimal) A, B-separator simply as a *(minimal) u, v-separator*. A *minimal vertex separator* in G is a minimal u, v-separator for some non-adjacent vertex pair u, v. Note that every minimal cutset of G is a minimal vertex separator, but not vice versa. The minimal cutsets are exactly the minimal u, v-separators that do not contain any other x, y-separator. The connection between the CD graphs and the derived hypergraphs and Boolean functions will be developed in Section 3 using the following characterization of CD sets due to Kanté et al. [38].

Proposition 2.1 (Kanté et al. [38]). *In every connected graph G that is not complete, a subset $D \subseteq V(G)$ is a CD set if and only if $D \cap S \neq \emptyset$ for every minimal cutset S in G.*

In other words, unless a connected graph G is complete, its CD sets are exactly the transversals of the cutset hypergraph of G (see Section 2.3 and Definition 3.2 for definitions of these notions).

A graph G is *chordal* if it does not contain any induced cycle of order at least 4, and *split* if it has a *split partition*, that is, a partition of its vertex set into a clique and an independent set. One of our proofs (the proof of Theorem 5.4) will rely on the following property of chordal graphs.

Lemma 2.2 (Kumar and Veni Madhavan [45]). *If S is a minimal cutset of a chordal graph G, then each connected component of $G - S$ has a vertex that is adjacent to all the vertices of S.*

For graph theoretic notions not defined above, see, e.g., [65].

2.2 Boolean functions

Let n be a positive integer. Given two vectors $x, y \in \{0,1\}^n$, we write $x \leq y$ if $x_i \leq y_i$ for all $i \in [n] := \{1, \ldots, n\}$. A Boolean function $f\colon \{0,1\}^n \to \{0,1\}$ is *positive* (or: *monotone*) if $f(x) \leq f(y)$ holds for every two vectors $x, y \in \{0,1\}^n$ such that $x \leq y$. A *literal* of f is either a variable, x_i, or the negation of a variable, denoted by $\overline{x_i}$. An *elementary conjunction* is an expression of the form $C = \big(\bigwedge_{i \in A} x_i\big) \wedge \big(\bigwedge_{j \in B} \overline{x_j}\big)$ where

$A \cap B = \emptyset$. An *implicant* of a Boolean function f is an elementary conjunction C such that $f(x) = 1$ for all $x \in \{0,1\}^n$ for which C takes value 1 (we also say that C *implies* f). An implicant is said to be *prime* if it is not implied by any other implicant. If f is positive, then none of the variables appearing in any of its prime implicants appears negated. Every n-variable positive Boolean function f can be expressed with its *complete DNF (disjunctive normal form)*, defined as the disjunction of all prime implicants of f.

A positive Boolean function f is said to be *threshold* if there exist non-negative real weights $w = (w_1, \ldots, w_n)$ and a non-negative real number t such that for every $x \in \{0,1\}^n$, $f(x) = 0$ if and only if $\sum_{i=1}^{n} w_i x_i \leq t$. Such a pair (w, t) is called a *separating structure* of f. Every threshold Boolean function admits an integral separating structure (see [22, Theorem 9.5]). A positive Boolean function $f(x_1, \ldots, x_n)$ is threshold if and only if its *dual function* $f^d(x) = \overline{f(\overline{x})}$ is threshold [22]; moreover, if $(w_1, \ldots, w_n, t)$ is an integral separating structure of f, then $(w_1, \ldots, w_n, \sum_{i=1}^{n} w_i - t - 1)$ is a separating structure of f^d.

Threshold Boolean functions have been characterized in [19] and [29], as follows. A *false point* of f is an input vector $x \in \{0,1\}^n$ such that $f(x) = 0$; a *true point* is defined analogously. For $k \geq 2$, a positive Boolean function $f\colon \{0,1\}^n \to \{0,1\}$ is said to be k-*summable* if, for some $r \in \{2, \ldots, k\}$, there exist r (not necessarily distinct) false points of f, say, $x^1, x^2, \ldots, x^r$, and r (not necessarily distinct) true points of f, say $y^1, y^2, \ldots, y^r$, such that $\sum_{i=1}^{r} x^i = \sum_{i=1}^{r} y^i$ (note that the sums are in $\mathbb{Z}^n$ and not in $\mathbb{Z}_2^n$, the n-dimensional vector space over $\mathrm{GF}(2)$). Function f is said to be k-*asummable* if it is not k-summable, and it is *asummable* if it is k-asummable for all $k \geq 2$.

Theorem 2.3 (Chow [19], Elgot [29], see also [22, Theorem 9.14]). *A positive Boolean function f is threshold if and only if it is asummable.*

The problem of determining whether a positive Boolean function given by its complete DNF is threshold is solvable in polynomial time, using dualization and linear programming (see [55] and [22, Theorem 9.16]). The algorithm tests if a polynomially sized derived linear program has a feasible solution, and in case of a yes instance, the solution found yields a separating structure of the given function. Using, e.g., Karmarkar's interior point method for linear programming [39], one can assure that a rational solution is found. This results in a rational separating structure, which can be easily turned into an integral one. We summarize this result as follows.

Theorem 2.4. *There exists a polynomial-time algorithm for recognizing threshold Boolean functions given by the complete DNF. In case of a yes instance, the algorithm also computes an integral separating structure of the given function.*

Remark 2.5. The existence of a "purely combinatorial" polynomial-time recognition algorithm for threshold Boolean functions (that is, one not relying on solving an auxiliary linear program) is an open problem [22].

A similar approach as the one outlined above shows that every connected-domishold graph has an integral CD structure; we will often use this fact in the paper. For further background on Boolean functions, we refer to the comprehensive monograph by Crama and Hammer [22].

2.3 Hypergraphs

A *hypergraph* is a pair $\mathcal{H} = (V, E)$ where V is a finite set of *vertices* and E is a set of subsets of V, called *hyperedges* [2]. When the vertex set or the hyperedge set of $\mathcal{H}$ will not be explicitly given, we will refer to them by $V(\mathcal{H})$ and $E(\mathcal{H})$, respectively. A *transversal* (or: *hitting set*) of $\mathcal{H}$ is a set $S \subseteq V$ such that $S \cap e \neq \emptyset$ for all $e \in E$. A hypergraph $\mathcal{H} = (V, E)$ is *threshold* if there exist a weight function $w\colon V \to \mathbb{R}_+$ and a threshold $t \in \mathbb{R}_+$ such that for all subsets $X \subseteq V$, it holds that $w(X) \leq t$ if and only if X contains no hyperedge of $\mathcal{H}$ [34]. Such a pair (w, t) is said to be a *separating structure* of $\mathcal{H}$.

To every hypergraph $\mathcal{H} = (V, E)$, we can naturally associate a positive Boolean function $f_{\mathcal{H}}\colon \{0,1\}^V \to \{0,1\}$, defined by the positive DNF expression

$$f_{\mathcal{H}}(x) = \bigvee_{e \in E} \bigwedge_{u \in e} x_u$$

for all $x \in \{0,1\}^V$. Conversely, to every positive Boolean function $f\colon \{0,1\}^n \to \{0,1\}$ given by a positive DNF $\phi = \bigvee_{j=1}^{m} \bigwedge_{i \in C_j} x_i$, we can associate a hypergraph $\mathcal{H}(\phi) = (V, E)$ as follows: $V = [n]$ and $E = \{C_1, \ldots, C_m\}$. It follows directly from the definitions that the thresholdness of hypergraphs and of Boolean functions are related as follows.

Proposition 2.6. *A hypergraph $\mathcal{H} = (V, E)$ is threshold if and only if the positive Boolean function $f_{\mathcal{H}}$ is threshold. A positive Boolean function given by a positive DNF $\phi = \bigvee_{j=1}^{m} \bigwedge_{i \in C_j} x_i$ is threshold if and only if the hypergraph $\mathcal{H}(\phi)$ is threshold.*

Applying Theorem 2.3 to the language of hypergraphs gives the following characterization of threshold hypergraphs. For $k \geq 2$, a hypergraph $\mathcal{H} = (V, E)$ is said to be *k-summable* if, for some $r \in \{2, \ldots, k\}$, there exist r (not necessarily distinct) subsets $A_1, \ldots, A_r$ of V such that each A_i contains a hyperedge of $\mathcal{H}$, and r (not necessarily distinct) subsets $B_1, \ldots, B_r$ of V such that each B_i does not contain a hyperedge of $\mathcal{H}$ and such that for every vertex $v \in V$, we have:

$$|\{i : v \in A_i\}| = |\{i : v \in B_i\}|. \tag{2.1}$$

We say that a hypergraph $\mathcal{H}$ is *k-asummable* if it is not k-summable and it is *asummable* if it is k-asummable for all $k \geq 2$.

Corollary 2.7. *A hypergraph $\mathcal{H}$ is threshold if and only if it is asummable.*

A hypergraph $\mathcal{H} = (V, E)$ is said to be *Sperner* (or: a *clutter*) if no hyperedge of $\mathcal{H}$ contains another hyperedge, that is, if for every two distinct hyperedges e and f of $\mathcal{H}$, it holds that $\min\{|e \setminus f|, |f \setminus e|\} \geq 1$. Chiarelli and Milanič defined in [16, 18] the notion of *dually Sperner hypergraphs* as the hypergraphs such that the inequality $\min\{|e \setminus f|, |f \setminus e|\} \leq 1$ holds for every pair of distinct hyperedges e and f of $\mathcal{H}$. It was proved in [16, 18] that dually Sperner hypergraphs are threshold; they were applied in the characterizations of total domishold graphs and their hereditary variant. Boros et al. introduced in [8] the following restriction of dually Sperner hypergraphs.

Definition 2.8 (Boros et al. [8]). A hypergraph $\mathcal{H} = (V, E)$ is said to be *1-Sperner* if for every two distinct hyperedges e and f of $\mathcal{H}$, it holds that $\min\{|e \setminus f|, |f \setminus e|\} = 1$.

Note that a hypergraph is 1-Sperner if and only if it is both Sperner and dually Sperner. In particular, for Sperner hypergraphs the notions of dually Sperner and 1-Sperner hypergraphs coincide. Since a hypergraph $\mathcal{H}$ is threshold if and only if the Sperner hypergraph obtained from $\mathcal{H}$ by keeping only its inclusion-wise minimal hyperedges is threshold, the fact that dually Sperner hypergraphs are threshold is equivalent to the following fact, proved constructively by Boros et al. in [8] using a composition result for 1-Sperner hypergraphs developed therein.

Theorem 2.9 (Chiarelli and Milanič [18], Boros et al. [8]). *Every* 1*-Sperner hypergraph is threshold.*

3 Connected-domishold graphs via hypergraphs and Boolean functions

In a previous work [18, Proposition 4.1 and Theorem 4.5], total domishold graphs were characterized in terms of thresholdness of a derived hypergraph and a derived Boolean function. In this section we give similar characterizations of connected-domishold graphs.

We first recall some relevant definitions and a result from [18]. A *total dominating set* in a graph G is a set $S \subseteq V(G)$ such that every vertex of G has a neighbor in S. Note that only graphs without isolated vertices have total dominating sets. A graph $G = (V, E)$ is said to be *total domishold* (*TD* for short) if there exists a pair (w, t) where $w\colon V \to \mathbb{R}_+$ is a weight function and $t \in \mathbb{R}_+$ is a threshold such that for every subset $S \subseteq V$, $w(S) := \sum_{x \in S} w(x) \geq t$ if and only if S is a total dominating set in G. A pair (w, t) as above will be referred to as a *total domishold (TD) structure* of G. The *neighborhood hypergraph* of a graph G is the hypergraph denoted by $\mathcal{N}(G)$ and defined as follows: the vertex set of $\mathcal{N}(G)$ is $V(G)$ and the hyperedge set consists precisely of the *minimal neighborhoods* in G, that is, of the inclusion-wise minimal sets in the family of neighborhoods $\{N(v) : v \in V(G)\}$.[3] Note that a set $S \subseteq V(G)$ is a total dominating set in G if and only if it is a transversal of $\mathcal{N}(G)$.

Proposition 3.1 (Chiarelli and Milanič [18]). *For a graph $G = (V, E)$, the following are equivalent:*

1. *G is total domishold.*
2. *Its neighborhood hypergraph $\mathcal{N}(G)$ is threshold.*

The constructions of the derived hypergraph and the derived Boolean function used in our characterizations of connected-domishold graphs in terms of their thresholdness are specified by Definitions 3.2 and 3.3.

Definition 3.2. Given a graph G, the *cutset hypergraph* of G is the hypergraph $\mathcal{C}(G)$ with vertex set $V(G)$ whose hyperedges are precisely the minimal cutsets in G.

Given a finite non-empty set V, we denote by $\{0, 1\}^V$ the set of all binary vectors with coordinates indexed by V. Given a graph $G = (V, E)$ and a binary vector $x \in \{0, 1\}^V$, its *support set* is the set denoted by $S(x)$ and defined by $S(x) = \{v \in V : x_v = 1\}$. In

[3] In [18], the neighborhood hypergraph of G was named *reduced neighborhood hypergraph* (of G) and denoted by $\mathcal{RN}(G)$. We changed the terminology in analogy with the term "cutset hypergraph", which will be introduced shortly.

the following definition, we associate a Boolean function to a given n-vertex graph G. In order to avoid fixing a bijection between its vertex set and the set $[n]$, we will consider the corresponding Boolean function as being defined on the set $\{0,1\}^V$, where $V = V(G)$. Accordingly, a separating structure of such a Boolean function can be seen as a pair (w,t) where $w\colon V \to \mathbb{R}^+$ and $t \in \mathbb{R}^+$ such that for every $x \in \{0,1\}^V$, we have $f(x) = 0$ if and only if $\sum_{v\in S(x)} w(v) \leq t$.

Definition 3.3. Given a graph $G = (V,E)$, its *cutset function* is the positive Boolean function $f_G^{cut}\colon \{0,1\}^V \to \{0,1\}$ that takes value 1 precisely on vectors $x \in \{0,1\}^V$ whose support set contains some minimal cutset of G.

The announced characterizations of connected-domishold graphs in terms of their cutset hypergraphs and cutset functions are given in the following proposition. The proof is based on two ingredients: the characterization of the connected dominating sets of a given connected (non-complete) graph given by Proposition 2.1 and the fact that threshold Boolean functions are closed under dualization.

Proposition 3.4. *For a connected graph $G = (V,E)$, the following are equivalent:*

1. *G is connected-domishold.*
2. *Its cutset hypergraph $\mathcal{C}(G)$ is threshold.*
3. *Its cutset function f_G^{cut} is threshold.*

Moreover, if G is not a complete graph and (w,t) is an integral separating structure of f_G^{cut} or of $\mathcal{C}(G)$, then $(w, w(V) - t)$ is a CD structure of G.

Proof. We consider two cases, depending on whether G is a complete graph or not.

Case 1: G is complete.

In this case all the three statements hold. Recall that every complete graph is CD (see Example 1.2). Since complete graphs have no cutsets, the set of hyperedges of the cutset hypergraph $\mathcal{C}(G)$ is empty. Hence the hypergraph $\mathcal{C}(G)$ is threshold. The absence of (minimal) cutsets also implies that the cutset function f_G^{cut} is constantly equal to 0 and hence threshold.

Case 2: G is not complete.

First we will show the equivalence between statements 1 and 3. Since a positive Boolean function f is threshold if and only if its dual function $f^d(x) = \overline{f(\overline{x})}$ is threshold, it suffices to argue that G is connected-domishold if and only if $(f_G^{cut})^d$ is threshold.

We claim that for every $x \in \{0,1\}^V$, we have $(f_G^{cut})^d(x) = 1$ if and only if $S(x)$, the support set of x, is a connected dominating set of G. Let $x \in \{0,1\}^V$ and let S be the support set of x. By definition, $(f_G^{cut})^d(x) = 1$ if and only if $f_G^{cut}(\overline{x}) = 0$, which is the case if and only if $V \setminus S$ does not contain any minimal cutset of G. This is in turn equivalent to the condition that S is a transversal of the cutset hypergraph of G, and, by Proposition 2.1, to the condition that S is a connected dominating set of G. Therefore, $(f_G^{cut})^d(x) = 1$ if and only if S is a connected dominating set of G, as claimed.

Now, if G is connected-domishold, then it has an integral connected-domishold structure, say (w,t), and $(w,t-1)$ is a separating structure of the dual function $(f_G^{cut})^d$, which implies that $(f_G^{cut})^d$ is threshold. Conversely, if the dual function is threshold, with an integral separating structure (w, t), then $(w, t + 1)$ is a connected-domishold structure of G. This establishes the equivalence between statements 1 and 3.

Next, we show the equivalence between statements 2 and 3. Note that the complete DNF of f_G^{cut}, the cutset function of G, is given by the expression $\bigvee_{S\in\mathcal{C}(G)}\bigwedge_{u\in S} x_u$. It now follows directly from the definitions of threshold Boolean functions and threshold hypergraphs that function $f_G^{cut}(x)$ is threshold if and only if cutset hypergraph $\mathcal{C}(G)$ is threshold.

Finally, if (w, t) is an integral separating structure of f_G^{cut}, then $(w, w(V) - t - 1)$ is a separating structure of $(f_G^{cut})^d$ and hence $(w, w(V) - t)$ is a connected-domishold structure of G. □

Recall that every 1-Sperner hypergraph is threshold (Theorem 2.9) and every threshold hypergraph is asummable (Corollary 2.7). Thus, in particular, every threshold hypergraph is 2-asummable. Applying these relations to the specific case of the minimal cutset hypergraphs, Proposition 3.4 leads to the following.

Corollary 3.5. *For every connected graph G, the following holds:*

1. *If the cutset hypergraph $\mathcal{C}(G)$ is 1-Sperner, then G is connected-domishold.*
2. *If G is connected-domishold, then its cutset hypergraph $\mathcal{C}(G)$ is 2-asummable.*

We will show in Section 4.1 that neither of the two statements in Corollary 3.5 can be reversed. On the other hand, in Section 5 we will prove that all the three properties become equivalent in the hereditary setting.

4 Connected-domishold split graphs

The following examples show that for general connected graphs, the CD and TD properties are incomparable:

- The path P_6 is connected-domishold (it has a unique minimal connected dominating set, formed by its internal vertices) but it is not total domishold (see, e.g., [18]).
- The graph in Figure 2 is TD but not CD.

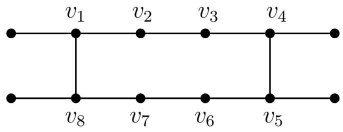


Figure 2: A TD graph that is not CD.

The graph is total domishold: it has a unique minimal total dominating set, namely $\{v_1, v_4, v_5, v_8\}$. On the other hand, the graph is not connected-domishold. This can be shown by observing that its cutset hypergraph is not 2-asummable and applying Corollary 3.5. To see that the cutset hypergraph of G is 2-summable, note that condition (2.1) is satisfied if we take $k = r = 2$ and $A_1 = \{v_2, v_7\}$, $A_2 = \{v_3, v_6\}$, $B_1 = \{v_2, v_3\}$, and $B_2 = \{v_6, v_7\}$.

Interestingly, we will show in Section 5 that if the CD and TD properties are required also for all connected induced subgraphs, then the corresponding graph classes become comparable (see Corollary 5.9). In the rest of this section, we will prove that the two properties coincide in the class of connected split graphs and examine some consequences of this result. Recall that a graph is split if and only if its vertex set has a partition into a clique and an independent set. Foldes and Hammer characterized split graphs as exactly the graphs that are $\{2K_2, C_4, C_5\}$-free [30]. In particular, this implies that a split graph can be disconnected only if it has an isolated vertex.

Lemma 4.1. *Let G be a connected graph and let G' be the graph obtained from G by adding to it a universal vertex. Then, G is connected-domishold if and only if G' is connected-domishold.*

Proof. Let $V(G') = V(G) \cup \{u\}$, where u is the added vertex. Suppose that G is connected-domishold and let (w, t) be a CD structure of G. Since the set of connected dominating sets of G' consists of all connected dominating sets of G together with all subsets of $V(G')$ containing u, we can obtain a CD structure, say (w', t'), of G' by setting $w'(x) = w(x)$ for all $x \in V(G)$, $w'(u) = t$, and $t' = t$. Therefore, G' is connected-domishold.

Conversely, if (w', t') is a CD structure of G', then (w, t) where $t = t'$ and w is the restriction of w' to $V(G)$ is a CD structure of G. This is because a set $X \subseteq V(G)$ is a connected dominating set of G if and only if it is a connected dominating set of G'. Therefore, if G' is connected-domishold then so is G. □

Recall that given a connected graph G, we denote by $\mathcal{C}(G)$ (resp., $\mathcal{N}(G)$) its cutset (resp., neighborhood) hypergraph.

Lemma 4.2. *Let G be a connected split graph without universal vertices. Then*

$$\mathcal{C}(G) = \mathcal{N}(G).$$

Proof. Fix a split partition of $V(G)$, say $V(G) = K \cup I$ where K is a clique, I is an independent set, and $K \cap I = \emptyset$. Clearly, the hypergraphs $\mathcal{C}(G)$ and $\mathcal{N}(G)$ have the same vertex set. We show that the hyperedge sets are also the same in two steps.

First, we show that $E(\mathcal{C}(G)) \subseteq E(\mathcal{N}(G))$, that is, that every minimal cutset is a minimal neighborhood. To this end, it suffices to show that every minimal cutset S in G *is* a neighborhood, that is, a set of the form $S = N(v)$ for some $v \in V(G)$. This is indeed enough, because if a minimal cutset S in G satisfies $S = N(v)$ for some $v \in V(G)$, but $N(v)$ properly contains some other neighborhood, say $N(u)$, then the fact that $N(u)$ is a cutset in G (for instance, it is a u, v-separator) would imply that S is not a minimal cutset.

Let S be a minimal cutset in G. Then, S is a minimal u, v-separator for some non-adjacent vertex pair u, v; in particular, $S \subseteq V(G) \setminus \{u, v\}$. We claim that $N(u) \subseteq S$ or $N(v) \subseteq S$. Suppose that this is not the case. Then, there exist a neighbor of u, say u', such that $u' \notin S$, and a neighbor of v, say v', such that $v' \notin S$. Since $\{u, v, u', v'\} \subseteq V(G) \setminus S$ and u and v are in different components of $G - S$, vertices u' and v' are distinct and non-adjacent. Thus, at least one of u' and v', say u', is in I. This implies that $u \in K$ and therefore $v \in I$, which implies that $v' \in K$ and hence (u, v', v) is a u, v-path in $G - S$, a contradiction. This shows that $N(u) \subseteq S$ or $N(v) \subseteq S$, as claimed. Since each of $N(u)$ and $N(v)$ is a u, v-separator, the fact that S is a minimal u, v-separator implies that $S \in \{N(u), N(v)\}$. This completes the proof of the inclusion $E(\mathcal{C}(G)) \subseteq E(\mathcal{N}(G))$.

It remains to show that $E(\mathcal{N}(G)) \subseteq E(\mathcal{C}(G))$. Let S be a minimal neighborhood in G. Then $S = N(v)$ for some $v \in V(G)$. Since v is not universal, the set $V(G) \setminus N[v]$ is non-empty. Therefore S is a v, w-separator for any $w \in V(G) \setminus N[v]$; in particular, S is a cutset in G. Suppose for a contradiction that S is not a minimal cutset in G. Then S properly contains some minimal cutset, say S', in G. By the first part of the proof, S' is of the form $S' = N(z)$ for some $z \in V(G)$. However, since $N(z)$ is a neighborhood properly contained in $S = N(v)$, this contradicts the fact that S is a minimal neighborhood. □

Theorem 4.3. *A connected split graph is connected-domishold if and only if it is total domishold.*

Proof. If G is complete, then G is both connected-domishold and total domishold. So we may assume that G is not complete. More generally, we show next that we may assume that G does not have any universal vertices. Suppose that G has a universal vertex, say u, and let $G' = G - u$. By [18, Proposition 3.3], G is TD if and only if G' is TD. If G' is not connected, then $\{u\}$ is the only minimal connected dominating set of G and hence G is connected-domishold in this case. Furthermore, G is also total domishold: since G' is a disconnected $2K_2$-free graph, G' has an isolated vertex. Therefore, by [18], G' is TD, and hence so is G. If G' is connected, then by Lemma 4.1, G is CD if and only if G' is CD. Therefore, the problem of verifying whether the CD and the TD properties are equivalent for G reduces to the same problem for G'. An iterative application of the above argument eventually reduces the graph to either a graph where both properties hold or to a connected graph without universal vertices.

Now, let G be a connected split graph without universal vertices. By Proposition 3.4, G is connected-domishold if and only if its cutset hypergraph $\mathcal{C}(G)$ is threshold. By Proposition 3.1, G is total domishold if and only if its neighborhood hypergraph $\mathcal{N}(G)$ is threshold. Therefore, to prove the theorem it suffices to show that $\mathcal{C}(G) = \mathcal{N}(G)$. But this was established in Lemma 4.2. □

Theorem 4.3 implies another relation between connected-domishold (split) graphs and threshold hypergraphs, one that in a sense reverses the one stated in Proposition 3.4. Given a hypergraph $\mathcal{H} = (V, E)$, the *split-incidence graph* of $\mathcal{H}$ (see, e.g., [38]) is the split graph G such that $V(G) = V \cup E$, V is a clique, E is an independent set, and $v \in V$ is adjacent to $e \in E$ if and only if $v \in e$.

Theorem 4.4. *Let $\mathcal{H} = (V, E)$ be a hypergraph with $\emptyset \notin E$. Then $\mathcal{H}$ is threshold if and only if its split-incidence graph is connected-domishold.*

Proof. Since $\emptyset \notin E$, the split-incidence graph of $\mathcal{H}$ is connected. It was shown in [18] that a hypergraph is threshold if and only if its split-incidence graph is total domishold. The statement of the theorem now follows from Theorem 4.3. □

It might be worth pointing out that in view of Remark 2.5 and Theorem 4.4, it is an open problem of whether there is a "purely combinatorial" polynomial-time algorithm for recognizing connected-domishold split graphs. Further issues regarding the recognition problem of CD graphs will be discussed in Section 6.1.

4.1 Examples related to Corollary 3.5

We now show that neither of the two statements in Corollary 3.5 can be reversed. First we exhibit an infinite family of CD split graphs whose cutset hypergraphs are not 1-Sperner.

Example 4.5. Let $n \geq 4$ and let $G = K_n^*$ be the graph obtained from the complete graph K_n by gluing a triangle on every edge. Formally,

$$\begin{aligned} V(G) &= \{u_1, \dots, u_n\} \cup \{v_{ij} : 1 \leq i < j \leq n\} \text{ and} \\ E(G) &= \{u_i u_j : 1 \leq i < j \leq n\} \cup \{u_i v_{jk} : 1 \leq j < k \leq n \text{ and } i \in \{j, k\}\}. \end{aligned}$$

The graph G is a CD graph: setting

$$w(x) = \begin{cases} 1, & \text{if } x \in \{u_1, \dots, u_n\}; \\ 0, & \text{otherwise.} \end{cases}$$

and $t = n - 1$ results in a CD structure of G. On the other hand, the cutset hypergraph of G is not 1-Sperner. Since every pair of the form $\{u_i, u_j\}$ with $1 \leq i < j \leq n$ is a minimal cutset of G, the cutset hypergraph contains $\{u_1, u_2\}$ and $\{u_3, u_4\}$ as hyperedges and is therefore not 1-Sperner.

Next, we argue that there exists a split graph G whose cutset hypergraph is 2-asummable but G is not CD. As observed already in [18], the fact that not every 2-asummable positive Boolean function is threshold can be used to construct split graphs G such that $\mathcal{N}(G)$ is 2-asummable and G is not total domishold. The existence of split graphs with claimed properties now follows from Theorem 4.3 and its proof. For the sake of self-containment, we describe an example of such a construction in Appendix A.

5 The hereditary case

In this section we present the main result of this paper, Theorem 5.4, which gives several characterizations of graphs all connected induced subgraphs of which are CD, and derive some of its consequences. The proof of Theorem 5.4 relies on a technical lemma about chordal graphs, which will be proved in Section 7.

We start with an example showing that, contrary to the classes of threshold and domishold graphs, the class of connected-domishold graphs is not hereditary. We assume notation from Example 1.3.

Example 5.1. The graph G obtained from C_4 by adding to it a new vertex, say v_5, and making it adjacent exactly to one vertex of the C_4, say to v_4, is CD: the (inclusion-wise) minimal CD sets of G are $\{v_1, v_4\}$ and $\{v_3, v_4\}$, hence a CD structure of G is given by $w(v_2) = w(v_5) = 0$, $w(v_1) = w(v_3) = 1$, $w(v_4) = 2$, and $t = 3$.

This motivates the following definition.

Definition 5.2. A graph G is said to be *hereditarily connected-domishold* (*hereditarily CD* for short) if every connected induced subgraph of G is connected-domishold.

In general, for a property Π of connected graphs, a graph is said to be *hereditarily* Π if every connected induced subgraph of it satisfies Π. Characterizations of classes of hereditarily Π graphs where Π denotes the property that the graph has a connected dominating set inducing a graph with a certain property Π' were given, for various choices of property Π', by Michalak in [49]. In [57], Pržulj et al. gave characterizations of hereditarily Π graphs where Π denotes the property that the graph has a dominating pair of vertices (that is, a pair of vertices such that every path between them is dominating). The class of hereditarily connected-domishold graphs corresponds to the case when Π is the property of being connected-domishold.

In order to state the technical lemma to be used in the proof of Theorem 5.4, we need some terminology. A *diamond* is a graph obtained from K_4 by deleting an edge. Given a diamond D, we will refer to its vertices of degree two as its *tips* and denote them as t and t', and to its vertices of degree three as its *centers* and denote them as c and c'. The respective vertex sets will be denoted by T and C. Similar notation will be used for diamonds denoted by D_1 or D_2.

Lemma 5.3 (Diamond Lemma). *Let G be a connected chordal graph. Suppose that G contains two induced diamonds $D_1 = (V_1, E_1)$ and $D_2 = (V_2, E_2)$ such that:*

(i) $C_1 \cap C_2 = \emptyset$.

(ii) If no vertex in C_1 is adjacent to a vertex in C_2, then there exists a C_1, C_2-separator in G of size one.

(iii) For each $j \in \{1, 2\}$ the tips (i.e., t_j, t_j') of D_j belong to different components of $G - C_j$.

(iv) For $j \in \{1, 2\}$ every component of $G - C_j$ has a vertex that dominates C_j.

Then G has an induced subgraph isomorphic to F_1, F_2, or H_i for some $i \geq 1$, where the graphs F_1, F_2, and a general member of the family $\{H_i\}$ are depicted in Figure 3.

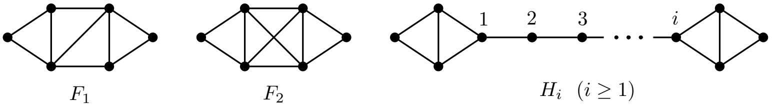


Figure 3: Graphs F_1, F_2, and H_i.

The proof of Lemma 5.3 is postponed to Section 7.

Theorem 5.4. *For every graph G, the following are equivalent:*

1. *G is hereditarily connected-domishold.*
2. *The cutset hypergraph of every connected induced subgraph of G is 1-Sperner.*
3. *The cutset hypergraph of every connected induced subgraph of G is threshold.*
4. *The cutset hypergraph of every connected induced subgraph of G is 2-asummable.*
5. *G is an $\{F_1, F_2, H_1, H_2, \ldots\}$-free chordal graph.*

Proof. The equivalence between items 1 and 3 follows from Proposition 3.4.

The implications $2 \Rightarrow 1 \Rightarrow 4$ follow from Corollary 3.5.

For the implication $4 \Rightarrow 5$, we only need to verify that the cutset hypergraph of none of the graphs in the set $\mathcal{F} := \{C_k : k \geq 4\} \cup \{F_1, F_2\} \cup \{H_i : i \geq 1\}$ is 2-asummable. Let $F \in \mathcal{F}$. Suppose first that F is a cycle C_k for some $k \geq 4$, let u_1, u_2, u_3, u_4 be four consecutive vertices on the cycle. Let $A_1 = \{u_1, u_3\}$, $A_2 = \{u_2, u_4\}$, $B_1 = \{u_1, u_2\}$ and $B_2 = \{u_3, u_4\}$. Then, A_1 and A_2 are minimal cutsets of F and thus hyperedges of the hypergraph $\mathcal{C}(F)$, while B_1 and B_2 do not contain any minimal cutset of F and are consequently independent sets in the hypergraph $\mathcal{C}(F)$. Since the sets A_1, A_2, B_1 and B_2 satisfy condition (2.1), this implies that the hypergraph $\mathcal{C}(F)$ is 2-summable. If $F \in \{F_1, F_2\} \cup \{H_i : i \geq 1\}$, then let a and b be the two vertices of degree 2 in F, let $N(a) = \{a_1, a_2\}$, $N(b) = \{b_1, b_2\}$, let $A_1 = N(a)$, $A_2 = N(b)$, $B_1 = \{a_1, b_1\}$ and $B_2 = \{a_2, b_2\}$. The rest of the proof is the same as above.

It remains to show the implication $5 \Rightarrow 2$. Suppose that the implication fails and let G be a minimal counterexample. That is, G is an $\{F_1, F_2, H_1, H_2, \ldots\}$-free chordal graph such that its cutset hypergraph is not 1-Sperner, but the cutset hypergraph of every $\{F_1, F_2, H_1, H_2, \ldots\}$-free chordal graph with fewer vertices than G is 1-Sperner. Since $\mathcal{C}(G)$ is not 1-Sperner, G has two minimal cutsets, say S and S', such that $\min\{|S \setminus S'|, |S' \setminus S|\} \geq 2$. The minimality of G implies that the empty set is not a minimal cutset, hence G is connected. Furthermore, the minimality ensures that S and S' are disjoint sets (otherwise one can remove $S \cap S'$ from G and have a smaller counterexample). Thus, $\min\{|S|, |S'|\} \geq 2$. The minimality also ensures that $|S| = |S'| = 2$. Indeed, removing a third vertex z, if present, from S does not affect the minimal cutset status of S. Since every minimal cutset in a chordal graph is a clique [25], removing a third vertex z, if present, from S will also not affect the minimal cutset status of S' since the entire S (which is a clique) is present in one component of $G - S'$.

The minimality also ensures that if there are no edges between S and S', then every minimal S, S'-separator T is of size one. Indeed, if this is not the case, then $|T| \geq 2$ since G is connected. Let X be a component of $G - S$ containing S' and let Y be any other component of $G - S$. The fact that T separates S from S' implies that T contains all vertices in $N(S) \cap V(X)$, which is a non-empty set due to the minimality of S. Since T is a minimal cutset in a chordal graph, it is a clique; in particular, it is fully contained in X. However, this implies that the sets S' and T are minimal cutsets in the graph $G - V(Y)$ such that $\min\{|S' \setminus T|, |T \setminus S'|\} \geq 2$, contrary to the minimality of G.

Let X, Y be two distinct components of $G - S$ and X', Y' two distinct components of $G - S'$. By Lemma 2.2, there exist vertices $x \in X$ and $y \in Y$ such that each of x and y dominates S and $x' \in X'$ and $y' \in Y'$ such that each of x' and y' dominates S'. Let D_1 be the subgraph of G induced by $S \cup \{x, y\}$ and let D_2 be the subgraph of G induced by $S' \cup \{x', y'\}$. The definitions of D_1 and D_2 and Lemma 2.2 imply that D_1 and D_2 are two induced diamonds in G satisfying the hypotheses of the Diamond Lemma (Lemma 5.3). Consequently, G has an induced subgraph isomorphic to F_1, F_2, or H_i for some $i \geq 1$, a contradiction. This completes the proof of the theorem. □

Remark 5.5. The cutset hypergraph of a disconnected graph H is equal to $(V(H), \{\emptyset\})$ and is clearly 1-Sperner (and therefore also threshold and 2-asummable). It follows that conditions from items 2–4 in Theorem 5.4 are equivalent to the analogous conditions in which the respective properties are imposed on cutset hypergraphs of all induced subgraphs of G (and not only of connected ones).

In the rest of this section, we examine some of the consequences of the forbidden induced subgraph characterization of hereditarily CD graphs given by Theorem 5.4. The *kite* (also known as the *co-fork* or the *co-chair*) is the graph depicted in Figure 4.

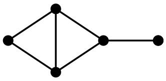


Figure 4: The kite.

The equivalence between items 1 and 5 in Theorem 5.4 implies that the class of hereditarily CD graphs is a proper generalization of the class of kite-free chordal graphs.

Corollary 5.6. *Every kite-free chordal graph is hereditarily CD.*

Corollary 5.6 further implies that the class of hereditarily CD graphs generalizes two well known classes of chordal graphs, the class of block graphs and the class of trivially perfect graphs. A graph is said to be a *block graph* if every block (maximal connected subgraph without cut vertices) of it is complete. The block graphs are well known to coincide with the diamond-free chordal graphs. A graph G is said to be *trivially perfect* [33] if for every induced subgraph H of G, it holds $\alpha(H) = |\mathcal{K}(H)|$, where $\alpha(H)$ denotes the *independence number* of H (that is, the maximum size of an independent set in H) and $\mathcal{K}(H)$ denotes the set of all maximal cliques of H. Trivially perfect graphs coincide with the so-called *quasi-threshold graphs* [67], and are exactly the $\{P_4, C_4\}$-free graphs [33].

Corollary 5.7. *Every block graph is hereditarily CD. Every trivially perfect graph is hereditarily CD.*

Another class of graphs contained in the class of hereditarily CD graphs is the class of graphs defined similarly as the hereditarily CD graphs but with respect to total dominating sets. These so-called *hereditarily total domishold graphs* (abbreviated *hereditarily TD graphs*) were studied in [18], where characterizations analogous to those given by Theorem 5.4 were obtained, including the following characterization in terms of forbidden induced subgraphs.

Theorem 5.8 (Chiarelli and Milanič [18]). *For every graph G, the following are equivalent:*

1. *G is hereditarily total domishold.*
2. *No induced subgraph of G is isomorphic to a graph in Figure 5.*

Theorems 5.4 and 5.8 imply the following.

Corollary 5.9. *Every hereditarily TD graph is hereditarily CD.*

Proof. It suffices to verify that each of the forbidden induced subgraphs for the class of hereditarily connected-domishold graphs contains one of the graphs from Figure 5 as induced subgraph. A cycle C_k with $k \geq 4$ contains (or is equal to) one of C_4, C_5, C_6, P_6. The graphs F_1 and F_2 are contained in both sets of forbidden induced subgraphs. Finally, each graph of the form H_i where $i \geq 1$ contains $2K_3$ as induced subgraph. □

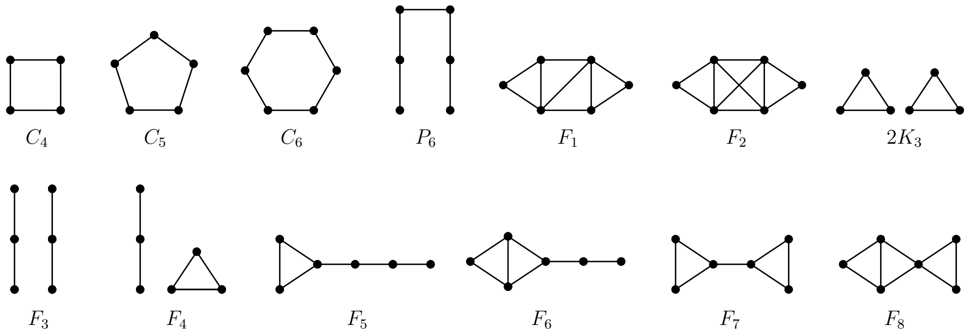


Figure 5: The set of forbidden induced subgraphs for the class of hereditarily total domishold graphs.

Since a graph is split if and only if it is $\{2K_2, C_4, C_5\}$-free and each of the forbidden induced subgraphs for the class of hereditarily total domishold graphs other than F_2 contains either $2K_2$, C_4, or C_5 as induced subgraph, Corollary 5.9 implies the following.

Corollary 5.10. *Every F_2-free split graph is hereditarily CD.*

Figure 6 shows a Hasse diagram depicting the inclusion relations among the class of hereditarily connected-domishold graphs and several other, well studied graph classes. All definitions of graph classes depicted in Figure 6 and the relations between them can be found in [23], with the exception of hereditarily CD and hereditarily TD graphs. The fact that every co-domishold graph is hereditarily TD and that every hereditarily TD graph is $(1, 2)$-polar chordal was proved in [18]. The remaining inclusion and non-inclusion relations can be easily verified using the forbidden induced subgraph characterizations of the depicted graph classes, see [10,23,34].

6 Algorithmic aspects via vertex separators

In this section, we build on the above results, together with some known results from the literature on connected dominating sets and minimal vertex separators in graphs, to study certain algorithmic aspects of the class of connected-domishold graphs and its hereditary variant.

6.1 The recognition problems

We start with computational complexity aspects of the problems of recognizing whether a given graph is CD, resp. hereditarily CD. For general graphs, the computational complexity of recognizing connected-domishold graphs is not known. However, we will now show that the hypergraph approach outlined in Section 3 leads to a sufficient condition for the problem to be polynomially solvable, capturing a large number of graph classes. The condition is expressed using the notion of minimal vertex separators. Recall that a u, v*-separator* (for a pair of non-adjacent vertices u, v) is a set $S \subseteq V(G) \setminus \{u, v\}$ such that u and v are in different components of $G - S$ and that a u, v-separator is minimal if it does not contain any other u, v-separator. Recall also that a *minimal vertex separator* in G is a minimal u, v-separator for some non-adjacent vertex pair u, v.

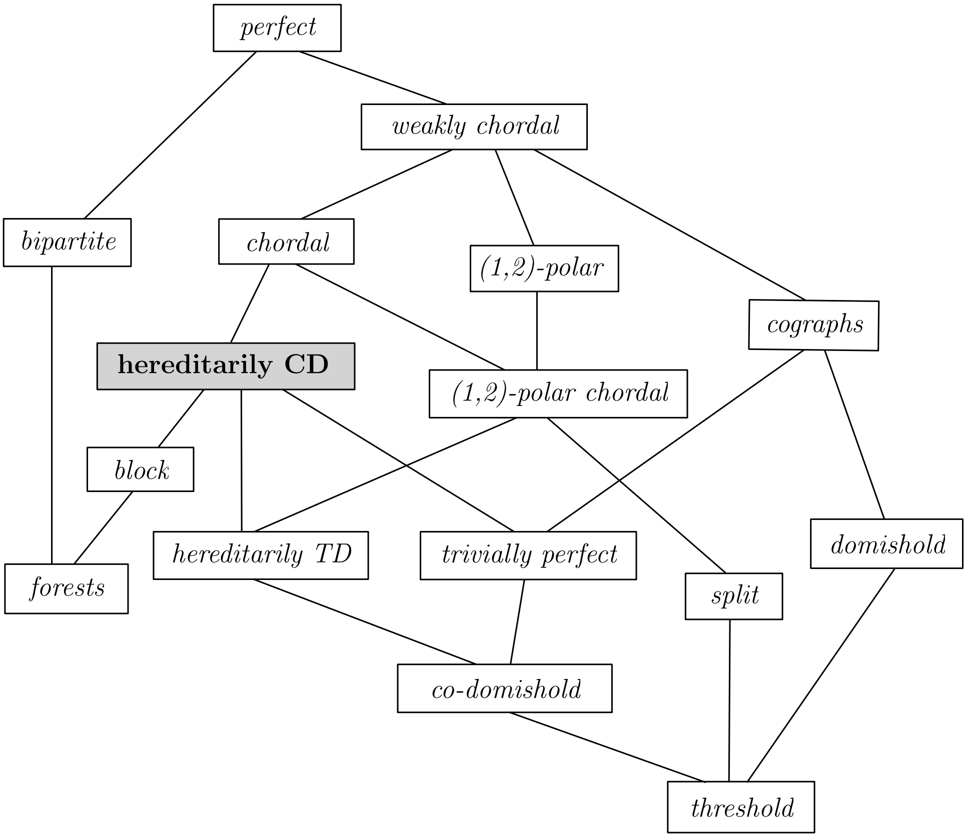


Figure 6: A Hasse diagram depicting the inclusion relations within several families of perfect graphs, focused around the class of hereditarily connected-domishold graphs.

A sufficient condition for the polynomial-time solvability of the recognition problem for CD graphs in a class of graphs $\mathcal{G}$ is that there exists a polynomial p such that every connected graph $G \in \mathcal{G}$ has at most $p(|V(G)|)$ minimal vertex separators. This is the case for chordal graphs, which have at most $|V(G)|$ minimal vertex separators [59], as well as for many other classes of graphs, including permutation graphs, circle graphs, circular-arc graphs, chordal bipartite graphs, trapezoid graphs, cocomparability graphs of bounded dimension, distance-hereditary graphs, and weakly chordal graphs (see, e.g., [9, 43, 51]). For a polynomial p, let $\mathcal{G}_p$ be the class of graphs with at most $p(|V(G)|)$ minimal vertex separators. Since every minimal cutset is a minimal vertex separator, every connected graph $G \in \mathcal{G}_p$ has at most $p(|V(G)|)$ minimal cutsets.

It is known that the set of all minimal vertex separators of a given connected n-vertex graph can be enumerated in output-polynomial time. More precisely, Berry et al. [3] have developed an algorithm solving this problem in time $\mathcal{O}(n^3|\Sigma|)$ where Σ is the set of all minimal vertex separators of G, improving on earlier (independently achieved) running times of $\mathcal{O}(n^5|\Sigma|)$ due to Shen and Liang [63] and Kloks and Kratsch [44]. Based on these results, we derive the following.

Theorem 6.1. *For every polynomial p there is a polynomial-time algorithm to determine whether a given connected graph $G \in \mathcal{G}_p$ is connected-domishold. In case of a yes instance, the algorithm also computes an integral CD structure of G.*

Proof. Let $G = (V, E) \in \mathcal{G}_p$ be a connected graph that is the input to the algorithm.

The algorithm proceeds as follows. If G is complete, then G is connected-domishold and an integral CD structure of G is returned, say (w, t) with $w(x) = 1$ for all $x \in V(G)$ and $t = 1$. Assume now that G is not complete. First, using the algorithm of Berry et al. [3], we compute in time $\mathcal{O}(|V(G)|^3 p(|V(G)|))$ the set Σ of all minimal vertex separators of G. Next, the cutset hypergraph, $\mathcal{C}(G)$, is computed by comparing each pair of sets in Σ and discarding the non-minimal ones. Since $\mathcal{C}(G)$ is Sperner, there is a bijective correspondence between the hyperedges of $\mathcal{C}(G)$ and the prime implicants of the cutset function f_G^{cut}; this yields the complete DNF of f_G^{cut}. Finally, we run the algorithm given by Theorem 2.4 on the complete DNF of f_G^{cut}. If f_G^{cut} is not threshold, then we conclude that G is not connected-domishold. Otherwise, the algorithm returned an integral separating structure, say (w, t), of f_G^{cut}. In this case we return $(w, w(V) - t)$ as a CD structure of G.

It is clear that the algorithm runs in polynomial time. Its correctness follows from Proposition 3.4. □

Let $\tilde{\mathcal{G}}$ be the largest hereditary graph class such that a connected graph $G \in \tilde{\mathcal{G}}$ is connected-domishold if and only if it is total domishold. By Theorem 4.3, class $\tilde{\mathcal{G}}$ is a generalization of the class of split graphs. Since there is a polynomial-time algorithm for recognizing total domishold graphs [16, 18], there is a polynomial-time algorithm to determine whether a given connected graph $G \in \tilde{\mathcal{G}}$ is connected-domishold. This motivates the following question (which we leave open).

Question. *What is the largest hereditary graph class $\tilde{\mathcal{G}}$ such that a connected graph $G \in \tilde{\mathcal{G}}$ is connected-domishold if and only if it is total domishold?*

A polynomial-time recognition algorithm for the class of hereditarily CD graphs can be derived from the characterization of hereditarily CD graphs in terms of forbidden induced subgraphs given by Theorem 5.4.

Proposition 6.2. *There exists a polynomial-time algorithm to determine whether a given graph G is hereditarily CD. In the case of a yes instance, an integral CD structure of G can be computed in polynomial time.*

Proof. One can verify in linear time that G is chordal [34] and verifying that G is also $\{F_1, F_2, H_1, H_2\}$-free can be done in time $\mathcal{O}(|V(G)|^8)$. Therefore, we only have to show that we can check in polynomial time that G does not contain an induced subgraph of the form H_i for each $i > 2$. Observe that for all $i > 2$ the graph H_i contains an induced subgraph isomorphic to $2D$, the union of two diamonds (see Figure 3 and Figure 4). In $\mathcal{O}(|V(G)^8|)$ time, we can enumerate all induced subgraphs F of G isomorphic to $2D$. For each such subgraph F we have to verify whether it can be extended to an induced subgraph of the form H_i, for some $i > 2$. We do this as follows. Let D_1 and D_2 be the connected components (diamonds) of F. Furthermore, let u_1, u_2 be the two vertices of degree 2 in D_1 and similarly let v_1, v_2 be the two vertices of degree 2 in D_2. Now we can verify that F is not contained in any induced subgraph of G isomorphic to H_i (for some $i > 2$) by checking for each pair u_i, v_j, with $i, j \in \{1, 2\}$, that u_i and v_j belong to different components of $G - (N_{G-u_i}[V(D_1) \setminus \{u_i\}] \cup N_{G-v_j}[V(D_2) \setminus \{v_j\}])$. This can be done in polynomial time and consequently the recognition of hereditarily CD graphs is a polynomially solvable problem.

The second part of the theorem follows from Theorem 6.1, since every hereditarily CD graph is chordal and chordal graphs are a subclass of $\mathcal{G}_p$ for the polynomial $p(n) = n$ [59]. □

It might seem conceivable that a similar approach as the one used in Theorem 6.1 could be used to develop an efficient algorithm for recognizing connected-domishold graphs in classes of graphs with only polynomially many minimal connected dominating sets. However, it is not known whether there exists an output-polynomial-time algorithm for the problem of enumerating minimal connected dominating sets. In fact, as shown by Kanté et al. [38], even when restricted to split graphs, this problem is equivalent to the well-known TRANS-ENUM problem in hypergraphs, the problem of enumerating the inclusion-minimal transversals of a given hypergraph. The TRANS-ENUM problem has been intensively studied but it is still open whether there exists an output-polynomial-time algorithm for the problem (see, e.g., the survey [28]).

6.2 The weighted connected dominating set problem

The WEIGHTED CONNECTED DOMINATING SET (WCDS) problem takes as input a connected graph G together with a cost function $c\colon V(G) \to \mathbb{R}^+$, and the task is to compute a connected dominating set of minimum total cost, where the cost of a set $S \subseteq V(G)$ is defined, as usual, as $c(S) = \sum_{v \in S} c(v)$. The WCDS problem has been studied extensively due to its many applications in networking (see, e.g., [6, 26, 66]). The problem is NP-hard not only for general graphs [36] but also for split graphs [46], chordal bipartite graphs [52], circle graphs [40], and cocomparability graphs [14]. Polynomial-time algorithms for the problem were developed for interval graphs [15] and more generally for trapezoid graphs [64] and circular-arc graphs [15, 37], as well as for distance-hereditary graphs [68].

In this section, we will identify further graph classes where the WCDS problem is polynomially solvable, including the class of F_2-free split graphs (see Figure 1). This result is interesting in view of the fact that for split graphs, the WCDS problem is not only NP-hard but also hard to approximate, even in the unweighted case. This can be seen as follows: Let $\mathcal{H} = (V, E)$ be a Sperner hypergraph with $\emptyset, V \notin E$ and let G be its split-incidence graph. Then G is a connected split graph without universal vertices, hence $\mathcal{C}(G) = \mathcal{N}(G)$ by Lemma 4.2. It can be seen that the hyperedge set of $\mathcal{N}(G)$ is exactly E, and therefore Proposition 2.1 implies that the problem of finding a minimum connected dominating set in G is equivalent to the HITTING SET problem in hypergraphs, the problem of finding a minimum transversal of a given hypergraph. This latter problem is known to be equivalent to the well-known SET COVER problem and hence inapproximable in polynomial time to within a factor of $(1 - \epsilon) \log |V|$, for any $\epsilon > 0$, unless $\mathsf{P} = \mathsf{NP}$ [24]. It follows that the WCDS problem is hard to approximate to within a factor of $(1 - \epsilon) \log |V(G)|$ in the class of split graphs.

We will show that the WCDS problem is polynomially solvable in the class of hereditarily CD graphs; the result for F_2-free split graphs will then follow. Our approach is based on connections with vertex separators and Boolean functions. First, we recall the following known results about: (i) the relation between the numbers of prime implicants of a threshold Boolean function and its dual, and (ii) the complexity of dualizing threshold Boolean functions. These results were proved in the more general context of regular Boolean functions (as well as for other generalizations, see, e.g., [7]).

Theorem 6.3. *Let f be an n-variable threshold Boolean function having exactly q prime implicants. Then:*

1. *(Bertolazzi and Sassano [5], Crama [21], see also [22, Theorem 8.29]) The dual function f^d has at most N prime implicants, where N is the total number of variables in the complete DNF of f.*

2. *(Crama and Hammer [22, Theorem 8.28] and Peled and Simeone [56]) There is an algorithm running in time $\mathcal{O}(n^2 q)$ that, given the complete DNF of f, computes the complete DNF of the dual function f^d.*

The algorithm by Crama and Hammer [22] is already presented as having time complexity $\mathcal{O}(n^2 q)$, while the one by Peled and Simeone [56] is claimed to run in time $\mathcal{O}(nq)$. However, since f^d can have $\mathcal{O}(nq)$ prime implicants, the total size of the output is of the order $\mathcal{O}(n^2 q)$. The time complexity $\mathcal{O}(nq)$ of the algorithm by Peled and Simeone relies on the assumption that the algorithm outputs the prime implicants of the dual function one by one, each time overwriting the previous prime implicant (with a constant number of operations per implicant on average).

The relation between the numbers of prime implicants of a threshold Boolean function and its dual given by Theorem 6.3 implies that classes of connected-domishold graphs with only polynomially many minimal cutsets are exactly the same as the classes of connected-domishold graphs with only polynomially many minimal connected dominating sets. More precisely:

Lemma 6.4. *Let $G = (V, E)$ be an n-vertex connected-domishold graph that is not complete. Let ν_c (resp. ν_s) denote the number of minimal connected dominating sets (resp. of minimal cutsets) of G. Then $\nu_s \leq (n - 2)\nu_c$ and $\nu_c \leq (n - 2)\nu_s$.*

Proof. By Proposition 3.4, the cutset function f_G^{cut} is threshold. Function f_G^{cut} is an n-variable function with exactly ν_s prime implicants in its complete DNF. Recall from the proof of Proposition 3.4 that the dual function $(f_G^{cut})^d$ takes value 1 precisely on the vectors $x \in \{0,1\}^V$ whose support is a connected dominating set of G. Therefore, the prime implicants of $(f_G^{cut})^d$ are in bijective correspondence with the minimal connected dominating sets of G and the number of prime implicants of $(f_G^{cut})^d$ is exactly ν_c. Since every minimal cutset of G has at most $n-2$ vertices, Theorem 6.3 implies that $\nu_c \leq (n-2)\nu_s$, as claimed.

Conversely, since $f_G^{cut} = ((f_G^{cut})^d)^d$, the inequality $\nu_s \leq (n-2)\nu_c$ can be proved by a similar approach, provided we show that every minimal connected dominating set of G has at most $n-2$ vertices. But this is true since if D is a connected dominating set of G with at least $n-1$ vertices, with $V(G) \setminus \{u\} \subseteq D$ for some $u \in V(G)$, then a smaller connected dominating set D' of G could be obtained by fixing an arbitrary spanning tree T of $G[D]$ and deleting from D an arbitrary leaf v of T such that $N_G(u) \neq \{v\}$. (Note that since G is connected but not complete, it has at least three vertices, hence T has at least two leaves.) This completes the proof. □

We now have everything ready to derive the main result of this section. Recall that for a polynomial p, we denote by $\mathcal{G}_p$ the class of graphs with at most $p(|V(G)|)$ minimal vertex separators.

Theorem 6.5. *For every nonzero polynomial p, the set of minimal connected dominating sets of an n-vertex connected-domishold graph from $\mathcal{G}_p$ has size at most $\mathcal{O}(n \cdot p(n))$ and can be computed in time $\mathcal{O}(n \cdot p(n) \cdot (n^2 + p(n)))$. In particular, the WCDS problem is solvable in polynomial time in the class of connected-domishold graphs from $\mathcal{G}_p$.*

Proof. Let p and G be as in the statement of the theorem and let $\mathcal{CD}(G)$ be the set of minimal connected dominating sets of G. If G is complete, then

$$\mathcal{CD}(G) = \{\{v\} : v \in V(G)\}$$

and thus $|\mathcal{CD}(G)| = n = \mathcal{O}(n \cdot p(n))$ (since the polynomial is nonzero). Otherwise, we can apply Lemma 6.4 to derive $|\mathcal{CD}(G)| \leq (n-2) \cdot p(n)$.

A polynomial-time algorithm to solve the WCDS problem for a given connected-domishold graph $G \in \mathcal{G}_p$ with respect to a cost function $c\colon V(G) \to \mathbb{R}^+$ can be obtained as follows. First, we may assume that G is not complete, since otherwise we can return a set $\{v\}$ where v is a vertex minimizing $c(v)$. We use a similar approach as in the proof of Theorem 6.1. Using the algorithm of Berry et al. [3], we compute in time $\mathcal{O}(n^3 p(n))$ the set Σ of all minimal vertex separators of G. We can assume that each minimal vertex separator has its elements listed according to some fixed order of $V(G)$ (otherwise, we can sort them in time $\mathcal{O}(n \cdot p(n))$ using, e.g., bucket sort). The cutset hypergraph, $\mathcal{C}(G)$, is then computed by comparing each pair of sets in Σ and discarding the non-minimal ones; this can be done in time $\mathcal{O}(n \cdot (p(n))^2)$. The cutset hypergraph directly corresponds to the complete DNF of the cutset function f_G^{cut}.

The next step is to compute the complete DNF of the dual function $(f_G^{cut})^d$. By Theorem 6.3, this can be done in time $\mathcal{O}(n^2 \cdot p(n))$. Since each term of the DNF is a prime implicant of $(f_G^{cut})^d$ and the prime implicants of $(f_G^{cut})^d$ are in bijective correspondence with the minimal connected dominating sets of G, we can read off from the DNF all the minimal connected dominating sets of G. The claimed time complexity follows.

Once the list of all minimal connected dominating sets is available, a polynomial-time algorithm for the WCDS problem on (G, c) follows immediately. □

In the case of chordal graphs, we can improve the running time by using one of the known linear-time algorithms for listing the minimal vertex separators of a given chordal graph due to Kumar and Veni Madhavan [45], Chandran and Grandoni [13], and Berry and Pogorelcnik [4].

Theorem 6.6. *Every n-vertex connected-domishold chordal graph has at most $\mathcal{O}(n^2)$ minimal connected dominating sets, which can be enumerated in time $\mathcal{O}(n^3)$. In particular, the WCDS problem is solvable in time $\mathcal{O}(n^3)$ in the class of connected-domishold chordal graphs.*

Proof. Let G be an n-vertex connected-domishold chordal graph. The theorem clearly holds for complete graphs, so we may assume that G is not complete. Since G is chordal, it has at most n minimal vertex separators [59]; consequently, G has at most n minimal cutsets. Since G is connected-domishold, it has at most $n(n-2)$ minimal connected dominating sets, by Lemma 6.4.

The minimal connected dominating sets of G can be enumerated as follows. First, we compute all the $\mathcal{O}(n)$ minimal vertex separators of G in time $\mathcal{O}(n+m)$ (where $m = |E(G)|$) using one of the known algorithms for this problem on chordal graphs [4, 13, 45]. Assuming again that each minimal vertex separator has its elements listed according to some fixed order of $V(G)$, we then eliminate those that are not minimal cutsets in time $\mathcal{O}(n^3)$, by directly comparing each of the $\mathcal{O}(n^2)$ pairs for inclusion.

The list of $\mathcal{O}(n)$ minimal cutsets of G yields its cutset function, f_G^{ms}. The list of minimal connected dominating sets of G can be obtained in time $\mathcal{O}(n^3)$ by dualizing f_G^{ms} using one of the algorithms given by Theorem 6.3. The WCDS problem can now be solved in time $\mathcal{O}(n^3)$ by evaluating the cost of each of the $\mathcal{O}(n^2)$ minimal connected dominating sets and outputting one of minimum cost. □

From Theorem 6.6 we derive two new polynomially solvable cases of the WCDS problem. Recall that the graphs F_1, F_2, and a general member of the family $\{H_i\}$ are depicted in Figure 3.

Corollary 6.7. *The WCDS problem is solvable in time $\mathcal{O}(n^3)$ in the class of $\{F_1, F_2, H_1, H_2, \ldots\}$-free chordal graphs and in particular in the class of F_2-free split graphs.*

Proof. By Theorem 5.4, every $\{F_1, F_2, H_1, H_2, \ldots\}$-free chordal graphs is (hereditarily) CD so Theorem 6.6 applies. The statement for F_2-free split graphs follows from Corollary 5.10. □

We conclude this section with two remarks, one related to Theorem 6.6 and one related to Theorems 6.1 and 6.5.

Remark 6.8. The bound $\mathcal{O}(n^2)$ given by Theorem 6.6 on the number of minimal connected dominating sets in an n-vertex connected-domishold chordal graph is sharp. There exist n-vertex connected-domishold chordal graphs with $\Theta(n^2)$ minimal connected dominating sets. For instance, let S_n be the split graph with $V(S_n) = K \cup I$ where $K = \{u_1, \ldots, u_n\}$ is a clique, $I = \{v_1, \ldots, v_n\}$ is an independent set, $K \cap I = \emptyset$, and for each $i \in [n]$, vertex u_i is adjacent to all vertices of I except v_i. Since every vertex in I has a unique non-neighbor in K, we infer that S_n is F_2-free. Therefore, by Corollary 5.10 graph S_n is a (hereditarily) connected-domishold graph. Note that every set of the form $\{u_i, u_j\}$ where $1 \leq i < j \leq n$ is a minimal connected dominating set of S_n. It follows that S_n has at least $\binom{n}{2} = \Theta(|V(S_n)|^2)$ minimal connected dominating sets.

Remark 6.9. Theorems 6.1 and 6.5 motivate the question of whether there is a polynomial p such that every connected CD graph G has at most $p(|V(G)|)$ minimal vertex separators. As shown by the following family of graphs, this is not the case. For $n \geq 2$, let G_n be the graph obtained from the disjoint union of n copies of the P_4, say (x_i, a_i, b_i, y_i) for $i = 1, \ldots, n$, by identifying all vertices x_i into a single vertex x, all vertices y_i into a single vertex y, and for each vertex z other than x or y, adding a new vertex z' and making it adjacent only to z. It is not difficult to see that G_n has exactly two minimal CD sets, namely $\{a_1, \ldots, a_n\} \cup \{b_1, \ldots, b_n\} \cup \{v\}$ for $v \in \{x, y\}$. A CD structure of G_n is given by (w, t) where $t = 4n + 1$, $w(x) = w(y) = 1$, $w(a_i) = w(b_i) = 2$ for all $i \in \{1, \ldots, n\}$ and $w(z) = 0$ for all other vertices z. Therefore, G_n is CD. However, G_n has $4n + 2$ vertices and 2^n minimal x, y-separators, namely all sets of the form $\{c_1, \ldots, c_n\}$ where $c_i \in \{a_i, b_i\}$ for all i.

7 Proof of Lemma 5.3 (Diamond Lemma)

In the proof of the Diamond Lemma, we use the following notation. We write $u \sim v$ (resp. $u \nsim v$) to denote the fact that two vertices u and v are adjacent (resp. non-adjacent). Given two vertex sets A and B in a graph G, we denote by $e(A, B)$ the number of edges with one endpoint in A and one endpoint in B. A *pattern* is a triple (V, E, F) where $G = (V, E)$ is a graph and F is a subset of non-adjacent vertex pairs of G. We say that a graph G' *realizes a pattern* (V, E, F) if $V(G') = V$ and $E \subseteq E(G') \subseteq E \cup F$.

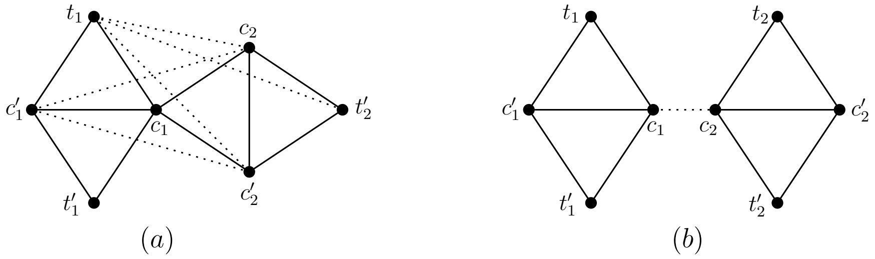


Figure 7: Two patterns (V, E, F) used in the proofs. Graphs (V, E) are depicted with solid lines. Possible additional edges (elements of F) are depicted with dotted lines.

We start with a lemma.

Lemma 7.1. *Let G be a connected chordal graph and let H be an induced subgraph of G that realizes the pattern in Figure 7(a). Moreover, suppose that:*

(1) vertices t_1 and t_1' are in different components of $G - \{c_1, c_1'\}$, and

(2) the component of $G - \{c_1, c_1'\}$ containing $\{c_2, c_2', t_2'\}$ has a vertex dominating $\{c_1, c_1'\}$.

Then G contains F_1 or F_2 as an induced subgraph.

Proof. By contradiction. Suppose that G and H satisfy the assumptions of the lemma, but G is $\{F_1, F_2\}$-free. We first show that none of the dotted edges can be present in H. We infer that $c_1' \nsim c_2$ and $c_1' \nsim c_2'$, for otherwise an induced F_1 or F_2 arises on the vertex set $V(H) \setminus \{t_1\}$, depending on whether one or both edges are present. Next, $t_1 \nsim t_2'$, since otherwise a 4-cycle arises on the vertex set $\{t_1, c_1, c_2', t_2'\}$ (if $t_1 \nsim c_2'$) or an induced F_1

arises on the vertex set $V(H)\setminus\{c_2\}$ (otherwise). Finally, we infer that $t_1 \nsim c_2$ and $t_1 \nsim c_2'$, for otherwise an induced F_1 or F_2 arises on the vertex set $V(H)\setminus\{t_1'\}$, depending whether one or both edges are present.

Let K be the component of $G - \{c_1, c_1'\}$ such that $V_2' = \{c_2, c_2', t_2'\} \subseteq V(K)$, and let $w \in V(K)$ be a vertex dominating $\{c_1, c_1'\}$ that is closest to V_2' in K. The preceding paragraph implies that $w \notin V_2'$. We will now show that $w \nsim v$ for any $v \in V_2'$. Suppose for a contradiction that $w \sim v$ for some $v \in V_2'$. Note that $w \notin \{t_1, t_1'\}$ since there are no edges between the sets $\{t_1, t_1'\}$ and V_2'. Furthermore, property (1) implies that there exists some $t \in \{t_1, t_1'\}$ such that $w \nsim t$. Suppose that $w \sim t_2'$. Then $w \sim c_2$, since otherwise a 4-cycle arises on the vertex set $\{w, c_1, c_2, t_2'\}$. But now the vertex set $\{t_2', c_2, w, c_1, c_1', t\}$ induces a copy of F_1 in G. Therefore $w \nsim t_2'$, and an induced F_1 or F_2 arises on the vertex set $V_2' \cup \{w, c_1, c_1'\}$, depending on whether w is adjacent to one or both vertices in $\{c_2, c_2'\}$. This contradiction shows that w has no neighbor in V_2'.

Let $P = (w = w_1, \ldots, w_k)$ with $w_k \in V_2'$ be a shortest w, V_2'-path in K. Note that $k \geq 3$ and the choice of P implies that for all $i \in \{1, \ldots, k-2\}$ vertex w_i is not adjacent to any vertex in V_2'. In order to avoid an induced cycle of length at least 4 within $V(P) \cup V_2' \cup \{c_1\}$, we infer that vertex c_1 must be adjacent to all the internal vertices of P (that is, to $w_2, \ldots, w_{k-1}$). Next we infer that $w_{k-1} \sim t_2'$, since otherwise the vertex set $V_2' \cup \{c_1, w_{k-1}, w_{k-2}\}$ induces a copy of F_1 or F_2 (depending on the number of edges between w_{k-1} and $\{c_2, c_2'\}$). Moreover, to avoid an induced 4-cycle on the vertex set $\{t_2', w_{k-1}, c_1, c_2\}$, we infer that $w_{k-1} \sim c_2$. But now an induced F_1 arises on the vertex set $\{t_2', c_2, c_1, w_{k-1}, w_{k-2}, w_{k-3}\}$ (where if $k = 3$ we define $w_0 = c_1'$). This last contradiction completes the proof of Lemma 7.1. □

Let us now recall Lemma 5.3.

Lemma 5.3 (Diamond Lemma). *Let G be a connected chordal graph. Suppose that G contains two induced diamonds $D_1 = (V_1, E_1)$ and $D_2 = (V_2, E_2)$ such that:*

(i) $C_1 \cap C_2 = \emptyset$.

(ii) If no vertex in C_1 is adjacent to a vertex in C_2, then there exists a C_1, C_2-separator in G of size one.

(iii) For each $j \in \{1, 2\}$ the tips (i.e., t_j, t_j') of D_j belong to different components of $G - C_j$.

(iv) For $j \in \{1, 2\}$ every component of $G - C_j$ has a vertex that dominates C_j.

Then G has an induced subgraph isomorphic to F_1, F_2, or H_i for some $i \geq 1$, where the graphs F_1, F_2, and a general member of the family $\{H_i\}$ are depicted in Figure 3.

Proof. We will prove the Diamond Lemma by contradiction through a series of claims. Let G be a connected chordal graph and let D_1 and D_2 be two induced diamonds with properties (i) – (iv) in G. Suppose for a contradiction that G is $\{F_1, F_2, H_1, H_2, \ldots\}$-free.

Claim 1. *For each $j \in \{1, 2\}$, there exists some $t \in T_j$ such that $N[t] \cap C_{3-j} = \emptyset$ (that is, each diamond has a tip that is not adjacent to any center of the other diamond).*

Proof. Suppose that each tip of D_j is adjacent to at least one vertex in C_{3-j}. Then T_j belongs to one component of $G - C_j$, contradicting property (iii). □

Claim 2. *If there exists some $t \in T_1 \cap T_2$, then $T_1 \cap T_2 = \{t\}$ and $T_j \cap C_{3-j} = \emptyset$ for $j \in \{1, 2\}$.*

Proof. Follows immediately from Claim 1 and property (iii). □

Claim 3. $|V_1 \cap V_2| \leq 1$.

Proof. First note that we have $|T_1 \cap V_2| \leq 1$, since otherwise $T_1 = T_2$, contradicting property (iii). Observe also that by property (i) we have $C_1 \cap V_2 \subseteq C_1 \cap T_2$, implying that $|C_1 \cap V_2| \leq 1$. Consequently $|V_1 \cap V_2| \leq 2$.

Now suppose for a contradiction that $|V_1 \cap V_2| = 2$. By property (i) and Claim 2 we may assume without loss of generality that $c_1 = t_2$ and $t_1' = c_2'$. To avoid an induced 4-cycle on the set $T_1 \cup T_2$ we infer that $t_1 \nsim t_2'$. Furthermore, property (iii) implies that $c_1' \nsim t_2'$ and $c_2 \nsim t_1$. But now the set $V_1 \cup V_2$ induces a copy of F_1 (if $c_1' \nsim c_2$) or a copy of F_2 (otherwise). □

Claim 4. *If $V_1 \cap V_2 = \{v\}$ then $v \in T_1 \cap T_2$.*

Proof. Suppose for a contradiction that $V_1 \cap V_2 = \{v\}$, and $v \notin T_1 \cap T_2$. Property (i) implies that $v \in T_j \cap C_{3-j}$ for some $j \in \{1, 2\}$, say $v = c_1 = t_2$. Claim 1 implies (without loss of generality) that $t_1' \nsim c_2$ and $t_1' \nsim c_2'$. Property (iii) implies that $c_1' \nsim t_2'$. Note that $t_1' \nsim t_2'$, for otherwise a 4-cycle arises on the vertex set $\{t_1', c_1, c_2, t_2'\}$. Now the subgraph of G induced by $V_1 \cup V_2$ realizes the pattern depicted in Figure $7(a)$ and we apply Lemma 7.1 to derive a contradiction. □

Claim 5. $V_1 \cap V_2 = \emptyset$.

Proof. Suppose for a contradiction that $V_1 \cap V_2 \neq \emptyset$. Claim 3 implies that $V_1 \cap V_2 = \{v\}$ and by Claim 4, $v \in T_1 \cap T_2$. Without loss of generality we may assume that $t_1 = t_2$. Claim 1 implies that there is no edge between t_1' and C_2 and between t_2' and C_1. Furthermore, we must have $t_1' \nsim t_2'$ since otherwise G contains an induced 4-cycle on the vertex set $\{t_1', c_1, c_2, t_2'\}$ (if $c_1 \sim c_2$) or an induced 5-cycle on the vertex set $\{t_1', c_1, t_1, c_2, t_2'\}$ (otherwise).

It remains to analyze the edges between C_1 and C_2. Clearly, $e(C_1, C_2) \in \{0, 1, \ldots, 4\}$. Notice that

$$e(C_1, C_2) = \begin{cases} 0 & \text{implies an induced } H_1 \text{ on the set } V_1 \cup V_2; \\ 1 & \text{implies an induced } F_1 \text{ on the vertex set } (V_1 \cup V_2) \setminus \{t_1'\}; \\ 3 & \text{implies an induced } F_1 \text{ on the vertex set } (V_1 \cup V_2) \setminus \{t_1\}; \\ 4 & \text{implies an induced } F_2 \text{ on the vertex set } (V_1 \cup V_2) \setminus \{t_1\}. \end{cases}$$

Consequently $e(C_1, C_2) = 2$, and without loss of generality, to avoid an induced 4-cycle, we may assume that $c_1 \sim c_2$ and $c_1 \sim c_2'$. But now an induced F_2 arises on the vertex set $(V_1 \cup V_2) \setminus \{t_1'\}$. □

In the rest of the proof of the Diamond Lemma we consider the edges between V_1 and V_2. By Claim 1 and property (iii) we may assume without loss of generality the following.

Assumption 1. $e(\{t_1'\}, V_2) = e(\{t_2'\}, V_1) = 0$.

Therefore, it remains to consider only the (non-)edges between $\{t_1\}$ and C_2, between $\{t_2\}$ and C_1, between C_1 and C_2, and between $\{t_1\}$ and $\{t_2\}$.

Claim 6. $e(C_1, C_2) \leq 1$.

Proof. Clearly, $e(C_1, C_2) \leq 4$. Note that if $e(C_1, C_2) \in \{3, 4\}$, then the vertex set $(V_1 \cup V_2) \setminus \{t_1, t_2\}$ induces either a copy of F_1 or a copy of F_2. Furthermore, if $e(C_1, C_2) = 2$, then, to avoid an induced 4-cycle, we may assume without loss of generality that $c_1 \sim c_2$ and $c_1 \sim c_2'$. Now the subgraph of G induced by $(V_1 \cup V_2) \setminus \{t_2\}$ realizes the pattern depicted in Figure $7(a)$ and we apply Lemma 7.1 to derive a contradiction. □

By Claim 6 we may assume without loss of generality the following.

Assumption 2. *$c_1' \nsim c_2$, $c_1' \nsim c_2'$, and $c_1 \nsim c_2'$.*

Claim 7. $e(\{t_j\}, C_{3-j}) \leq 1$ *for* $j \in \{1, 2\}$.

Proof. Suppose for a contradiction that $e(t_j, C_{3-j}) = 2$. To avoid an induced H_1 on the vertex set $(V_1 \cup V_2) \setminus \{t_{3-j}\}$, we must have an edge between C_1 and C_2. By Claim 6 and Assumption 2, we have $c_1 \sim c_2$, but now an induced F_1 arises on the vertex set $V_j \cup C_{3-j}$. □

Claim 8. *We may assume without loss of generality that $t_j \nsim c_{3-j}'$ for $j \in \{1, 2\}$.*

Proof. Let $j \in \{1, 2\}$. By Claim 7, we have that either $t_j \nsim c_{3-j}$ or $t_j \nsim c_{3-j}'$. If both edges are missing, then there is nothing to show. Suppose now that $e(t_j, C_{3-j}) = 1$. To see that we may assume that $t_j \sim c_{3-j}$, note that this can be achieved by swapping c_{3-j} and c_{3-j}' (if necessary) when $c_1 \nsim c_2$, while if $c_1 \sim c_2$, then $t_j \sim c_{3-j}$, since otherwise the vertex set $\{t_j, c_1, c_2, c_{3-j}'\}$ induces a 4-cycle in G. □

Claim 8 yields the following.

Assumption 3. *$t_1 \nsim c_2'$ and $t_2 \nsim c_1'$.*

Claim 9. $t_1 \nsim t_2$.

Proof. Suppose for a contradiction that $t_1 \sim t_2$. First we will show that $c_1 \sim t_2$ or $c_2 \sim t_1$. Suppose for a contradiction that $c_1 \nsim t_2$, and $c_2 \nsim t_1$. Then an induced H_2 arises on the set $V_1 \cup V_2$ (if $c_1 \nsim c_2$) or an induced 4-cycle on the vertex set $\{c_1, t_1, t_2, c_2\}$ (otherwise).

Without loss of generality we may assume that $c_1 \sim t_2$. By Assumption 3 we have $t_2 \nsim c_1'$, and to avoid an induced H_1 on the vertex set $(V_1 \cup V_2) \setminus \{t_1'\}$, we must have an edge between t_1 and C_2 or $c_1 \sim c_2$. If $t_1 \sim c_2$, then the vertex set $C_1 \cup C_2 \cup \{t_1, t_2\}$ induces a copy of F_1 or F_2 (depending on whether $c_1 \sim c_2$ or not). Consequently $t_1 \nsim c_2$. Therefore the only edge we can have is $c_1 \sim c_2$, but now an induced F_1 arises on the vertex set $C_1 \cup C_2 \cup \{t_1, t_2\}$. □

Claim 10. *$t_1 \nsim c_2$ and $t_2 \nsim c_1$.*

Proof. By symmetry, it suffices to show that $c_1 \nsim t_2$. Suppose for a contradiction that $c_1 \sim t_2$. Claim 9 implies that $t_1 \nsim t_2$. Recall that by Assumption 1 we have $t_2 \nsim t_1'$. Furthermore $e(\{t_1\}, C_2) = 0$, since otherwise $t_1 \sim c_2$ (by Assumption 3) and either the vertex set $\{t_1, c_1, t_2, c_2\}$ induces a 4-cycle (if $c_1 \nsim c_2$) or the vertex set $C_1 \cup C_2 \cup \{t_1, t_2\}$ induces an F_1 (otherwise).

Let K be the component of $G - C_1$ such that $V_2 \subseteq V(K)$. By property (iv) there exists a vertex in $V(K)$ that dominates C_1. Let $w \in V(K)$ be a vertex that dominates C_1 and is closest to V_2 in K. Clearly, $w \notin V_2$. Property (iii) implies that there exists some $t \in T_1$ such that $w \neq t$ and $w \nsim t$. Note that $c_1 \nsim c_2$, since otherwise the subgraph of G induced by $C_1 \cup C_2 \cup \{w, t, t_2\}$ realizes the pattern depicted in Figure $7(a)$ and we apply Lemma 7.1 to derive a contradiction. Furthermore, $w \nsim t_2'$, since otherwise t_2 and t_2' would belong to the same component of $G - C_2$, contradicting property (iii). Next, we have that $w \nsim c_2$, since otherwise either the vertex set $\{w, c_1, t_2, c_2\}$ induces a 4-cycle (if $w \nsim t_2$) or the vertex set $C_1 \cup \{t, w, t_2, c_2\}$ induces an F_1 (otherwise). By symmetry, $w \nsim c_2'$. Consequently, $w \nsim t_2$, for otherwise a copy of H_1 arises on the vertex set $C_1 \cup V_2 \cup \{w\}$.

Let $P = (w = w_1, \ldots, w_k)$ with $w_k \in V_2$ be a shortest w, V_2-path in K. Note that $k \geq 3$ and that the choice of P implies that for all $i \in \{1, \ldots, k-2\}$ vertex w_i is not adjacent to any vertex in V_2. Furthermore, $w_{k-1} \nsim t_2'$, since otherwise t_2 and t_2' would belong to the same component of $G - C_2$, contradicting property (iii). In order to avoid an induced cycle of length at least 4 within $V(P) \cup V_2 \cup \{c_1\}$, we infer that vertex c_1 must be adjacent to all the internal vertices of P (that is, $w_2, \ldots, w_{k-1}$). If $w_{k-1} \nsim t_2$, then $w_k \in C_2$, which yields an induced 4-cycle on the vertex set $\{c_1, t_2, w_k, w_{k-1}\}$. Therefore, $w_{k-1} \sim t_2$. But now either an induced H_1 arises on the vertex set $V_2 \cup \{w_{k-1}, w_{k-2}, c_1\}$ (if $e(\{w_{k-1}\}, C_2) = 0$) or an induced F_1 or F_2 arises on the vertex set $V_2 \cup \{w_{k-1}, c_1\}$ (otherwise). □

Assumptions 1 – 3 and Claims 7, 9, and 10 imply the following.

Claim 11. *The only possible edge between V_1 and V_2 is the edge c_1c_2.*

Let H be the subgraph of G induced by $V_1 \cup V_2$. By Claim 11, H realizes the pattern in Figure $7(b)$. Let K_2^{-1} be the component of $G - C_1$ containing V_2 and let U^{-1} be the set of vertices in K_2^{-1} that dominate C_1. By property (iv), set U^{-1} is non-empty. Let u^{-1} be a vertex in U^{-1} that is closest in K_2^{-1} to C_2. Graph K_1^{-2} and vertex u^{-2} are defined analogously.

By property (iii) we may assume without loss of generality the following.

Assumption 4. $t_1' \notin V(K_2^{-1})$ *and* $t_2' \notin V(K_1^{-2})$.

Claim 12. $\{u^{-1}, u^{-2}\} \cap \{t_1', t_2'\} = \emptyset$ *and* $e(\{u^{-1}, u^{-2}\}, \{t_1', t_2'\}) = 0$.

Proof. Since $u^{-1} \in V(K_2^{-1})$ and $t_1' \notin V(K_2^{-1})$, the definition of K_2^{-1} implies that $u^{-1} \neq t_1'$ and $u^{-1} \nsim t_1'$. By symmetry, we also have $u^{-2} \neq t_2'$ and $u^{-2} \nsim t_2'$.

We next show that $u^{-1} \neq t_2'$ and $u^{-1} \nsim t_2'$ (and then the remaining inequality $u^{-2} \neq t_1'$ and non-adjacency $u^{-2} \nsim t_1'$ will follow by symmetry). First note that $u^{-1} \neq t_2'$ since u^{-1} dominates C_1 and $e(\{t_2'\}, C_1) = 0$ by Assumption 1. Suppose for a contradiction that $u^{-1} \sim t_2'$. This implies that $u^{-1} \nsim t_2$, since otherwise t_2 and t_2' would belong to the same component of $G - C_2$, contradicting property (iii). But now, either an induced H_2 arises on the vertex set $V_2 \cup C_1 \cup \{u^{-1}, t_1'\}$ (if $e(\{u^{-1}\}, C_2) = 0$), or an induced H_1 arises either on the vertex set $C_1 \cup C_2 \cup \{u^{-1}, t_1', t_2'\}$ (if $e(\{u^{-1}\}, C_2) = 1$) or on the vertex set $C_1 \cup C_2 \cup \{u^{-1}, t_1', t_2\}$ (otherwise). □

Claim 13. *Vertices u^{-1} and u^{-2} are distinct and non-adjacent, and at least one of the sets $N(u^{-1}) \cap V_2$, $N(u^{-2}) \cap V_1$ is empty.*

Proof. First we prove that $u^{-1} \nsim c_2$ or $u^{-1} \nsim c_2'$. Suppose for a contradiction that $e(\{u^{-1}\}, C_2) = 2$. Then either an induced F_1 arises on the vertex set $C_1 \cup C_2 \cup \{u^{-1}, t_2'\}$ (if $c_1 \sim c_2$) or an induced H_1 arises on the vertex set $C_1 \cup C_2 \cup \{u^{-1}, t_1', t_2'\}$ (otherwise). Therefore, $u^{-1} \nsim c_2$ or $u^{-1} \nsim c_2'$, as claimed.

Since u^{-2} dominates C_2 but u^{-1} does not, we infer that $u^{-1} \neq u^{-2}$.

Next we prove that $u^{-1} \nsim u^{-2}$. Suppose for a contradiction that $u^{-1} \sim u^{-2}$. We claim that $u^{-1} \sim c_2$ or $u^{-2} \sim c_1$. Suppose to the contrary that $u^{-1} \nsim c_2$ and $u^{-2} \nsim c_1$. Then $c_1 \nsim c_2$, since otherwise an induced 4-cycle arises on the vertex set $\{c_1, c_2, u^{-2}, u^{-1}\}$. Furthermore, $u^{-1} \sim c_2'$ or $u^{-2} \sim c_1'$, since otherwise an induced H_2 arises on the vertex set $C_1 \cup C_2 \cup \{t_1', u^{-1}, u^{-2}, t_2'\}$. If only one of the edges $u^{-1}c_2'$ and $u^{-2}c_1'$ is present, say $u^{-1}c_2'$, then an induced H_1 arises on the vertex set $C_1 \cup C_2 \cup \{t_1', u^{-1}, u^{-2}\}$. If both edges $u^{-1}c_2'$ and $u^{-2}c_1'$ are present, then an induced F_1 arises on the vertex set $C_1 \cup C_2 \cup \{u^{-1}, u^{-2}\}$. Both cases lead to a contradiction, thus $u^{-1} \sim c_2$ or $u^{-2} \sim c_1$, as claimed. We may assume without loss of generality that $u^{-1} \sim c_2$. Now we must have $c_1 \nsim c_2$ and $c_1 \nsim u^{-2}$, since otherwise an induced F_1 or F_2 arises on the vertex set $C_1 \cup C_2 \cup \{u^{-1}, u^{-2}\}$, depending on whether one or both edges are present. But now an induced H_1 arises on the vertex set $C_1 \cup C_2 \cup \{t_1', u^{-1}, u^{-2}\}$, a contradiction.

To complete the proof, we consider the two cases depending on whether c_1 is adjacent to c_2 or not. Suppose first that $c_1 \sim c_2$. Then $u^{-1} \nsim c_2'$, for otherwise $u^{-1} \nsim c_2$ and G contains an induced 4-cycle on the vertex set $\{u^{-1}, c_2', c_2, c_1\}$. By symmetry, we also have $u^{-2} \nsim c_1'$. If $u^{-1} \sim c_2$ and $u^{-2} \sim c_1$, then an induced F_1 arises on the vertex set $C_1 \cup C_2 \cup \{u^{-1}, u^{-2}\}$. It follows that H contains at most one of the edges $u^{-1}c_2$ and $u^{-2}c_1$. By symmetry, we may assume without loss of generality that $u^{-1} \nsim c_2$. We infer that $u^{-1} \nsim t_2$, since otherwise G contains an induced C_4 on the vertex set $\{u^{-1}, c_1, c_2, t_2\}$. It follows that the set $N(u^{-1}) \cap V_2$ is empty.

Finally, suppose that $c_1 \nsim c_2$. Then either $e(\{u^{-1}\}, C_2) = 0$ or $e(\{u^{-2}\}, C_1) = 0$, for otherwise G contains an induced 4-cycle on the vertex set $\{u^{-1}, x, u^{-2}, y\}$ where $x \in N(u^{-1}) \cap C_2$ and $y \in N(u^{-2}) \cap C_1$. By symmetry, we may assume without loss of generality that $e(\{u^{-1}\}, C_2) = 0$. We infer that $u^{-1} \nsim t_2$, since otherwise G contains an induced H_2 on the vertex set $C_1 \cup V_2 \cup \{u^{-1}, t_1'\}$. It follows that the set $N(u^{-1}) \cap V_2$ is empty. □

By Claim 13 we may assume without loss of generality the following.

Assumption 5. $e(\{u^{-1}\}, V_2) = 0$.

Claim 14. $c_1 \nsim c_2$.

Proof. Suppose for a contradiction that $c_1 \sim c_2$ and consider K_2^{-1}, u^{-1}, K_1^{-2}, and u^{-2}. Clearly, $u^{-1} \notin C_1 \cup C_2 \cup \{t_2'\}$. Moreover, by Claim 12 we have we have $u^{-1} \neq t_1'$ and $u^{-1} \nsim t_1'$. Also, by symmetry, $u^{-2} \notin C_1 \cup C_2 \cup \{t_1'\}$, $u^{-2} \neq t_2'$ and $u^{-2} \nsim t_2'$. Furthermore, by Assumption 5 we have $N(u^{-1}) \cap V_2 = \emptyset$.

Let $P^{-1} = (u^{-1} = u_1, u_2, \ldots, u_k)$, with $u_k \in V_2' = C_2 \cup \{t_2'\}$ be a shortest u^{-1}, V_2'-path in K_2^{-1}, and similarly, let $P^{-2} = (u^{-2} = v_1, v_2, \ldots, v_\ell)$, with $v_\ell \in V_1' =$

$C_1 \cup \{u^{-1}, t_1'\}$ be a shortest u^{-2}, V_1'-path in $V(K_1^{-2})$. The fact that $N(u^{-1}) \cap V_2 = \emptyset$ implies that $k \geq 3$. Furthermore, Claims 11 and 13 imply that $u^{-2} \notin V_1 \cup \{u^{-1}\}$. Therefore, $\ell \geq 2$.

Since $u^{-1} \nsim c_2$, we infer that vertex c_1 must be adjacent to all the internal vertices of P^{-1}, for otherwise G would contain an induced cycle of length at least 4. Consequently, the definition of u^{-1} implies that $u_j \nsim c_1'$ for all $j \in \{2, \ldots, k-1\}$.

Suppose that $u_{k-1} \sim c_2'$. To avoid an induced 4-cycle on the vertex set $\{c_1, c_2', c_2, u_{k-1}\}$, we infer that $u_{k-1} \sim c_2$. We must have $k = 3$ since if $k \geq 4$, then the vertex set $C_2 \cup \{c_1, u_{k-1}, u_{k-2}, u_{k-3}\}$ induces a copy of F_1. But now, since $c_1' \nsim u_2$, an induced copy of F_1 arises on the vertex set $C_1 \cup C_2 \cup \{u_1, u_2\}$, a contradiction. Therefore, $u_{k-1} \nsim c_2'$.

Suppose that $u_{k-1} \sim t_2'$. To avoid an induced 4-cycle on the vertex set $\{c_1, c_2, t_2', u_{k-1}\}$, we must have $u_{k-1} \sim c_2$. But now, the vertex set $V_2' \cup \{u_{k-1}, u_{k-2}, c_1\}$ induces a copy of F_1, a contradiction. Therefore, $u_{k-1} \nsim t_2'$. Consequently, $u_k = c_2$.

Suppose that $u^{-2} \sim c_1$. If in addition $u^{-2} \nsim u_{k-1}$, then also $u^{-2} \nsim u_{k-2}$ (since otherwise the vertex set $\{u_{k-2}, u_{k-1}, c_2, u^{-2}\}$ would induce a 4-cycle), but now, the vertex set $\{u_{k-2}, u_{k-1}, c_1, c_2, c_2', u^{-2}\}$ induces a copy of F_1, a contradiction. Therefore, $u^{-2} \sim u_{k-1}$. Let u_i be the neighbor of u^{-2} on P^{-1} minimizing i. Since $u_1 \nsim u^{-2}$, we have $i \geq 2$. Moreover, since $u^{-2} \sim u_{k-1}$, we have $i \leq k-1$. But now, the vertex set $C_2 \cup \{u_{i-1}, c_1, u_i, u^{-2}\}$ induces either a copy of F_1 (if $u_i \nsim c_2$) or of F_2 (otherwise), a contradiction. Therefore, $u^{-2} \nsim c_1$.

Note that $N(u^{-2}) \cap V_1 = \emptyset$, for otherwise if there is a vertex $x \in N(u^{-2}) \cap V_1$, then $x \neq c_1$ and G contains an induced 4-cycle on the vertex set $\{u^{-2}, c_2, c_1, x\}$, a contradiction. Since $N(u^{-2}) \cap V_1 = \emptyset$, we can now apply symmetric arguments as for P^{-1} to deduce that $\ell \geq 3$, vertex c_2 is adjacent to all the internal vertices of P^{-2}, and $v_\ell = c_1$.

Suppose first that $V(P^{-1}) \cap V(P^{-2}) = \emptyset$. To avoid an induced 4-cycle on the vertex set $\{u_{k-2}, c_2, c_1, v_{\ell-2}\}$, we infer that $u_{k-2} \nsim v_{\ell-2}$. Suppose that $u_{k-1} \nsim v_{\ell-1}$. Then also $u_{k-1} \nsim v_{\ell-2}$ (since otherwise we would have an induced 4-cycle on the vertex set $\{u_{k-1}, v_{\ell-2}, v_{\ell-1}, c_1\}$) and by a symmetric argument also $u_{k-2} \nsim v_{\ell-1}$. But now, we have an induced F_1 on the vertex set $\{u_{k-2}, c_1, u_{k-1}, c_2, v_{\ell-1}, v_{\ell-2}\}$. Thus, $u_{k-1} \sim v_{\ell-1}$. Moreover, we have either $u_{k-2} \sim v_{\ell-1}$ or $v_{\ell-2} \sim u_{k-1}$, since otherwise an induced F_2 arises on the vertex set $\{c_1, v_{\ell-1}, v_{\ell-2}, c_2, u_{k-1}, u_{k-2}\}$. Without loss of generality, assume that $u_{k-2} \sim v_{\ell-1}$. But now, setting $v_0 = c_2'$ if $\ell = 3$, either an induced 4-cycle arises on the vertex set $\{u_{k-2}, v_{\ell-1}, v_{\ell-2}, v_{\ell-3}\}$ (if $u_{k-2} \sim v_{\ell-3}$) or an induced copy of F_1 arises on the vertex set $\{u_{k-2}, c_1, v_{\ell-1}, v_{\ell-2}, c_2, v_{\ell-3}\}$ (otherwise). This contradiction shows that $V(P^{-1}) \cap V(P^{-2}) \neq \emptyset$.

Since $v_\ell = c_1$ and due to the minimality of P^{-2}, we have $N(c_1) \cap V(P^{-2}) = \{v_{\ell-1}\}$. On the other hand, since c_1 dominates P^{-1}, we have $N(c_1) \cap V(P^{-1}) = V(P^{-1})$. Therefore

$$\begin{aligned}\emptyset \neq V(P^{-2}) \cap V(P^{-1}) &= V(P^{-2}) \cap \big(N(c_1) \cap V(P^{-1})\big) \\ &= \big(N(c_1) \cap V(P^{-2})\big) \cap V(P^{-1}) = \{v_{\ell-1}\} \cap V(P^{-1}) \subseteq \{v_{\ell-1}\},\end{aligned}$$

which yields $V(P^{-1}) \cap V(P^{-2}) = \{v_{\ell-1}\}$. A symmetric argument implies that $V(P^{-1}) \cap V(P^{-2}) = \{u_{k-1}\}$; in particular, $v_{\ell-1} = u_{k-1}$. To avoid an induced 4-cycle on the vertex set $\{u_{k-2}, c_1, c_2, v_{\ell-2}\}$, we infer that $u_{k-2} \nsim v_{\ell-2}$. But now, an induced copy of F_1 arises on the vertex set $\{u_{k-3}, u_{k-2}, c_1, u_{k-1}, c_2, v_{\ell-2}\}$ (where if $k = 3$ we define $u_0 = c_1'$). This contradiction completes the proof of Claim 14. □

By Claim 5, we have $V_1 \cap V_2 = \emptyset$. By Assumptions 1 and 2 and Claims 9, 10, and 14 we have $e(V_1, V_2) = 0$. However, since G is connected, there exists a path connecting the two diamonds D_1 and D_2. In particular, we will again consider K_2^{-1}, u^{-1}, K_1^{-2}, and u^{-2}, and analyze the possible interrelations between two particular paths to produce a forbidden induced subgraph.

Recall that by Assumption 4 we have $t_1' \notin V(K_2^{-1})$ and $t_2' \notin V(K_1^{-2})$. Furthermore, since $e(V_1, V_2) = 0$, we have $u^{-1} \notin V_2$ and $u^{-2} \notin V_1$. Recall also that Claim 13 implies that $u^{-1} \neq u^{-2}$, $u^{-1} \nsim u^{-2}$.

Let $P^{-1} = (u^{-1} = u_1, u_2, \dots, u_k)$, with $u_k \in C_2$, be a shortest u^{-1}, C_2-path in K_2^{-1}, and let $P^{-2} = (u^{-2} = v_1, v_2, \dots, v_\ell)$, with $v_\ell \in C_1$, be a shortest u^{-2}, C_1-path in K_1^{-2}. We may assume that $u_k = c_2$ and $v_\ell = c_1$. The fact that $N(u^{-1}) \cap V_2 = \emptyset$ implies that $k \geq 3$ and since $u^{-2} \notin C_1$, we have $\ell \geq 2$.

Claim 15. $\ell \geq 3$.

Proof. Suppose that $\ell = 2$. Then, $u^{-2} \sim c_1$. Moreover, we have that $u^{-2} \nsim c_1'$ since otherwise u^{-2} would be a vertex in U^{-1} closer in K_2^{-1} to C_2 than u^{-1}, which is impossible due to the definition of u^{-1}.

We first show that $u^{-2} \neq u_{k-1}$. Suppose that $u^{-2} = u_{k-1}$. Then $u_{k-1} \sim c_1$ and $u_{k-1} \sim c_2'$. Hence, in order to avoid an induced cycle of length at least 4 within $V(P^{-1}) \cup \{c_1\}$, we infer that vertex c_1 must be adjacent to all the internal vertices of P^{-1}. By Assumption 4, vertex t_2' has no neighbors in the set $V(K_1^{-2})$; in particular, t_2' has no neighbors in the set $V(P^{-1}) \cup C_1$. Therefore, G contains an induced H_1 on the vertex set $C_2 \cup \{t_2', c_1, u_{k-1}, u_{k-2}, u_{k-3}\}$ (where if $k = 3$ we define $u_0 = c_1'$), a contradiction.

Suppose that $u_{k-1} \sim c_1$. In particular, $u_{k-1} \neq t_2'$. To avoid an induced 4-cycle on the vertex set $\{c_1, u_{k-1}, c_2, u^{-2}\}$, we infer that $u_{k-1} \sim u^{-2}$. Moreover, $u_{k-1} \sim t_2'$ since otherwise the vertex set $C_2 \cup \{t_2', u^{-2}, u_{k-1}, c_1\}$ induces a copy of either F_1 (if $u_{k-1} \nsim c_2'$) or F_2 (otherwise). But now u^{-2} and t_2' are in the same component of $G - C_2$, contradicting the fact that $u^{-2} \in V(K_1^{-2})$ and $t_2' \notin V(K_1^{-2})$. This contradiction implies that $u_{k-1} \nsim c_1$.

Let $j \in \{1, \dots, k\}$ be the maximum index such that $c_1 \sim u_j$. Then $j \leq k-2$. To avoid a long induced cycle, we infer that $c_1 \sim u_{j'}$ for all $j' \in \{1, \dots, j\}$. Let $i \in \{1, \dots, k\}$ be the minimum index such that $u^{-2} \sim u_i$. Note that $i > 1$ since $u_1 = u^{-1} \nsim u^{-2}$. To avoid a long induced cycle, we infer that $i \leq j$ and that $u^{-2} \sim u_{i'}$ for all $i' \in \{i, \dots, k\}$. Note that if $i < j$, then $(u^{-1} = u_1, u_2, \dots, u_i, u^{-2}, u_k = c_2)$ is a u^{-1}, V_2'-path in K_2^{-1} strictly shorter than P^{-1}, contradicting the minimality of P^{-1}. Therefore, $i = j$. But now, the vertex set $\{u_{j-1}, u_j, u_{j+1}, u_{j+2}, u^{-2}, c_1\}$ induces a copy of F_1. This contradiction implies that $\ell \geq 3$. □

Claim 16. $u_{k-1} \neq v_1$ *and* $v_{\ell-1} \neq u_1$.

Proof. Suppose for a contradiction that $u_{k-1} = v_1$. Recall that $v_1 = u^{-2}$. By the minimality of P^{-1}, we have $c_2 \nsim u_j$ and $c_2' \nsim u_j$ for every $j \in \{1, \dots, k-2\}$. Furthermore, since $u_1 = u^{-1} \nsim u^{-2} = u_{k-1}$, we have $k \geq 4$. Since u^{-2} and t_2' are in different components of $G - C_2$, we infer that $t_2' \nsim u_j$ for all $j \in \{1, \dots, k-2\}$. If $c_1 \sim u_3$, then we obtain an induced copy of H_i for some $i \geq 1$ on the vertex set

$$C_2 \cup \{t_2', v_1 = u_{k-1}, u_{k-2}, \dots, u_j, u_{j-1}, u_{j-2}, c_1\},$$

where $j \in \{3, \dots, k\}$ is the maximum index such that $c_1 \sim u_j$. (Note that $j \leq k-2$ since $c_1 \nsim c_2 = u_k$ and $c_1 = v_\ell \nsim v_1 = u_{k-1}$ by Claim 15.)

Therefore, $c_1 \nsim u_3$, and to avoid a long induced cycle, also $c_1 \nsim u_j$ for $j \geq 4$. A similar argument shows that $c_1' \nsim u_j$ for $j \geq 3$. If $c_1 \nsim u_2$ and $c_1' \nsim u_2$, then we obtain an induced copy of some H_i on the vertex set $V(P^{-1}) \cup C_1 \cup C_2 \cup \{t_1', t_2'\}$. If $c_1 \sim u_2$ and $c_1' \nsim u_2$ (or vice-versa), then an induced copy of some H_i arises on the vertex set $V(P^{-1}) \cup C_1 \cup C_2 \cup \{t_2'\}$, and if $c_1 \sim u_2$ and $c_1' \sim u_2$, then an induced copy of some H_i arises on the vertex set $(V(P^{-1}) \setminus \{u_1\}) \cup C_1 \cup C_2 \cup \{t_1', t_2'\}$. This contradiction shows that $u_{k-1} \neq v_1$.

Similar arguments as above imply that $v_{\ell-1} \neq u_1$. □

Property (ii) in the statement of the Diamond Lemma implies the following.

Claim 17. $V(P^{-1}) \cap V(P^{-2}) \neq \emptyset$.

We are now ready to complete the proof of the Diamond Lemma. Let $r \in \{1, \ldots, k\}$ be the minimum index such that $u_r \in V(P^{-2})$. Note that $r < k$, since $u_k \in C_2$ and $C_2 \cap V(P^{-2}) = \emptyset$. Let $s \in \{1, \ldots, \ell\}$ be the index such that $u_r = v_s$. If $r = 1$, then $u_1 = v_{\ell-1}$, contradicting Claim 16. Therefore, $r \geq 2$. Similarly, if $s = 1$, then $v_1 = u_{k-1}$, again contradicting Claim 16. Therefore, $s \geq 2$.

Consider the path $Q = (u_1, \ldots, u_r = v_s, v_{s-1}, \ldots, v_1)$. Let D and D' be the subgraphs of G induced by $\{t_1', c_1, c_1', u_1\}$ and $\{t_2', c_2, c_2', v_1\}$, respectively. Notice that D and D' are diamonds. We will refer to tips u_1 and v_1 as the *roots* of D and D', respectively. Then, Q is a path connecting the two roots. Moreover, by Assumption 4 we have $t_1' \notin V(K_2^{-1})$ and $V(Q) \subseteq V(K_2^{-1})$, we infer that t_1' has no neighbors on Q. Similarly, t_2' has no neighbors on Q.

We may also assume that Q is an induced path; otherwise, we replace Q with a shortest u_1, v_1-path in $G[V(Q)]$. To complete the proof, we will show that G is not $\{F_1, F_2, H_1, H_2, \ldots\}$-free. We say that an induced subgraph H of G is a *weakly induced* H_n if H has a spanning subgraph H_n with $n \geq 1$ consisting of two diamonds and a path connecting them such that, assuming notation from Figure 8, the following holds:

(i) each of the two diamonds is induced in G,

(ii) there are no edges in G connecting a vertex from one diamond with a vertex from another diamond, except perhaps edges incident with their roots (if $n = 1$) or the unique edge on the path connecting the two roots (if $n = 2$),

(iii) the path connecting the two diamonds is induced in G, and

(iv) vertices x_1 and z_1 do not have any neighbors on the path.

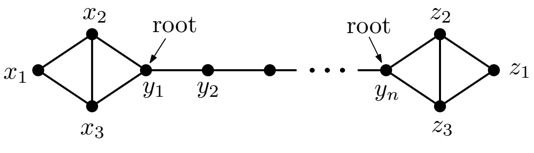


Figure 8: A weakly induced H_n.

Note, in particular, that for $n \in \{1, 2\}$ every weakly induced H_n is isomorphic to H_n.

The above considerations show that the subgraph of G induced by

$$V(D) \cup V(D') \cup V(Q)$$

contains a weakly induced H_n. Choose one such induced subgraph, say H, with minimum value of n, and let F be the corresponding spanning subgraph of H isomorphic to H_n. To complete the proof, we will now show that either H equals F or G contains an induced F_1 or F_2. Suppose that this is not the case. The only possible edges that can be present in H but not in F are those connecting one of the vertices x_2, x_3, z_2, z_3 with one of the vertices in the set $\{y_2, \dots, y_{n-1}\}$.

Let us first show that for each $i \in \{2, \dots, n-1\}$, at most one of x_2 and x_3 is adjacent to y_i. Suppose that $x_2 \sim y_i$ and $x_3 \sim y_i$ for some $i \in \{2, \dots, n-1\}$. Then $y_i \sim z_2$ or $y_i \sim z_3$, since otherwise the subgraph of G induced by $\{x_1, x_2, x_3, y_i, \dots, y_n, z_1, z_2, z_3\}$ would be a weakly induced H_{n-i+1}, contradicting the minimality of H. If $y_i \sim z_2$ and $y_i \sim z_3$, then the vertex set $\{x_1, x_2, x_3, y_i, z_1, z_2, z_3\}$ induces an H_1 in G. We may thus assume that y_i is adjacent only to one of z_2, z_3, say to z_3. If $i = n-1$, then the vertex set $\{x_1, x_2, x_3, y_{n-1}, y_n, z_2, z_3\}$ induces an H_1 in G. If $i \leq n-2$, then the fact that G is chordal implies that $z_3 \sim y_j$ for all $j \in \{i, \dots, n\}$, and the vertex set $\{x_1, x_2, x_3, y_i, y_{i+1}, y_{i+2}, z_3\}$ induces an H_1 in G. This contradiction shows that for each $i \in \{2, \dots, n-1\}$, at most one of x_2 and x_3 is adjacent to y_i.

Next, we argue that at least one of x_2 and x_3 is not adjacent to any vertex y_i with $i \in \{2, \dots, n-1\}$. Indeed, if $x_2 \sim y_r$ and $x_3 \sim y_s$, with $2 \leq r \leq s \leq n-1$ (say), then $r < s$ and the fact that G is chordal implies that $x_3 \sim y_j$ for all $j \in \{2, \dots, s\}$, contradicting the fact that at most one of x_2 and x_3 is adjacent to y_r. Therefore, we may assume without loss of generality that x_2 has no neighbors in the set $\{y_2, \dots, y_{n-1}\}$. Similarly, we may assume that z_2 has no neighbors in the set $\{y_2, \dots, y_{n-1}\}$.

Let $r \in \{1, \dots, n-1\}$ be the maximum index such that $x_3 \sim y_r$. Similarly, let $s \in \{2, \dots, n\}$ be the minimum index such that $z_3 \sim y_s$. If $r = 1$ and $s = n$, then $H = F$ and we are done. Thus, we may assume without loss of generality that $r \geq 2$. In particular, this implies that $x_3 \sim y_2$ (since G is chordal). If $y_2 \nsim z_3$, then the subgraph of G induced by $\{x_2, x_3, y_1, \dots, y_n, z_1, z_2, z_3\}$ is a weakly induced H_{n-1}, contradicting the minimality of n. Therefore, $y_2 \sim z_3$, or, equivalently, $s = 2$. A similar argument shows that $r = n-1$. Now, if $n = 3$, then the vertex set $\{x_2, x_3, y_1, y_2, y_3, z_2, z_3\}$ induces an H_1 in G, and if $n \geq 4$, then the vertex set $\{x_2, x_3, y_1, y_2, y_3, z_3\}$ induces an F_1 in G. This contradiction completes the proof of the Diamond Lemma. □

References

[1] C. Benzaken and P. L. Hammer, Linear separation of dominating sets in graphs, *Ann. Discrete Math.* **3** (1978), 1–10, doi:10.1016/s0167-5060(08)70492-8.

[2] C. Berge, *Hypergraphs*, volume 45 of *North-Holland Mathematical Library*, North-Holland, Amsterdam, 1989.

[3] A. Berry, J.-P. Bordat and O. Cogis, Generating all the minimal separators of a graph, *Internat. J. Found. Comput. Sci.* **11** (2000), 397–403, doi:10.1142/s0129054100000211.

[4] A. Berry and R. Pogorelcnik, A simple algorithm to generate the minimal separators and the maximal cliques of a chordal graph, *Inform. Process. Lett.* **111** (2011), 508–511, doi:10.1016/j.ipl.2011.02.013.

[5] P. Bertolazzi and A. Sassano, An $O(mn)$ algorithm for regular set-covering problems, *Theoret. Comput. Sci.* **54** (1987), 237–247, doi:10.1016/0304-3975(87)90131-9.

[6] J. Blum, M. Ding, A. Thaeler and X. Cheng, Connected dominating set in sensor networks and MANETs, in: D.-Z. Du and P. M. Pardalos (eds.), *Handbook of Combinatorial Optimization, Supplement Volume B*, Springer, New York, pp. 329–369, 2005, doi:10.1007/0-387-23830-1_8.

[7] E. Boros, *Dualization of Aligned Boolean Functions*, RUTCOR Research Report 9-94, Rutgers Center for Operations Research, 1994, `http://rutcor.rutgers.edu/pub/rrr/reports94/09.ps`.

[8] E. Boros, V. Gurvich and M. Milanič, Decomposing 1-Sperner hypergraphs, 2018, `arXiv:1510.02438 [math.CO]`.

[9] V. Bouchitté and I. Todinca, Treewidth and minimum fill-in: grouping the minimal separators, *SIAM J. Comput.* **31** (2001), 212–232, doi:10.1137/s0097539799359683.

[10] A. Brandstädt, V. B. Le and J. P. Spinrad, *Graph Classes: A Survey*, SIAM Monographs on Discrete Mathematics and Applications, SIAM, Philadelphia, Pennsylvania, 1999, doi: 10.1137/1.9780898719796.

[11] S. Butenko, S. Kahruman-Anderoglu and O. Ursulenko, On connected domination in unit ball graphs, *Optim. Lett.* **5** (2011), 195–205, doi:10.1007/s11590-010-0211-0.

[12] L. S. Chandran, A. Das, D. Rajendraprasad and N. M. Varma, Rainbow connection number and connected dominating sets, *J. Graph Theory* **71** (2012), 206–218, doi:10.1002/jgt.20643.

[13] L. S. Chandran and F. Grandoni, A linear time algorithm to list the minimal separators of chordal graphs, *Discrete Math.* **306** (2006), 351–358, doi:10.1016/j.disc.2005.12.010.

[14] M.-S. Chang, Weighted domination of cocomparability graphs, *Discrete Appl. Math.* **80** (1997), 135–148, doi:10.1016/s0166-218x(97)80001-7.

[15] M.-S. Chang, Efficient algorithms for the domination problems on interval and circular-arc graphs, *SIAM J. Comput.* **27** (1998), 1671–1694, doi:10.1137/s0097539792238431.

[16] N. Chiarelli and M. Milanič, Linear separation of total dominating sets in graphs, in: K. Jansen and R. Reischuk (eds.), *Graph-Theoretic Concepts in Computer Science*, Springer, Heidelberg, volume 8165 of *Lecture Notes in Computer Science*, pp. 165–176, 2013, doi:10.1007/978-3-642-45043-3_15, revised papers from the 39th International Workshop (WG 2013) held in Lübeck, June 19 – 21, 2013.

[17] N. Chiarelli and M. Milanič, Linear separation of connected dominating sets in graphs (extended abstract), in: *International Symposium on Artificial Intelligence and Mathematics (ISAIM 2014)*, 2014, held in Fort Lauderdale, Florida, USA, January 6 – 8, 2014, `https://www.cs.uic.edu/pub/Isaim2014/WebPreferences/ISAIM2014_Boolean_Chiarelli_Milanic.pdf`.

[18] N. Chiarelli and M. Milanič, Total domishold graphs: a generalization of threshold graphs, with connections to threshold hypergraphs, *Discrete Appl. Math.* **179** (2014), 1–12, doi:10.1016/j.dam.2014.09.001.

[19] C. K. Chow, Boolean functions realizable with single threshold devices, *Proc. IRE* **49** (1961), 370–371, doi:10.1109/jrproc.1961.287827.

[20] V. Chvátal and P. L. Hammer, Aggregation of inequalities in integer programming, in: P. L. Hammer, E. L. Johnson, B. H. Korte and G. L. Nemhauser (eds.), *Studies in Integer Programming*, North-Holland, Amsterdam, volume 1 of *Annals of Discrete Mathematics*, 1977 pp. 145–162, doi:10.1016/s0167-5060(08)70731-3, proceedings of the Workshop on Integer Programming held in Bonn, September 8 – 12, 1975.

[21] Y. Crama, Dualization of regular Boolean functions, *Discrete Appl. Math.* **16** (1987), 79–85, doi:10.1016/0166-218x(87)90056-4.

[22] Y. Crama and P. L. Hammer, *Boolean Functions: Theory, Algorithms, and Applications*, volume 142 of *Encyclopedia of Mathematics and its Applications*, Cambridge University Press, Cambridge, 2011, doi:10.1017/cbo9780511852008.

[23] H. N. de Ridder et al., Information System on Graph Classes and their Inclusions (ISGCI), 2001–2014, `http://www.graphclasses.org/` (accessed on 30 August 2016).

[24] I. Dinur and D. Steurer, Analytical approach to parallel repetition, in: D. Shmoys (ed.), *STOC'14: Proceedings of the 2014 ACM Symposium on Theory of Computing*, ACM Press, New York, pp. 624–633, 2014, held in New York, May 31 – June 3, 2014.

[25] G. A. Dirac, On rigid circuit graphs, *Abh. Math. Sem. Univ. Hamburg* **25** (1961), 71–76, doi:10.1007/bf02992776.

[26] D.-Z. Du and P.-J. Wan, *Connected Dominating Set: Theory and Applications*, volume 77 of *Springer Optimization and Its Applications*, Springer, New York, 2013, doi:10.1007/978-1-4614-5242-3.

[27] W. Duckworth and B. Mans, Connected domination of regular graphs, *Discrete Math.* **309** (2009), 2305–2322, doi:10.1016/j.disc.2008.05.029.

[28] T. Eiter, K. Makino and G. Gottlob, Computational aspects of monotone dualization: a brief survey, *Discrete Appl. Math.* **156** (2008), 2035–2049, doi:10.1016/j.dam.2007.04.017.

[29] C. C. Elgot, Truth functions realizable by single threshold organs, in: *Switching Circuit Theory and Logical Design*, IEEE Computer Society, 1961 pp. 225–245, doi:10.1109/focs.1961.39, papers from the 1st Annual Symposium held in Chicago, Illinois, USA, October 9 – 14, 1960.

[30] S. Foldes and P. L. Hammer, Split graphs, in: F. Hoffman, R. C. Mullin, K. B. Reid and R. G. Stanton (eds.), *Proceedings of the Eighth Southeastern Conference on Combinatorics, Graph Theory and Computing*, Utilitas Mathematica Publishing, Winnipeg, Manitoba, 1977 pp. 311–315, held at Louisiana State University, Baton Rouge, Louisiana, February 28 – March 3, 1977.

[31] F. V. Fomin, F. Grandoni and D. Kratsch, Solving connected dominating set faster than 2^n, *Algorithmica* **52** (2008), 153–166, doi:10.1007/s00453-007-9145-z.

[32] I. J. Gabelman, *The Functional Behavior of Majority (Threshold) Elements*, Ph.D. thesis, Department of Electrical Engineering, Syracuse University, New York, 1961.

[33] M. C. Golumbic, Trivially perfect graphs, *Discrete Math.* **24** (1978), 105–107, doi:10.1016/0012-365x(78)90178-4.

[34] M. C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, volume 57 of *Annals of Discrete Mathematics*, Elsevier, Amsterdam, 2nd edition, 2004, doi:10.1016/c2013-0-10739-8.

[35] T. W. Haynes, S. T. Hedetniemi and P. J. Slater (eds.), *Domination in Graphs: Advanced Topics*, volume 209 of *Monographs and Textbooks in Pure and Applied Mathematics*, Marcel Dekker, New York, 1998.

[36] T. W. Haynes, S. T. Hedetniemi and P. J. Slater, *Fundamentals of Domination in Graphs*, volume 208 of *Monographs and Textbooks in Pure and Applied Mathematics*, Marcel Dekker, New York, 1998.

[37] R.-W. Hung and M.-S. Chang, A simple linear algorithm for the connected domination problem in circular-arc graphs, *Discuss. Math. Graph Theory* **24** (2004), 137–145, doi:10.7151/dmgt.1220.

[38] M. M. Kanté, V. Limouzy, A. Mary and L. Nourine, On the enumeration of minimal dominating sets and related notions, *SIAM J. Discrete Math.* **28** (2014), 1916–1929, doi:10.1137/120862612.

[39] N. Karmarkar, A new polynomial-time algorithm for linear programming, *Combinatorica* **4** (1984), 373–395, doi:10.1007/bf02579150.

[40] J. M. Keil, The complexity of domination problems in circle graphs, *Discrete Appl. Math.* **42** (1993), 51–63, doi:10.1016/0166-218x(93)90178-q.

[41] H. Kellerer, U. Pferschy and D. Pisinger, *Knapsack Problems*, Springer-Verlag, Berlin, 2004, doi:10.1007/978-3-540-24777-7.

[42] C. J. Klivans, Threshold graphs, shifted complexes, and graphical complexes, *Discrete Math.* **307** (2007), 2591–2597, doi:10.1016/j.disc.2006.11.018.

[43] T. Kloks, H. Bodlaender, H. Müller and D. Kratsch, Computing treewidth and minimum fill-in: all you need are the minimal separators, in: T. Lengauer (ed.), *Algorithms – ESA '93*, Springer, Berlin, volume 726 of *Lecture Notes in Computer Science*, 1993 pp. 260–271, doi:10.1007/3-540-57273-2_61, proceedings of the First Annual European Symposium held in Bad Honnef, September 30 – October 2, 1993.

[44] T. Kloks and D. Kratsch, Listing all minimal separators of a graph, *SIAM J. Comput.* **27** (1998), 605–613, doi:10.1137/s009753979427087x.

[45] P. S. Kumar and C. E. Veni Madhavan, Minimal vertex separators of chordal graphs, *Discrete Appl. Math.* **89** (1998), 155–168, doi:10.1016/s0166-218x(98)00123-1.

[46] R. Laskar and J. Pfaff, *Domination and Irredundance in Split Graphs*, Technical Report 430, Department of Mathematical Sciences, Clemson University, Clemson, South Carolina, 1983.

[47] V. E. Levit, M. Milanič and D. Tankus, On the recognition of k-equistable graphs, in: M. C. Golumbic, M. Stern, A. Levy and G. Morgenstern (eds.), *Graph-Theoretic Concepts in Computer Science*, Springer, Heidelberg, volume 7551 of *Lecture Notes in Computer Science*, pp. 286–296, 2012, doi:10.1007/978-3-642-34611-8_29, revised selected papers from the 38th International Workshop (WG 2012) held in Jerusalem, June 26 – 28, 2012.

[48] N. V. R. Mahadev and U. N. Peled, *Threshold Graphs and Related Topics*, volume 56 of *Annals of Discrete Mathematics*, North-Holland, Amsterdam, 1995.

[49] D. Michalak, Hereditarily dominated graphs, *Discrete Math.* **307** (2007), 952–957, doi:10.1016/j.disc.2005.11.044.

[50] M. Milanič, J. Orlin and G. Rudolf, Complexity results for equistable graphs and related classes, *Ann. Oper. Res.* **188** (2011), 359–370, doi:10.1007/s10479-010-0720-3.

[51] P. Montealegre and I. Todinca, On distance-d independent set and other problems in graphs with "few" minimal separators, in: P. Heggernes (ed.), *Graph-Theoretic Concepts in Computer Science*, Springer, Berlin, volume 9941 of *Lecture Notes in Computer Science*, pp. 183–194, 2016, doi:10.1007/978-3-662-53536-3_16, revised selected papers from the 42nd International Workshop (WG 2016) held at Boğaziçi University, Istanbul, June 22 – 24, 2016.

[52] H. Müller and A. Brandstädt, The NP-completeness of steiner tree and dominating set for chordal bipartite graphs, *Theoret. Comput. Sci.* **53** (1987), 257–265, doi:10.1016/0304-3975(87)90067-3.

[53] S. Muroga, *Threshold Logic and its Applications*, Wiley-Interscience (John Wiley & Sons), New York–London–Sydney, 1971.

[54] C. Payan, A class of threshold and domishold graphs: equistable and equidominating graphs, *Discrete Math.* **29** (1980), 47–52, doi:10.1016/0012-365x(90)90286-q.

[55] U. N. Peled and B. Simeone, Polynomial-time algorithms for regular set-covering and threshold synthesis, *Discrete Appl. Math.* **12** (1985), 57–69, doi:10.1016/0166-218x(85)90040-x.

[56] U. N. Peled and B. Simeone, An $O(nm)$-time algorithm for computing the dual of a regular Boolean function, *Discrete Appl. Math.* **49** (1994), 309–323, doi:10.1016/0166-218x(94)90215-1.

[57] N. Pržulj, D. G. Corneil and E. Köhler, Hereditary dominating pair graphs, *Discrete Appl. Math.* **134** (2004), 239–261, doi:10.1016/s0166-218x(03)00304-4.

[58] J. Reiterman, V. Rödl, E. Šiňajová and M. Tůma, Threshold hypergraphs, *Discrete Math.* **54** (1985), 193–200, doi:10.1016/0012-365x(85)90080-9.

[59] D. J. Rose, R. E. Tarjan and G. S. Lueker, Algorithmic aspects of vertex elimination on graphs, *SIAM J. Comput.* **5** (1976), 266–283, doi:10.1137/0205021.

[60] O. Schaudt, A note on connected dominating sets of distance-hereditary graphs, *Discrete Appl. Math.* **160** (2012), 1394–1398, doi:10.1016/j.dam.2012.02.003.

[61] O. Schaudt, On graphs for which the connected domination number is at most the total domination number, *Discrete Appl. Math.* **160** (2012), 1281–1284, doi:10.1016/j.dam.2011.12.025.

[62] O. Schaudt and R. Schrader, The complexity of connected dominating sets and total dominating sets with specified induced subgraphs, *Inform. Process. Lett.* **112** (2012), 953–957, doi:10.1016/j.ipl.2012.09.002.

[63] H. Shen and W. Liang, Efficient enumeration of all minimal separators in a graph, *Theoret. Comput. Sci.* **180** (1997), 169–180, doi:10.1016/s0304-3975(97)83809-1.

[64] Y.-T. Tsai, Y.-L. Lin and F. R. Hsu, Efficient algorithms for the minimum connected domination on trapezoid graphs, *Inform. Sci.* **177** (2007), 2405–2417, doi:10.1016/j.ins.2007.02.001.

[65] D. B. West, *Introduction to Graph Theory*, Prentice Hall, Upper Saddle River, New Jersey, 2nd edition, 2001.

[66] J. Wu and H. Li, A dominating-set-based routing scheme in ad hoc wireless networks, *Telecommun. Syst.* **18** (2001), 13–36, doi:10.1023/a:1016783217662.

[67] J.-H. Yan, J.-J. Chen and G. J. Chang, Quasi-threshold graphs, *Discrete Appl. Math.* **69** (1996), 247–255, doi:10.1016/0166-218x(96)00094-7.

[68] H.-G. Yeh and G. J. Chang, Weighted connected domination and Steiner trees in distance-hereditary graphs, *Discrete Appl. Math.* **87** (1998), 245–253, doi:10.1016/s0166-218x(98)00060-2.

Appendix A A non-connected-domishold split graph whose cutset hypergraph is 2-asummable

Based on an example due to Gabelman [32], Crama and Hammer proposed in the proof of [22, Theorem 9.15] an example of a 9-variable 2-asummable positive Boolean function f that is not threshold. From this function we can derive a split graph $G = (V, E)$ on 71 vertices, as follows. Let $V = K \cup I$ where $K = \{v_1, \ldots, v_9\}$ is a clique and $I = V(G) - K$ is an independent set. To define the edges between K and I, we first associate a non-negative integer weight to each vertex, as follows: $w(v_1) = 14$, $w(v_2) = 18$, $w(v_3) = 24$, $w(v_4) = 26$, $w(v_5) = 27$, $w(v_6) = 30$, $w(v_7) = 31$, $w(v_8) = 36$, $w(v_9) = 37$, and $w(v) = 0$ for all $v \in I$. Let $\mathcal{S}$ be the set of all subsets S of K such that $w(S) \geq 82$ and let $S_1 = \{v_1, v_6, v_9\}$, $S_2 = \{v_2, v_5, v_8\}$, and $S_3 = \{v_3, v_4, v_7\}$. (Note that $w(S_i) = 81$ for all $i \in [3]$.) Let $\mathcal{H}$ be the hypergraph with vertex set K and hyperedge set given by the inclusion-wise minimal sets in $\mathcal{S} \cup \{S_1, S_2, S_3\}$. It can be verified that $\mathcal{H}$ has precisely 62 hyperedges (including S_1, S_2, and S_3).[4] The edges of G between vertices of I and K are defined so that set of the neighborhoods of the 62 vertices of I is exactly the set of hyperedges of $\mathcal{H}$.

To show that G is not CD, it suffices, by Proposition 3.4, to show that the cutset hypergraph is not threshold. In the proof of Theorem 9.15 in [22] it is shown that the function f is not threshold, by showing that f is 3-summable. This corresponds to the fact that the cutset hypergraph of G is 3-summable, as can be observed by noticing that condition (2.1) is satisfied for $k = r = 3$ and for the sets $A_i = S_i$ for all $i \in [3]$ and $B_1 = \{v_1, v_7, v_8\}$, $B_2 = \{v_2, v_4, v_9\}$, and $B_3 = \{v_3, v_5, v_6\}$. On the other hand, the fact that f is 2-asummable implies that the cutset hypergraph of G is 2-asummable.

[4] The following is the list of sets (omitting commas and brackets) of indices of the elements of the 62 inclusion-wise minimal hyperedges of $\mathcal{H}$: 169, 179, 189, 258, 259, 268, 269, 278, 279, 289, 347, 348, 349, 357, 358, 359, 367, 368, 369, 378, 379, 389, 456, 457, 458, 459, 467, 468, 469, 478, 479, 489, 567, 568, 569, 578, 579, 589, 678, 679, 689, 789, 1234, 1235, 1236, 1237, 1238, 1239, 1245, 1246, 1247, 1248, 1249, 1256, 1257, 1267, 1345, 1346, 1356, 2345, 2346, 2356.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 527–547
https://doi.org/10.26493/1855-3974.1475.3d3
(Also available at http://amc-journal.eu)

Smooth skew morphisms of dihedral groups[*]

Na-Er Wang , Kan Hu

Department of Mathematics, Zhejiang Ocean University,
Zhoushan, Zhejiang 316022, P.R. China and
Key Laboratory of Oceanographic Big Data Mining & Application of Zhejiang Province,
Zhoushan, Zhejiang 316022, P.R. China

Kai Yuan

School of Mathematics, Capital Normal University, Beijing 100037, P.R. China

Jun-Yang Zhang

School of Mathematical Sciences, Chongqing Normal University,
Chongqing 401331, P.R. China

Received 4 September 2017, accepted 17 January 2019, published online 28 March 2019

Abstract

A skew morphism φ of a finite group A is a permutation on A fixing the identity element of A and for which there exists an integer-valued function π on A such that $\varphi(ab) = \varphi(a)\varphi^{\pi(a)}(b)$ for all $a, b \in A$. In the case where $\pi(\varphi(a)) = \pi(a)$, for all $a \in A$, the skew morphism is smooth. The concept of smooth skew morphism is a generalization of that of t-balanced skew morphism. The aim of this paper is to develop a general theory of smooth skew morphisms. As an application we classify smooth skew morphisms of dihedral groups.

Keywords: Cayley map, skew morphism, smooth subgroup.

Math. Subj. Class.: 05E18, 20B25, 05C10

[*]The authors are grateful to the anonymous referees for their helpful comments and suggestions which have improved the content and presentation of the paper. This research was supported by the following grants: Zhejiang Provincial Natural Science Foundation of China (No. LY16A010010, LQ17A010003); National Natural Science Foundation of China (No. 11801507, 11671276); Teacher Professional Development Program of Zhejiang Provincial Education Department (No. FX2017029); Basic Research and Frontier Exploration Project of Chongqing (No. cstc2018jcyjAX0010); Science and Technology Research Program of Chongqing Municipal Education Commission (No. KJQN201800512); Natural Science Foundation of Fujian (No. 2016J01027).

E-mail addresses: wangnaer@zjou.edu.cn (Na-Er Wang), hukan@zjou.edu.cn (Kan Hu), pktide@163.com (Kai Yuan), jyzhang@cqnu.edu.cn (Jun-Yang Zhang)

1 Introduction

Throughout the paper all groups considered are finite, unless stated otherwise. A *skew morphism* φ of a finite group A is a bijection on the underlying set of A fixing the identity element of A and for which there exists an integer-valued function $\pi\colon A \to \mathbb{Z}$ such that $\varphi(ab) = \varphi(a)\varphi^{\pi(a)}(b)$, for all $a, b \in A$. Note that π is not uniquely determined by φ, however, as a permutation if φ has order n, then π can be viewed as a function $\pi\colon A \to \mathbb{Z}_n$. In this sense the function π is uniquely determined by φ, and it will be called *the power function* of φ.

Jajcay and Širáň introduced the concept of skew morphism as an algebraic tool to investigate regular Cayley maps [10]. Conder, Jajcay and Tucker have shown in [5] that skew morphisms are also closely related to group factorisations with a cyclic complement. Thus the study of skew morphisms is important for both combinatorics and algebra.

Let X be a generating set of a group A such that $1 \notin X$ and $X = X^{-1}$, let P be a cyclic permutation of X. A *Cayley map* $M = \mathrm{CM}(A, X, P)$ is a 2-cell embedding of the Cayley graph $\mathrm{Cay}(A, X)$ into an orientable closed surface such that the local cyclic orientation of the arcs (g, x) emanating from any vertex g induced by the orientation of the supporting surface agrees with the prescribed cyclic permutation P of X. An automorphism of M is an automorphism of the underlying Cayley graph which extends to an orientation-preserving self-homeomorphism of the supporting surface. It is well known that the automorphism group $\mathrm{Aut}(M)$ of M acts semi-regularly on the arcs of M. In the case where this action is transitive, and hence regular, the map M is called a *regular Cayley map*. The left regular representation of A induces a subgroup of map automorphisms which acts transitively on the vertices of M. It follows that M is regular if and only if M admits an automorphism which fixes a vertex, say the identity vertex 1, and maps the arc $(1, x)$ to $(1, P(x))$. It is a nontrivial result proved by Jajcay and Širáň that a Cayley map $\mathrm{CM}(A, X, P)$ is regular if and only if there is a skew morphism φ of A such that the restriction $\varphi\restriction_X$ of φ to X is equal to P [10, Theorem 1]. A skew morphism of A will be called a *Cayley skew morphism* if it has an inverse-closed generating orbit. Thus the study of regular Cayley maps of a group A is equivalent to the study of Cayley skew morphisms of A.

Among the variety of problems considered with regard to skew morphisms the most important seems to be the classification of regular Cayley maps for given families of groups. This problem is completely settled for cyclic groups [6], and only partial results are known for other abelian groups [4, 5, 23]. For dihedral groups D_n of order $2n$, if n is odd this problem was solved in [14], whereas if n is even only partial classification is at hand [11, 12, 17, 21, 24, 25]. For other non-abelian groups the interested reader is referred to [18, 20, 21].

Although skew morphisms are usually investigated along with regular Cayley maps, they also deserve to be studied independently in a purely algebraic setting. Let $G = AC$ be a group factorisation, where A and C are subgroups of G with $A \cap C = 1$. If $C = \langle c \rangle$ is cyclic, then the commuting rule $ca = \varphi(a)c^{\pi(a)}$, for all $a \in A$, determines a skew morphism φ of A with the associated power function π. Conversely, each skew morphism φ of A determines a group factorisation $L_A\langle\varphi\rangle$ with $L_A \cap \langle\varphi\rangle = 1$, where L_A denotes the left regular representation of A [5, Proposition 3.1]. Thus, there is a correspondence between skew morphisms and group factorisations with cyclic complements.

Let φ be a skew morphism of a group A. A subgroup N of A is *φ-invariant* if $\varphi(N) = N$. Note that the restriction of φ to N is a skew morphism of N, so it is important to study φ-invariant subgroups. The first important φ-invariant subgroup is $\mathrm{Fix}\,\varphi$, the subgroup consisting of fixed points of φ [10]. Later, Zhang discovered in [25] another important φ-invariant subgroup, called the *core* of φ and denoted by $\operatorname{Core}\varphi$. This is a normal subgroup of A, so φ induces a skew morphism $\bar{\varphi}$ of the quotient group $\bar{A} := A/\operatorname{Core}\varphi$ in a natural way. As a consequence, we obtain a new φ-invariant subgroup $\operatorname{Smooth}\varphi = \{a \in A \mid \bar{a} \in \operatorname{Fix}\bar{\varphi}\}$ by means of coverings of skew morphisms; see Section 3.

Section 4 is devoted to a study of the extremal case where $\operatorname{Smooth}\varphi = A$. In this case the skew morphism φ is termed *smooth*. We prove that a skew morphism φ of A is smooth if and only if $\pi(\varphi(a)) = \pi(a)$ for all $a \in A$. It follows that the power function of a smooth skew morphism takes constant value on orbits of φ, so smooth skew morphisms may be viewed as a generalization of t-balanced Cayley skew morphisms studied in [4]. Note that for abelian groups smooth skew morphisms are identical with the *coset-preserving* skew morphisms studied by Bachratý and Jajcay in [1]. We establish in Theorems 4.5 and 4.9 an unexpected relationship between smooth skew morphisms and kernel-preserving skew morphisms. Note that a skew morphism φ of A is *kernel-preserving* if its kernel $\operatorname{Ker}\varphi$ is a φ-invariant subgroup of A.

Kovács and Kwon [13] have recently announced a complete classification of regular Cayley maps of dihedral groups. Thus, to complete the classification of skew morphisms of dihedral groups, it remains to determine the non-Cayley skew morphisms. As shown in [8], every non-Cayley skew morphism of dihedral groups is smooth. Our last aim of this paper is to employ the newly-developed theory to give a classification of smooth skew morphisms of the dihedral groups, see Section 5.

2 Preliminaries

In this section we summarize some basic results concerning skew morphisms which will be used throughout the paper.

Let φ be a skew morphism of a group A, let π be the power function of φ, and let n be the order of φ. As already mentioned above, the sets

$$\operatorname{Ker}\varphi = \{a \in A \mid \pi(a) = 1\}, \qquad \operatorname{Fix}\varphi = \{a \in A \mid \varphi(a) = a\}$$

and

$$\operatorname{Core}\varphi = \bigcap_{i=1}^{n} \varphi^i(\operatorname{Ker}\varphi)$$

form subgroups of A. Note that, for any two elements $a, b \in A$, $\pi(a) = \pi(b)$ if and only if $ab^{-1} \in \operatorname{Ker}\varphi$. Thus, the index $|A : \operatorname{Ker}\varphi|$ is equal to the number of distinct values of the power function. This number is called the *skew type* of φ, and it is strictly less than n if φ is not trivial. Clearly, φ is an automorphism of A if and only if it has skew type 1. If φ is not an automorphism, then it will be termed *proper*. On the other hand, $\operatorname{Core}\varphi$ is the largest φ-invariant subgroup contained in $\operatorname{Ker}\varphi$, and in particular, it is normal in A [25].

Lemma 2.1 ([10]). *Let φ be a skew morphism of a group A, let π be the power function of φ, and let n be the order of φ. Then, for any $a, b \in A$,*

$$\varphi^k(ab) = \varphi^k(a)\varphi^{\sigma(a,k)}(b) \qquad \textit{and} \qquad \pi(ab) \equiv \sigma(b, \pi(a)) \pmod{n},$$

where k is an arbitrary positive integer and $\sigma(a,k) = \sum_{i=1}^{k} \pi(\varphi^{i-1}(a))$.

Lemma 2.2 ([7]). *Let φ be a skew morphism of a group A, let π be the power function of φ. Then for any automorphism γ of A, $\psi = \gamma^{-1}\varphi\gamma$ is a skew morphism of A with power function $\pi_\psi = \pi\gamma$. Moreover,* $\operatorname{Ker}\psi = \gamma^{-1}(\operatorname{Ker}\varphi)$ *and* $\operatorname{Core}\psi = \gamma^{-1}(\operatorname{Core}\varphi)$.

Proof. Since γ is an automorphism of A, for any $a, b \in A$, we have

$$\begin{aligned}\psi(ab) &= \gamma^{-1}\varphi\gamma(ab) = \gamma^{-1}\varphi(\gamma(a)\gamma(b)) = \gamma^{-1}\big(\varphi(\gamma(a))\varphi^{\pi\gamma(a)}(\gamma(b))\big)\\ &= \gamma^{-1}\varphi\gamma(a)\gamma^{-1}\varphi^{\pi\gamma(a)}\gamma(b) = \psi(a)\psi^{\pi\gamma(a)}(b).\end{aligned}$$

Thus, ψ is a skew morphism of A with power function $\pi_\psi = \pi\gamma$. Since $|\psi| = |\varphi|$, we have

$$\begin{aligned}a \in \operatorname{Ker}\psi \quad &\Longleftrightarrow \quad \pi_\psi(a) \equiv 1 \pmod{|\psi|} \quad \Longleftrightarrow\\ &\pi\gamma(a) \equiv 1 \pmod{|\varphi|} \quad \Longleftrightarrow \quad a \in \gamma^{-1}(\operatorname{Ker}\varphi).\end{aligned}$$

Therefore, $\operatorname{Ker}\psi = \gamma^{-1}(\operatorname{Ker}\varphi)$. Similarly, $\operatorname{Core}\psi = \gamma^{-1}(\operatorname{Core}\varphi)$. □

Lemma 2.3 ([1, 5]). *Let φ be a skew morphism of a group A, let π be the power function of φ, and let n be the order of φ. Then for any positive integer k, $\mu = \varphi^k$ is a skew morphism of A if and only if the congruences*

$$kx \equiv \sigma(a, k) \pmod{n} \tag{2.1}$$

are solvable for all $a \in A$. Moreover, if μ is a skew morphism of A, then it has order $m = n/\gcd(n, k)$ and for each $a \in A$, $\pi_\mu(a)$ is the solution of the equation (2.1) *in $\mathbb{Z}_m$.*

Lemma 2.4 ([5]). *Let φ be a skew morphism of a group A. If A is nontrivial, then $|\varphi| \le |A|$ and* $|\operatorname{Ker}\varphi| > 1$.

Lemma 2.5 ([9]). *Let φ be a skew morphism of a group A, and let O_a denote the orbit of φ containing the element $a \in A$. Then for each $a \in A$, $O_{a^{-1}} = O_a^{-1}$, where $O_a^{-1} = \{g^{-1} \mid g \in O_a\}$.*

The following result was partially obtained for Cayley skew morphisms in [4].

Lemma 2.6 ([7]). *Let φ be a skew morphism of a group A, and let π the power function of φ, and let n be the order of φ. Then for any $a \in A$,*

$$\sigma(a, m) \equiv 0 \pmod{m},$$

where $m = |O_a|$ is length of the orbit O_a containing a. Moreover, $\sigma(a, n) \equiv 0 \pmod{n}$.

Proof. By Lemma 2.1, we have

$$1 = \varphi^m(aa^{-1}) = \varphi^m(a)\varphi^{\sigma(a,m)}(a^{-1}) = a\varphi^{\sigma(a,m)}(a^{-1}),$$

so $\varphi^{\sigma(a,m)}(a^{-1}) = a^{-1}$. By Lemma 2.5, $m = |O_{a^{-1}}|$. Thus, $\sigma(a, m) \equiv 0 \pmod{m}$. Since m divides n, we obtain

$$\sigma(a, n) = \sum_{i=1}^{n} \pi(\varphi^{i-1}(a)) = \frac{n}{m}\sigma(a, m) \equiv 0 \pmod{n},$$

as required. □

Lemma 2.7 ([7]). *Let φ be a skew morphism of a group A. Then for any $a, b \in A$, $|O_{ab}|$ divides* $\operatorname{lcm}(|O_a|, |O_b|)$.

Proof. Denote $c = |O_a|$, $d = |O_b|$ and $\ell = \operatorname{lcm}(|O_a|, |O_b|)$. Then $\ell = cp = dq$ for some positive integers p, q. By Lemma 2.1, we have $\varphi^\ell(ab) = \varphi^\ell(a)\varphi^{\sigma(a,\ell)}(b) = a\varphi^{\sigma(a,\ell)}(b)$. By Lemma 2.6,

$$\sigma(a,\ell) = \sum_{i=1}^{\ell} \pi(\varphi^{i-1}(a)) = p\sum_{i=1}^{c} \pi(\varphi^{i-1}(a)) = p\sigma(a,c) \equiv 0 \pmod{\ell}.$$

Thus, $\varphi^\ell(ab) = ab$, and consequently, $|O_{ab}|$ divides ℓ. □

Lemma 2.8. *Let φ be a skew morphism of a group A, and let π the power function of φ, and let n be the order of φ. If $A = \langle a_1, \ldots, a_r\rangle$, then $n = \operatorname{lcm}(|O_{a_1}|, \ldots, |O_{a_r}|)$. Moreover, for any $g \in A$, $\varphi(g)$ and $\pi(g)$ are completely determined by the action of φ and the values of π on the generating orbits $O_{a_1}, \ldots, O_{a_r}$.*

Proof. The first part was first proved in [26, Lemma 3.1]. The reader is invited to give an alternative proof using Lemma 2.7 (and induction on the length of words in the generators).

To prove the second part we use induction on the length k of g in the generators. If $k = 1$ then g is a generator of A, the assertion is trivially true. Assume that the assertion is true for words of length k. Then, for a word g of length $k + 1$, we have $g = ha$, where h is a word of length k and $a \in \{a_1, \ldots, a_r\}$. By Lemma 2.1, we have

$$\varphi(g) = \varphi(ha) = \varphi(h)\varphi^{\pi(h)}(a) \quad \text{and} \quad \pi(g) \equiv \pi(ha) \equiv \sum_{i=1}^{\pi(h)} \pi(\varphi^{i-1}(a)) \pmod{n}.$$

Since $\varphi(h)$ and $\pi(h)$ are completely determined by the action of φ and the values of π on the generating orbits, so are $\varphi(g)$ and $\pi(g)$, as required. □

Lemma 2.9. *Let φ be a skew morphism of a group A, let π the power function of φ, and let n be the order of φ. If N is a φ-invariant normal subgroup of A, then*

(a) φ induces a skew morphism $\bar\varphi$ of $\bar A = A/N$ by defining $\bar\varphi$ as $\bar\varphi(\bar a) = \overline{\varphi(a)}$ and the power function $\bar\pi\colon \bar A \to \mathbb{Z}_m$ associated with $\bar\varphi$ is determined by $\bar\pi(\bar a) \equiv \pi(a) \pmod{m}$ where $m = |\bar\varphi|$,

(b) $\operatorname{Ker}\varphi N/N \le \operatorname{Ker}\bar\varphi$, $\operatorname{Core}\varphi N/N \le \operatorname{Core}\bar\varphi$ *and* $\operatorname{Fix}\varphi N/N \le \operatorname{Fix}\bar\varphi$.

Proof. The proof of (a) can be found in [26, Lemma 3.3] while (b) is obvious. □

3 Invariant subgroups

In this section, we introduce covering techniques to the study of skew morphisms and define several new invariant subgroups.

Proposition 3.1. *Let φ be a skew morphism of a group A. If M and N are φ-invariant subsets of A, so are $M \cap N$ and MN.*

Proof. For any $y \in \varphi(M \cap N)$, there exists $x \in M \cap N$ such that $y = \varphi(x)$. Since M and N are both φ-invariant, $\varphi(x) \in M$ and $\varphi(x) \in N$, so $y \in M \cap N$, whence $\varphi(M \cap N) = M \cap N$. Therefore $M \cap N$ is also φ-invariant. Similarly for any $y \in \varphi(MN)$, there exist $u \in M$ and $v \in N$ such that $y = \varphi(uv)$. We have $y = \varphi(uv) = \varphi(u)\varphi^{\pi(u)}(v) \in \varphi(M)\varphi(N) = MN$, so $\varphi(MN) = MN$, whence MN is also φ-invariant. □

Let Π be a finite set of primes, a positive integer k will be called a Π-*number* if all prime factors of k belong to Π. For instance, if $\Pi = \{2, 3\}$, then $2, 6, 9$ are Π-numbers, whereas $5, 10, 30$ are not. We define 1 to be a Π-number for any set Π of primes.

Now let φ be a skew morphism of a group A. An orbit of φ will be called a Π-*orbit* if its length is a Π-number. Define $\operatorname{Orbit}^{\Pi} \varphi$ to be the union of all Π-orbits of φ, namely,

$$\operatorname{Orbit}^{\Pi} \varphi = \{a \in A \mid |O_a| \text{ is a } \Pi\text{-number}\}.$$

Proposition 3.2. *Let φ be a skew morphism of A, and let Π be a finite set of primes, then $\operatorname{Orbit}^{\Pi} \varphi$ is a φ-invariant subgroup of A containing $\operatorname{Fix} \varphi$.*

Proof. By definition, all fixed points of φ belong to $\operatorname{Orbit}^{\Pi} \varphi$, so $\operatorname{Orbit}^{\Pi} \varphi$ is not empty. Moreover, for any $a, b \in \operatorname{Orbit}^{\Pi} \varphi$, $|O_a|$ and $|O_b|$ are Π-numbers, so $\operatorname{lcm}(|O_x|, |O_y|)$ is also a Π-number. By Lemma 2.7, $|O_{ab}|$ divides $\operatorname{lcm}(|O_a|, |O_b|)$. It follows that $|O_{ab}|$ is also a Π-number. Hence, $ab \in \operatorname{Orbit}^{\Pi} \varphi$. Therefore, $\operatorname{Orbit}^{\Pi} \varphi$ is a subgroup of A, which is clearly φ-invariant. □

Example 3.3. Consider the skew morphism of the cyclic group $\mathbb{Z}_{21}$ defined by

$$\varphi = (0)\,(1, 2, 4, 8, 16, 11)\,(3, 6, 12)\,(5, 10, 20, 19, 17, 13)\,(7, 14)\,(9, 18, 15).$$

This is an automorphism of $\mathbb{Z}_{21}$. We have

$$\operatorname{Orbit}^{\{2\}} \varphi = \langle 7 \rangle, \quad \operatorname{Orbit}^{\{3\}} \varphi = \langle 3 \rangle, \quad \operatorname{Orbit}^{\{5\}} \varphi = \langle 0 \rangle, \quad \text{and} \quad \operatorname{Orbit}^{\{2,3\}} \varphi = \mathbb{Z}_{21}.$$

Now we introduce covering techniques to the study of skew morphisms.

Definition 3.4. Let φ_i be skew morphisms of finite groups A_i, $i = 1, 2$. If there is an epimorphism $\theta \colon A_1 \to A_2$ such that the identity

$$\theta\varphi_1(a) = \varphi_2\theta(a)$$

holds for all $a \in A_1$, then φ_1 will be called a *covering* (or a lift) of φ_2, and φ_2 will be called a *projection* (or a quotient) of φ_1. The covering will be denoted by $\varphi_1 \to \varphi_2$, and the epimorphism $\theta \colon A_1 \to A_2$ will be said to be associated with the covering.

Lemma 3.5. *Let $\varphi_1 \to \varphi_2$ be a covering between skew morphisms φ_i of groups A_i, $i = 1, 2$, and let $\theta \colon A_1 \to A_2$ be the associated epimorphism. Then*

(a) every φ_1-invariant subgroup M of A_1 projects to a φ_2-invariant subgroup $\theta(M)$ of A_2,

(b) every φ_2-invariant subgroup N of A_2 lifts to a φ_1-invariant subgroup $\theta^{-1}(N)$ of A_1.

Proof. (a): For any $y \in \theta(M)$, $y = \theta(x)$ for some $x \in M$. Since M is φ_1-invariant, $\varphi_1(x) \in M$, so $\varphi_2(y) = \varphi_2\theta(x) = \theta\varphi_1(x) \in \theta(M)$, whence $\theta(M)$ is φ_1-invariant.

(b): For any $x \in \theta^{-1}(N)$, $y = \theta(x) \in N$. Since N is φ_2-invariant, $\varphi_2(y) \in N$, so $\theta\varphi_1(x) = \varphi_2\theta(x) = \varphi_2(y) \in N$. Hence $\varphi_1(x) \in \theta^{-1}(N)$. □

Since $\{1\}$, $\operatorname{Fix}\varphi_2$ and $\operatorname{Core}\varphi_2$ are all φ_2-invariant subgroups of A_2, by Lemma 3.5, $\operatorname{Ker}\theta = \theta^{-1}(1)$, $\theta^{-1}(\operatorname{Fix}\varphi_2)$ and $\theta^{-1}(\operatorname{Core}\varphi_2)$ are all φ_1-invariant subgroups of A_1. In particular, both $\operatorname{Ker}\theta$ and $\theta^{-1}(\operatorname{Core}\varphi_2)$ are normal in A_1.

Now we are ready to introduce another new φ-invariant subgroup for skew morphisms. Let φ be a skew morphism of a group A, and let π be the power function of φ. Recall that $\operatorname{Core}\varphi$ is a normal φ-invariant subgroup of A. Let $\operatorname{Smooth}\varphi$ be a subset of A defined by

$$\operatorname{Smooth}\varphi = \{a \in A \mid \varphi(a) \equiv a \pmod{\operatorname{Core}\varphi}\}.$$

Proposition 3.6. *Let φ be a skew morphism of a group A, let π be the power function of φ, and let $\bar{\varphi}$ be the φ-induced skew morphism of $\bar{A} = A/\operatorname{Core}\varphi$. Then, for any $a \in A$, the following are equivalent:*

(a) $a \in \operatorname{Smooth}\varphi$,

(b) $\pi(\varphi^i(a)) = \pi(a)$ *for all positive integers i,*

(c) $\bar{a} \in \operatorname{Fix}\bar{\varphi}$.

Proof. (a) $\Longrightarrow$ (b): Since $a \in \operatorname{Smooth}\varphi$, by definition, $\varphi(a) = ua$ for some $u \in \operatorname{Core}\varphi$, and so $\varphi^i(a) = \varphi^{i-1}(u)\cdots\varphi(u)ua$ for all positive integers i. Since $\operatorname{Core}\varphi$ is a φ-invariant subgroup, we have $\varphi^{i-1}(u)\cdots\varphi(u)u \in \operatorname{Core}\varphi$. Therefore, $\pi(\varphi^i(a)) = \pi(a)$.

(b) $\Longrightarrow$ (c): Since $\pi(\varphi(a)) = \pi(a)$, we have $\varphi(a) = ua$ for some $u \in \operatorname{Ker}\varphi$ and then $\varphi^2(a) = \varphi(ua) = \varphi(u)\varphi(a) = \varphi(u)ua$. Since $\pi(\varphi^2(a)) = \pi(a)$, we get $\varphi(u)u \in \operatorname{Ker}\varphi$ and hence $\varphi(u) \in \operatorname{Ker}\varphi$. Repeating the above process, we get $\varphi^i(u) \in \operatorname{Ker}\varphi$ for all positive integers i. Consequently, $u \in \operatorname{Core}\varphi$ and hence $\bar{\varphi}(\bar{a}) = \bar{a}$, that is, $\bar{a} \in \operatorname{Fix}\bar{\varphi}$.

(c) $\Longrightarrow$ (a): Since $\bar{a} \in \operatorname{Fix}\bar{\varphi}$, we have $\bar{\varphi}(\bar{a}) = \bar{a}$ and so $\varphi(a) = ua$ for some $u \in \operatorname{Core}\varphi$. Since $\operatorname{Core}\varphi \trianglelefteq A$, we obtain $a \in \operatorname{Smooth}\varphi$. □

The following result is a direct corollary of Proposition 3.6.

Corollary 3.7. *Suppose that φ, A, $\bar{\varphi}$ and $\bar{A}$ are defined as Proposition 3.6. Then*

$$\operatorname{Fix}\overline{\varphi} = \overline{\operatorname{Smooth}\varphi}$$

and $\operatorname{Smooth}\varphi$ *is a φ-invariant subgroup of A. In particular,*

(a) $\operatorname{Smooth}\varphi = \operatorname{Core}\varphi$ *if and only if* $\operatorname{Fix}\bar{\varphi} = \bar{1}$,

(b) $\operatorname{Smooth}\varphi = A$ *if and only if* $\operatorname{Fix}\bar{\varphi} = \bar{A}$*, and*

(c) $\operatorname{Smooth}\varphi = \operatorname{Fix}\varphi$ *if* $\operatorname{Core}\varphi = 1$.

Example 3.8 ([22])**.** Consider a skew morphism of the cyclic group $\mathbb{Z}_{18}$ defined by

$$\varphi = (0)\,(1,15,17,7,3,5,13,9,11)\,(2,14,8)\,(4,10,16)\,(6)\,(12),$$
$$\pi = [1]\,[2,5,8,2,5,8,2,5,8]\,[7,7,7]\,[4,4,4]\,[1]\,[1].$$

Then $\operatorname{Core}\varphi = \operatorname{Ker}\varphi = \langle 6\rangle$, so $\bar{\varphi} = (\bar{0})\,(\bar{1},\bar{3},\bar{5})\,(\bar{2})\,(\bar{4})$ and $\operatorname{Smooth}\varphi = \langle 2\rangle$.

The following example is due to Conder, as mentioned in [1],

Example 3.9. Consider a skew morphism

$$\begin{aligned}\varphi &= (1)\,(a, a^2)\,(b, bc, c)\,(ab, a^2bc, ac, a^2b, abc, a^2c),\\ \pi &= [\,1\,]\,[\,1, 1\,]\,[\,1, 4, 4\,]\,[\,1, 4, 4, 1, 4, 4\,]\end{aligned}$$

of the non-abelian group $A = D_3 \times C_2$, where

$$D_3 = \langle a, b \mid a^3 = b^2 = (ab)^2 = 1\rangle \qquad \text{and} \qquad C_2 = \langle c \mid c^2 = 1\rangle.$$

We have $\operatorname{Ker}\varphi = \langle a, b\rangle$ and $\operatorname{Core}\varphi = \langle a\rangle$. Thus, $\bar{\varphi} = (\bar{1})\,(\bar{b}, \bar{b}\bar{c}, \bar{c})$, and hence

$$\operatorname{Smooth}\varphi = \operatorname{Core}\varphi.$$

4 Smooth skew morphisms

In this section we establish a relationship between kernel-preserving skew morphisms and smooth skew morphisms.

In general, the kernel $\operatorname{Ker}\varphi$ of a skew morphism φ does not have to be a φ-invariant subgroup. However, as we already mentioned above, a skew morphism φ will be called *kernel-preserving* if $\operatorname{Ker}\varphi$ is φ-invariant. Clearly, φ is kernel-preserving if and only if $\operatorname{Core}\varphi = \operatorname{Ker}\varphi$. It is well known that every skew morphism φ of an abelian group is kernel-preserving [4, Lemma 5.1]. For non-abelian groups, there do exist skew morphisms which are not kernel-preserving, see Example 3.9.

Kernel-preserving skew morphisms have many interesting properties.

Lemma 4.1. *Let φ be a skew morphism of a group A, π be the power function of φ, and let n be the order of φ. If φ is kernel-preserving, then*

(a) $\operatorname{Ker}\varphi$ *is a normal subgroup of A, and φ restricted to* $\operatorname{Ker}\varphi$ *is an automorphism of* $\operatorname{Ker}\varphi$*,*

(b) for some positive integer k, if $\mu = \varphi^k$ is a skew morphism of A, then $\operatorname{Ker}\varphi \le \operatorname{Ker}\mu$*,*

(c) for any automorphism γ of A, $\gamma^{-1}\varphi\gamma$ is a kernel-preserving skew morphism of A,

(d) for any pair of elements $a \in A$ and $u \in$ $\operatorname{Ker}\varphi$ *there is a unique element $v \in$* $\operatorname{Ker}\varphi$ *such that $au = va$ and $\varphi(a)\varphi^{\pi(a)}(u) = \varphi(v)\varphi(a)$. In particular, if A is abelian then* $\pi(a) \equiv 1 \pmod{m}$ *where m is the order of the restriction of φ to* $\operatorname{Ker}\varphi$*.*

Proof. (a): Since φ is kernel-preserving, $\operatorname{Ker}\varphi = \operatorname{Core}\varphi$, which is a normal subgroup of A. Moreover, for all $a, b \in \operatorname{Ker}\varphi$, we have $\varphi(ab) = \varphi(a)\varphi(b)$, so φ restricted to $\operatorname{Ker}\varphi$ is an automorphism of $\operatorname{Ker}\varphi$.

(b): For any $a \in \operatorname{Ker}\varphi = \operatorname{Core}\varphi$, $\pi(\varphi^{i-1}(a)) = 1$, $i = 1, 2, \dots, n$. By Lemma 2.3, the power function π_μ of μ is determined by the the congruence $k\pi_\mu(a) \equiv \sigma(a, k) = k \pmod{n}$, so $\pi_\mu(a) \equiv 1 \pmod{n/\gcd(n, k)}$, which implies that $a \in \operatorname{Ker}\mu$.

(c): This is an immediate consequence of Lemma 2.2.

(d): Since $\operatorname{Ker}\varphi \trianglelefteq A$, for any pair (a, u) of elements $a \in A$ and $u \in \operatorname{Ker}\varphi$, there is a unique element $v \in \operatorname{Ker}\varphi$ such that $au = va$. Then $\varphi(a)\varphi^{\pi(a)}(u) = \varphi(au) = \varphi(va) = \varphi(v)\varphi(a)$. In particular, if A is abelian, then $u = v$ and $\varphi^{\pi(a)}(u) = \varphi(u)$ for all $u \in \operatorname{Ker}\varphi$, so $\pi(a) \equiv 1 \pmod{m}$, where m is the order of the restriction of φ to $\operatorname{Ker}\varphi$. □

Proposition 4.2. *Every kernel-preserving skew morphism of a non-abelian simple group A is an automorphism of A.*

Proof. If φ is not an automorphism of A, then $1 < \operatorname{Ker}\varphi < A$ by Lemma 2.4. Since φ is kernel-preserving, by Lemma 4.1(a) $\operatorname{Ker}\varphi \trianglelefteq A$, a contradiction. □

Let φ be a skew morphism of a group A. Recall that $\operatorname{Smooth}\varphi$ consists of elements $a \in A$ such that $\varphi(a) \equiv a \pmod{\operatorname{Core}\varphi}$. If $\operatorname{Smooth}\varphi = A$, then φ will be called a *smooth* skew morphism. The concept of smooth skew morphism was first introduced by Hu in the unpublished manuscript [7]. Bachratý and Jajcay rediscovered it under the name of *coset-preserving* skew morphisms [1].

Lemma 4.3. *Let φ be a skew morphism of a group A. If φ is smooth, then every subgroup of A containing $\operatorname{Core}\varphi$ is φ-invariant; in particular, φ is kernel-preserving.*

Proof. Suppose that φ is a smooth skew morphism of A. By Proposition 3.6, the induced skew morphism $\bar{\varphi}$ of $\bar{A} = A/\operatorname{Core}\varphi$ is the identity permutation on $\bar{A}$, so every subgroup of $\bar{A}$ is $\bar{\varphi}$-invariant. Therefore, by Lemma 3.5, every subgroup of A containing $\operatorname{Core}\varphi$ is φ-invariant. In particular, since $\operatorname{Core}\varphi \le \operatorname{Ker}\varphi$, $\varphi(\operatorname{Ker}\varphi) = \operatorname{Ker}\varphi$. □

The following lemma characterizes smooth skew morphisms in terms of their power functions.

Lemma 4.4. *Let φ be a skew morphism of a group A, and let π be the power function of φ. Then φ is smooth if and only if $\pi(\varphi(a)) = \pi(a)$, for all $a \in A$.*

Proof. If φ is smooth, then, by Proposition 3.6, $\pi(\varphi(a)) = \pi(a)$, for all $a \in A$. Conversely, suppose that, for any $a \in A$, $\pi(\varphi(a)) = \pi(a)$. Then $\varphi(a) = ua$ for some $u \in \operatorname{Ker}\varphi$. By the assumption, we have $\pi(\varphi^{n-1}(u)) = \cdots = \pi(\varphi(u)) = \pi(u) = 1$, where $n = |\varphi|$, so $u \in \operatorname{Core}\varphi$. Therefore, $\varphi(a) \equiv a \pmod{\operatorname{Core}\varphi}$, that is, φ is smooth. □

The smallest positive integer d such that $\pi(\varphi^d(a)) \equiv \pi(a) \pmod{|\varphi|}$, for all $a \in A$, is called the *period* of φ. It is easily seen that d is a divisor of n uniquely determined by φ. Bachratý and Jajcay proved that if A is abelian, then $\mu = \varphi^d$ is a smooth skew morphism of A; in particular, if φ is nontrivial and contains a generating orbit, then d is a proper divisor of n [1]. In what follows we present a generalization.

Theorem 4.5. *Let φ be a skew morphism of a group A, let d be the period of φ, and let $\bar{\varphi}$ be the φ-induced skew morphism of $\bar{A} = A/\operatorname{Core}\varphi$. Then the following hold true:*

(a) d is equal to the order of $\bar{\varphi}$,

(b) $\sigma(a, d) \equiv 0 \pmod{d}$ for all $a \in A$,

(c) $\mu = \varphi^d$ is a smooth skew morphism of A,

(d) $\mu = \varphi^d$ is an automorphism of A if and only if $\sigma(a, d) \equiv d \pmod{n}$ for all $a \in A$.

Proof. Denote $n = |\varphi|$ and $m = |\bar{\varphi}|$.

(a): By the assumption, for any $a \in A$, we have $\pi(\varphi^d(a)) = \pi(a)$, and so $\varphi^d(a) = ua$ for some $u \in \operatorname{Ker}\varphi$. Thus,

$$\pi(\varphi^{d+1}(a)) = \pi(\varphi(ua)) = \pi(\varphi(u)\varphi(a)).$$

Since $\pi(\varphi^{d+1}(a)) = \pi(\varphi(a))$, we obtain $\varphi(u) \in \operatorname{Ker}\varphi$. Repeating this process we get $\varphi^{i-1}(u) \in \operatorname{Ker}\varphi$, $i = 1, 2, \ldots, n$. Thus, $u \in \operatorname{Core}\varphi$, and consequently, $\bar{\varphi}^d(\bar{a}) = \bar{a}$.

Therefore, $m \leq d$. On the other hand, since $|\bar{\varphi}| = m$, $\bar{\varphi}^m(\bar{a}) = \bar{a}$ for any $a \in A$, so $\varphi^m(a) = ua$ for some $u \in \operatorname{Core}\varphi$. Thus, $\pi(\varphi^m(a)) = \pi(ua) = \pi(a)$. The minimality of d then implies that $d \leq m$.

(b): For each $a \in A$, by (a) we have

$$\sigma(a,n) = \sum_{i=1}^{n} \pi(\varphi^{i-1}(a)) = \frac{n}{d}\sum_{i=1}^{d} \pi(\varphi^{i-1}(a)) = \frac{n}{d}\sigma(a,d) \pmod{n}.$$

By Lemma 2.6, $\sigma(a,n) = 0 \pmod{n}$ and hence, $\sigma(a,d) \equiv 0 \pmod{d}$.

(c): By (b) and Lemma 2.3, $\mu = \varphi^d$ is a skew morphism of A with its power function determined by $\pi_\mu(a) \equiv \sigma(a,d)/d \pmod{n/d}$. Since $\pi(\mu(a)) = \pi(\varphi^d(a)) \equiv \pi(a) \pmod{n}$, we obtain $\pi_\mu(\mu(a)) \equiv \pi_\mu(a) \pmod{n/d}$. Therefore, μ is smooth by Proposition 4.4.

(d): Since $\pi_\mu(a) \equiv \sigma(a,d)/d \pmod{n/d}$, μ is an automorphism if and only if $\sigma(a,d) \equiv d \pmod{n}$. □

Corollary 4.6. *Let φ be a kernel-preserving skew morphism of a group A, and let n be the order of φ. If φ is nontrivial, then the period d of φ is a proper divisor of n, and so $\mu = \varphi^d$ is a nontrivial smooth skew morphism of A.*

Proof. If φ is nontrivial, then $|A : \operatorname{Ker}\varphi| < |\varphi| = n$. By Lemma 2.4, $d = |\bar{\varphi}| \leq |\bar{A}| = |A : \operatorname{Ker}\varphi|$. Thus, d is a proper divisor of n and therefore, φ^d is a nontrivial smooth skew morphism by Theorem 4.5. □

Example 4.7 ([22])**.** Consider the skew morphism of the cyclic group $\mathbb{Z}_{18}$ given by

$$\begin{aligned} \varphi &= (0)\,(1,5,13,11,7,17)\,(2,16,8,10,14,4)\,(3,5)\,(6,12)\,(9),\\ \pi &= [\,1\,]\,[\,3,5,3,5,3,5\,]\,[\,5,3,5,3,5,3\,]\,[\,1,1\,]\,[\,1,1\,]\,[\,1\,]. \end{aligned}$$

Then $\operatorname{Ker}\varphi = \operatorname{Core}\varphi = \langle 3\rangle$ and $\bar{\varphi} = (\bar{0})\,(\bar{1},\bar{2})$. Note that φ has period 2, which is precisely the order of $\bar{\varphi}$. Since $\sigma(x,2) \equiv 0 \pmod{2}$, for all $x \in \mathbb{Z}_{18}$, by Theorem 4.5(c), $\mu = \varphi^2$ is an automorphism of A.

Let us revisit the skew morphism φ of the non-abelian group $D_3 \times C_2$ considered in Example 3.9. It has period 3, which is a proper divisor of the order of φ. As we already mentioned, the skew morphism is not kernel-preserving. This leads us to pose the following problem.

Problem 4.8. Let d be the period of a nontrivial skew morphism φ of a group A. If φ is not kernel-preserving, under what condition is $\mu = \varphi^d$ nontrivial?

We close this section with some important properties of smooth skew morphisms, see also [1, 7].

Theorem 4.9. *Let φ be a skew morphism of A, let π be the power function of φ, and let n be the order of φ. If φ is smooth, then*

(a) *$\pi\colon A \to \mathbb{Z}_n^*$ is a group homomorphism from A to the multiplicative group $\mathbb{Z}_n^*$ with* $\operatorname{Ker}\pi = \operatorname{Ker}\varphi$,

(b) *for any φ-invariant normal subgroup N of A, the induced skew morphism $\bar{\varphi}$ on A/N is also smooth; in particular, if $N = \operatorname{Ker}\varphi$ then $\bar{\varphi}$ is the identity permutation,*

(c) for any positive integer k, $\mu = \varphi^k$ is a smooth skew morphism,

(d) for any automorphism γ of A, $\psi = \gamma^{-1}\varphi\gamma$ is a smooth skew morphism of A.

Proof. (a): Since φ is smooth, $\pi(a) = \pi(\varphi(a)) = \cdots = \pi(\varphi^{n-1}(a))$. By Lemma 2.1, we have

$$\pi(ab) \equiv \sum_{i=1}^{\pi(a)} \pi(\varphi^{i-1}(b)) \equiv \pi(a)\pi(b) \pmod{n}.$$

Since $1 \equiv \pi(ab^{-1}) = \pi(a)\pi(a^{-1}) \pmod{n}$, $\pi(a) \in \mathbb{Z}_n^*$. Therefore, π is a group homomorphism from A to the multiplicative group $\mathbb{Z}_n^*$.

(b): Since φ is smooth, for any $a \in A$, we have $\pi(\varphi(a)) = \pi(a)$, and so $\bar{\pi}(\bar{\varphi}(\bar{a})) = \bar{\pi}(\bar{a}) \pmod{m}$, where $m = |\bar{\varphi}|$. By Lemma 4.4, $\bar{\varphi}$ is smooth.

(c): For any positive integer k, since $\pi(\varphi^{i-1}(a)) = \pi(a)$, $i = 1, 2, \dots, k$, we have

$$\sigma(a, k) = \sum_{i=1}^{k} \pi(\varphi^{i-1}(a)) \equiv k\pi(a) \pmod{n}.$$

It follows that the equations $kx \equiv \sigma(a, k) \pmod{n}$ are solvable for all $a \in A$. Thus, by Lemma 2.3, $\mu = \varphi^k$ is a skew morphism of A and the associated power function $\pi_\mu \colon A \to \mathbb{Z}_m$ is determined by $\pi_\mu(a) \equiv \pi(a) \pmod{m}$, where $m = n/\gcd(n, k)$ is the order of μ. Since $\pi_\mu(\mu(a)) \equiv \pi(\varphi^k(a)) \equiv \pi(a) \equiv \pi_\mu(a) \pmod{m}$, by Lemma 4.4, μ is also smooth.

(d): By Lemma 2.2, $\psi = \gamma^{-1}\varphi\gamma$ is a skew morphism with $\operatorname{Core}\psi = \gamma^{-1}(\operatorname{Core}\varphi)$. For any $a \in A$, since φ is smooth, $\varphi(\gamma(a)) \equiv \gamma(a) \pmod{\operatorname{Core}\varphi}$, or equivalently, $\gamma^{-1}\varphi\gamma(a) \equiv a \pmod{\gamma^{-1}(\operatorname{Core}\varphi)}$. Thus, $\psi(a) \equiv a \pmod{\operatorname{Core}\psi}$ and hence, ψ is smooth. □

5 Smooth skew morphisms of dihedral groups

Throughout this section, D_n will denote the dihedral group of order $2n$ with presentation

$$D_n = \langle a, b \mid a^n = b^2 = 1, b^{-1}ab = a^{-1}\rangle, \quad n \geq 3. \tag{5.1}$$

Moreover, for positive integers u and k, $\tau(u, k)$ and $\rho(u, k)$ are functions defined by

$$\tau(u, k) = \sum_{i=1}^{k} u^{k-1} \qquad \text{and} \qquad \rho(u, k) = \sum_{i=1}^{k} (-u)^{k-1}. \tag{5.2}$$

If k is even, we use $\lambda(u, k)$ to denote the function defined by

$$\lambda(u, k) = \sum_{i=1}^{k/2} u^{2(i-1)}. \tag{5.3}$$

The following result on normal subgroups of D_n is well known.

Lemma 5.1 ([16, Section 1.6, Exercise 8]). *Let K be a proper normal subgroup of D_n, $n \geq 3$.*

(a) if n is odd then $K = \langle a^u \rangle$, where u divides n,

(b) if n is even, then either $K = \langle a^2, b\rangle$, $K = \langle a^2, ab\rangle$ or $K = \langle a^u\rangle$, where u divides n.

Lemma 5.2 ([5]). *Let φ be a skew morphism of D_n, $n \geq 3$, then* $\operatorname{Ker}\varphi \neq \langle a\rangle$.

Lemma 5.3. *Let φ be a smooth skew morphism of D_n, $n \geq 3$. If n is odd, then φ is an automorphism of A, whereas if n is even and φ is not an automorphism of D_n, then* $\operatorname{Ker}\varphi = \langle a^2\rangle$, $\operatorname{Ker}\varphi = \langle a^2, ab\rangle$ *or* $\operatorname{Ker}\varphi = \langle a^2, b\rangle$. *Moreover, the involutory automorphism of D_n taking $a \mapsto a^{-1}, b \mapsto ab$ transposes the smooth skew morphisms of D_n with kernels $\langle a^2, b\rangle$ and $\langle a^2, ab\rangle$.*

Proof. Assume that φ is not an automorphism of D_n, then $1 < \operatorname{Ker}\varphi < D_n$. Since φ is smooth, by Theorem 4.9(a), the power function $\pi\colon D_n \to \mathbb{Z}^*_{|\varphi|}$ is a group homomorphism with $\operatorname{Ker}\pi = \operatorname{Ker}\varphi$. It follows that $\operatorname{Ker}\varphi$ is a proper normal subgroup of A. Since $\mathbb{Z}^*_{|\varphi|}$ is abelian, $D_n' \leq \operatorname{Ker}\varphi$, where D_n' is the derived subgroup of D_n.

If n is odd then $D_n' = \langle a\rangle$, which is a maximal subgroup of D_n. By Lemma 5.2 $\operatorname{Ker}\varphi \neq \langle a\rangle$, so $\operatorname{Ker}\varphi = D_n$, and hence φ is automorphism of D_n, a contradiction.

On the other hand, if n is even, then $D_n' = \langle a^2\rangle$, so $\langle a^2\rangle \leq \operatorname{Ker}\varphi$. By Lemma 5.1, one of the following three cases may happen: $\operatorname{Ker}\varphi \leq \langle a\rangle$, $\operatorname{Ker}\varphi = \langle a^2, b\rangle$, or $\operatorname{Ker}\varphi = \langle a^2, ab\rangle$. For the first case, by Lemma 5.2, we have $\operatorname{Ker}\varphi \neq \langle a\rangle$, so $\operatorname{Ker}\varphi = \langle a^2\rangle$.

Finally, by Theorem 4.9(d), the automorphism of D_n taking $a \mapsto a^{-1}, b \mapsto ab$ transposes the smooth skew morphisms of D_n with kernels $\langle a^2, b\rangle$ and $\langle a^2, ab\rangle$. □

The following result classifies smooth skew morphisms of the dihedral groups D_n with $\operatorname{Ker}\varphi = \langle a^2\rangle$ for even integers $n \geq 4$.

Theorem 5.4. *Let $D_n = \langle a, b\rangle$ be the dihedral group of order $2n$, where $n \geq 4$ is an even number. Then every smooth skew morphism φ of D_n with* $\operatorname{Ker}\varphi = \langle a^2\rangle$ *is defined by*

$$\begin{cases} \varphi(a^{2i}) = a^{2iu}, \\ \varphi(a^{2i+1}) = a^{2iu+2r+1}, \\ \varphi(a^{2i}b) = a^{2iu+2s}b, \\ \varphi(a^{2i+1}b) = a^{2iu+2r+2s\tau(u,e)+1}b \end{cases} \quad \text{and} \quad \begin{cases} \pi(a^{2i}) = 1, \\ \pi(a^{2i+1}) = e, \\ \pi(a^{2i}b) = f, \\ \pi(a^{2i+1}b) = ef, \end{cases} \tag{5.4}$$

where r, s, u, e, f are nonnegative integers satisfying the following conditions

(a) $r, s \in \mathbb{Z}_{n/2}$ and $u \in \mathbb{Z}^*_{n/2}$,

(b) the order of φ is the smallest positive integer k such that $r\tau(u,k) \equiv 0 \pmod{n/2}$ and $s\tau(u,k) \equiv 0 \pmod{n/2}$,

(c) $e, f \in \mathbb{Z}^*_k$ generate the Klein four group,

(d) $u^{e-1} \equiv 1 \pmod{n/2}$ and $u^{f-1} \equiv 1 \pmod{n/2}$,

(e) $r\tau(u, e-1) \equiv u - 2r - 1 \pmod{n/2}$ and $s\tau(u, f-1) \equiv 0 \pmod{n/2}$,

(f) $r\tau(u, f-1) + s\tau(u, e-1) \equiv u - 2r - 1 \pmod{n/2}$.

Proof. First suppose that φ is a smooth skew morphism of D_n with $\operatorname{Ker}\varphi = \langle a^2\rangle$. Then by Theorem 4.9(b), the induced skew morphism $\bar{\varphi}$ on $D_n/\operatorname{Ker}\varphi$ is the identity permutation, so there exist integers $r, s \in \mathbb{Z}_{n/2}$ such that

$$\varphi(a) = a^{1+2r} \qquad \text{and} \qquad \varphi(b) = a^{2s}b.$$

Since φ is kernel-preserving, the restriction of φ to $\operatorname{Ker}\varphi = \langle a^2\rangle$ is an automorphism, so $\varphi(a^2) = a^{2u}$ where $u \in \mathbb{Z}^*_{n/2}$. Assume that $\pi(a) \equiv e \pmod k$ and $\pi(b) \equiv f \pmod k$, where $k = |\varphi|$.

From the above identities we derive the following formulae by induction:

$$\varphi^j(a) = a^{1+2r\tau(u,j)} \qquad \text{and} \qquad \varphi^j(b) = a^{2s\tau(u,j)}b,$$

where j is a positive integer and $\tau(u,j) = \sum_{i=1}^{j} u^{i-1}$. Since $D_n = \langle a, b\rangle$, by Lemma 2.8, the order $k = |\varphi|$ is equal to $\operatorname{lcm}(|O_a|, |O_b|)$, the least common multiple of the lengths of the orbits containing a and b. That is, k is the smallest positive integer such that $\varphi^k(a) = a$ and $\varphi^k(b) = b$. Using the above formulae we then deduce that k is the smallest positive integer such that $r\tau(u,k) \equiv 0 \pmod{n/2}$ and $s\tau(u,k) \equiv 0 \pmod{n/2}$.

Now we determine the skew morphism and the associated power function. By the assumption we have

$$\begin{aligned}\varphi(a^{2i}) &= (a^{2u})^i = a^{2iu},\\ \varphi(a^{2i}b) &= \varphi(a^{2i})\varphi(b) = a^{2iu+2s}b.\end{aligned}$$

Similarly, we have

$$\begin{aligned}\varphi(a^{2i+1}) &= \varphi(a^{2i}a) = \varphi(a^{2i})\varphi(a) = a^{1+2r+2iu},\\ \varphi(a^{2i+1}b) &= \varphi(a^{2i})\varphi(a)\varphi^e(b) = a^{2iu+1+2r+2s\tau(u,e)}.\end{aligned}$$

Since $\pi\colon D_n \to \mathbb{Z}^*_k$ is a group homomorphism, we have $e^2 \equiv \pi(a)^2 = \pi(a^2) \equiv 1 \pmod k$ and $f^2 \equiv \pi(b)^2 \equiv \pi(b^2) \equiv 1 \pmod k$, so $e^2 \equiv 1 \pmod k$ and $f^2 \equiv 1 \pmod k$. Hence, $\pi(a^{2i}) \equiv 1$, $\pi(a^{2i+1}) \equiv e$, $\pi(a^{2i}b) \equiv f$, $\pi(a^{2i+1}b) \equiv ef$. In particular, since $|D_n : \operatorname{Ker}\varphi| = 4$, $\langle e, f\rangle \le \mathbb{Z}^*_k$ is the Klein four group. Therefore φ and π have the claimed form (5.4).

Moreover, we have

$$\begin{aligned}a^{1+2r+2u^e} &= \varphi(a)\varphi^e(a^2) = \varphi(a)\varphi^{\pi(a)}(a^2) = \varphi(aa^2) = \varphi(a^2a)\\ &= \varphi(a^2)\varphi(a) = a^{1+2r+2u},\end{aligned}$$

and so $u^{e-1} \equiv 1 \pmod{n/2}$. Similarly, since

$$\varphi(b)\varphi^f(a^2) = \varphi(b)\varphi^{\pi(b)}(a^2) = \varphi(ba^2) = \varphi(a^{-2}b) = \varphi(a^{-2})\varphi(b),$$

we have

$$a^{2s-2u^f}b = a^{2s}ba^{2u^f} = \varphi(b)\varphi^f(a^2) = \varphi(a^{-2})\varphi(b) = a^{2s-2u}b.$$

Thus, $u^{f-1} \equiv 1 \pmod{n/2}$.

Furthermore, since

$$a^{2u} = \varphi(a^2) = \varphi(a)\varphi^{\pi(a)}(a) = \varphi(a)\varphi^e(a) = a^{2+2r+2r\tau(u,e)},$$

we get

$$r(1+\tau(u,e)) \equiv u-1 \pmod{n/2}. \tag{5.5}$$

Similarly,

$$1 = \varphi(b^2) = \varphi(b)\varphi^{\pi(b)}(b) = \varphi(b)\varphi^f(b) = a^{2s}ba^{2s\tau(u,f)}b = a^{2s-2s\tau(u,f)},$$

we obtain

$$s\tau(u,f) \equiv s \pmod{n/2}. \tag{5.6}$$

Employing induction it is easy to deduce that $\varphi^j(a^{-1}) = a^{1-2u^j+2r\tau(u,j)}$, where j is an arbitrary positive integer. Then

$$\varphi(a)\varphi^e(b) = \varphi(ab) = \varphi(ba^{-1}) = \varphi(b)\varphi^f(a^{-1}).$$

Upon substitution we get

$$\begin{aligned} a^{1+2r+2s\tau(u,e)}b = \varphi(a)\varphi^e(b) = \varphi(b)\varphi^f(a^{-1}) = a^{2s}ba^{1-2u^f+2r\tau(u,f)} \\ = a^{2s-1+2u^f-2r\tau(u,f)}b. \end{aligned}$$

Hence,

$$r\tau(u,f) + s\tau(u,e) \equiv s + u^f - r - 1 \pmod{n/2}.$$

Since $u^f \equiv u \pmod{n/2}$, the congruence is reduced to

$$r\tau(u,f) + s\tau(u,e) \equiv s + u - r - 1 \pmod{n/2}. \tag{5.7}$$

Recall that $u^{e-1} \equiv 1 \pmod{n/2}$ and $u^{f-1} \equiv 1 \pmod{n/2}$, so

$$\begin{aligned} \tau(u,e) &\equiv \tau(u,e-1) + 1 \pmod{n/2}, \\ \tau(u,f) &\equiv \tau(u,f-1) + 1 \pmod{n/2}. \end{aligned}$$

Upon substitution the congruences (5.5), (5.6) and (5.7) are reduced to the numerical conditions in (e) and (f).

Conversely, for a quintuple (r,s,u,e,f) of nonnegative integers satisfying the stated numerical conditions, we verify that φ given by (5.4) is a smooth skew morphism of D_n with $\operatorname{Ker}\varphi = \langle a^2 \rangle$ and the function π is the associated power function. It is evident that φ is a bijection on D_n and $\varphi(1) = 1$.

It remains to verify the identity $\varphi(xy) = \varphi(x)\varphi^{\pi(x)}(y)$ for all $x, y \in D_n$. By Lemma 2.8, it suffices to verify this for $x, y \in O_a \cup O_b$, where O_a and O_b are the generating orbits of φ of the form

$$\begin{aligned} O_a &= (a, a^{1+2r\tau(u,1)}, a^{1+2r\tau(u,2)}, \dots, a^{1+2r\tau(u,i)}, \dots), \\ O_b &= (b, a^{2s\tau(u,1)}b, a^{2s\tau(u,2)}b, \dots, a^{2s\tau(u,j)}b, \dots). \end{aligned}$$

It follows that one of the following four cases may happen:

(i) $x, y \in O_a$;

(ii) $x, y \in O_b$;

(iii) $x \in O_a$, $y \in O_b$ or

(iv) $x \in O_b$, $y \in O_a$.

We shall demonstrate the verification for the first case, and leave other cases to the reader.

If $x, y \in O_a$, then $x = a^{1+2r\tau(u,i)}$ and $y = a^{1+2r\tau(u,j)}$ for some i, j. We have

$$\varphi(x)\varphi(y) = \varphi(a^{2r(\tau(u,i)+\tau(u,j))+2}) = a^{2ru(\tau(u,i)+\tau(u,j))+2u}$$

and

$$\varphi(x)\varphi^{\pi(x)}(y) = \varphi(a^{1+2r\tau(u,i)})\varphi^e(a^{1+2r\tau(u,j)}) = a^{2r(\tau(u,i+1)+\tau(u,j+e))+2}.$$

By the numerical conditions (d) and (e), we have

$$\begin{aligned} r(\tau(u,i+1) &+ \tau(u,j+e)) + 1 - (ru(\tau(u,i)+\tau(u,j)) + u) \\ &= r\Big((\tau(u,i+1) - u\tau(u,i)) + (\tau(u,j+e) - u\tau(u,j))\Big) + 1 - u \\ &\overset{\text{(d)}}{\equiv} r\Big(1 + (\tau(u,j+e) - u^e\tau(u,j))\Big) + 1 - u \\ &\equiv r(2 + \tau(u,e-1)) + 1 - u \\ &\overset{\text{(e)}}{\equiv} 0 \pmod{n/2}. \end{aligned}$$

Therefore, $\varphi(xy) = \varphi(x)\varphi^{\pi(x)}(y)$.

Finally, from the choices of the parameters it is easily seen that distinct quintuples (r, s, u, e, f) give rise to different skew morphisms of D_n, as required. □

Remark 5.5. In Theorem 5.4, consider the particular case where $u = 1$. By Condition (b) we have

$$k = \operatorname{lcm}\left(\frac{n/2}{\gcd(r, n/2)}, \frac{n/2}{\gcd(s, n/2)}\right).$$

The numerical conditions are reduced to

$$\begin{cases} r(e+1) \equiv 0 \pmod{n/2}, \\ s(f-1) \equiv 0 \pmod{n/2}, \\ r(f+1) + s(e-1) \equiv 0 \pmod{n/2}, \end{cases}$$

where $r, s \in \mathbb{Z}_{n/2}$ and $\langle e, f\rangle \le \mathbb{Z}_k^*$ is the Klein four group. If $n = 8m$, where $m \ge 3$ is an odd number, then it can be easily verified that the quintuple $(r, s, u, e, f) = (m + 4, m, 1, 4m - 1, 2m - 1)$ fulfills the numerical conditions. Therefore, we obtain an infinite family of skew morphisms of D_{8m} of order $4m$ with $\operatorname{Ker}\varphi = \langle a^2\rangle$. This example was first discovered by Zhang and Du in [26, Example 1.4].

Example 5.6. By computations using the MAGMA system we found that the smallest n for which there is a smooth skew morphism φ of D_n with $\operatorname{Ker}\varphi = \langle a^2\rangle$ is the number 24. In this case, all such skew morphisms have order 12, and the corresponding quintuples (r, s, u, e, f) are listed below:

$$\begin{aligned} (r, s, u, e, f) = {} & (1,3,1,11,5), (1,4,1,11,7), (1,9,1,11,5), (1,10,1,11,7), \\ & (5,2,1,11,7), (5,3,1,11,5), (5,8,1,11,7), (5,9,1,11,5), \\ & (7,3,1,11,5), (7,4,1,11,7), (7,9,1,11,5), (7,10,1,11,7), \\ & (11,,2,1,11,7), (11,3,1,11,5), (11,8,1,11,7), (11,9,1,11,5). \end{aligned}$$

Note that in each case we have $u = 1$, so the restriction of φ to $\operatorname{Ker}\varphi$ is the identity automorphism of $\operatorname{Ker}\varphi$. However, further computations show that, for other n, there do exist examples with $u \neq 1$.

For even numbers n, by Lemma 5.3, the involutory automorphism γ of D_n taking $a \mapsto a^{-1}$, $b \mapsto ab$ transposes the smooth skew morphisms of D_n with kernels $\langle a^2, b\rangle$ or $\langle a^2, ab\rangle$. Thus, to complete the classification of smooth skew morphisms of D_n, it suffices to determine the smooth skew morphisms of D_n with kernel $\operatorname{Ker}\varphi = \langle a^2, b\rangle$.

Theorem 5.7. *Let D_n be the dihedral group of order $2n$, where $n \geq 8$ is an even number. If φ is a smooth skew morphism of D_n with $\operatorname{Ker}\varphi = \langle a^2, b\rangle$, then φ belongs to one of the following two families of skew morphisms:*

(I) skew morphisms of order k defined by

$$\begin{cases}\varphi(a^{2i}) = a^{2iu},\\ \varphi(a^{2i+1}) = a^{2iu+2r+1},\\ \varphi(ba^{2i}) = ba^{2iu+2s},\\ \varphi(ba^{2i+1}) = ba^{2r+2s+2iu+1}\end{cases} \quad \text{and} \quad \begin{cases}\pi(a^{2i}) = 1,\\ \pi(a^{2i+1}) = e,\\ \pi(ba^{2i}) = 1,\\ \pi(ba^{2i+1}) = e,\end{cases} \tag{5.8}$$

where r, s, u, k, e are nonnegative integers satisfying the following conditions

(a) $r, s \in \mathbb{Z}_{n/2}$ and $u \in \mathbb{Z}^*_{n/2}$,
(b) k is the smallest positive integer such that $r\tau(u,k) \equiv 0 \pmod{n/2}$ and $s\tau(u,k) \equiv 0 \pmod{n/2}$,
(c) $e \in \mathbb{Z}^*_k$ such that $e \not\equiv 1 \pmod k$ and $e^2 \equiv 1 \pmod k$,
(d) $u^{e-1} \equiv 1 \pmod{n/2}$,
(e) $r\tau(u, e-1) \equiv u - 2r - 1 \pmod{n/2}$ and $s\tau(u, e-1) \equiv -u + 2r + 1 \pmod{n/2}$.

(II) skew morphisms of order $2(e-1)$ defined by

$$\begin{cases}\varphi(a^{2i}) = a^{2iu},\\ \varphi(a^{2i+1}) = ba^{2r-2iu+1},\\ \varphi(ba^{2i}) = ba^{2s+2iu},\\ \varphi(ba^{2i+1}) = a^{2r-2s-2iu+1}\end{cases} \quad \text{and} \quad \begin{cases}\pi(a^{2i}) = 1,\\ \pi(a^{2i+1}) = e,\\ \pi(ba^{2i}) = 1,\\ \pi(ba^{2i+1}) = e,\end{cases} \tag{5.9}$$

where r, s, u, e are nonnegative integers satisfying the following conditions

(a) $r, s \in \mathbb{Z}_{n/2}$, $u \in \mathbb{Z}^*_{n/2}$ and $e > 1$ is an odd number,
(b) $u^{e-1} \equiv -1 \pmod{n/2}$,
(c) $s\tau(u, e-1) \equiv u + 2r + 1 \pmod{n/2}$,
(d) $r\rho(u, e-1) \equiv s\lambda(u, e-1) - 1 \pmod{n/2}$.

Proof. First suppose that φ is a smooth skew morphism of D_n with $\operatorname{Ker}\varphi = \langle a^2, b\rangle$. By Theorem 4.9, the induced skew morphism $\bar{\varphi}$ of $D_n/\operatorname{Ker}\varphi$ is the identity permutation and the restriction of φ to $\operatorname{Ker}\varphi = \langle a^2, b\rangle$ is an automorphism of $\operatorname{Ker}\varphi$. It follows that there exist integers $r, s, u \in \mathbb{Z}_{n/2}$ and $\ell \in \mathbb{Z}_2$ such that

$$\varphi(a) = b^\ell a^{1+2r}, \qquad \varphi(b) = ba^{2s} \qquad \text{and} \qquad \varphi(a^2) = a^{2u}.$$

Assume that $\pi(a) \equiv e \pmod{k}$, where $k = |\varphi|$ denotes the order of φ. Since $b \in \operatorname{Ker}\varphi$, $\pi(b) \equiv 1 \pmod{k}$. By Theorem 4.9, the power function $\pi\colon D_n \to \mathbb{Z}_k$ is a group homomorphism from D_n to the multiplicative group $\mathbb{Z}_k^*$, so

$$e^{-1} \equiv \pi(a^{-1}) \equiv \pi(b^{-1}ab) \equiv \pi(a) \equiv e \pmod{k},$$

and hence $e^2 \equiv 1 \pmod{k}$. It follows that $\pi(a^{2i}) \equiv \pi(a^{2i}b) \equiv 1$ and $\pi(a^{2i+1}) \equiv \pi(a^{2i+1}b) \equiv e$. Since φ has skew type 2, $e \not\equiv 1 \pmod{k}$. To proceed we distinguish two cases:

Case (I): $\ell = 0$.

In this case, we have

$$\varphi(a) = a^{1+2r}, \qquad \varphi(b) = ba^{2s} \qquad \text{and} \qquad \varphi(a^2) = a^{2u}.$$

Then

$$\begin{aligned} \varphi(a^{2i}) &= \varphi(a^2)^i = a^{2iu}, \\ \varphi(ba^{2i}) &= \varphi(b)\varphi(a^2)^i = ba^{2iu+2s}. \end{aligned}$$

Similarly,

$$\begin{aligned} \varphi(a^{2i+1}) &= \varphi(a^{2i}a) = \varphi(a^2)^i\varphi(a) = a^{2iu+2r+1}, \\ \varphi(ba^{2i+1}) &= \varphi(ba^{2i}a) = \varphi(b)\varphi(a^{2i})\varphi(a) = ba^{2r+2s+2iu+1}. \end{aligned}$$

Hence, the skew morphism has the form given by (5.8).

Using induction it is easy to prove that

$$\varphi^j(a) = a^{1+2r\tau(u,j)} \qquad \text{and} \qquad \varphi^j(b) = ba^{2s\tau(u,j)},$$

where j is a positive integer and $\tau(u,j) = \sum_{i=1}^{j} u^{i-1}$. Since $D_n = \langle a, b\rangle$, $k = |\varphi|$ is the smallest positive integer such that $\varphi^k(a) = a$ and $\varphi^k(b) = b$, which implies that

$$r\tau(u,k) \equiv 0 \pmod{n/2} \qquad \text{and} \qquad s\tau(u,k) \equiv 0 \pmod{n/2}.$$

Moreover, we have

$$a^{1+2r+2u^e} = \varphi(a)\varphi^e(a^2) = \varphi(aa^2) = \varphi(a^2a) = \varphi(a^2)\varphi(a) = a^{1+2r+2u},$$

so $u^{e-1} \equiv 1 \pmod{n/2}$.

Furthermore, since

$$a^{2u} = \varphi(a^2) = \varphi(a)\varphi^e(a) = a^{1+2r}a^{1+2r\tau(u,e)} = a^{2+2r+2r\tau(u,e)},$$

we obtain

$$r\big(\tau(u,e) + 1\big) \equiv u - 1 \pmod{n/2}. \tag{5.10}$$

Similarly,

$$\varphi(a)\varphi^e(b) = \varphi(ab) = \varphi(ba^{-1}) = \varphi(b)\varphi(a^{-1}) = \varphi(b)\varphi(a^{-2}a) = \varphi(b)\varphi(a^{-2})\varphi(a).$$

By the above formula we have

$$\varphi(a)\varphi^e(b) = a^{1+2r}ba^{2s\tau(u,e)} = ba^{-1-2r+2s\tau(u,e)}$$

and

$$\varphi(b)\varphi(a^{-2})\varphi(a) = ba^{1+2r+2s-2u}.$$

Consequently, upon substitution we obtain

$$s(\tau(u,e)-1) \equiv -u+2r+1 \pmod{n/2}. \tag{5.11}$$

Recall that $u^{e-1} \equiv 1 \pmod{n/2}$, so

$$\tau(u,e) = \tau(u,e-1) + u^{e-1} \equiv \tau(u,e-1)+1 \pmod{n/2}.$$

Upon substitution the equations (5.10) and (5.11) are reduced to

$$\begin{aligned} r\tau(u,e-1) &\equiv u-2r-1 \pmod{n/2}, \\ s\tau(u,e-1) &\equiv -u+2r+1 \pmod{n/2}. \end{aligned}$$

Case (II): $\ell = 1$.

In this case we have

$$\varphi(a) = ba^{1+2r}, \qquad \varphi(b) = ba^{2s} \qquad \text{and} \qquad \varphi(a^2) = a^{2u}.$$

Then

$$\begin{aligned} \varphi(a^{2i}) &= a^{2iu}, \\ \varphi(ba^{2i}) &= \varphi(b)\varphi(a^{2i}) = ba^{2s+2iu}. \end{aligned}$$

Similarly,

$$\begin{aligned} \varphi(a^{2i+1}) &= \varphi(a^{2i}a) = a^{2iu}ba^{1+2r} = ba^{2r-2iu+1}, \\ \varphi(ba^{2i+1}) &= \varphi(b)\varphi(a^{2i})\varphi(a) = a^{2r-2s-2iu+1}. \end{aligned}$$

Hence φ has the form (5.9).

Using induction it is easy to derive the following formula

$$\varphi^j(b) = ba^{2s\tau(u,j)} \qquad \text{and} \qquad \varphi^j(a) = \begin{cases} a^{2r\rho(u,j)-2s\lambda(u,j)+1}, & \text{if } j \text{ is even,} \\ ba^{2r\rho(u,j)+2su\lambda(u,j-1)+1}, & \text{if } j \text{ is odd,} \end{cases}$$

where τ, ρ and λ are the functions defined by (5.2) and (5.3). Since $\varphi(a) = ba^{1+2r}$ and $D_n = \langle a, ba^{1+2r}\rangle$, $k = |\varphi| = |O_a|$. Thus, k is the smallest positive integer such that

$$r\rho(u,k) \equiv s\lambda(u,k) \pmod{n/2}.$$

In particular, since elements from the cosets $\langle a\rangle$ and $b\langle a\rangle$ alternate in the orbit O_a, k is even, and hence e is odd. Thus,

$$a^{2u} = \varphi(a^2) = \varphi(a)\varphi^e(a) = \varphi(a)\varphi^e(a) = a^{2r\rho(u,e)-2r+2su\lambda(u,e-1)}.$$

Consequently, we obtain

$$r\rho(u,e) + su\lambda(u,e-1) \equiv r + u \pmod{n/2}. \tag{5.12}$$

Furthermore, we have

$$\begin{aligned} ba^{1+2r+2u^e} &= \varphi(a)\varphi^e(a^2) = \varphi(aa^2) = \varphi(a^2a) \\ &= \varphi(a^2)\varphi(a) = a^{2u}ba^{1+2r} = ba^{2r-2u+1}, \end{aligned}$$

so $u^{e-1} \equiv -1 \pmod{n/2}$. Similarly

$$\begin{aligned} a^{-1-2r+2s\tau(u,e)} &= \varphi(a)\varphi^e(b) = \varphi(ab) = \varphi(ba^{-2}a) \\ &= \varphi(b)\varphi(a^{-2})\varphi(a) = a^{1+2r-2s+2u}. \end{aligned}$$

Hence

$$s\tau(u,e) \equiv 1 + 2r + u - s \pmod{n/2}. \tag{5.13}$$

Recall that $u^{e-1} \equiv -1 \pmod{n/2}$, so

$$\begin{aligned} \tau(u,e) &\equiv \tau(u,e-1) - 1 \pmod{n/2}, \\ \rho(u,e) &\equiv \rho(u,e-1) - 1 \pmod{n/2}. \end{aligned}$$

Upon substitution the equations (5.12) and (5.13) are reduced to

$$r\rho(u,e-1) + su\lambda(u,e-1) \equiv 2r + u \pmod{n/2}, \tag{5.14}$$

$$s\tau(u,e-1) \equiv 2r + u + 1 \pmod{n/2}. \tag{5.15}$$

Subtracting we then get

$$r\rho(u,e-1) \equiv s\lambda(u,e-1) - 1 \pmod{n/2}.$$

Finally, note that

$$\begin{aligned} \rho(u,2(e-1)) &= \sum_{i=1}^{2(e-1)} (-u)^{2(e-1)} \\ &= \sum_{i=1}^{e-1} (-u)^{i-1} + u^{e-1}\sum_{i=1}^{e-1}(-u)^{i-1} \equiv 0 \pmod{n/2}, \end{aligned}$$

and

$$\begin{aligned} \lambda(u,2(e-1)) &= \sum_{i=1}^{e-1} u^{2i} \\ &= \sum_{i=1}^{(e-1)/2} u^{2(i-1)} + u^{e-1}\sum_{i=1}^{(e-1)/2} u^{2(i-1)} \equiv 0 \pmod{n/2}, \end{aligned}$$

Hence,

$$r\rho(u, 2(e-1)) \equiv s\lambda(u, 2(e-1)) \pmod{n/2}.$$

The minimality of k yields $k \mid 2(e-1)$. But $e - 1 < k$, which forces $k = 2(e-1)$.

Conversely, in each case for any quadruple (r, s, u, e) satisfying the numerical conditions, it is straightforward to verify that φ of the given form is a smooth skew morphism of D_n with $\operatorname{Ker}\varphi = \langle a^2, b\rangle$ and π is the associated power function. In particular, from the choices of the parameters it is easily seen that distinct quadruples (r, s, u, e) give rise to different skew morphisms of D_n, as required. □

Remark 5.8. Let φ be any skew morphism from (II) of Theorem 5.7. Note that the orbit of φ containing a also contains ba^{2r+1}, so the orbit O_a generates D_n. Clearly, it is closed under taking inverses. Therefore, such skew morphisms give rise to the e-balanced regular Cayley maps of D_n, which were first classified by Kwak, Kwon and Feng [17].

References

[1] M. Bachratý and R. Jajcay, Powers of skew-morphisms, in: J. Širáň and R. Jajcay (eds.), *Symmetries in Graphs, Maps, and Polytopes*, Springer, Cham, volume 159 of *Springer Proceedings in Mathematics & Statistics*, 2016 pp. 1–25, doi:10.1007/978-3-319-30451-9_1, papers from the 5th SIGMAP Workshop held in West Malvern, July 7 – 11, 2014.

[2] M. Bachratý and R. Jajcay, Classification of coset-preserving skew-morphisms of finite cyclic groups, *Australas. J. Combin.* **67** (2017), 259–280, `https://ajc.maths.uq.edu.au/pdf/67/ajc_v67_p259.pdf`.

[3] M. Conder, R. Jajcay and T. Tucker, Regular Cayley maps for finite abelian groups, *J. Algebraic Combin.* **25** (2007), 259–283, doi:10.1007/s10801-006-0037-0.

[4] M. Conder, R. Jajcay and T. Tucker, Regular t-balanced Cayley maps, *J. Comb. Theory Ser. B* **97** (2007), 453–473, doi:10.1016/j.jctb.2006.07.008.

[5] M. D. E. Conder, R. Jajcay and T. W. Tucker, Cyclic complements and skew morphisms of groups, *J. Algebra* **453** (2016), 68–100, doi:10.1016/j.jalgebra.2015.12.024.

[6] M. D. E. Conder and T. W. Tucker, Regular Cayley maps for cyclic groups, *Trans. Amer. Math. Soc.* **366** (2014), 3585–3609, doi:10.1090/s0002-9947-2014-05933-3.

[7] K. Hu, Theory of skew morphisms, 2012, preprint.

[8] K. Hu and Y. S. Kwon, Regular Cayley maps and skew morphisms of dihedral groups: a survey, in preparation.

[9] R. Jajcay and R. Nedela, Half-regular Cayley maps, *Graphs Combin.* **31** (2015), 1003–1018, doi:10.1007/s00373-014-1428-y.

[10] R. Jajcay and J. Širáň, Skew-morphisms of regular Cayley maps, *Discrete Math.* **244** (2002), 167–179, doi:10.1016/s0012-365x(01)00081-4.

[11] I. Kovács and Y. S. Kwon, Regular Cayley maps on dihedral groups with the smallest kernel, *J. Algebraic Combin.* **44** (2016), 831–847, doi:10.1007/s10801-016-0689-3.

[12] I. Kovács and Y. S. Kwon, Classification of reflexible Cayley maps for dihedral groups, *J. Comb. Theory Ser. B* **127** (2017), 187–204, doi:10.1016/j.jctb.2017.06.002.

[13] I. Kovács and Y. S. Kwon, private communication, 2018.

[14] I. Kovács, D. Marušič and M. Muzychuk, On G-arc-regular dihedrants and regular dihedral maps, *J. Algebraic Combin.* **38** (2013), 437–455, doi:10.1007/s10801-012-0410-0.

[15] I. Kovács and R. Nedela, Decomposition of skew-morphisms of cyclic groups, *Ars Math. Contemp.* **4** (2011), 329–349, doi:10.26493/1855-3974.157.fc1.

[16] H. Kurzweil and B. Stellmacher, *The Theory of Finite Groups: An Introduction*, Universitext, Springer-Verlag, New York, 2004, doi:10.1007/b97433.

[17] J. H. Kwak, Y. S. Kwon and R. Feng, A classification of regular t-balanced Cayley maps on dihedral groups, *European J. Combin.* **27** (2006), 382–393, doi:10.1016/j.ejc.2004.12.002.

[18] J. H. Kwak and J.-M. Oh, A classification of regular t-balanced Cayley maps on dicyclic groups, *European J. Combin.* **29** (2008), 1151–1159, doi:10.1016/j.ejc.2007.06.023.

[19] Y. S. Kwon, A classification of regular t-balanced Cayley maps for cyclic groups, *Discrete Math.* **313** (2013), 656–664, doi:10.1016/j.disc.2012.12.012.

[20] J.-M. Oh, Regular t-balanced Cayley maps on semi-dihedral groups, *J. Comb. Theory Ser. B* **99** (2009), 480–493, doi:10.1016/j.jctb.2008.09.006.

[21] Y. Wang and R. Q. Feng, Regular balanced Cayley maps for cyclic, dihedral and generalized quaternion groups, *Acta Math. Sin.* **21** (2005), 773–778, doi:10.1007/s10114-004-0455-7.

[22] K. Yuan, Y. Wang and J. H. Kwak, Enumeration of skew-morphisms of cyclic groups of small orders and their corresponding Cayley maps, *Adv. Math. (China)* **45** (2016), 21–36.

[23] J.-Y. Zhang, Regular Cayley maps of skew-type 3 for abelian groups, *European J. Combin.* **39** (2014), 198–206, doi:10.1016/j.ejc.2014.01.006.

[24] J.-Y. Zhang, A classification of regular Cayley maps with trivial Cayley-core for dihedral groups, *Discrete Math.* **338** (2015), 1216–1225, doi:10.1016/j.disc.2015.01.036.

[25] J.-Y. Zhang, Regular Cayley maps of skew-type 3 for dihedral groups, *Discrete Math.* **338** (2015), 1163–1172, doi:10.1016/j.disc.2015.01.038.

[26] J.-Y. Zhang and S. Du, On the skew-morphisms of dihedral groups, *J. Group Theory* **19** (2016), 993–1016, doi:10.1515/jgth-2016-0027.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 549–561
https://doi.org/10.26493/1855-3974.1657.d75
(Also available at http://amc-journal.eu)

A q-queens problem
VI. The bishops' period

Seth Chaiken [*]

Computer Science Department, The University at Albany (SUNY),
Albany, NY 12222, U.S.A.

Christopher R. H. Hanusa [†]

Department of Mathematics, Queens College (CUNY),
65-30 Kissena Blvd., Queens, NY 11367-1597, U.S.A.

Thomas Zaslavsky

Department of Mathematical Sciences, Binghamton University (SUNY),
Binghamton, NY 13902-6000, U.S.A.

Received 27 March 2018, accepted 22 August 2018, published online 31 March 2019

Abstract

The number of ways to place q nonattacking queens, bishops, or similar chess pieces on an $n \times n$ square chessboard is essentially a quasipolynomial function of n (by Part I of this series). The period of the quasipolynomial is difficult to settle. Here we prove that the empirically observed period 2 for three to ten bishops is the exact period for every number of bishops greater than 2. The proof depends on signed graphs and the Ehrhart theory of inside-out polytopes.

Keywords: Nonattacking chess pieces, Ehrhart theory, inside-out polytope, arrangement of hyperplanes, signed graph.

Math. Subj. Class.: 05A15, 00A08, 05C22, 52C07, 52C35

[*] Chaiken and Zaslavsky thank the very hospitable Isaac Newton Institute for facilitating their work on this project.

[†] Hanusa gratefully acknowledges support from PSC-CUNY Research Awards PSCOOC-40-124, PSCREG-41-303, TRADA-42-115, TRADA-43-127, and TRADA-44-168.

E-mail addresses: sdc@cs.albany.edu (Seth Chaiken), chanusa@qc.cuny.edu (Christopher R. H. Hanusa), zaslav@math.binghamton.edu (Thomas Zaslavsky)

1 Introduction

The famous n-Queens Problem is to count the number of ways to place n nonattacking queens on an $n \times n$ chessboard. That problem has been solved only for small values of n; there is no real hope for a complete solution. In this series of papers we treat a more general problem wherein we place q identical pieces like the queen or bishop on an $n \times n$ square board and we seek a formula for $u(q; n)$, the number of ways to place them so that none attacks another. The piece may be any one of a large class of traditional and fairy chess pieces called “riders”, which are distinguished by the fact that their moves have unlimited distance. We proved in Part I [4] that in each such problem the number of solutions, times a factor of $q!$, is a quasipolynomial function of n; that is, $q!u(q; n)$ is given by a cyclically repeating sequence of polynomials in n and q, the exact polynomial depending on the residue class of n modulo some number p called the *period* of the function; and furthermore that each coefficient of the quasipolynomial is a polynomial function of q. Here we prove that for three or more bishops the period is always exactly 2.[1] This period was previously observed by Kotěšovec for $3 \leq q \leq 10$ as a result of his extensive computations for five to ten bishops, added to previous work by Fabel for three and four bishops (see [10, pp. 126–129] for $q \leq 6$ and [11, pp. 234–241, 254–257] for $q \leq 10$).

The number of nonattacking placements of q unlabelled bishops on an $n \times n$ board is denoted by $u_{\mathbb{B}}(q; n)$. The number for labelled bishops is therefore $q!u_{\mathbb{B}}(q; n)$.

Theorem 1.1. *For $q \geq 3$, the quasipolynomial $q!u_{\mathbb{B}}(q; n)$ involved in counting the nonattacking positions of q bishops on an $n \times n$ board has period equal to 2. For $q < 3$ the period is 1.*

To get our results we treat non-attacking configurations as integral lattice points $\mathbf{z} := (z_1, \ldots, z_q)$, $z_i = (x_i, y_i)$, in a $2q$-dimensional inside-out polytope (see Section 2). The Ehrhart theory of inside-out polytopes (from [3]) implies quasipolynomiality with polynomials of degree $2q$ and that the period divides the least common multiple of the denominators of the coordinates of vertices of the inside-out polytope. We find the structure of these coordinates explicitly: in Lemma 4.4 we show that a vertex of the bishops' inside-out polytope has each $z_i \in \{0, 1\}^2$ or $z_i = (\frac{1}{2}, \frac{1}{2})$. From that, along with a formula from Part III [6] for the coefficient of n^{2q-6} that implies the period is even if $q \geq 3$, Theorem 1.1 follows directly.

One reason to want the period is a computational method for discovering $u(q; n)$. To find it (for a fixed number q of pieces) one can count solutions as n ranges from 1 up to some upper limit N and interpolate the counting quasipolynomial from the resulting data. That can be done if one knows the degree of the quasipolynomial, which is $2q$ by [4, Lemma 2.1], and the period, for which there is no known general formula; then $N = 2qp$ suffices (since the leading term is $n^{2q}/q!$ by general Ehrhart theory; see [4, Lemma 2.1]). Evidently, knowing the period is essential to knowing the right value of N, if the formula is to be considered proven. In general, for a particular rider piece and number q it is very hard to find the period; its value is known only for trivial pieces or very small values of q. In contrast, Theorem 1.1 gives the exact period for bishops, and it follows that to find the exact number of placements of q bishops it suffices to compute only $4q$ values of the counting function.

The reader may ask why we do not seek the complete formula for bishops placements in terms of both n and q. Remarkably, there is a simple such formula, due in essence to

[1] This paper was originally announced as Part V, in Parts I and II.

Arshon in a nearly forgotten paper [2] and completed by Kotěšovec [11, pp. 244, 254–257]. We restate this expression in Part V [8]. The trouble is that it is not in the form of a quasipolynomial, so that, for instance, we could not use it to obtain the number of combinatorial types of nonattacking configuration, which by [4, Theorem 5.3] is its evaluation at $n = -1$. We cannot even deduce the period from the Arshon formula.[2] So there is reason to seek the general quasipolynomial $q!u_{\mathbb{B}}(q;n)$ for every number q. The simple reason we do not seek to do so is that we have not found a way to do it. That remains an open problem whose solution would reveal the full character of the dependence of $u_{\mathbb{B}}(q;n)$ on its two arguments. This has not yet been discovered for any rider—other than the mathematically trivial rook.

After necessary mathematical background in the next two sections, we prove Theorem 1.1 in Section 4, applying the geometry of the inside-out polytope for bishops and the properties of signed graphs, which we introduce in Sections 2 and 3, respectively. We conclude with research questions. For the benefit of the authors and readers, we append a dictionary of the notation in this paper.

2 Essentials from Parts I and II

We build upon the counting theory of previous parts as it applies to the square board, from Part II [5]. We summarize essential aspects here. First, we specialize our notation to q nonattacking bishops $\mathbb{B}$ on a square board. We assume that $q > 0$.

The full expression for the number of nonattacking configurations of unlabelled bishops is

$$u_{\mathbb{B}}(q;n) = \gamma_0(n)n^{2q} + \gamma_1(n)n^{2q-1} + \gamma_2(n)n^{2q-2} + \cdots + \gamma_{2q}(n)n^0,$$

where each coefficient $\gamma_i(n)$ varies periodically with n, and for labelled pieces the number is $o_{\mathbb{B}}(q;n)$, which equals $q!u_{\mathbb{B}}(q;n)$. (The coefficients also depend on q but we suppress that in the notation because only the variation with n concerns us here.)

The $n \times n$ board consists of the integral points in the interior $(n+1)(0,1)^2$ of an integral multiple $(n+1)[0,1]^2$ of the unit square $\mathcal{B} = [0,1]^2 \subset \mathbb{R}^2$, or equivalently, the $1/(n+1)$-fractional points in $(0,1)^2$. Thus, the board consists of the points $z = (x,y)$ for integers $x, y = 1, 2, \dots, n$.

A *move* is the difference between a new position and the original position. The bishop has moves given by all integral multiples of the vectors $(1,1)$ and $(1,-1)$, which are called the *basic moves*. (Note that for a move $m = (c,d)$, the slope d/c contains all necessary information and can be specified instead of m itself.) A bishop in position $z = (x,y)$ may move to any location $z + \mu m$ with $\mu \in \mathbb{Z}$ and a basic move m, provided that location is on the board.

A *configuration* is the vector $(z_1, z_2, \dots, z_q)$ of positions of all q bishops. The constraint on a configuration is that no two pieces may attack each other, or to say it mathematically, when there are pieces at positions z_i and z_j, then $z_j - z_i$ is not a multiple of any basic move m.

The object on which our theory relies is the *inside-out polytope* $(\mathcal{P}, \mathscr{A}_{\mathbb{B}})$, where $\mathcal{P}$ is the hypercube $[0,1]^{2q}$ and $\mathscr{A}_{\mathbb{B}}$ is the *move arrangement* for bishops. The move arrangement is a finite set of hyperplanes whose members are the *move hyperplanes* or *attack hyperplanes*,

$$\mathcal{H}^{\pm}_{ij} := \{\mathbf{z} \in \mathbb{R}^{2q} : (y_j - y_i) = \pm(x_j - x_i)\}.$$

[2] Stanley in [12, Solution to Exercise 4.42] says the period is easily obtained from Arshon's formula, which has one form for even n and another for odd n; but we think it is not that easy.

Each attack hyperplane contains the configuration points $z = (z_1, z_2, \ldots, z_q) \in \mathbb{Z}^{2q}$ in which bishops i and j attack each other. (The pieces in a configuration are labelled 1 through q to enable effective description.) The *intersection lattice* of $\mathscr{A}_{\mathbb{B}}$ is the set of all intersections of subsets of the move arrangement, ordered by reverse inclusion. These intersection subspaces are the heart of our method.

3 Signed graphs

The signed graph we employ to describe an intersection subspace efficiently is a special case of the slope graph from [4, Section 3.3]. The fact that the bishops' two slopes are ±1 makes it possible to apply the well-developed theory of signed graphs.

A graph is $\Gamma = (N, E)$, with node set N and edge set E. It may have multiple edges but not loops. A 1*-forest* is a graph in which each component consists of a tree together with one more edge; thus, each component contains exactly one circle. A spanning 1-forest is a spanning subgraph (it contains all nodes) that is a 1-forest. The notation e_{ij} means the edge has end nodes v_i and v_j.

A *signed graph*, $\Sigma = (N, E, \sigma)$, is a graph in which each edge e is labelled $\sigma(e) = +$ or $-$. In a signed graph, a circle (cycle, circuit) is called *positive* or *negative* according to the product of its edge signs. A *signed circuit* is either a positive circle or a connected subgraph that contains exactly two circles, both negative. A node v is *homogeneous* if all incident edges have the same sign. We generally write $q := |N|$ because the nodes correspond to the bishops in a configuration.

Let $c(\Sigma)$ denote the number of components of a signed (or unsigned) graph and $\xi(\Sigma) := |E| - |N| + c(\Sigma)$, the cyclomatic number of the underlying unsigned graph.

The *incidence matrix* of Σ is the $|N| \times |E|$ matrix $\mathsf{H}(\Sigma)$ (H is "Eta") such that, in the column indexed by edge e, the elements are $\eta(v, e) = \pm 1$ if v is an endpoint of e and $= 0$ if it is not, with the signs chosen so that, if v_i and v_j are the endpoints, then $\eta(v_i, e)\eta(v_j, e) = -\sigma(e)$ [13, Section 8A]. That is, in the column of a positive edge there are one $+1$ and one -1, while in the column of a negative edge there are two $+1$'s or two -1's. The *rank* of Σ is the rank of its incidence matrix. From [13, Theorem 5.1(j)] we know a formula for the rank: $\operatorname{rk}(\Sigma) = |N| - b(\Sigma)$, where $b(\Sigma)$ is the number of components in which there is no negative circle. This rank function applied to spanning subgraphs makes a matroid $G(\Sigma)$ on the edge set of Σ [13]. An unsigned graph Γ acts as if it is an all-positive signed graph; therefore its incidence matrix has rank $\operatorname{rk}(\Gamma) = |N| - c(\Gamma)$ where $c(\Gamma)$ is the number of components and the corresponding matroid $G(\Gamma) := G(+\Gamma)$ is the cycle matroid of Γ.

From this and [13, Theorem 8B.1] we also know that $\mathsf{H}(\Sigma)$ has full column rank if and only if Σ contains no signed circuit and it has full row rank if and only if every component of Σ contains a negative circle. A signed graph that has both of these properties is necessarily a 1-forest in which every circle is negative.

A *positive clique* in Σ is a maximal set of nodes that are connected by positive edges; equivalently, it is the node set of a connected component of the spanning subgraph Σ^+ formed by the positive edges. A *negative clique* is similar. Either kind of set is called a *signed clique*. We call them "cliques" (in a slight abuse of terminology) because the signed cliques of a graph do not change if we complete the induced positive subgraph on a positive clique, and similarly for a negative clique. A homogeneous node v gives rise to a singleton signed clique with the sign not represented by an edge at v; if v is isolated it gives rise to two singleton cliques, one of each sign.

The number of positive cliques in Σ is $c(\Sigma^+)$ and the number of negative cliques is $c(\Sigma^-)$. Let $\mathsf{A}(\Sigma) := \{A_1, \dots, A_{c(\Sigma^+)}\}$ and $\mathsf{B}(\Sigma) := \{B_1, \dots, B_{c(\Sigma^-)}\}$ (read "Alpha" and "Beta") be the sets of positive and negative cliques, respectively. Since each node of Σ is in precisely one positive and one negative clique, we can define a bipartite graph $C(\Sigma)$, called the *clique graph* of Σ, whose node set is $\mathsf{A}(\Sigma) \cup \mathsf{B}(\Sigma)$ and whose edge set is N, the endpoints of the edge v_i being the cliques $A \in \mathsf{A}(\Sigma)$ and $B \in \mathsf{B}(\Sigma)$ such that $v_i \in A \cap B$.

Let us call an edge e *redundant* if $\Sigma \backslash e$ (Σ with e deleted) has the same signed cliques as does Σ, and call Σ *irredundant* if it has no redundant edges, in other words, if each signed clique has just enough edges of its sign to connect its nodes. A signed graph is irredundant if and only if both Σ^+ and Σ^- are forests. For example, a signed forest is irredundant. Any signed graph can be reduced to irredundancy with the same signed cliques by pruning redundant edges one by one.

Lemma 3.1. *If Σ is a signed graph with q nodes, then*

$$|\mathsf{A}(\Sigma)| + |\mathsf{B}(\Sigma)| = 2q - [\mathrm{rk}(\Sigma^+) + \mathrm{rk}(\Sigma^-)].$$

If Σ is irredundant, then

$$|\mathsf{A}(\Sigma)| + |\mathsf{B}(\Sigma)| = 2q - |E| = q + c(\Sigma) - \xi(\Sigma).$$

In particular, a signed tree has $q + 1$ signed cliques.

Proof. The first formula follows directly from the general formula for the rank of a graph.

If Σ is irredundant, Σ^+ is a forest with $|\mathsf{A}(\Sigma)|$ components and Σ^- is a forest with $|\mathsf{B}(\Sigma)|$ components. Therefore, $|\mathsf{A}(\Sigma)| + |\mathsf{B}(\Sigma)| = 2q - |E| = q - \xi(\Sigma) + c(\Sigma)$.

A more entertaining proof is by induction on the number of inhomogeneous nodes. Define $g(\Sigma) := |\mathsf{A}(\Sigma)| + |\mathsf{B}(\Sigma)| - 2q + |E| = |\mathsf{A}(\Sigma)| + |\mathsf{B}(\Sigma)| - q - c(\Sigma) + \xi(\Sigma)$. If all nodes are homogeneous, obviously $g(\Sigma) = 0$. Otherwise, let v be an inhomogeneous node. Split v into two nodes, v^+ and v^-, incident respectively to all the positive or negative edges at v. The new graph has one less inhomogeneous node, two more signed cliques (a positive clique $\{v^-\}$ and a negative clique $\{v^+\}$), one more node, and the same number of edges, hence the same value of g as does Σ. Thus, by induction, $g \equiv 0$. □

4 Proof of the bishops' period

We are now prepared to prove Theorem 1.1. We already proved in [6, Theorem 3.1] that the coefficients $\gamma_i(n)$ are constant (as functions of n) for $i < 6$ and that $\gamma_6(n)$ has period 2. Thus it will suffice to prove that the denominator of the inside-out polytope $(\mathcal{B}, \mathscr{A}_{\mathbb{B}})$ for q bishops divides 2. (In fact, what we prove is the stronger result stated in Lemma 4.4.) To do this, we find the denominators of all vertices explicitly by analyzing all sets of $2q$ equations that determine a point. We use the polytope $[0, 1]^{2q}$ for the boundary inequalities and the move arrangement $\mathscr{A}_{\mathbb{B}}$ for the equations of attack.

We use a fundamental fact from linear algebra.

Lemma 4.1. *The coordinates $z_i = (x_i, y_i)$ belong to a vertex of the inside-out polytope if and only if there are k attack equations and $2q - k$ boundary equations that uniquely determine those coordinates.*

We assume the q bishops are labelled $\mathbb{B}_1, \dots, \mathbb{B}_q$. A configuration of bishops is described by a point $\mathbf{z} = (z_1, z_2, \dots, z_q) \in \mathbb{R}^{2q}$, where $z_i = (x_i, y_i)$ is the normalized plane coordinate vector of the ith bishop $\mathbb{B}_i$; that is, $x_i, y_i \in (0,1)$ and the position of $\mathbb{B}_i$ is $(n+1)z_i$. The bishops constraints are that $\mathbf{z}$ should not lie in any of the $q(q-1)$ *bishops hyperplanes*,

$$\mathcal{H}^+_{ij} : x_i - y_i = x_j - y_j, \qquad \mathcal{H}^-_{ij} : x_i + y_i = x_j + y_j, \tag{4.1}$$

where $i \neq j$. The corresponding equations are the *bishops equations* and a subspace $\mathcal{U}$ defined by a set of bishops equations is a *bishops subspace*. The boundary equations of $[0,1]^{2q}$ have the form $x_i = 0$ or 1 and $y_i = 0$ or 1. We generalize the boundary constraints; we call any equation of the form $x_i = c_i \in \mathbb{Z}$ or $y_i = d_i \in \mathbb{Z}$ a *fixation*. We call any point of $\mathbb{R}^{2q}$ determined by m bishops equations and $2q - m$ fixations a *lattice vertex*. (The term "fixation" was used in Part IV [7] only for a boundary equation; here it means any equation that fixes one coordinate at an integral value.)

The first step is to find the dimension of a bishops subspace. We do so by means of a signed graph $\Sigma_{\mathbb{B}}$ with node set $N := \{v_1, v_2, \dots, v_q\}$ corresponding to the bishops $\mathbb{B}_i$ and their plane coordinates $z_i = (x_i, y_i)$ and with edges corresponding to the bishops hyperplanes. For a hyperplane $\mathcal{H}^+_{ij}$ we have a *positive edge* e^+_{ij} and for a hyperplane $\mathcal{H}^-_{ij}$ we have a *negative edge* e^-_{ij}. Thus, $\Sigma_{\mathbb{B}}$ is a complete signed link graph: it has all possible edges (barring loops, of which we have no need) of both signs. For each bishops subspace $\mathcal{U}$ we have a spanning subgraph $\Sigma(\mathcal{U})$ whose edges correspond to the bishops hyperplanes that contain $\mathcal{U}$. (This is nothing other than the slope graph defined in [4, Section 3.3], except that it has extra nodes to make up a total of q.) Then $\mathcal{U}$ is the intersection of all the hyperplanes whose corresponding edges are in $\Sigma(\mathcal{U})$.

Lemma 4.2. *For any $\mathscr{S} \subseteq \mathscr{A}_{\mathbb{B}}$, with corresponding signed graph $\Sigma \subseteq \Sigma_{\mathbb{B}}$,*

$$\operatorname{codim} \bigcap \mathscr{S} = \operatorname{rk}(\Sigma^+) + \operatorname{rk}(\Sigma^-).$$

For a bishops subspace $\mathcal{U}$,

$$\dim \mathcal{U} = |\mathsf{A}(\Sigma(\mathcal{U}))| + |\mathsf{B}(\Sigma(\mathcal{U}))| \quad \textit{and} \quad \operatorname{codim} \mathcal{U} = \operatorname{rk}(\Sigma(\mathcal{U})^+) + \operatorname{rk}(\Sigma(\mathcal{U})^-).$$

Proof. We begin with $\mathscr{S}$ by looking at a single sign. Adjacent edges $e^\varepsilon_{ij}, e^\varepsilon_{jk}$ of sign ε in Σ, corresponding to $\mathcal{H}^\varepsilon_{ij}$ and $\mathcal{H}^\varepsilon_{jk}$, imply the third positive edge because the hyperplanes' equations imply that of $\mathcal{H}^\varepsilon_{ik}$. Consequently we may replace $E(\Sigma^\varepsilon)$ by a spanning tree of each ε-signed clique without changing the intersection subspace. Call the revised graph Σ'. Being irredundant, it has $2q - (|\mathsf{A}(\Sigma)| + |\mathsf{B}(\Sigma)|)$ edges by Lemma 3.1. As each hyperplane reduces the dimension of the intersection by at most 1, we conclude that $\operatorname{codim} \bigcap \mathscr{S} \leq 2q - (|\mathsf{A}(\Sigma)| + |\mathsf{B}(\Sigma)|)$.

On the other hand it is clear that $\mathscr{A}_{\mathbb{B}}$ intersects in the subspace $\{(z, z, \dots, z) : z \in \mathbb{R}^2\}$; thus, $2q - 2 = \operatorname{codim} \bigcap \mathscr{A}_{\mathbb{B}}$. The corresponding signed graph $\Sigma_{\mathbb{B}}$, when reduced to irredundancy, consists of a spanning tree of each sign; in other words, it has $2q - 2$ edges. One can choose the irredundant reduction of $\Sigma_{\mathbb{B}}$ to contain Σ'; it follows that every hyperplane of $\mathscr{S}$ must reduce the dimension of the intersection by exactly 1 in order for the reduced $\Sigma_{\mathbb{B}}$ to correspond to a 2-dimensional subspace of $\mathbb{R}^2$. Therefore, $\operatorname{codim} \bigcap \mathscr{S} = |E(\Sigma')| = 2q - (|\mathsf{A}(\Sigma)| + |\mathsf{B}(\Sigma)|) = \operatorname{rk}(\Sigma^+) + \operatorname{rk}(\Sigma^-)$.

The dimension formula for $\mathcal{U}$ follows by taking $\mathscr{S} := \{H \in \mathscr{A}_{\mathbb{B}} : H \supseteq \mathcal{U}\}$. □

Defining the rank of an arrangement $\mathscr{A}$ of hyperplanes to be the codimension of its intersection yields a matroid whose ground set is $\mathscr{A}$. The matroid's rank function encodes the linear dependence structure of the bishops arrangement $\mathscr{A}_{\mathbb{B}}$. The complete graph of order q is K_q.

Proposition 4.3. *The matroid of the hyperplane arrangement $\mathscr{A}_{\mathbb{B}}$ is isomorphic to*

$$G(K_q) \oplus G(K_q).$$

Proof. The rank of $\mathscr{S} \subseteq \mathscr{A}_{\mathbb{B}}$, corresponding to $\Sigma \subseteq \Sigma_{\mathbb{B}}$, is the codimension of $\bigcap \mathscr{S}$, which by Lemma 4.2 equals $\operatorname{rk}(\Sigma^+) + \operatorname{rk}(\Sigma^-)$. The matroid this implies on $E(\Sigma_{\mathbb{B}})$ is the direct sum of $G(\Sigma_{\mathbb{B}}^+)$ and $G(\Sigma_{\mathbb{B}}^-)$. Both $\Sigma_{\mathbb{B}}^+$ and $\Sigma_{\mathbb{B}}^-$ are unsigned complete graphs. The proposition follows. □

Now we return to the analysis of a lattice vertex $\mathbf{z}$. A point is *strictly half integral* if its coordinates have least common denominator 2; it is *weakly half integral* if its coordinates have least common denominator 1 or 2. A *weak half integer* is an element of $\frac{1}{2}\mathbb{Z}$; a *strict half integer* is a fraction that, in lowest terms, has denominator 2.

Lemma 4.4. *A point $\mathbf{z} = (z_1, z_2, \dots, z_q) \in \mathbb{R}^{2q}$, determined by a total of $2q$ bishops equations and fixations, is weakly half integral. Furthermore, in each z_i, either both coordinates are integers or both are strict half integers.*

Consequently, a vertex of the bishops' inside-out polytope $([0,1]^{2q}, \mathscr{A}_{\mathbb{B}})$ has each $z_i \in \{0,1\}^2$ or $z_i = (\frac{1}{2}, \frac{1}{2})$.

Proof. For the lattice vertex $\mathbf{z}$, find a bishops subspace $\mathcal{U}$ such that $\mathbf{z}$ is determined by membership in $\mathcal{U}$ together with $\dim \mathcal{U}$ fixations.

Suppose $v_i, v_j \in A_k$, a positive clique in $\Sigma(\mathcal{U})$; then $x_i - y_i = x_j - y_j$; thus, the value of $x_i - y_i$ is a constant a_k on A_k. Similarly, $x_i + y_i$ is a constant b_l on each negative clique B_l.

Now replace $\Sigma(\mathcal{U})$ by an irredundant subgraph Σ with the same positive and negative cliques. The edges of Σ within each clique are a tree. The total number of edges is $2q - (|\mathsf{A}(\Sigma(\mathcal{U}))| + |\mathsf{B}(\Sigma(\mathcal{U}))|)$; this is the number of bishops equations in the set determining $\mathbf{z}$. The remaining $|\mathsf{A}(\Sigma(\mathcal{U}))| + |\mathsf{B}(\Sigma(\mathcal{U}))|$ equations are fixations.

Write $C_{\mathcal{U}}$ for the clique graph $C(\Sigma) = C(\Sigma(\mathcal{U}))$. Let $\mp C_{\mathcal{U}}$ be the graph $C_{\mathcal{U}}$ with each edge v_i replaced by two edges called v_i^x and v_i^y. If we (arbitrarily) regard x as $-$ and y as $+$, this is a signed graph.

We defined a_k and b_l in terms of the x_i and y_i. We now reverse the viewpoint, treating the a's and b's as independent variables and the x's and y's as dependent variables. This is possible because, if A_k, B_l are the endpoints of v_i in $C_{\mathcal{U}}$, then $x_i - y_i = a_k$ and $x_i + y_i = b_l$, so

$$x_i = \frac{1}{2}(a_k + b_l) \quad \text{and} \quad y_i = \frac{1}{2}(-a_k + b_l);$$

in matrix form,

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \mathsf{H}(-C_{\mathcal{U}})^{\mathrm{T}} \\ \mathsf{H}(+C_{\mathcal{U}})^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \frac{1}{2} \mathsf{H}(\mp C_{\mathcal{U}})^{\mathrm{T}} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \tag{4.2}$$

where $\mathbf{x} = \big[x_i\big]_{i=1}^{q}$, $\mathbf{y} = \big[y_j\big]_{j=1}^{q}$, $\mathbf{a} = \big[a_k\big]_{k=1}^{|\mathsf{A}(\Sigma(\mathcal{U}))|}$, and $\mathbf{b} = \big[b_l\big]_{l=1}^{|\mathsf{B}(\Sigma(\mathcal{U}))|}$ are column vectors and $\mathsf{H}(\varepsilon C_{\mathcal{U}})$ is the incidence matrix of $C_{\mathcal{U}}$ with, respectively, all edges positive for

$\varepsilon = +$ (with $-$ and $+$ at the A and B ends) and all edges negative for $\varepsilon = -$ (with $+$ at both ends). Thus, the first coefficient matrix is the transposed incidence matrix of $\mp C_{\mathcal{U}}$ written with a particular ordering of the edges.

Fixing a total of $|\mathsf{A}(\Sigma(\mathcal{U}))| + |\mathsf{B}(\Sigma(\mathcal{U}))|$ variables $x_{i_1}, \dots$ and $y_{j_1}, \dots$ should determine all the values $x_1, y_1, \dots, x_q, y_q$. The fixations of $\mathbf{z}$ correspond to edges in $\mp C_{\mathcal{U}}$ so we may treat a choice of fixations as a choice of edges of $\mp C_{\mathcal{U}}$, where fixing x_i or y_i corresponds to choosing the edge v_i^x or v_i^y. We need to know what kind of edge set the fixations correspond to. Let $\Psi_{\mathbf{z}}$ denote the spanning subgraph of $\mp C_{\mathcal{U}}$ whose edges are the chosen edges. The fixation equations can be written in matrix form as

$$M^{\mathrm{T}} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = 2 \begin{bmatrix} x_{i_1} \\ \vdots \\ y_{j_1} \\ \vdots \end{bmatrix} = 2 \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}, \tag{4.3}$$

where the fixation edges are $v_{i_1}^x, \dots$ with endpoints $A_{k_1}, B_{l_1}, \dots$ and $v_{j_1}^y, \dots$ with endpoints $A_{k_1'}, B_{l_1'}, \dots$; the fixations are $x_{i_r} = c_r$ and $y_{j_s} = d_s$; $\mathbf{c} = \big[c_r\big]_{r=1}^{\bar{r}}$ and $\mathbf{d} = \big[d_s\big]_{s=1}^{\bar{s}}$ are column vectors (with $\bar{r} + \bar{s} = |\mathsf{A}(\Sigma(\mathcal{U}))| + |\mathsf{B}(\Sigma(\mathcal{U}))|$, the total number of fixations); and M is an $(|\mathsf{A}(\Sigma(\mathcal{U}))| + |\mathsf{B}(\Sigma(\mathcal{U}))|) \times (|\mathsf{A}(\Sigma(\mathcal{U}))| + |\mathsf{B}(\Sigma(\mathcal{U}))|)$ matrix representing the relationships between the a's and b's and the fixed variables:

$$M := \begin{array}{cc} \begin{array}{cccc} x_{i_1} & & y_{j_1} & \end{array} & \\ \left[\begin{array}{cccc} 1 & \cdots & 0 & \cdots \\ \vdots & \ddots & \vdots & \ddots \\ 0 & \cdots & -1 & \cdots \\ \vdots & \ddots & \vdots & \ddots \\ \hline 1 & \cdots & 0 & \cdots \\ \vdots & \ddots & \vdots & \ddots \\ 0 & \cdots & 1 & \cdots \\ \vdots & \ddots & \vdots & \ddots \end{array}\right] & \begin{array}{l} \\ \mathsf{A}(\Sigma(\mathcal{U})) \\ \\ \\ \\ \mathsf{B}(\Sigma(\mathcal{U})) \\ \\ \\ \end{array} \end{array}.$$

The rows of M are indexed by the signed cliques and the columns are indexed by the fixations. The column of a fixation involving a node v_i, whose endpoints in $C_{\mathcal{U}}$ are A_k and B_l, has exactly two nonzero entries, one in row A_k and one in row B_l, whose values are, respectively, $1, 1$ for an x-fixation and $-1, 1$ for a y-fixation. Thus, each column has exactly two nonzero elements, each of which is ± 1.

Consequently, M is the incidence matrix of a signed graph, in fact, $M = \mathsf{H}(\Psi_{\mathbf{z}})$. M must be nonsingular since the fixed x's and y's uniquely determine the a's and b's (because they determine $\mathbf{z}$). It follows (see Section 3) that the fixation equations for $\mathbf{z}$ are a set corresponding to a spanning 1-forest in $\mp C_{\mathcal{U}}$ in which every circle is negative. This 1-forest is $\Psi_{\mathbf{z}}$. There is choice in the selection of $\Psi_{\mathbf{z}}$ but it is not completely arbitrary. Let $J_{\mathbf{z}}$ be the set of nodes v_i such that z_i is integral; consider $J_{\mathbf{z}}$ as a subset of $E(C_{\mathcal{U}})$. As fixations must be integral, $E(\Psi_{\mathbf{z}})$ must be a subset of $\pm J_{\mathbf{z}}$. As fixations are arbitrary integers, $\Psi_{\mathbf{z}}$ may be any spanning 1-forest of $\mp C_{\mathcal{U}}$ that is contained in $\pm J_{\mathbf{z}}$ and whose circles are negative. Thus we have found the graphical form of the equations of a lattice vertex.

Example 4.5. For an example, suppose there are three positive and four negative cliques, so $\mathsf{A}(\Sigma(\mathcal{U})) = \{A_1, A_2, A_3\}$ and $\mathsf{B}(\Sigma(\mathcal{U})) = \{B_1, B_2, B_3, B_4\}$, and eight nodes, $N = \{v_1, \dots, v_8\}$, with the clique graph $C_{\mathcal{U}}$ shown in Figure 1.

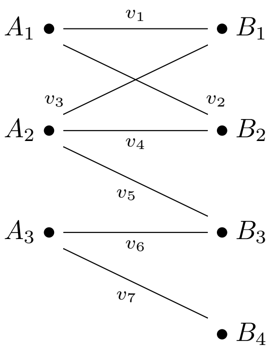


Figure 1: The clique graph $C_{\mathcal{U}}$.

An example of a suitable 1-forest $\Psi_{\mathbf{z}} \subseteq \mp C_{\mathcal{U}}$ is shown in Figure 2. It corresponds to fixations

$$x_1 = c_1, \quad y_2 = d_1, \quad x_3 = c_2, \quad x_4 = c_3, \quad y_5 = d_2, \quad x_7 = c_4, \quad y_7 = d_3.$$

The incidence matrix is

$$M := \mathsf{H}(\Psi_{\mathbf{z}}) = \begin{array}{c} \begin{array}{ccccccc} x_1 & x_3 & x_4 & x_7 & y_2 & y_5 & y_7 \end{array} \\ \left[\begin{array}{ccccccc} 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ \hline 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \end{array} \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ B_1 \\ B_2 \\ B_3 \\ B_4 \end{array} .$$

Every column has two nonzeros. The equations of the fixations in matrix form are

$$M^{\mathrm{T}} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = 2 \begin{bmatrix} x_1 \\ x_3 \\ x_4 \\ x_7 \\ y_2 \\ y_5 \\ y_7 \end{bmatrix} = 2 \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ d_1 \\ d_2 \\ d_3 \end{bmatrix},$$

where the c_i's and d_j's are any integers we wish in the lemma (but in the application to

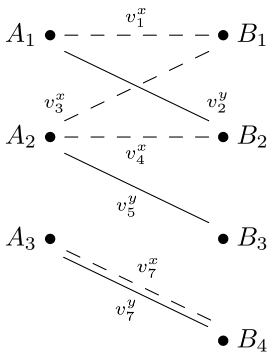


Figure 2: A suitable 1-forest.

Theorem 1.1 they will be 0's and 1's). The solution is

$$\begin{aligned}
a_1 &= x_1 - x_3 + x_4 - y_2 = c_1 - c_2 + c_3 - d_1,\\
a_2 &= -x_1 + x_3 + x_4 - y_2 = -c_1 + c_2 + c_3 - d_1,\\
a_3 &= x_7 - y_7 = c_4 - d_3,\\
b_1 &= x_1 + x_3 - x_4 + y_2 = c_1 + c_2 - c_3 + d_1,\\
b_2 &= x_1 - x_3 + x_4 + y_2 = c_1 - c_2 + c_3 + d_1,\\
b_3 &= -x_1 + x_3 + x_4 - y_2 + 2y_5 = -c_1 + c_2 + c_3 - d_1 + 2d_2,\\
b_4 &= x_7 + y_7 = c_4 + d_3,
\end{aligned}$$

and the unfixed variables are

$$\begin{aligned}
x_2 &= \frac{a_1 + b_2}{2} = c_1 - c_2 + c_3,\\
x_5 &= \frac{a_2 + b_3}{2} = -c_1 + c_2 + c_3 - d_1 + d_2,\\
x_6 &= \frac{a_3 + b_3}{2} = \frac{-c_1 + c_2 + c_3 + c_4 - d_1 + 2d_2 - d_3}{2},\\
y_1 &= \frac{-a_1 + b_1}{2} = c_2 - c_3 + d_1,\\
y_3 &= \frac{-a_2 + b_1}{2} = c_1 - c_3 + d_1,\\
y_4 &= \frac{-a_2 + b_2}{2} = c_1 - c_2 + d_1,\\
y_6 &= \frac{-a_3 + b_3}{2} = \frac{-c_1 + c_2 + c_3 - c_4 - d_1 + 2d_2 + d_3}{2}.
\end{aligned}$$

Observe that x_6 and y_6 are the only possibly fractional coordinates and their difference, $x_6 - y_6 = a_3 = c_4 - d_3$, is integral; therefore, either z_6 is integral, or both x_6 and y_6 are half integers and $z_6 = (\frac{1}{2}, \frac{1}{2})$ if $\mathbf{z} \in [0,1]^{2q}$.

We are now prepared to prove Lemma 4.4. We need a result from (e.g.) [9], which can be stated:

Lemma 4.6. *The solution of a linear system with integral constant terms, whose coefficient matrix is the transpose of a nonsingular signed-graph incidence matrix, is weakly half-integral.*

Proof. The way in which this is contained in [9] is explained in [1, p. 197]. □

Since M is the incidence matrix of a signed graph, and since the constant terms in Equation (4.3), being twice the fixed values, are even integers, the a's and b's are integers by Lemma 4.6. The remaining x's and y's are halves of sums or differences of a's and b's, so they are weak half-integers. The exact formula is obtained by substituting Equation (4.3) into Equation (4.2):

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} = \mathsf{H}(\mp C_{\mathcal{U}})^{\mathrm{T}}(M^{-1})^{\mathrm{T}}\begin{bmatrix}\mathbf{c}\\ \mathbf{d}\end{bmatrix}. \tag{4.4}$$

□

Theorem 1.1 is an immediate corollary of Lemma 4.4.

5 Open questions

5.1 Coefficient periods

We proved that $\gamma_6(n)$ is the first coefficient that depends on n, having period 2. We guess that every coefficient after $\gamma_6(n)$ also has period 2.

5.2 Subspace structure

We have not been able to find a complete formula for all q. By our method, that would need a general structural analysis of all subspaces, which is too complicated for now. We propose the following problem: Give a complete description of all subspaces, for all q, in terms of signed graphs. That is, we ask for the slope matroid (see [4, Section 7.3]). The signed-graphic frame matroid $G(\Sigma)$ ([13, Theorem 5.1], corrected and generalized in [15, Theorem 2.1]), while simpler than the slope matroid, perhaps could help find a description of the latter.

5.3 Similar two-move riders

The slope matroid for the bishop is simple compared to those for other riders. We wonder if riders with two slopes that are related by negation (that is, the basic moves are symmetrical under reflection in an axis), or negation and inversion (that is, the basic moves are perpendicular), may be amenable to an analysis that uses the bishops analysis as a guide.

5.4 Other two-move riders

We expect that finding formulas for any rider with only two basic moves is intrinsically easier than for riders with more than two and can be done for all such riders in a comprehensive though complicated manner.

Dictionary of notation

$b(\Sigma)$	# of signed-graph components with no negative circle (p. 552)
$c(\Gamma), c(\Sigma)$	# of components of a graph (p. 552)
$c(\Sigma^{\pm})$	# of positive or negative cliques (p. 553)
d/c	slope of a line or move (p. 551)
(c, d)	coordinates of a move vector m (p. 551)
c_i, d_i	fixation equation constants (p. 554)
e	edge of a (signed) graph (p. 552)
e^{ε}_{ij}	edge of a signed graph with sign ε and end nodes v_i, v_j (p. 554)
$g(\Sigma)$	function on a signed graph (p. 553)
k, l	indices in the clique graph (p. 555)
$m = (c, d)$	basic move (p. 551)
n	size of a square board
$o_{\mathbb{B}}(q; n)$	# of nonattacking labelled configurations (p. 551)
p	period of a quasipolynomial (p. 550)
q	# of pieces on a board (p. 550)
q	# of nodes in a (signed) graph (p. 552)
r, s	indices of fixations (p. 556)
$u_{\mathbb{B}}(q; n)$	# of nonattacking unlabelled configurations (p. 550)
v	node in a signed graph (p. 552)
$z = (x, y)$, $z_i = (x_i, y_i)$	piece positions (p. 551)
$\mathbf{a}$, $\mathbf{b}$	clique vectors (p. 555)
$\mathbf{c}$, $\mathbf{d}$	fixation vectors (p. 556)
$\mathbf{x}$, $\mathbf{y}$	x, y coordinate vectors of a configuration (p. 555)
$\mathbf{z} = (z_1, \dots, z_q)$	a configuration in $\mathbb{R}^{2q}$ (p. 554)
$\gamma_i(n)$	coefficient of $u_{\mathbb{B}}$ (p. 551)
ε	sign of an edge (p. 554)
ξ	cyclomatic number (p. 552)
σ	sign function of the signed graph Σ (p. 552)
rk	rank of the incidence matrix of a (signed) graph (p. 552)
A_k, B_l	positive, negative cliques (p. 552)
$C(\Sigma)$	clique graph (p. 553)
$C_{\mathcal{U}} = C(\Sigma(\mathcal{U}))$	clique graph (p. 555)
E	edge set of a graph (p. 552)
G	matroid on ground set E (p. 552)
$J_{\mathbf{z}}$	set of vertices z_i in the configuration $\mathbf{z}$ such that z_i is integral (p. 556)
K_q	complete graph (p. 555)
M	incidence matrix $\mathsf{H}(\Psi_{\mathbf{z}})$ (p. 556)
N	node set of a graph (p. 552)
$\mathscr{A}_{\mathbb{B}}$	move arrangement of bishops $\mathbb{B}$ (p. 551)
$\mathcal{B}$	board polygon $[0, 1]^q$ (p. 551)
$\mathcal{H}^{\pm}_{ij}$	bishops hyperplane (p. 550)
$(\mathcal{P}, \mathscr{A})$	inside-out polytope (p. 551)
$\mathscr{S}$	subarrangement (p. 554)
$\mathcal{U}$	subspace in the intersection lattice of an arrangement (p. 554)

$\mathbb{R}$	real numbers
$\mathbb{Z}$	integers
$\mathbb{B}$	bishop (p. 551)
$\mathsf{A}(\Sigma), \mathsf{B}(\Sigma)$	sets of positive, negative cliques (p. 552)
Γ	graph (p. 552)
H	incidence matrix (read "Eta") of a (signed) graph (p. 552)
Σ	signed graph (p. 552)
$\Sigma(\mathcal{U})$	signed graph of the bishops subspace $\mathcal{U}$ (p. 554)
$\Psi_{\mathbf{z}}$	subgraph for a vertex $\mathbf{z}$ (p. 556)

References

[1] G. Appa and B. Kotnyek, A bidirected generalization of network matrices, *Networks* **47** (2006), 185–198, doi:10.1002/net.20108.

[2] S. E. Arshon, Resheniye odnoy kombinatornoy zadachi, *Mat. Prosveshchenie Ser. 1* **8** (1936), 24–29, `http://mi.mathnet.ru/eng/mp694`.

[3] M. Beck and T. Zaslavsky, Inside-out polytopes, *Adv. Math.* **205** (2006), 134–162, doi:10.1016/j.aim.2005.07.006.

[4] S. Chaiken, C. R. H. Hanusa and T. Zaslavsky, A q-queens problem. I. General theory, *Electron. J. Combin.* **21** (2014), #P3.33 (28 pages), `https://www.combinatorics.org/ojs/index.php/eljc/article/view/v21i3p33`.

[5] S. Chaiken, C. R. H. Hanusa and T. Zaslavsky, A q-queens problem. II. The square board, *J. Algebraic Combin.* **41** (2015), 619–642, doi:10.1007/s10801-014-0547-0.

[6] S. Chaiken, C. R. H. Hanusa and T. Zaslavsky, A q-queens problem. III. Nonattacking partial queens, submitted, `arXiv:1402.4886` [math.CO].

[7] S. Chaiken, C. R. H. Hanusa and T. Zaslavsky, A q-queens problem. IV. Attacking configurations and their denominators, submitted, `arXiv:1807.04741` [math.CO].

[8] S. Chaiken, C. R. H. Hanusa and T. Zaslavsky, A q-queens problem. V. Some of our favorite pieces: Queens, bishops, rooks, and nightriders, under revision, `arXiv:1609.00853` [math.CO].

[9] D. S. Hochbaum, N. Megiddo, J. Naor and A. Tamir, Tight bounds and 2-approximation algorithms for integer programs with two variables per inequality, *Math. Programming Ser. B* **62** (1993), 69–83, doi:10.1007/bf01585160.

[10] V. Kotěšovec, *Non-Attacking Chess Pieces*, self-published online book, 3rd edition, 2011, `http://www.kotesovec.cz/math.htm`.

[11] V. Kotěšovec, *Non-Attacking Chess Pieces*, self-published online book, 6th edition, 2013, `http://www.kotesovec.cz/math.htm`.

[12] R. P. Stanley, *Enumerative Combinatorics, Volume 1*, volume 49 of *Cambridge Studies in Advanced Mathematics*, Cambridge University Press, Cambridge, 2nd edition, 2012.

[13] T. Zaslavsky, Signed graphs, *Discrete Appl. Math.* **4** (1982), 47–74, doi:10.1016/0166-218x(82)90033-6.

[14] T. Zaslavsky, Erratum: "Signed graphs", *Discrete Appl. Math.* **5** (1983), 248, doi:10.1016/0166-218x(83)90047-1.

[15] T. Zaslavsky, Biased graphs. II. The three matroids, *J. Comb. Theory Ser. B* **51** (1991), 46–72, doi:10.1016/0095-8956(91)90005-5.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 563–583
https://doi.org/10.26493/1855-3974.793.2a9
(Also available at http://amc-journal.eu)

Classification of some reflexible edge-transitive embeddings of complete bipartite graphs

Jin Ho Kwak
Mathematics, Beijing Jiaotong University, Beijing, 100044, P.R. China

Young Soo Kwon
Mathematics, Yeungnam University, Kyeongsan, 712-749 Republic of Korea

Received 8 January 2015, accepted 28 February 2019, published online 21 April 2019

Abstract

In this paper, we classify some reflexible edge-transitive orientable embeddings of complete bipartite graphs. As a by-product, we classify groups Γ such that (i) $\Gamma = XY$ for some cyclic groups $X = \langle x\rangle$ and $Y = \langle y\rangle$ with $X \cap Y = \{1_\Gamma\}$ and (ii) there exists an automorphism of Γ which sends x and y to x^{-1} and y^{-1}, respectively.

Keywords: Complete bipartite graphs, reflexible edge-transitive embedding.

Math. Subj. Class.: 05C10, 05C30

1 Preliminaries

A *map* is a 2-cell embedding of a graph G in a compact, connected surface. A map is called *orientable* or *nonorientable* according to whether the supporting surface is orientable or nonorientable. In this paper, we only consider orientable maps.

For a simple connected graph G, an *arc* of G is an ordered pair (u, v) of adjacent vertices in G. The set of all arcs in G is denoted by $D(G)$. An orientable map $\mathcal{M}$ can be described by a pair $(G; R)$, where G is the underlying graph of $\mathcal{M}$ and R is a permutation of the arc set $D(G)$ whose orbits coincide with the sets of arcs emanating from the same vertex. The permutation R is called the *rotation* of the map $\mathcal{M}$.

For given two maps $\mathcal{M}_1 = (G_1; R_1)$ and $\mathcal{M}_2 = (G_2; R_2)$, a *map isomorphism* $\phi\colon \mathcal{M}_1 \to \mathcal{M}_2$ is a graph isomorphism $\phi\colon G_1 \to G_2$ such that $\phi R_1(u, v) = R_2\phi(u, v)$ for any arc (u, v) in G_1. Furthermore if $\mathcal{M}_1 = \mathcal{M}_2 = \mathcal{M}$, ϕ is called a *map automorphism* of $\mathcal{M}$. The set of all map automorphisms of $\mathcal{M}$ denoted by $\mathrm{Aut}(\mathcal{M})$ is a group under the composition operation, and it is called the *automorphism group* of $\mathcal{M}$. For a map $\mathcal{M} = (G; R)$,

E-mail addresses: jinkwak@postech.ac.kr (Jin Ho Kwak), ysookwon@ynu.ac.kr (Young Soo Kwon) the group $\mathrm{Aut}(\mathcal{M})$ acts semi-regularly on the arc set $D(G)$, so $|\mathrm{Aut}(\mathcal{M})| \le 2|E(G)|$. If this bound is attained, then $\mathrm{Aut}(\mathcal{M})$ acts regularly on the arc set, and the map is called a *regular map* or a *regular embedding*. The map $\mathcal{M}$ is said to be *vertex-transitive* or *edge-transitive* if $\mathrm{Aut}(\mathcal{M})$ acts transitively on $V(G)$ or $E(G)$, respectively. For an orientable embedding $\mathcal{M}$ of a bipartite graph G, if the set of partite set preserving map automorphisms acts transitively on $E(G)$ then we call $\mathcal{M}$ an *edge-transitive map* or an *edge-transitive embedding* satisfying the Property (P) in this paper. For a map $\mathcal{M} = (G; R)$, if $\mathcal{M}$ and $\mathcal{M}^{-1} = (G; R^{-1})$ are isomorphic, $\mathcal{M}$ is called *reflexible*.

Classifying highly symmetric embeddings of graphs in a given class is an interesting problem in topological graph theory. In recent years, there has been particular interest in the regular embeddings of complete bipartite graphs $K_{n,n}$ by several authors [1, 2, 4, 5, 6, 7, 8, 10]. The reflexible regular embeddings and self-Petrie dual regular embeddings of $K_{n,n}$ have been classified by the authors [7]. Recently, G. Jones has completed the classification of regular embeddings of $K_{n,n}$ [5] and the authors have classified nonorientable regular embeddings of $K_{n,n}$ [8]. In [3], Graver and Watkins classified edge-transitive maps on closed surfaces into fourteen types. In this paper, we classify reflexible edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P) which correspond to types 1 or 2 among 14 types. The following theorem is the main result in this paper.

Theorem 1.1. *For any integers*

$$\begin{aligned} m &= 2^a p_1^{a_1} \cdots p_\ell^{a_\ell} p_{\ell+1}^{a_{\ell+1}} \cdots p_{\ell+f}^{a_{\ell+f}} \quad \textit{and} \\ n &= 2^b p_1^{b_1} \cdots p_\ell^{b_\ell} q_{\ell+1}^{a_{\ell+1}} \cdots q_{\ell+g}^{b_{\ell+g}} \quad \textit{(prime decompositions)} \end{aligned}$$

with $\gcd(m,n) = 2^c p_1^{c_1} \cdots p_\ell^{c_\ell}$ *and* $a \le b$*, the number (up to isomorphism) of reflexible edge-transitive embeddings of* $K_{m,n}$ *satisfying the Property (P) is* 1 *if both* m *and* n *are odd;* $2^f(1+p_1^{c_1})\cdots(1+p_\ell^{c_\ell})$ *if exactly one of* m *and* n *is even, namely, only* n *is even;* $A(a,b)2^{f+g+\ell}(1+p_1^{c_1})\cdots(1+p_\ell^{c_\ell})$ *if both* m *and* n *are even, where*

$$A(a,b) = \begin{cases} 1 & \textit{if } (a,b) = (1,1), \\ 2 & \textit{if } (a,b) = (1,2), \\ 4 & \textit{if } (a,b) = (2,2) \textit{ or } (1,k) \textit{ with } k \ge 3, \\ 10 & \textit{if } (a,b) = (2,3), \\ 12 & \textit{if } (a,b) = (2,k) \textit{ with } k \ge 4, \\ 28 & \textit{if } (a,b) = (3,3), \\ 40 & \textit{if } (a,b) = (3,4), \\ 36 & \textit{if } (a,b) = (3,k) \textit{ with } k \ge 5, \\ 20(1+2^{a-2}) & \textit{if } a = b \ge 4, \\ 20 + 18\cdot 2^{a-2} & \textit{if } b-1 = a \ge 4, \\ 20 + 16\cdot 2^{a-2} & \textit{if } b-2 \ge a \ge 4. \end{cases}$$

Our paper is organized as follows. In the next section, we consider some relations between edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P) and products of two cyclic groups. In Section 3, we classify reflexible edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P) when at least one of m and n is odd. In Section 4, for even integers m and n, the classification of reflexible edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P) is given. In the final section, we classify groups Γ satisfying the conditions:

(i) $\Gamma = XY$ for some cyclic groups $X = \langle x\rangle$ and $Y = \langle y\rangle$ with $X \cap Y = \{1_\Gamma\}$ and

(ii) there exists an automorphism of Γ which sends x and y to x^{-1} and y^{-1}.

2 (m, n)-bicyclic triples in $\mathrm{Aut}(K_{m,n})$

Regular embeddings of the complete bipartite graphs $K_{n,n}$ are related to groups Γ with two generators satisfying some conditions [4]. Using this relation, G. Jones classify regular embeddings of $K_{n,n}$ [5]. Similarly, we aim to find a relation between edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P) and groups with two generators satisfying some conditions in this section.

In [4], G. Jones et al. showed that any finite group Γ is isomorphic to $\mathrm{Aut}(\mathcal{M})$ for some regular embedding $\mathcal{M}$ of $K_{n,n}$ if and only if Γ has cyclic subgroups $X = \langle x\rangle$ and $Y = \langle y\rangle$ of order n such that:

(i) $\Gamma = XY$

(ii) $X \cap Y = \{1_\Gamma\}$ and

(iii) there is an automorphism α of Γ transposing x and y.

They call the triple (Γ, x, y) satisfying these conditions the *n-isobicyclic* triple. In this relation, x and y correspond to rotations of $\mathcal{M}$ around two fixed adjacent vertices u and v, respectively. The automorphism α corresponds to the half-turn reversing the edge uv. For two n-isobicyclic triples (Γ_1, x_1, y_1) and (Γ_2, x_2, y_2), two corresponding regular embeddings $\mathcal{M}_1$ and $\mathcal{M}_2$ are isomorphic if and only if there exists a group isomorphism from Γ_1 to Γ_2 given by $x_1 \mapsto x_2$ and $y_1 \mapsto y_2$. Using this, one can show that the regular embedding $\mathcal{M}$ induced by n-isobicyclic triple (Γ, x, y) is reflexible if and only if there exists an automorphism β of Γ which sends x and y to x^{-1} and y^{-1}, respectively. (For more information, the reader is referred to [4].)

Note that one can define an embedding of $K_{n,n}$ by using the first and second conditions of n-isobicyclic triple, and the induced map is edge-transitive map satisfying the Property (P) even though the third condition of n-isobicyclic triple is not satisfied. Conversely, any edge transitive embedding of $K_{n,n}$ satisfying the Property (P) is isomorphic to some induced map by such a triple (Γ, x, y). One can show that for different positive integers m and n, an edge-transitive embedding of $K_{m,n}$ satisfying the Property (P) can also be represented by a similar triple. For a group Γ containing cyclic subgroups $X = \langle x\rangle$ of order n and $Y = \langle y\rangle$ of order m, the triple (Γ, x, y) is called *(m, n)-bicyclic* if it satisfies:

(i) $\Gamma = XY$ and

(ii) $X \cap Y = \{1_\Gamma\}$.

For any (m, n)-bicyclic triple (Γ, x, y), one can define an embedding of $K_{m,n}$ by a similar way to define an embedding of $K_{n,n}$ using n-isobicyclic triple. We denote this embedding by $\mathcal{M}(\Gamma, x, y)$. One can see that $\mathcal{M}(\Gamma, x, y)$ is an edge-transitive embedding of $K_{m,n}$ satisfying the Property (P). Furthermore the following result holds.

Lemma 2.1 ([9]). *Let m, n be two positive integers (not necessarily distinct).*

(1) Any edge-transitive embedding $\mathcal{M}$ of $K_{m,n}$ satisfying the Property (P) is isomorphic to $\mathcal{M}(\Gamma, x, y)$ for some (m, n)-bicyclic triple (Γ, x, y).

(2) *For two (m,n)-bicyclic triples (Γ_1, x_1, y_1) and (Γ_2, x_2, y_2), two edge-transitive embeddings $\mathcal{M}(\Gamma_1, x_1, y_1)$ and $\mathcal{M}(\Gamma_2, x_2, y_2)$ are isomorphic if and only if there exists a group isomorphism from Γ_1 to Γ_2 given by $x_1 \mapsto x_2$ and $y_1 \mapsto y_2$.*

For any (m,n)-bicyclic triple (Γ, x, y), there exists a subgroup H of the automorphism group $\operatorname{Aut}(K_{m,n})$ such that:

(i) H is isomorphic to Γ and

(ii) x and y in Γ correspond to elements in H which cyclically permute vertices in the partite sets of size n and m, respectively.

Hence it suffices to deal with such (m,n)-bicyclic triples in $\operatorname{Aut}(K_{m,n})$ to classify edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P).

For any positive integer m, denote the set $\{0, 1, \ldots, m-1\}$ by $[m]$. Let

$$V = \{0, 1, \ldots, (m-1)\} \cup \{0', 1', \ldots, (n-1)'\} = [m] \cup [n]'$$

be the vertex set of $K_{m,n}$ as partite sets, and let

$$D = \{(i, j'), (j', i) : 0 \le i \le m-1 \text{ and } 0 \le j \le n-1\}$$

be the arc set, where (i, j') is the arc emanating from i to j' and (j', i) denotes its inverse. We denote the symmetric group on $[m]$ and $[n]'$ by S and S', respectively. Let S_0 and S'_0 be their stabilizers of 0 and $0'$, respectively. Note that $\operatorname{Aut}(K_{m,n})$ is isomorphic to $S \times S'$ when $m \neq n$; $S \wr \mathbb{Z}_2$ when $m = n$. We identify integers $0, 1, 2, \ldots$ with their residue classes modulo m or n according to the context.

Let (Γ, x, y) be an (m,n)-bicyclic triple such that Γ is a subgroup of $\operatorname{Aut}(K_{m,n})$. Now there exists an automorphism $\phi \in \operatorname{Aut}(K_{m,n})$ such that

$$x^\phi = \phi^{-1} x \phi = \alpha(0'\ 1'\ \cdots\ (n-1)') \quad \text{and} \quad y^\phi = \phi^{-1} y \phi = \beta(0\ 1\ \cdots\ m-1),$$

where $\alpha \in S_0$ and $\beta \in S'_0$. For any $\alpha \in S_0$ and $\beta \in S'_0$, let

$$x_\alpha = \alpha(0'\ 1'\ \cdots\ (n-1)') \quad \text{and} \quad y_\beta = \beta(0\ 1\ \cdots\ m-1).$$

From now on, we only consider triples $(\langle x_\alpha, y_\beta \rangle, x_\alpha, y_\beta)$ as candidates of (m,n)-bicyclic triples.

Lemma 2.2 ([9]). *For any $\alpha \in S_0$ and $\beta \in S'_0$,*

1. *the group $\langle x_\alpha, y_\beta \rangle$ acts transitively on the edge set of $K_{m,n}$ and*
2. *the triple $(\langle x_\alpha, y_\beta \rangle, x_\alpha, y_\beta)$ is (m,n)-bicyclic if and only if $|\langle x_\alpha, y_\beta \rangle| = mn$.*

By Lemma 2.2, we need to characterize $\alpha \in S_0$ and $\beta \in S'_0$ satisfying $|\langle x_\alpha, y_\beta \rangle| = mn$ to classify edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P). To do this, we denote

$$\mathrm{ET}_{m,n} = \{(\alpha, \beta) : \alpha \in S_0,\ \beta \in S'_0 \text{ and } |\langle x_\alpha, y_\beta \rangle| = mn\}.$$

Note that for any $(\alpha, \beta) \in \mathrm{ET}_{m,n}$, $(\langle x_\alpha, y_\beta \rangle, x_\alpha, y_\beta)$ is an (m,n)-bicyclic triple and hence $\mathcal{M}(\langle x_\alpha, y_\beta \rangle, x_\alpha, y_\beta)$ is an edge-transitive embedding of $K_{m,n}$ satisfying the Property (P). Conversely for any edge-transitive embedding $\mathcal{M}$ of $K_{m,n}$ satisfying the Property (P), there exists $(\alpha, \beta) \in \mathrm{ET}_{m,n}$ such that $\mathcal{M}$ is isomorphic to $\mathcal{M}(\langle x_\alpha, y_\beta \rangle, x_\alpha, y_\beta)$.

Remark 2.3.

(1) For any $(\alpha, \beta) \in \mathrm{ET}_{m,n}$,

$$\langle x_\alpha, y_\beta\rangle = \{x_\alpha^i y_\beta^j \mid i \in [n],\ j \in [m]\} = \{y_\beta^j x_\alpha^i \mid i \in [n],\ j \in [m]\}.$$

Hence in many cases, if α satisfies some properties then β also satisfies the same properties and vice versa.

(2) Note that for different positive integers m and n and for an orientable embedding $\mathcal{M}$ of $K_{m,n}$, any automorphism of $\mathcal{M}$ is partite set preserving. Let $m = n$ be odd and let $\mathcal{M}$ be an orientable edge-transitive embedding of $K_{n,n}$. If a subgroup Γ of $\mathrm{Aut}(\mathcal{M})$ acts regularly on the edge set then $|\Gamma| = m^2$ is odd and hence there exists no partite set reversing element in Γ. Hence for odd n, every edge-transitive embedding of $K_{n,n}$ is an edge-transitive embedding of $K_{n,n}$ satisfying the Property (P). On the other hand, for even n, we do not know whether the above statement is true or not.

The next lemma shows that for different $(\alpha_1, \beta_1), (\alpha_2, \beta_2) \in \mathrm{ET}_{m,n}$, two induced edge-transitive embeddings are non-isomorphic.

Lemma 2.4 ([9])**.** *For any $(\alpha_1, \beta_1), (\alpha_2, \beta_2) \in \mathrm{ET}_{m,n}$, the induced edge-transitive embeddings $\mathcal{M}(\langle x_{\alpha_1}, y_{\beta_1}\rangle, x_{\alpha_1}, y_{\beta_1})$ and $\mathcal{M}(\langle x_{\alpha_2}, y_{\beta_2}\rangle, x_{\alpha_2}, y_{\beta_2})$ are isomorphic if and only if $(\alpha_1, \beta_1) = (\alpha_2, \beta_2)$.*

By Lemma 2.4, distinct pairs in $\mathrm{ET}_{m,n}$ give non-isomorphic edge-transitive embeddings of $K_{m,n}$ and the number of edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P) equals to the cardinality $|\,\mathrm{ET}_{m,n}\,|$. But for distinct pairs $(\alpha_1, \beta_1), (\alpha_2, \beta_2) \in \mathrm{ET}_{m,n}$, two groups $\langle x_{\alpha_1}, y_{\beta_1}\rangle$ and $\langle x_{\alpha_2}, y_{\beta_2}\rangle$ may possibly be isomorphic. We do not know a necessary and sufficient condition for $\langle x_{\alpha_1}, y_{\beta_1}\rangle \simeq \langle x_{\alpha_2}, y_{\beta_2}\rangle$. So we propose the following problem.

Problem 2.5. For any positive integers m and n and for any $(\alpha_1, \beta_1), (\alpha_2, \beta_2) \in \mathrm{ET}_{m,n}$, find a necessary and sufficient condition for $\langle x_{\alpha_1}, y_{\beta_1}\rangle \simeq \langle x_{\alpha_2}, y_{\beta_2}\rangle$.

From now on, we aim to characterize the set $\mathrm{ET}_{m,n}$. Note that for any $(\alpha, \beta) \in \mathrm{ET}_{m,n}$, the stabilizers $\langle x_\alpha, y_\beta\rangle_0$ and $\langle x_\alpha, y_\beta\rangle_{0'}$ are cyclic groups $\langle x_\alpha\rangle$ of order n and $\langle y_\beta\rangle$ of order m, respectively.

Lemma 2.6. *For any $(\alpha, \beta) \in \mathrm{ET}_{m,n}$, $\langle\alpha\rangle$ and $\langle\beta\rangle$ are cyclic groups of order $|\{\alpha^i(1) : i \in [n]\}|$ and $|\{\beta^i(1') : i \in [m]\}|$, the lengths of the orbit containing 1 and $1'$, respectively. Furthermore they are divisors of n and m, respectively.*

Proof. Let $d_1 = |\{\alpha^i(1) : i \in [n]\}|$ and $d_2 = |\{\beta^i(1') : i \in [m]\}|$. Now d_1 and d_2 are divisors of the orders $|\langle x_\alpha\rangle| = n$ and $|\langle y_\beta\rangle| = m$, respectively. Note that

$$\alpha^{d_1}(1) = 1 \quad \text{and} \quad y_\beta^{-1} x_\alpha^{d_1} y_\beta(0) = 0,$$

which implies that, as a conjugate of $x_\alpha^{d_1}$, $y_\beta^{-1} x_\alpha^{d_1} y_\beta$ belongs to the vertex stabilizer $\langle x_\alpha, y_\beta\rangle_0 = \langle x_\alpha\rangle$. Since d_1 is a divisor of n, $y_\beta^{-1} x_\alpha^{d_1} y_\beta = x_\alpha^{rd_1}$ for some $r \in [n]$ such that $\gcd(r, \frac{n}{d_1}) = 1$, where $\gcd(r, \frac{n}{d_1})$ is the greatest common divisor of r and $\frac{n}{d_1}$. Now, suppose to the contrary that $|\langle\alpha\rangle| \neq d_1$. Then there exists $k \in [m]$ such that $\alpha^{d_1}(k) \neq k$. Let q be the largest element in $[m]$ such that $\alpha^{d_1}(q) \neq q$. On the other hand,

$$\alpha^{rd_1}(q) = x_\alpha^{rd_1}(q) = y_\beta^{-1} x_\alpha^{d_1} y_\beta(q) = y_\beta^{-1} x_\alpha^{d_1}(q+1) = y_\beta^{-1}(q+1) = q,$$

contradictory to $\alpha^{rd_1}(q) \neq q$. Therefore $|\langle\alpha\rangle| = d_1$. Similarly, one can show that $|\langle\beta\rangle| = d_2$. □

For any $(\alpha, \beta) \in \mathrm{ET}_{m,n}$, it follows from Lemma 2.6 that the length of each cycle in α (β, resp.) is a divisor of the length d_1 (d_2, resp.) of the cycle containing 1 ($1'$, resp.).

From now on we denote i', $[n]'$ and $\beta(i')$ simply by i, $[n]$ and $\beta(i)$ for any $i' \in [n]'$, respectively. The following lemma is related to a characterization of the set $\mathrm{ET}_{m,n}$.

Lemma 2.7 ([9]). *Let $\alpha \in S_0$ and $\beta \in S_0'$. Then $(\alpha, \beta) \in \mathrm{ET}_{m,n}$ if and only if for each $i \in [n]$, there exist $a(i) \in [n]$ and $b(i) \in [m]$ such that $\alpha^i(k) = \alpha^{a(i)}(k + b(i)) - 1$ for all $k \in [m]$ and $\beta(t+i) = \beta^{b(i)}(t) + a(i)$ for all $t \in [n]$. In this case, we have $a(i) = \beta(i)$ and $b(i) = -\alpha^{-i}(-1)$.*

Note that the equations in Lemma 2.7 is equivalent to $y_\beta x_\alpha^i = x_\alpha^{a(i)} y_\beta^{b(i)}$. The next lemma gives a characterization of $(\alpha, \beta) \in \mathrm{ET}_{m,n}$ whose induced edge-transitive embedding contains a partite set preserving reflection.

Lemma 2.8 ([9]). *For any $(\alpha, \beta) \in \mathrm{ET}_{m,n}$, $\mathcal{M}(\langle x_\alpha, y_\beta\rangle, x_\alpha, y_\beta)$ contains a partite set preserving reflection if and only if $\alpha^{-1}(-k) = -\alpha(k)$ for any $k \in [m]$ and $\beta^{-1}(-t) = -\beta(t)$ for any $t \in [n]$.*

For our convenience, we denote

$$\begin{aligned}\mathrm{RET}_{m,n} = \{(\alpha, \beta) \in \mathrm{ET}_{m,n} : \alpha^{-1}(-k) = -\alpha(k) &\text{ for any } k \in [m] \text{ and} \\ \beta^{-1}(-t) = -\beta(t) &\text{ for any } t \in [n]\}.\end{aligned}$$

We call an edge-transitive embedding of $K_{m,n}$ satisfying the Property (P) which also contains a partite set preserving reflection a *reflexible edge-transitive embedding of $K_{m,n}$ satisfying the Property (P)*. By Lemmas 2.4 and 2.8, the number (up to isomorphism) of reflexible edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P) equals to the cardinality $|\,\mathrm{RET}_{m,n}\,|$. Note that if $\alpha \in S$ and $\beta \in S'$ are the identity permutations, then (α, β) belongs to $\mathrm{RET}_{m,n}$ by Lemma 2.8. So for any two positive integers m and n, there exists at least one reflexible edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P).

By Lemma 2.8, for any $(\alpha, \beta) \in \mathrm{RET}_{m,n}$ and for any $j \in [m]$ and $i \in [n]$

$$\alpha^{-i}(-j) = \alpha^{-i+1}(-\alpha(j)) = \alpha^{-i+2}(-\alpha^2(j)) = \cdots = \alpha^{-1}(-\alpha^{i-1}(j)) = -\alpha^i(j)$$

and similarly $\beta^{-j}(-i) = -\beta^j(i)$.

Lemma 2.9. *For any $(\alpha, \beta) \in \mathrm{RET}_{m,n}$ and for any $j \in [m]$ and $i \in [n]$,*

$$y_\beta^j x_\alpha^i = x_\alpha^{\beta^j(i)} y_\beta^{\alpha^i(j)}.$$

Proof. Since $\langle x_\alpha, y_\beta\rangle = \langle x_\alpha\rangle\langle y_\beta\rangle$, for any $j \in [m]$ and $i \in [n]$, there exist $a(i,j) \in [n]$ and $b(i,j) \in [m]$ such that $y_\beta^j x_\alpha^i = x_\alpha^{a(i,j)} y_\beta^{b(i,j)}$. By taking their values of $k \in [m]$ and $t \in [n]$, we have

$$\alpha^i(k) + j = \alpha^{a(i,j)}(k + b(i,j)) \quad \text{and} \quad \beta^j(t+i) = \beta^{b(i,j)}(t) + a(i,j).$$

Inserting $k = -b(i,j)$ and $t = 0$ to the equation $\alpha^i(k) + j = \alpha^{a(i,j)}(k + b(i,j))$ and $\beta^j(t+i) = \beta^{b(i,j)}(t) + a(i,j)$, respectively, we have

$$b(i,j) = -\alpha^{-i}(-j) = \alpha^i(j) \quad \text{and} \quad a(i,j) = \beta^j(i). \qquad \square$$

Lemma 2.10. *Let* $(\alpha,\beta) \in \mathrm{RET}_{m,n}$ *and let* $d_1 = |\langle\alpha\rangle|$ *and* $d_2 = |\langle\beta\rangle|$. *It holds that* $\alpha(k) \equiv -k \pmod{d_2}$ *for any* $k \in [m]$ *and* $\beta(t) \equiv -t \pmod{d_1}$ *for any* $t \in [n]$.

Proof. By Lemma 2.7, for each $i \in [n]$, there exist $a(i) \in [n]$ and $b(i) \in [m]$ such that $\alpha^i(k) = \alpha^{a(i)}(k + b(i)) - 1$ for all $k \in [m]$ and $\beta(t+i) = \beta^{b(i)}(t) + a(i)$ for all $t \in [n]$. Furthermore $a(i) = \beta(i)$ and $b(i) = -\alpha^{-i}(-1) = \alpha^i(1)$. Inserting $k = 0$ to the equation $\alpha^i(k) = \alpha^{a(i)}(k + b(i)) - 1$, we have $b(i) = \alpha^{-a(i)}(1) = \alpha^{-\beta(i)}(1)$. Hence $\alpha^i(1) = \alpha^{-\beta(i)}(1)$ for any $i \in [n]$. Since the order of α equals to the length of the orbit containing 1 by Lemma 2.6, $\beta(i) \equiv -i \pmod{d_1}$. By symmetry between α and β, it also holds that $\alpha(k) \equiv -k \pmod{d_2}$ for any $k \in [m]$. □

By Lemmas 2.7 and 2.10, $b(i) = -\alpha^{-i}(-1) = \alpha^i(1) \equiv (-1)^i \pmod{d_2}$. Hence for any $(\alpha,\beta) \in \mathrm{RET}_{m,n}$ with $d_1 = |\langle\alpha\rangle|$ and $d_2 = |\langle\beta\rangle|$, we have

$$\beta(t+i) = \beta^{b(i)}(t) + a(i) = \beta^{\alpha^i(1)}(t) + \beta(i) = \beta^{(-1)^i}(t) + \beta(i)$$

for all $i, t \in [n]$. By symmetry, it also holds $\alpha(k+j) = \alpha^{(-1)^j}(k) + \alpha(j)$ for all $j, k \in [m]$.

Lemma 2.11. *Let* $(\alpha,\beta) \in \mathrm{RET}_{m,n}$ *and let* $d_1 = |\langle\alpha\rangle|$ *and* $d_2 = |\langle\beta\rangle|$. *Now*

(1) if one of d_1 *and* d_2 *is 1, say* $d_1 = 1$, *then either* $d_2 = 1$ *or (*m *is even and* $d_2 = 2$*);*

(2) if one of d_1 *and* d_2 *is at least 3, say* $d_1 \geq 3$, *then both* m *and* d_2 *are even;*

(3) if m*(*n*, resp.) is even then* α *(*β*, resp.) is parity preserving. Furthermore there exists* $s, t \in [m]$ *such that* $\alpha(2k) = 2kt$, $\alpha(2k+1) = 2kt + 2s + 1$ *and* $2t^2 = 2$*;*

(4) if both d_1 *and* d_2 *are at least 3 then they are divisors of* $\gcd(m,n)$.

Proof. (1): Let $d_1 = 1$ and $d_2 \geq 2$. By Lemma 2.10, $\alpha(1) \equiv -1 \pmod{d_2}$. Since α is the identity, $1 \equiv -1 \pmod{d_2}$. By the assumption $d_2 \geq 2$, $d_2 = 2$. By Lemma 2.6, d_2 is a divisor of m, and hence m is even.

(2): Let $d_1 \geq 3$. By lemma 2.10, $\beta(k) \equiv -k \pmod{d_1}$, which implies that the order d_2 of β is even. Since d_2 is a divisor of m, m is also even.

(3): Let m be even. If $d_1 = 1$ then α is the identity and hence α is parity preserving. If $d_1 = 2$ then $\alpha^{-1} = \alpha$ and

$$\alpha(k) = \alpha(k-1+1) = \alpha(k-1) + \alpha(1) = \alpha(k-2) + 2\alpha(1) = \cdots = k\alpha(1)$$

for all $k \in [m]$. Since $\alpha^2(1) = \alpha(\alpha(1)) = (\alpha(1))^2 = 1$ and m is even, $\alpha(1)$ should be odd. Hence α is parity preserving. Assume that $d_1 \geq 3$. Then, d_2 is even by (2). Since $\alpha(k) \equiv -k \pmod{d_2}$, α is parity preserving.

For any $2k \in [m]$,

$$\alpha(2k) = \alpha(2(k-1)) + \alpha(2) = \alpha(2(k-2)) + 2\alpha(2) = \cdots = k\alpha(2) \quad \text{and}$$
$$\alpha(2k+1) = \alpha(2(k-1)+1) + \alpha(2) = \cdots = \alpha(1) + k\alpha(2).$$

Let $\alpha(1) = 2s+1$ and $\alpha(2) = 2t$. Now $\alpha(2k) = k\alpha(2) = 2kt$ and $\alpha(2k+1) = k\alpha(2) + \alpha(1) = 2kt + 2s + 1$. Note that for any $2k \in [m]$, $\alpha(1) + \alpha(2k) = \alpha(2k + 1) = \alpha^{-1}(2k) + \alpha(1)$. Hence $\alpha^{-1}(2k) = \alpha(2k)$, namely, $\alpha^2(2k) = 2k$. So we have $\alpha^2(2) = \alpha(2t) = 2t^2 = 2$.

(4): Let $d_1, d_2 \geq 3$. Now all of d_1, d_2, m and n are even by (2). Hence there exist $s, t \in [m]$ such that $\alpha(2k) = 2kt$, $\alpha(2k+1) = 2kt + 2s + 1$ and $2t^2 = 2$ by (3). Since d_1 is even and

$$\alpha^{2i}(1) = \alpha^{2i-1}(2s+1) = \alpha^{2i-2}(2st + 2s + 1) = \cdots = 2is(t+1) + 1,$$

d_1 is the smallest positive integer such that $d_1 s(t+1) \equiv 0 \pmod{m}$ by Lemma 2.6. Hence d_1 is a divisor of m and consequently a divisor of $\gcd(m,n)$. Similarly d_2 is a divisor of $\gcd(m,n)$. □

3 At least one of m and n is odd

In this section, we classify reflexible edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P) when at least one of m and n is odd. Note that when at least one of m and n is odd, any orientable edge-transitive embedding of $K_{m,n}$ is an edge-transitive embedding satisfying the Property (P). In [9], the second author counted $|\operatorname{RET}_{m,n}|$ when both m and n are odd as follows.

Theorem 3.1 ([9]). *If both m and n are odd then $|\operatorname{RET}_{m,n}| = 1$, namely, there exists only one reflexible edge-transitive embedding of $K_{m,n}$ satisfying the Property (P) up to isomorphism.*

In the next theorem, we count $|\operatorname{RET}_{m,n}|$ when exactly one of m and n is odd. By symmetry, we assume that m is odd.

Theorem 3.2. *Let*

$$m = p_1^{a_1} \cdots p_\ell^{a_\ell} p_{\ell+1}^{a_{\ell+1}} \cdots p_{\ell+f}^{a_{\ell+f}} \quad \textit{(prime factorization)}$$

be odd and

$$n = 2^b p_1^{b_1} \cdots p_\ell^{b_\ell} q_{\ell+1}^{b_{\ell+1}} \cdots q_{\ell+g}^{b_{\ell+g}} \quad \textit{(prime factorization)}$$

be even. Let $\gcd(m,n) = p_1^{c_1} \cdots p_\ell^{c_\ell}$ with $c_i \geq 1$ for any $i = 1, \ldots, \ell$. Now

$$|\operatorname{RET}_{m,n}| = 2^f(1 + p_1^{c_1}) \cdots (1 + p_\ell^{c_\ell}),$$

namely, there exist $2^f(1 + p_1^{c_1}) \cdots (1 + p_\ell^{c_\ell})$ reflexible edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P) up to isomorphism.

Proof. Let $(\alpha, \beta) \in \operatorname{RET}_{m,n}$ and let $d_1 = |\langle\alpha\rangle|$ and $d_2 = |\langle\beta\rangle|$. Suppose that $d_1 \geq 3$. Then both d_2 and m are even by Lemma 2.11(2), which is a contradiction. Hence $d_1 = 1$ or 2. Furthermore for any $k \in [m]$,

$$\alpha(k) = \alpha^{-1}(k-1) + \alpha(1) = \alpha(k-1) + \alpha(1) = \cdots = k\alpha(1).$$

Let $\alpha(1) = r$. Now $\alpha(k) = rk$ and $\alpha^2(1) = \alpha(r) = r^2 \equiv 1 \pmod{m}$.

Since n is even, β is parity preserving and there exists $s, t \in [n]$ such that $\beta(2k) = 2kt$, $\beta(2k+1) = 2kt + 2s + 1$ and $2t^2 = 2$ for any $2k \in [n]$ by Lemma 2.11(3). If $2t \neq 2$ then the length of the orbit containing 2 is 2 and hence d_2 is even. But it can not happen because m is odd. Hence for any $2k \in [n]$, $\beta(2k) = 2k$, $\beta(2k+1) = 2k + 2s + 1$ and for any $i \in [m]$,

$$\beta^i(1) = \beta^{i-1}(2s+1) = \beta^{i-2}(2s+2s+1) = \cdots = 2is + 1.$$

Therefore d_2 is the smallest positive integer such that $2d_2 s \equiv 0 \pmod{n}$, which implies that d_2 is a divisor of n, and hence d_2 is a divisor of $\gcd(m,n) = p_1^{c_1} \cdots p_\ell^{c_\ell}$.

If $r \equiv 1 \pmod{p_i^{a_i}}$ for some $i = 1, 2, \ldots, \ell$, then the fact $\alpha(1) = r \equiv -1 \pmod{d_2}$ implies that p_i can not be a divisor of d_2. Hence $p_i^{b_i}$ should divide s, namely, $s \equiv 0 \pmod{p_i^{b_i}}$. If $r \equiv -1 \pmod{p_j^{a_j}}$ for some $j = 1, 2, \ldots, \ell$, then $s \equiv x \cdot p_j^{b_j - c_j} \pmod{p_j^{b_j}}$ for some x with $0 \leq x \leq p_j^{c_j} - 1$ because d_2 is a divisor of $\gcd(m,n)$. Therefore, for any $j = 1, \ldots, \ell$, the pair $(r \pmod{p_j^{a_j}}, s \pmod{p_j^{b_j}})$ is $(1, 0)$ or $(-1, x \cdot p_j^{b_j - c_j})$ for some x with $0 \leq x \leq p_j^{c_j} - 1$.

Because $d_2 \mid \gcd(m,n)$, we have $2s \equiv 0 \pmod{2^b}$ and for any $k = 1, 2, \ldots, g$, $s \equiv 0 \pmod{q_{\ell+k}^{b_{\ell+k}}}$. Since $r^2 \equiv 1 \pmod{m}$, $r \equiv \pm 1 \pmod{p_{\ell+j}^{a_{\ell+j}}}$ for any $j = 1, 2, \ldots f$.

Conversely for any $r \in [m]$ and $s \in [n]$ satisfying the conditions

(i) for any $j = 1, \ldots, \ell$, the pair $(r \pmod{p_j^{a_j}}), s \pmod{p_j^{b_j}}))$ is $(1, 0)$ or $(-1, x \cdot p_j^{b_j - c_j})$ for some integer x with $0 \leq x \leq p_j^{c_j} - 1$,

(ii) $2s \equiv 0 \pmod{2^b q_{\ell+1}^{b_{\ell+1}} \cdots q_{\ell+g}^{b_{\ell+g}}}$ and

(iii) for any $j = 1, 2, \ldots f$, $r \equiv \pm 1 \pmod{p_{\ell+j}^{a_{\ell+j}}}$,

define $\alpha(k) = rk$ for any $k \in [m]$ and $\beta(2t) = 2t$, $\beta(2t+1) = 2t + 2s + 1$ for any $2t \in [n]$. Note that $\alpha \in S_0$ and $\beta \in S_0'$. Let $d_1' = |\langle \alpha \rangle|$ and $d_2' = |\langle \beta \rangle|$. Now $d_1' = 1$ or 2 depending on the value of r and d_2' is the smallest positive integer satisfying $2d_2' s \equiv 0 \pmod{n}$. Note that d_2' divides $\gcd(m,n)$ and $r \equiv -1 \pmod{d_2'}$. For any $i \in [n]$, let $a(i) = \beta(i)$ and $b(i) = \alpha^i(1) = r^i$. For the first case, let i be even. Now $a(i) = \beta(i) = i$ and $b(i) = \alpha^i(1) = 1$. For any $2t \in [n]$,

$$\beta(2t+i) = 2t + i \quad \text{and}$$
$$\beta^{b(i)}(2t) + a(i) = \beta(2t) + \beta(i) = 2t + i$$

and

$$\beta(2t+1+i) = 2t + i + 2s + 1 \quad \text{and}$$
$$\beta^{b(i)}(2t+1) + a(i) = \beta(2t+1) + \beta(i) = 2t + 2s + 1 + i.$$

Hence $\beta(t+i) = \beta^{b(i)}(t) + a(i)$ for any $t \in [n]$. For any $k \in [m]$,

$$\alpha^i(k) = k \quad \text{and}$$
$$\alpha^{a(i)}(k + b(i)) - 1 = k.$$

Hence $\alpha^i(k) = \alpha^{a(i)}(k + b(x)) - 1$ for any $k \in [m]$.

For the remaining case, let i be odd. Now $a(i) = \beta(i) = i + 2s$ and $b(i) = \alpha^i(1) = r \equiv -1 \pmod{d_2'}$. For any $2t \in [n]$,

$$\beta(2t + i) = 2t + i + 2s \quad \text{and}$$
$$\beta^{b(i)}(2t) + a(i) = \beta^{-1}(2t) + \beta(i) = 2t + i + 2s$$

and

$$\beta(2t + 1 + i) = 2t + i + 1 \quad \text{and}$$
$$\beta^{b(i)}(2t + 1) + a(i) = \beta^{-1}(2t + 1) + \beta(i) = 2t + 1 - 2s + i + 2s = 2k + i + 1.$$

Hence $\beta(t + i) = \beta^{b(i)}(t) + a(i)$ for any $t \in [n]$. For any $k \in [m]$,

$$\alpha^i(k) = rk \quad \text{and}$$
$$\alpha^{a(i)}(k + b(i)) - 1 = \alpha(k + r) - 1 = rk + r^2 - 1 = rk.$$

Hence $\alpha^i(k) = \alpha^{a(i)}(k + b(i)) - 1$ for any $k \in [m]$. By Lemma 2.7, $(\alpha, \beta) \in \mathrm{ET}_{m,n}$. Furthermore one can easily check that $\alpha^{-1}(-k) = -\alpha(k)$ for any $k \in [m]$ and $\beta^{-1}(-t) = -\beta(t)$ for any $t \in [n]$. Hence $(\alpha, \beta) \in \mathrm{RET}_{m,n}$ by Lemma 2.8.

Therefore

$$|\,\mathrm{RET}_{m,n}\,| = 2^f(1 + p_1^{c_1}) \cdots (1 + p_\ell^{c_\ell}). \qquad \square$$

4 Both m and n are even

In this section, we classify reflexible edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P) when both m and n are even, and consequently prove Theorem 1.1. For the classification, we give the following lemma.

Lemma 4.1. *Let m and n be even and let $\alpha \in S_0$ and $\beta \in S_0'$ with $d_1 = |\langle\alpha\rangle|$ and $d_2 = |\langle\beta\rangle|$. Now $(\alpha, \beta) \in \mathrm{RET}_{m,n}$ if and only if α and β are defined by*

$$\alpha(2k) = 2kt_1 \quad \textit{and}$$
$$\alpha(2k + 1) = 2kt_1 + 2s_1 + 1$$

for any $2k \in [m]$ and

$$\beta(2k) = 2kt_2 \quad \textit{and}$$
$$\beta(2k + 1) = 2kt_2 + 2s_2 + 1$$

for any $2k \in [n]$ for some quadruple $(s_1, t_1; s_2, t_2) \in [\frac{m}{2}] \times [\frac{m}{2}] \times [\frac{n}{2}] \times [\frac{n}{2}]$ satisfying the following conditions;

(i) $d_1 \mid \gcd(m, n)$ *and* $d_2 \mid \gcd(m, n)$*;*

(ii) $2t_1^2 \equiv 2 \pmod{m}$ *and* $2t_2^2 \equiv 2 \pmod{n}$*;*

(iii) $2(s_1 + 1) \equiv 0 \pmod{d_2}$*,* $2(t_1 + 1) \equiv 0 \pmod{d_2}$*,*
$2(s_2 + 1) \equiv 0 \pmod{d_1}$*, and* $2(t_2 + 1) \equiv 0 \pmod{d_1}$*;*

(iv) $2(s_1 + 1)(t_1 - 1) \equiv 0 \pmod{m}$ *and* $2(s_2 + 1)(t_2 - 1) \equiv 0 \pmod{n}$*.*

Proof. ($\Leftarrow$): Assume that $2t_1 = 2$, namely, $t_1 = 1$. Then $\alpha(2k) = 2k$ and $\alpha(2k+1) = 2k+2s_1+1$ for any $2k \in [m]$. Since for any $i \in [n]$, $\alpha^i(2k+1) = 2k+2is_1+1$, d_1 is the smallest positive integer such that $2d_1s_1 \equiv 0 \pmod{m}$. Now assume that $2t_1 \neq 2$. Then d_1 should be even because $\alpha^2(2) = 2t_1^2 = 2$. Since for any $2i \in [n]$ and for any $2k \in [m]$, $\alpha^{2i}(2k+1) = 2k + 2is_1(t_1+1)+1$, d_1 is the smallest positive even integer such that $d_1s_1(t_1+1) \equiv 0 \pmod{m}$. Similarly one can show that d_2 is the smallest positive integer such that $2d_2s_2 \equiv 0 \pmod{n}$ if $t_2 = 1$; and the smallest positive even integer such that $d_2s_2(t_2+1) \equiv 0 \pmod{n}$ if $t_2 \neq 1$.

For any $i \in [n]$, let $a(i) = \beta(i)$ and $b(i) = \alpha^i(1)$. For the first case, let i be even. Then $a(i) = \beta(i) = it_2 \equiv -i \pmod{d_1}$ and $b(i) = \alpha^i(1) = is_1(t_1+1)+1 \equiv 1 \pmod{d_2}$. For any $2k \in [n]$,

$$\begin{aligned} \beta(2k+i) &= 2kt_2 + it_2 \quad \text{and} \\ \beta^{b(i)}(2k) + a(i) = \beta(2k) + \beta(i) &= 2kt_2 + it_2 \end{aligned}$$

and

$$\begin{aligned} \beta(2k+1+i) &= 2kt_2 + it_2 + 2s_2 + 1 \quad \text{and} \\ \beta^{b(i)}(2k+1) + a(i) = \beta(2k+1) + \beta(i) &= 2kt_2 + 2s_2 + 1 + it_2. \end{aligned}$$

Hence $\beta(k+i) = \beta^{b(i)}(k) + a(i)$ for any $k \in [n]$. For any $2k \in [m]$,

$$\begin{aligned} \alpha^i(2k) &= 2k \quad \text{and} \\ \alpha^{a(i)}(2k+b(i)) - 1 &= \alpha^{-i}(2k + is_1(t_1+1)+1) - 1 \\ &= (2k + is_1(t_1+1) - is_1(t_1+1)+1) - 1 = 2k \end{aligned}$$

and

$$\begin{aligned} \alpha^i(2k+1) &= 2k + is_1(t_1+1)+1, \quad \text{and} \\ \alpha^{a(i)}(2k+1+b(i)) - 1 &= \alpha^{-i}(2k + is_1(t_1+1)+2) - 1 \\ &= (2k + is_1(t_1+1)+2) - 1 = 2k + is_1(t_1+1)+1. \end{aligned}$$

Hence $\alpha^i(k) = \alpha^{a(i)}(k+b(i)) - 1$ for any $k \in [m]$.

For the remaining case, let i be odd. Now $a(i) = \beta(i) = (i-1)t_2 + 2s_2 + 1 \equiv -i \pmod{d_1}$ and $b(i) = \alpha^i(1) = (i-1)s_1(t_1+1) + 2s_1 + 1 \equiv -1 \pmod{d_2}$. For any $2k \in [n]$,

$$\begin{aligned} \beta(2k+i) &= 2kt_2 + (i-1)t_2 + 2s_2 + 1 \quad \text{and} \\ \beta^{b(i)}(2k) + a(i) = \beta^{-1}(2k) + \beta(i) &= 2kt_2 + (i-1)t_2 + 2s_2 + 1 \end{aligned}$$

and

$$\begin{aligned} \beta(2k+1+i) &= (2k+i+1)t_2 \quad \text{and} \\ \beta^{b(i)}(2k+1) + a(i) &= \beta^{-1}(2k+1) + \beta(i) \\ &= (2kt_2 - 2s_2t_2 + 1) + (i-1)t_2 + 2s_2 + 1 \\ &= (2k+i+1)t_2 - 2(s_2+1)(t_2-1) = (2k+i+1)t_2. \end{aligned}$$

Hence $\beta(k+i) = \beta^{b(i)}(k) + a(i)$ for any $k \in [n]$. For any $2k \in [m]$,

$$\begin{aligned} \alpha^i(2k) &= 2kt_1 \quad \text{and} \\ \alpha^{a(i)}(2k+b(i)) - 1 &= \alpha^{-i}(2k + (i-1)s_1(t_1+1) + 2s_1 + 1) - 1 \\ &= (2k + (i-1)s_1(t_1+1) + 2s_1)t_1 - (i+1)s_1(t_1+1) + 2s_1 \\ &= 2kt_1 - 2s_1(t_1+1) + 2s_1t_1 + 2s_1 = 2kt_1 \end{aligned}$$

and

$$\begin{aligned} \alpha^i(2k+1) &= 2kt_1 + (i-1)s_1(t_1+1) + 2s_1 + 1 \quad \text{and} \\ \alpha^{a(i)}(2k+1+b(i)) - 1 &= \alpha^{-i}(2k + (i-1)s_1(t_1+1) + 2s_1 + 2) - 1 \\ &= (2k + (i-1)s_1(t_1+1) + 2s_1 + 2)t_1 - 1 \\ &= 2kt_1 + (i-1)s_1(t_1+1) + 2s_1 + 1 + 2(s_1+1)(t_1-1) \\ &= 2kt_1 + (i-1)s_1(t_1+1) + 2s_1 + 1. \end{aligned}$$

Hence $\alpha^i(k) = \alpha^{a(i)}(k + b(i)) - 1$ for any $k \in [m]$. By Lemma 2.7, $(\alpha, \beta) \in \mathrm{ET}_{m,n}$. Furthermore one can easily check that $\alpha^{-1}(-k) = -\alpha(k)$ for any $k \in [m]$ and $\beta^{-1}(-k) = -\beta(k)$ for any $k \in [n]$. Hence $(\alpha, \beta) \in \mathrm{RET}_{m,n}$ by Lemma 2.8.

($\Rightarrow$): Since m and n are even, both α and β are parity preserving. For any $2k \in [m]$,

$$\begin{aligned} \alpha(2k) &= \alpha(2(k-1)) + \alpha(2) \\ &= \alpha(2(k-2)) + 2\alpha(2) = \cdots = k\alpha(2) \quad \text{and} \\ \alpha(2k+1) &= \alpha(2(k-1)+1) + \alpha(2) \\ &= \alpha(2(k-2)+1) + 2\alpha(2) = \cdots = \alpha(1) + k\alpha(2). \end{aligned}$$

Let $\alpha(1) = 2s_1 + 1$ and $\alpha(2) = 2t_1$ for some $s_1, t_1 \in [\frac{m}{2}]$. Then $\alpha(2k) = 2kt_1$ and $\alpha(2k+1) = 2kt_1 + 2s_1 + 1$ for any $2k \in [m]$. Note that for any $2k \in [m]$, $\alpha(1) + \alpha(2k) = \alpha(2k+1) = \alpha^{-1}(2k) + \alpha(1)$. Hence $\alpha^{-1}(2k) = \alpha(2k)$, namely, $\alpha^2(2k) = 2k$. It implies that $\alpha^2(2) = \alpha(2t_1) = 2t_1^2 \equiv 2 \pmod{m}$. Assume that $2t_1 = 2$, namely, $t_1 = 1$. Then by Lemma 2.6, the order $|\langle\alpha\rangle|$ is the smallest positive integer d_1 such that

$$\alpha^{d_1}(1) = \alpha^{d_1-1}(2s_1+1) = \alpha^{d_1-2}(2s_1 + 2s_1 + 1) = \cdots = 2d_1s_1 + 1 \equiv 1.$$

Now assume that $2t_1 \neq 2$. Then the order $|\langle\alpha\rangle|$ is even and it is the smallest positive even integer d_1 such that

$$\begin{aligned} \alpha^{d_1}(1) &= \alpha^{d_1-1}(2s_1+1) = \alpha^{d_1-2}(2s_1t_1 + 2s_1 + 1) = \alpha^{d_1-3}(2s_1t_1 + 4s_1 + 1) \\ &= \alpha^{d_1-4}(4s_1t_1 + 4s_1 + 1) = \cdots = d_1s_1(t_1+1) + 1 \equiv 1. \end{aligned}$$

Hence d_1 is a divisor of m and consequently a divisor of $\gcd(m,n)$.

By a similar reason, there exist $s_2, t_2 \in [\frac{n}{2}]$ such that $\beta(2k) = 2kt_2$ and $\beta(2k+1) = 2kt_2 + 2s_2 + 1$ for any $2k \in [n]$. Furthermore $2t_2^2 \equiv 2 \pmod{n}$ and d_2 is a divisor of $\gcd(m,n)$. By Lemma 2.10, $\alpha(1) = 2s_1 + 1 \equiv -1 \pmod{d_2}$, namely, $2(s_1+1) \equiv 0 \pmod{d_2}$ and $\alpha(2) = 2t_1 \equiv -2 \pmod{d_2}$, namely, $2(t_1+1) \equiv 0 \pmod{d_2}$. Similarly it holds that $2(s_2+1) \equiv 2(t_2+1) \equiv 0 \pmod{d_1}$. Note that

$$2t_1 = \alpha(2) = \alpha^{-1}(1) + \alpha(1) = (-2s_1t_1 + 1) + 2s_1 + 1.$$

Hence $2(s_1+1)(t_1-1) \equiv 0 \pmod{m}$. By a similar reason, it holds that $2(s_2+1)(t_2-1) \equiv 0 \pmod{n}$. □

For even m and n, let $\mathcal{Q}(m,n)$ be the set of quadruples $(s_1,t_1;s_2,t_2)\in[\frac{n}{2}]\times[\frac{n}{2}]\times[\frac{m}{2}]\times[\frac{m}{2}]$ satisfying the conditions in Lemma 4.1. By Lemma 4.1, the classification of reflexible edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P) is equivalent to the classification of $\mathcal{Q}(m,n)$, and the number $|\operatorname{RET}_{m,n}|$ equals to the cardinality $|\mathcal{Q}(m,n)|$.

In this section, let

$$\begin{aligned} m &= 2^a p_1^{a_1}\cdots p_\ell^{a_\ell} p_{\ell+1}^{a_{\ell+1}}\cdots p_{\ell+f}^{a_{\ell+f}} \quad \text{and}\\ n &= 2^b p_1^{b_1}\cdots p_\ell^{b_\ell} q_{\ell+1}^{a_{\ell+1}}\cdots q_{\ell+g}^{b_{\ell+g}} \quad \text{(prime decompositions)}\end{aligned}$$

and let $\gcd(m,n)=2^c p_1^{c_1}\cdots p_\ell^{c_\ell}$ with $c_i\geq 1$ for any $i=1,\ldots,\ell$. Without any loss of generality, assume that $a\leq b$, namely, $a=c$. By Chinese Remainder Theorem, it suffices to consider quadruples $(s_1,t_1;s_2,t_2)$ modulo prime powers dividing m and n, respectively. So we have the following lemma.

Lemma 4.2. *For a quadruple* $(s_1,t_1;s_2,t_2)\in[\frac{n}{2}]\times[\frac{n}{2}]\times[\frac{m}{2}]\times[\frac{m}{2}]$, $(s_1,t_1;s_2,t_2)$ *belongs to* $\mathcal{Q}(m,n)$ *if and only if:*

(1) for $i=1,\ldots,\ell$, $(s_1 \pmod{p_i^{a_i}}, t_1 \pmod{p_i^{a_i}}; s_2 \pmod{p_i^{b_i}}, t_2 \pmod{p_i^{b_i}})$ *is one of* $(-1,-1;-1,-1)$, $(-1,-1;y\cdot p_i^{b_i-c_i},1)$, $(x\cdot p_i^{a_i-c_i},1;-1,-1)$ *and* $(0,1;0,1)$, *where* $x,y=0,1,\ldots,p_i^{c_i}-1$;

(2) for any $j=1,2,\ldots,f$, $(s_1 \pmod{p_{\ell+j}^{a_{\ell+j}}}, t_1 \pmod{p_{\ell+j}^{a_{\ell+j}}})$ *is* $(0,1)$ *or* $(-1,-1)$;

(3) for any $k=1,2,\ldots,g$, $(s_2 \pmod{q_{\ell+k}^{b_{\ell+k}}}, t_2 \pmod{q_{\ell+k}^{b_{\ell+k}}})$ *is* $(0,1)$ *or* $(-1,-1)$;

(4) $(s_1 \pmod{2^a}, t_1 \pmod{2^a}; s_2 \pmod{2^b}, t_2 \pmod{2^b})$ *belongs to* $\mathcal{Q}(2^a,2^b)$.

Proof. Assume that $(s_1,t_1;s_2,t_2)$ belongs to $\mathcal{Q}(m,n)$. Then $t_1^2\equiv 1 \pmod{\frac{m}{2}}$ and $t_2^2\equiv 1 \pmod{\frac{n}{2}}$.

(1): First let us consider the quadruple modulo $p_i^{a_i}$ and $p_i^{b_i}$ for $i=1,\ldots,\ell$. Note that $t_1\equiv\pm1 \pmod{p_i^{a_i}}$ and $t_2\equiv\pm1 \pmod{p_i^{b_i}}$.

If $t_1\equiv -1 \pmod{p_i^{a_i}}$, then s_1 should be -1 modulo $p_i^{a_i}$ to satisfy

$$2(s_1+1)(t_1-1)\equiv 0 \pmod{p_i^{a_i}}.$$

By similar reason, if $t_2\equiv -1 \pmod{p_i^{b_i}}$, then $s_2\equiv -1 \pmod{p_i^{b_i}}$.

Let $(s_1,t_1)\equiv(-1,-1) \pmod{p_i^{a_i}}$. Since d_1 is the smallest positive even integer satisfying $d_1s_1(t_1+1)\equiv 0 \pmod{m}$, p_i does not divide d_1. If $t_2\equiv -1 \pmod{p_i^{b_i}}$ then s_2 should be -1 modulo $p_i^{b_i}$. If $t_2\equiv 1 \pmod{p_i^{b_i}}$, then $s_2\equiv y\cdot p_i^{b_i-c_i} \pmod{p_i^{b_i}}$ for some $y=0,1,\ldots,p_i^{c_i}-1$ because $d_2\mid\gcd(m,n)$. By a similar reason, one can say that if $(s_2,t_2)\equiv(-1,-1) \pmod{p_i^{b_i}}$, then $(s_1,t_1)\equiv(-1,-1)$ or $(x\cdot p_i^{a_i-c_i},1) \pmod{p_i^{a_i}}$ for some $x=0,1,\ldots,p_i^{c_i}-1$.

Let $(s_1,t_1)\equiv(0,1) \pmod{p_i^{a_i}}$. By the condition (iii) in Lemma 4.1, p_i does not divide d_2. Note that if $t_2=1$ then d_2 is the smallest positive integer satisfying $2d_2s_2\equiv 0 \pmod{n}$, and if $t_2\neq 1$ then d_2 is the smallest positive even integer such that $d_2s_2(t_2+1)\equiv 0 \pmod{n}$. Hence $s_2=0$ or $t_2=-1$ modulo $p_i^{b_i}$, which implies that $(s_2,t_2)\equiv(0,1)$ or $(-1,-1) \pmod{p_i^{b_i}}$.

Let $t_1\equiv 1 \pmod{p_i^{a_i}}$ and $s_1\not\equiv 0 \pmod{p_i^{a_i}}$. One can see that p_i divides d_1. By the condition (iii) in Lemma 4.1, $t_2\equiv -1 \pmod{p_i^{b_i}}$ and $s_2\equiv -1 \pmod{p_i^{b_i}}$.

Therefore

$$\begin{aligned}(s_1 \pmod{p_i^{a_i}}, t_1 \pmod{p_i^{a_i}}; s_2 \pmod{p_i^{b_i}}, t_2 \pmod{p_i^{b_i}}) =\\ (-1,-1;-1,-1), (-1,-1; y\cdot p_i^{b_i-c_i}, 1), (x\cdot p_i^{a_i-c_i}, 1; -1,-1) \text{ or } (0,1;0,1),\end{aligned}$$

where $x, y = 0, 1, \ldots, p_i^{c_i} - 1$.

(2): For any $j = 1, 2, \ldots, f$, $t_1 \equiv \pm 1 \pmod{p_{\ell+j}^{a_{\ell+j}}}$. If $t_1 \equiv 1 \pmod{p_{\ell+j}^{a_{\ell+j}}}$ then $s_1 \equiv 0 \pmod{p_{\ell+j}^{a_{\ell+j}}}$ because $p_{\ell+j}$ does not divide d_1. If $t_1 \equiv -1 \pmod{p_{\ell+j}^{a_{\ell+j}}}$ then $s_1 \equiv -1 \pmod{p_{\ell+j}^{a_{\ell+j}}}$ to satisfy $2(s_1+1)(t_1-1) \equiv 0 \pmod{p_{\ell+j}^{a_{\ell+j}}}$.

(3): By the similar reason with (2), for any $k = 1, 2, \ldots, g$, $(s_2 \pmod{q_{\ell+k}^{b_{\ell+k}}}, t_2 \pmod{q_{\ell+k}^{b_{\ell+k}}})$ is $(0,1)$ or $(-1,-1)$.

(4): If a quadruple $(s_1, t_1; s_2, t_2) \in [\frac{n}{2}] \times [\frac{n}{2}] \times [\frac{m}{2}] \times [\frac{m}{2}]$ satisfies all conditions in Lemma 4.1, then it also satisfies these conditions modulo 2^a and 2^b. Hence

$$(s_1 \pmod{2^a}, t_1 \pmod{2^a}; s_2 \pmod{2^b}, t_2 \pmod{2^b}) \in \mathcal{Q}(2^a, 2^b).$$

By Chinese Remainder Theorem, one can show that if (1), (2), (3) and (4) hold, then $(s_1, t_1; s_2, t_2) \in \mathcal{Q}(m, n)$. □

Corollary 4.3. *The number of reflexible edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P) up to isomorphism is $2^{f+g+\ell}(1+p_1^{c_1})\cdots(1+p_\ell^{c_\ell})|\mathcal{Q}(2^a, 2^b)|$.*

Proof. By Lemma 4.2, the number of reflexible edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P) up to isomorphism is

$$\begin{aligned}(2+2p_1^{c_1})\cdots(2+2p_\ell^{c_\ell})2^f 2^g|\mathcal{Q}(2^a, 2^b)| =\\ 2^{f+g+\ell}(1+p_1^{c_1})\cdots(1+p_\ell^{c_\ell})|\mathcal{Q}(2^a, 2^b)|.\end{aligned}$$ □

By Lemma 4.2, it suffices to classify $\mathcal{Q}(2^a, 2^b)$ to classify reflexible edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P). Let $\mathcal{P}(2) = \{(0,1)\}$ and for a 2-power 2^a $(a > 1)$, let $\mathcal{P}(2^a)$ be the set of all pairs $(s,t) \in [2^{a-1}] \times [2^{a-1}]$ satisfying the conditions:

(i) $2t^2 \equiv 2 \pmod{2^a}$ and

(ii) $2(s+1)(t-1) \equiv 0 \pmod{2^a}$.

For any $(s,t) \in \mathcal{P}(2^a) \backslash \{(0,1)\}$, let $d(s,t)$ be the smallest positive even number d such that $ds(t+1) \equiv 0 \pmod{2^a}$ and let $e(s,t)$ be the largest number 2^j with $2^j \le 2^a$ satisfying $2(s+1) \equiv 0 \pmod{2^j}$ and $2(t+1) \equiv 0 \pmod{2^j}$. Let $d(0,1) = 1$ and $e(0,1) = 2$. Now we have the following lemma.

Lemma 4.4. *For 2-powers 2^a $(a \ge 1)$ and 2^b $(b \ge 1)$, a quadruple $(s_1, t_1; s_2, t_2)$ belongs to $\mathcal{Q}(2^a, 2^b)$ if and only if $(s_1, t_1; s_2, t_2)$ satisfies the conditions*

(a) $(s_1, t_1) \in \mathcal{P}(2^a)$ *and* $(s_2, t_2) \in \mathcal{P}(2^b)$,

(b) $d(s_1, t_1) \le e(s_2, t_2)$ *and* $d(s_2, t_2) \le e(s_1, t_1)$.

Proof. The conditions (i) and (ii) in the definition of $\mathcal{P}(2^a)$ correspond to the conditions (ii) and (iv) in Lemma 4.1.

Suppose that $d(s_1,t_1) \le e(s_2,t_2)$ and $d(s_2,t_2) \le e(s_1,t_1)$. Since $d(s_1,t_1) \le 2^a$ and $e(s_2,t_2) \le 2^b$, $d(s_1,t_1)$ divides $\gcd(2^a,2^b)$, the minimum of 2^a and 2^b. Similarly $d(s_2,t_2)$ also divides $\gcd(2^a,2^b)$. Furthermore it holds that

$$\begin{aligned}
2(s_1+1) &\equiv 0 \pmod{d(s_2,t_2)},\\
2(t_1+1) &\equiv 0 \pmod{d(s_2,t_2)},\\
2(s_2+1) &\equiv 0 \pmod{d(s_1,t_1)} \quad \text{and}\\
2(t_2+1) &\equiv 0 \pmod{d(s_1,t_1)}.
\end{aligned}$$

Therefore the conditions (i) and (iii) in Lemma 4.1 hold, and hence $(s_1,t_1;s_2,t_2)$ belongs to $\mathcal{Q}(2^a,2^b)$.

Let $(s_1,t_1;s_2,t_2)$ belong to $\mathcal{Q}(2^a,2^b)$. Now the condition (iii) in Lemma 4.1 is equivalent to the condition $d(s_1,t_1) \le e(s_2,t_2)$ and $d(s_2,t_2) \le e(s_1,t_1)$. □

By Lemma 4.4, the calculation of $d(s,t)$ and $e(s,t)$ for each $(s,t) \in \mathcal{P}(2^a)$ is helpful to calculate $|\mathcal{Q}(2^a,2^b)|$. The following lemma gives full list of $(s,t) \in \mathcal{P}(2^a)$ and corresponding $d(s,t)$ and $e(s,t)$.

Lemma 4.5. *For a* 2*-power* 2^a *(*$a>1$*), the set* $\{(s,t,d(s,t),e(s,t)) : (s,t) \in \mathcal{P}(2^a)\}$ *is the following:*

$$\begin{cases}
\{(0,1,1,2),(1,1,2,4)\}, & \textit{if } a=2\\
\{(0,1,1,2),(1,1,4,4),(2,1,2,2),(3,1,4,4),(1,3,2,4),(3,3,2,8)\}, & \textit{if } a=3\\
\{(0,1,1,2),(2^{a-2}-1,2^{a-2}-1,4,2^{a-1}),(2^{a-1}-1,2^{a-2}-1,4,2^{a-1}), & \\
(2^{a-2}-1,2^{a-1}-1,2,2^{a-1}),(2^{a-1}-1,2^{a-1}-1,2,2^a)\} & \\
\cup\{(x,1,2^{a-1},4),(x,2^{a-2}+1,2^{a-1},4) : x=1,3,\ldots,2^{a-1}-1\} & \\
\cup\{(2^iy,1,2^{a-i-1},2) : i=1,\ldots,a-2,\ \ y=1,3,\ldots,2^{a-i-1}-1\} & \textit{if } a\ge 4.
\end{cases}$$

Proof. Let $(s,t) \in \mathcal{P}(2^a)$.

For $a=2$, t should be 1 and both $s=0$ and $s=1$ satisfy the conditions for $(s,t) \in \mathcal{P}(2^a)$. Hence $(s,t,d(s,t),e(s,t)) = (0,1,1,2)$ or $(1,1,2,4)$. Let $a=3$. Then $t=1$ and $t=3$. If $t=1$, then $s=i$ for some $i=0,1,2,3$. If $t=3$, then $s=1$ or $s=3$. In any possible pair (s,t), one can easily calculate $d(s,t)$ and $e(s,t)$.

Now assume that $a \ge 4$. Then $t = 1, 2^{a-2}-1, 2^{a-2}+1$ or $2^{a-1}-1$. For $t=1$, any number $0,1,2,\ldots,2^{a-1}-1$ is possible for s to satisfy the condition (ii) in the definition of $\mathcal{P}(2^a)$. Note that if $(s,t)=(0,1)$, then $(d(0,1),e(0,1))=(1,2)$. One can easily show that if $(s,t)=(x,1)$ for any $x=1,3,\ldots,2^{a-1}-1$ then $(d(s,t),e(s,t))=(2^{a-1},4)$. If $(s,t)=(2^iy,1)$ for any $i=1,\ldots,a-2$ and for any $y=1,3,\ldots,2^{a-i-1}-1$, then $(d(s,t),e(s,t))=(2^{a-i-1},2)$.

For $t=2^{a-2}-1$, both $s=2^{a-2}-1$ and $s=2^{a-1}-1$ satisfy the conditions for $(s,t) \in \mathcal{P}(2^a)$. If $(s,t)=(2^{a-2}-1,2^{a-2}-1)$ or $(2^{a-1}-1,2^{a-2}-1)$ then we have $(d(s,t),e(s,t))=(4,2^{a-1})$.

Let $t=2^{a-2}+1$. Then any number $s=1,3,\ldots,2^{a-1}-1$ satisfies the condition (ii) in the definition of $\mathcal{P}(2^a)$. For any $(s,t)=(x,2^{a-2}+1)$ with $x=1,3,\ldots,2^{a-1}-1$, we have $(d(s,t),e(s,t))=(2^{a-1},4)$.

For the final case, let $t = 2^{a-1}-1$. Then $s = 2^{a-2}-1$ or $2^{a-1}-1$. If $(s,t) = (2^{a-2}-1, 2^{a-1}-1)$ then we have $(d(s,t), e(s,t)) = (2, 2^{a-1})$; if $(s,t) = (2^{a-1}-1, 2^{a-1}-1)$ then $(d(s,t), e(s,t)) = (2, 2^a)$. □

Theorem 4.6. *For any 2-powers 2^a and 2^b with $a \leq b$, the number $|\mathcal{Q}(2^a, 2^b)|$ of reflexible edge-transitive embeddings of $K_{m,n}$ satisfying the Property (P) up to isomorphism is the following:*

$$|\mathcal{Q}(2^a,2^b)| = \begin{cases} 1 & \text{if } (a,b) = (1,1), \\ 2 & \text{if } (a,b) = (1,2), \\ 4 & \text{if } (a,b) = (2,2) \text{ or } (1,k) \text{ with } k \geq 3, \\ 10 & \text{if } (a,b) = (2,3), \\ 12 & \text{if } (a,b) = (2,k) \text{ with } k \geq 4, \\ 28 & \text{if } (a,b) = (3,3), \\ 40 & \text{if } (a,b) = (3,4), \\ 36 & \text{if } (a,b) = (3,k) \text{ with } k \geq 5, \\ 20(1+2^{a-2}) & \text{if } a = b \geq 4, \\ 20 + 18 \cdot 2^{a-2} & \text{if } b-1 = a \geq 4, \\ 20 + 16 \cdot 2^{a-2} & \text{if } b-2 \geq a \geq 4. \end{cases}$$

Proof. By Lemma 4.4, it suffices to find all $(s_1, t_1; s_2, t_2)$ satisfying the conditions

(a) $(s_1, t_1) \in \mathcal{P}(2^a)$ and $(s_2, t_2) \in \mathcal{P}(2^b)$,

(b) $d(s_1, t_1) \leq e(s_2, t_2)$ and $d(s_2, t_2) \leq e(s_1, t_1)$.

By Lemma 4.5, one can get all the lists of $(s_1, t_1; s_2, t_2)$ satisfying the conditions as Table 1. □

Proof of Theorem 1.1. For odd m and n, the number $|\operatorname{RET}_{m,n}|$ of reflexible edge-transitive embeddings of $K_{m,n}$ up to isomorphism is 1 by Theorem 3.1. When exactly one of m and n is odd, then the number $|\operatorname{RET}_{m,n}|$ is counted in Theorem 3.2.

Assume that both m and n are even. Let

$$\begin{aligned} m &= 2^a p_1^{a_1} p_2^{a_2} \cdots p_\ell^{a_\ell} p_{\ell+1}^{a_{\ell+1}} \cdots p_{\ell+f}^{a_{\ell+f}} \quad \text{and} \\ n &= 2^b p_1^{b_1} p_2^{b_2} \cdots p_\ell^{b_\ell} q_{\ell+1}^{a_{\ell+1}} \cdots q_{\ell+g}^{b_{\ell+g}} \quad \text{(prime decompositions)} \end{aligned}$$

and let $\gcd(m,n) = 2^c p_1^{c_1} p_2^{c_2} \cdots p_\ell^{c_\ell}$ with $c_i \geq 1$ for any $i = 1, \ldots, \ell$. Without any loss of generality, assume that $a \leq b$, namely, $a = c$. By Corollary 4.3, the number $|\operatorname{RET}_{m,n}| = |\mathcal{Q}(m,n)|$ is

$$2^{f+g+\ell}(1+p_1^{c_1}) \cdots (1+p_\ell^{c_\ell}) |\mathcal{Q}(2^a, 2^b)|.$$

Theorem 4.6 completes the proof. □

Table 1: All lists of $\mathcal{Q}(2^a, 2^b)$.

(a, b)	$\mathcal{Q}(2^a, 2^b)$
$(1, 1)$	$(0, 1; 0, 1)$
$(1, 2)$	$(0, 1; 0, 1)$, $(0, 1; 1, 1)$
$(1, \geq 3)$	$(0, 1; 0, 1)$, $(0, 1; 2^{b-2}, 1)$, $(0, 1; 2^{b-2} - 1, 2^{b-1} - 1)$, $(0, 1; 2^{b-1} - 1, 2^{b-1} - 1)$
$(2, 2)$	$(0, 1; 0, 1)$, $(0, 1; 1, 1)$, $(1, 1; 0, 1)$, $(1, 1; 1, 1)$
$(2, 3)$	$(0, 1; 0, 1)$, $(0, 1; 2, 1)$, $(0, 1; 1, 3)$, $(0, 1; 3, 3)$, $(1, 1; 0, 1)$, $(1, 1; 1, 1)$, $(1, 1; 2, 1)$, $(1, 1; 3, 1)$, $(1, 1; 1, 3)$, $(1, 1; 3, 3)$
$(2, \geq 4)$	$(0, 1; 0, 1)$, $(0, 1; 2^{b-2}, 1)$, $(0, 1; 2^{b-2} - 1, 2^{b-1} - 1)$, $(0, 1; 2^{b-1} - 1, 2^{b-1} - 1)$, $(1, 1; 0, 1)$, $(1, 1; 2^{b-3}, 1)$, $(1, 1; 2^{b-2}, 1)$, $(1, 1; 3 \cdot 2^{b-3}, 1)$, $(1, 1; 2^{b-2} - 1, 2^{b-2} - 1)$, $(1, 1; 2^{b-1} - 1, 2^{b-2} - 1)$, $(1, 1; 2^{b-2} - 1, 2^{b-1} - 1)$, $(1, 1; 2^{b-1} - 1, 2^{b-1} - 1)$
$(3, 3)$	$(0 \text{ or } 2, 1; 0, 1)$, $(0 \text{ or } 2, 1; 2, 1)$, $(0 \text{ or } 2, 1; 1, 3)$, $(0 \text{ or } 2, 1; 3, 3)$, $(1 \text{ or } 3, 1; 1, 1)$, $(1 \text{ or } 3, 1; 3, 1)$, $(1 \text{ or } 3, 1; 1, 3)$, $(1 \text{ or } 3, 1; 3, 3)$, $(1 \text{ or } 3, 3; 0, 1)$, $(1 \text{ or } 3, 3; 1, 1)$, $(1 \text{ or } 3, 3; 2, 1)$, $(1 \text{ or } 3, 3; 3, 1)$, $(1 \text{ or } 3, 3; 1, 3)$, $(1 \text{ or } 3, 3; 3, 3)$
$(3, 4)$	$(0 \text{ or } 2, 1; 0, 1)$, $(0 \text{ or } 2, 1; 4, 1)$, $(0 \text{ or } 2, 1; 3, 7)$, $(0 \text{ or } 2, 1; 7, 7)$, $(1 \text{ or } 3, 1; 3, 3)$, $(1 \text{ or } 3, 1; 7, 3)$, $(1 \text{ or } 3, 1; 3, 7)$, $(1 \text{ or } 3, 1; 7, 7)$; $(1, 3; x, 1), x = 0, 2, 4, 6;$ $(3, 3; s_2, t_2), (s_2, t_2) \in \mathcal{P}(2^4)$
$(3, \geq 5)$	$(0 \text{ or } 2, 1; 0, 1)$, $(0 \text{ or } 2, 1; 2^{b-2}, 1)$; $(0 \text{ or } 2, 1; x, 2^{b-1} - 1), x = 2^{b-2} - 1 \text{ or } 2^{b-1} - 1;$ $(1 \text{ or } 3, 1; x, y), x, y = 2^{b-2} - 1 \text{ or } 2^{b-1} - 1;$ $(1, 3; i \cdot 2^{b-3}, 1), i = 0, 1, 2, 3;$ $(1, 3; x, y), x, y = 2^{b-2} - 1 \text{ or } 2^{b-1} - 1;$ $(3, 3; i \cdot 2^{b-4}, 1), i = 0, 1, \ldots, 7;$ $(3, 3; x, y), x, y = 2^{b-2} - 1 \text{ or } 2^{b-1} - 1$
$(\geq 4, \geq a)$	$(0 \text{ or } 2^{a-2}, 1; x, y)$, $(x, y) = (0, 1), (2^{b-2}, 1), (2^{b-2} - 1, 2^{b-1} - 1) \text{ or } (2^{b-1} - 1, 2^{b-1} - 1);$ $(2x, 1; 2^{b-2} - 1, 2^{b-1} - 1), (2x, 1; 2^{b-1} - 1, 2^{b-1} - 1),$ $x = 1, 2, \ldots, 2^{a-2} - 1 \ (x \neq 2^{a-3});$ $(x, 1 \text{ or } 2^{a-2} + 1; y, z),$ $x = 1, 3, \ldots, 2^{a-1} - 1, y, z = 2^{b-2} - 1 \text{ or } 2^{b-1} - 1;$ $(2^{a-2} - 1 \text{ or } 2^{a-1} - 1, 2^{a-2} - 1 \text{ or } 2^{a-1} - 1; x, y),$ $x, y = 2^{b-2} - 1 \text{ or } 2^{b-1} - 1;$ $(2^{a-2} - 1, 2^{a-1} - 1; i \cdot 2^{b-a}, 1), i = 0, 1, \ldots, 2^{a-1} - 1;$ Only when $a = b$: $(2^{a-2}-1 \text{ or } 2^{a-1}-1, 2^{a-2}-1; x, 1 \text{ or } 2^{b-2}+1), x = 1, 3, \ldots, 2^{b-1}-1;$ Only when $a = b$: $(2^{a-2} - 1, 2^{a-1} - 1; x, 2^{b-2} + 1), x = 1, 3, \ldots, 2^{b-1} - 1;$ Only when $a = b$ or $b = a + 1$: $(2^{a-1} - 1, 2^{a-1} - 1; x, 1), x = 0, 1, \ldots, 2^{b-1} - 1;$ Only when $a = b$ or $b = a + 1$: $(2^{a-1} - 1, 2^{a-1} - 1; x, 2^{b-2} + 1), x = 1, 3, \ldots, 2^{b-1} - 1;$ Only when $b \geq a + 2$: $(2^{a-1} - 1, 2^{a-1} - 1; i \cdot 2^{b-a-1}, 1), i = 0, 1, \ldots, 2^a - 1$

5 Classification of some groups

In this section, we aim to consider a presentation of the group $\langle x_\alpha, y_\beta\rangle$ for any $(\alpha,\beta)\in \mathrm{RET}_{m,n}$. And we give some sufficient conditions and necessary conditions for $\langle x_{\alpha_1}, y_{\beta_1}\rangle$ and $\langle x_{\alpha_2}, y_{\beta_2}\rangle$ to be isomorphic for any $(\alpha_1,\beta_1),(\alpha_2,\beta_2)\in \mathrm{RET}_{m,n}$. For any positive integers m and n, a group Γ such that

(i) $\Gamma = XY$ for some cyclic groups $X=\langle x\rangle$ of order n and $Y=\langle y\rangle$ of order m with $X\cap Y=\{1_\Gamma\}$ and

(ii) there exists an automorphism of Γ which sends x and y to x^{-1} and y^{-1}, respectively,

is isomorphic to $\langle x_\alpha, y_\beta\rangle$ for some $(\alpha,\beta)\in\mathrm{RET}_{m,n}$. For our convenience, call a group Γ satisfying the conditions (i) and (ii) in the above sentence a *reflexible product of two cyclic groups* of order m and n. Now to classify reflexible products of two cyclic groups of order m and n, it suffices to consider $\langle x_\alpha, y_\beta\rangle$, where $(\alpha,\beta)\in\mathrm{RET}_{m,n}$. Note that for any integers i,j and for any $(\alpha,\beta)\in\mathrm{RET}_{m,n}$,

$$y_\beta^i x_\alpha^j = x_\alpha^{\beta^i(j)} y_\beta^{\alpha^j(i)}.$$

For example, $y_\beta x_\alpha = x_\alpha^{\beta(1)} y_\beta^{\alpha(1)}$ and $y_\beta x_\alpha^2 = x_\alpha^{\beta(2)} y_\beta^{\alpha^2(1)}$.

For odd integers m and n, since $\mathrm{RET}_{m,n}=\{(\mathrm{id},\mathrm{id})\}$, there is a unique reflexible product of two cyclic groups of order m and n up to isomorphism, namely, an abelian group $\mathbb{Z}_m\times\mathbb{Z}_n$.

Let

$$m = p_1^{a_1}p_2^{a_2}\cdots p_\ell^{a_\ell}p_{\ell+1}^{a_{\ell+1}}\cdots p_{\ell+f}^{a_{\ell+f}} \quad \text{(prime factorization)}$$

be odd and

$$n = 2^b p_1^{b_1}p_2^{b_2}\cdots p_\ell^{b_\ell}q_{\ell+1}^{b_{\ell+1}}\cdots q_{\ell+g}^{b_{\ell+g}} \quad \text{(prime factorization)}$$

be even. Let $\gcd(m,n) = p_1^{c_1}p_2^{c_2}\cdots p_\ell^{c_\ell}$ with $c_i\ge 1$ for any $i=1,\dots,\ell$. Now $|\,\mathrm{RET}_{m,n}\,| = 2^f(1+p_1^{c_1})\cdots(1+p_\ell^{c_\ell})$ by Theorem 3.2. Note that for any $(\alpha,\beta)\in\mathrm{RET}_{m,n}$ and for any integer k, $\alpha(k)=rk$, $\beta(2k)=2k$, $\beta(2k+1)=2k+1+2s$ for some integers $r\in[m]$ and $s\in[n]$ satisfying $r^2\equiv 1 \pmod m$, $2s\equiv 0 \pmod{2^b q_{\ell+1}^{b_{\ell+1}}\cdots q_{\ell+g}^{b_{\ell+g}}}$ and for any $j=1,2,\dots,\ell$, $s\equiv 0 \pmod{p_j^{b_j}}$ if $r\equiv 1 \pmod{p_j^{a_j}}$; $s\equiv z\cdot p_j^{b_j-c_j} \pmod{p_j^{b_j}}$ for some integer z with $0\le z\le p_j^{c_j}-1$ if $r\equiv -1 \pmod{p_j^{a_j}}$. Let us denote such α and β by α_r and β_s. Considering commuting rule

$$y_\beta^i x_\alpha^j = x_\alpha^{\beta^i(j)} y_\beta^{\alpha^j(i)},$$

one can check that the centralizer of $\langle x_{\alpha_r}, y_{\beta_s}\rangle$ is

$$\{x_{\alpha_r}^{2i} y_{\beta_s}^j : i\in\left[\frac{n}{2}\right], j(r-1)\equiv 0 \pmod m\} = \langle x_{\alpha_r}^2, y_{\beta_s}^k\rangle,$$

where k is the smallest positive integer j satisfying $j(r-1)\equiv 0 \pmod m$. This implies that for any $(\alpha_{r_1},\beta_{s_1}),(\alpha_{r_2},\beta_{s_2})\in\mathrm{RET}_{m,n}$, if two groups $\langle x_{\alpha_{r_1}}, y_{\beta_{s_1}}\rangle$ and $\langle x_{\alpha_{r_2}}, y_{\beta_{s_2}}\rangle$ are isomorphic, then $r_1=r_2$. Note that

$$y_{\beta_s}x_{\alpha_r} = x_{\alpha_r}^{\beta_s(1)} y_{\beta_s}^{\alpha_r(1)} = x_{\alpha_r}^{2s+1} y_{\beta_s}^r \quad\text{and}$$

$$y_{\beta_s}x_{\alpha_r}^2 = x_{\alpha_r}^{\beta_s(2)} y_{\beta_s}^{\alpha_r^2(1)} = x_{\alpha_r}^2 y_{\beta_s}.$$

In fact, the above two equations determine the whole commuting rules. For any $u \in [m]$ and $v \in [n]$, if v is even, then $y_{\beta_s}^u x_{\alpha_r}^v = x_{\alpha_r}^v y_{\beta_s}^u$, and if v is odd, then

$$\begin{aligned} y_{\beta_s}^u x_{\alpha_r}^v &= x_{\alpha_r}^{v-1} y_{\beta_s}^u x_{\alpha_r} = x_{\alpha_r}^{v-1} y_{\beta_s}^{u-1} x_{\alpha_r}^{2s+1} y_{\beta_s}^r \\ &= x_{\alpha_r}^{v-1+2s} y_{\beta_s}^{u-1} x_{\alpha_r} y_{\beta_s}^r = x_{\alpha_r}^{v-1+2s} y_{\beta_s}^{u-2} x_{\alpha_r}^{2s+1} y_{\beta_s}^{2r} \\ &= x_{\alpha_r}^{v-1+4s} y_{\beta_s}^{u-2} x_{\alpha_r} y_{\beta_s}^{2r} = \cdots = x_{\alpha_r}^{v+2us} y_{\beta_s}^{ur}. \end{aligned}$$

For any $v \in [n]$ with $\gcd(v, n) = 1$,

$$y_{\beta_s} x_{\alpha_r}^v = x_{\alpha_r}^{\beta_s(v)} y_{\beta_s}^{\alpha_r^v(1)} = x_{\alpha_r}^{v+2s} y_{\beta_s}^r = x_{\alpha_r}^{v(2v^{-1}s+1)} y_{\beta_s}^r$$

because v is odd, where v^{-1} is an integer satisfying $vv^{-1} \equiv 1 \pmod{n}$. For any $s_1, s_2 \in [\frac{n}{2}]$ with $\gcd(s_1, n) = \gcd(s_2, n)$, one can choose $v \in [n]$ satisfying that $\gcd(v, n) = 1$ and $v^{-1}s_1 \equiv s_2 \pmod{n}$. Therefore for any $(\alpha_{r_1}, \beta_{s_1}), (\alpha_{r_2}, \beta_{s_2}) \in \mathrm{RET}_{m,n}$, if $r_1 = r_2$ and $\gcd(s_1, n) = \gcd(s_2, n)$ then $\langle x_{\alpha_{r_1}}, y_{\beta_{s_1}} \rangle$ is isomorphic to $\langle x_{\alpha_{r_2}}, y_{\beta_{s_2}} \rangle$. This means that the number of non-isomorphic reflexible product of two cyclic groups of order m and n is at most $2^f(2 + c_1) \cdots (2 + c_\ell)$. So any reflexible product of two cyclic groups of order m and n is isomorphic to

$$\langle x, y \mid x^n = y^m = 1,\ yx = x^{2s+1} y^r,\ yx^2 = x^2 y \rangle$$

for some $r \in [m]$ and $s \in [n]$ satisfying $r^2 \equiv 1 \pmod{m}$, $2s \equiv 0 \pmod{2^b q_{\ell+1}^{b_{\ell+1}} \cdots q_{\ell+g}^{b_{\ell+g}}}$ and for any $j = 1, 2, \ldots, \ell$, $s \equiv 0 \pmod{p_j^{b_j}}$ if $r \equiv 1 \pmod{p_j^{a_j}}$; $s \equiv p_j^{b_j - c_j + z} \pmod{p_j^{b_j}}$ for some integer $z = 0, 1, \ldots, c_j$ if $r \equiv -1 \pmod{p_j^{a_j}}$.

Conversely, assume that for some $(\alpha_{r_1}, \beta_{s_1}), (\alpha_{r_2}, \beta_{s_2}) \in \mathrm{RET}_{m,n}$, $\langle x_{\alpha_{r_1}}, y_{\beta_{s_1}} \rangle$ is isomorphic to $\langle x_{\alpha_{r_2}}, y_{\beta_{s_2}} \rangle$. Let $\psi\colon \langle x_{\alpha_{r_1}}, y_{\beta_{s_1}} \rangle \to \langle x_{\alpha_{r_2}}, y_{\beta_{s_2}} \rangle$ be an isomorphism such that $\psi(x_{\alpha_{r_1}}^u) = x_{\alpha_{r_2}}$ and $\psi(y_{\beta_{s_1}}^v) = y_{\beta_{s_2}}$.

For the remaining case, let

$$\begin{aligned} m &= 2^a p_1^{a_1} p_2^{a_2} \cdots p_\ell^{a_\ell} p_{\ell+1}^{a_{\ell+1}} \cdots p_{\ell+f}^{a_{\ell+f}} \quad \text{and} \\ n &= 2^b p_1^{b_1} p_2^{b_2} \cdots p_\ell^{b_\ell} q_{\ell+1}^{a_{\ell+1}} \cdots q_{\ell+g}^{b_{\ell+g}} \quad \text{(prime decompositions)} \end{aligned}$$

with $\gcd(m, n) = 2^c p_1^{c_1} p_2^{c_2} \cdots p_\ell^{c_\ell}$, where $1 \le a \le b$ and $c_i \ge 1$ for any $i = 1, \ldots, \ell$. For any $(\alpha, \beta) \in \mathrm{RET}_{m,n}$ and for any integer k,

$$\begin{aligned} \alpha(2k) &= 2kt_1, \\ \alpha(2k+1) &= 2kt_1 + 2s_1 + 1, \\ \beta(2k) &= 2kt_2 \quad \text{and} \\ \beta(2k+1) &= 2kt_2 + 2s_2 + 1 \end{aligned}$$

for some $(s_1, t_1; s_2, t_2) \in \mathcal{Q}(m, n)$. Let α and β be such permutations. Note that

$$\begin{aligned} y_\beta x_\alpha &= x_\alpha^{\beta(1)} y_\beta^{\alpha(1)} = x_\alpha^{2s_2+1} y_\beta^{2s_1+1}, \\ y_\beta x_\alpha^2 &= x_\alpha^{\beta(2)} y_\beta^{\alpha^2(1)} = x_\alpha^{2t_2} y_\beta^{2s_1(t_1+1)+1}, \\ y_\beta^2 x_\alpha &= x_\alpha^{\beta^2(1)} y_\beta^{\alpha(2)} = x_\alpha^{2s_2(t_2+1)+1} y_\beta^{2t_1} \quad \text{and} \\ y_\beta^2 x_\alpha^2 &= x_\alpha^{\beta^2(2)} y_\beta^{\alpha^2(2)} = x_\alpha^2 y_\beta^2. \end{aligned}$$

In fact, the above four equations determine the whole commuting rules as follows. For any $i \in [m]$ and $j \in [n]$,

$$\begin{aligned}
y_\beta^{2i} x_\alpha^{2j} &= x_\alpha^{2j} y_\beta^{2i} \\
y_\beta^{2i} x_\alpha^{2j+1} &= x_\alpha^{2j} y_\beta^{2i} x_\alpha = x_\alpha^{2j} y_\beta^{2(i-1)} x_\alpha^{2s_2(t_2+1)+1} y_\beta^{2t_1} \\
&= x_\alpha^{2j+2s_2(t_2+1)} y_\beta^{2(i-1)} x_\alpha y_\beta^{2t_1} = \cdots = x_\alpha^{2j+2is_2(t_2+1)+1} y_\beta^{2it_1} \\
y_\beta^{2i+1} x_\alpha^{2j} &= y_\beta x_\alpha^{2j} y_\beta^{2i} = x_\alpha^{2t_2} y_\beta^{2s_1(t_1+1)+1} x_\alpha^{2(j-1)} y_\beta^{2i} \\
&= x_\alpha^{2t_2} y_\beta x_\alpha^{2(j-1)} y_\beta^{2i+2s_1(t_1+1)} = \cdots = x_\alpha^{2jt_2} y_\beta^{2i+2js_1(t_1+1)+1} \\
y_\beta^{2i+1} x_\alpha^{2j+1} &= y_\beta^{2i} y_\beta x_\alpha x_\alpha^{2j} = y_\beta^{2i} x_\alpha^{2s_2+1} y_\beta^{2s_1+1} x_\alpha^{2j} = x_\alpha^{2s_2} y_\beta^{2i} x_\alpha y_\beta x_\alpha^{2j} y_\beta^{2s_1} \\
&= x_\alpha^{2s_2} (x_\alpha^{2is_2(t_2+1)+1} y_\beta^{2it_1})(x_\alpha^{2jt_2} y_\beta^{2js_1(t_1+1)+1}) y_\beta^{2s_1} \\
&= x_\alpha^{2jt_2+2is_2(t_2+1)+2s_2+1} y_\beta^{2it_1+2js_1(t_1+1)+2s_1+1}.
\end{aligned}$$

So any reflexible product of two cyclic groups of order m and n is isomorphic to

$$\begin{aligned}
\langle x, y \mid x^n = y^m = 1,\ yx = x^{2s_2+1} y^{2s_1+1},\ yx^2 = x^{2t_2} y^{2s_1(t_1+1)+1}, \\
y^2 x = x^{2s_2(t_2+1)+1} y^{2t_1},\ y^2 x^2 = x^2 y^2 \rangle
\end{aligned}$$

for some $(s_1, t_1; s_2, t_2) \in \mathcal{Q}(m, n)$. In summary, we have the following theorem.

Theorem 5.1. *For any positive integers m and n, let Γ be a group such that $\Gamma = XY$ for some cyclic groups $X = \langle x \rangle$ of order n and $Y = \langle y \rangle$ of order m with $X \cap Y = \{1_\Gamma\}$ and there exists an automorphism of Γ which sends x and y to x^{-1} and y^{-1}, respectively.*

(1) If both m and n are odd, Γ is isomorphic to the abelian group $\mathbb{Z}_m \times \mathbb{Z}_n$.

(2) Let

$$m = p_1^{a_1} \cdots p_\ell^{a_\ell} p_{\ell+1}^{a_{\ell+1}} \cdots p_{\ell+f}^{a_{\ell+f}} \quad \textit{(prime factorization)}$$

be odd and let

$$n = 2^b p_1^{b_1} \cdots p_\ell^{b_\ell} q_{\ell+1}^{b_{\ell+1}} \cdots q_{\ell+g}^{b_{\ell+g}} \quad \textit{(prime factorization)}$$

be even with $\gcd(m, n) = p_1^{c_1} \cdots p_\ell^{c_\ell}$, where $c_i \geq 1$ for any $i = 1, \ldots, \ell$. Then Γ is isomorphic to

$$\langle x, y \mid x^n = y^m = 1,\ yx = x^{2s+1} y^r,\ yx^2 = x^2 y \rangle$$

for some $r \in [m]$ and $s \in [\frac{n}{2}]$ satisfying

$$r^2 \equiv 1 \pmod{m}, \qquad 2s \equiv 0 \pmod{2^b q_{\ell+1}^{b_{\ell+1}} \cdots q_{\ell+g}^{b_{\ell+g}}},$$

and for any $j = 1, 2, \ldots, \ell$, $s \equiv 0 \pmod{p_j^{b_j}}$ if

$$r \equiv 1 \pmod{p_j^{a_j}}, \qquad s \equiv p_j^{b_j - c_j + z} \pmod{p_j^{b_j}}$$

for some $z = 0, 1, \ldots, c_j$ if $r \equiv -1 \pmod{p_j^{a_j}}$.

(3) Let

$$m = 2^a p_1^{a_1} \cdots p_\ell^{a_\ell} p_{\ell+1}^{a_{\ell+1}} \cdots p_{\ell+f}^{a_{\ell+f}} \quad \textit{and}$$
$$n = 2^b p_1^{b_1} \cdots p_\ell^{b_\ell} q_{\ell+1}^{a_{\ell+1}} \cdots q_{\ell+g}^{b_{\ell+g}} \quad \textit{(prime factorization)}$$

with $\gcd(m,n) = 2^c p_1^{c_1} p_2^{c_2} \cdots p_\ell^{c_\ell}$*, where* $1 \le a \le b$ *and* $c_i \ge 1$ *for any* $i = 1, \ldots, \ell$*. Now* Γ *is isomorphic to*

$$\langle x, y \mid x^n = y^m = 1,\ yx = x^{2s_2+1} y^{2s_1+1},\ yx^2 = x^{2t_2} y^{2s_1(t_1+1)+1},$$
$$y^2 x = x^{2s_2(t_2+1)+1} y^{2t_1},\ y^2x^2 = x^2y^2\rangle$$

for some $(s_1, t_1; s_2, t_2) \in \mathcal{Q}(m,n)$*.*

For any positive integers m and n and for any $(\alpha, \beta), (\alpha', \beta') \in \mathrm{RET}_{m,n}$, we do not know a necessary and sufficient condition for $\langle x_\alpha, y_\beta \rangle \simeq \langle x_{\alpha'}, y_{\beta'} \rangle$. So we propose the following problem.

Problem 5.2. For any positive integers m and n and for any $(\alpha, \beta), (\alpha', \beta') \in \mathrm{RET}_{m,n}$, find a necessary and sufficient condition for $\langle x_\alpha, y_\beta \rangle \simeq \langle x_{\alpha'}, y_{\beta'} \rangle$. Consequently calculate the number of reflexible products of two cyclic groups of order m and n up to isomorphism.

References

[1] S.-F. Du, G. Jones, J. H. Kwak, R. Nedela and M. Škoviera, Regular embeddings of $K_{n,n}$ where n is a power of 2. I: Metacyclic case, *European J. Combin.* **28** (2007), 1595–1609, doi:10.1016/j.ejc.2006.08.012.

[2] S.-F. Du, G. Jones, J. H. Kwak, R. Nedela and M. Škoviera, Regular embeddings of $K_{n,n}$ where n is a power of 2. II: The non-metacyclic case, *European J. Combin.* **31** (2010), 1946–1956, doi:10.1016/j.ejc.2010.01.009.

[3] J. E. Graver and M. E. Watkins, Locally finite, planar, edge-transitive graphs, *Mem. Amer. Math. Soc.* **126** (1997), no. 601, doi:10.1090/memo/0601.

[4] G. Jones, R. Nedela and M. Škoviera, Complete bipartite graphs with a unique regular embedding, *J. Comb. Theory Ser. B* **98** (2008), 241–248, doi:10.1016/j.jctb.2006.07.004.

[5] G. A. Jones, Regular embeddings of complete bipartite graphs: classification and enumeration, *Proc. Lond. Math. Soc.* **101** (2010), 427–453, doi:10.1112/plms/pdp061.

[6] G. A. Jones, R. Nedela and M. Škoviera, Regular embeddings of $K_{n,n}$ where n is an odd prime power, *European J. Combin.* **28** (2007), 1863–1875, doi:10.1016/j.ejc.2005.07.021.

[7] J. H. Kwak and Y. S. Kwon, Regular orientable embeddings of complete bipartite graphs, *J. Graph Theory* **50** (2005), 105–122, doi:10.1002/jgt.20097.

[8] J. H. Kwak and Y. S. Kwon, Classification of nonorientable regular embeddings of complete bipartite graphs, *J. Comb. Theory Ser. B* **101** (2011), 191–205, doi:10.1016/j.jctb.2011.03.003.

[9] Y. S. Kwon, Classification of reflexible edge-transitive embeddings of $K_{m,n}$ for odd m, n, *East Asian Math. J.* **25** (2009), 533–541, `https://ynmath.jams.or.kr/jams/download/KCI_FI001404071.pdf`.

[10] R. Nedela, M. Škoviera and A. Zlatoš, Regular embeddings of complete bipartite graphs, *Discrete Math.* **258** (2002), 379–381, doi:10.1016/s0012-365x(02)00539-3.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 585–608
https://doi.org/10.26493/1855-3974.1651.e79
(Also available at http://amc-journal.eu)

A Carlitz type result for linearized polynomials*

Bence Csajbók †

*MTA–ELTE Geometric and Algebraic Combinatorics Research Group,
ELTE Eötvös Loránd University, Budapest, Hungary,
Department of Geometry, 1117 Budapest, Pázmány P. stny. 1/C, Hungary*

Giuseppe Marino

Dipartimento di Matematica e Fisica, Università degli Studi della Campania "Luigi Vanvitelli", Viale Lincoln 5, I-81100 Caserta, Italy and
Dipartimento di Matematica e Applicazioni "Renato Caccioppoli", Università degli Studi di Napoli "Federico II", Via Cintia, Monte S.Angelo I-80126 Napoli, Italy

Olga Polverino

Dipartimento di Matematica e Fisica, Università degli Studi della Campania "Luigi Vanvitelli", Viale Lincoln 5, I-81100 Caserta, Italy

Received 21 March 2018, accepted 9 October 2018, published online 30 April 2019

Abstract

For an arbitrary q-polynomial f over $\mathbb{F}_{q^n}$ we study the problem of finding those q-polynomials g over $\mathbb{F}_{q^n}$ for which the image sets of $f(x)/x$ and $g(x)/x$ coincide. For $n \leq 5$ we provide sufficient and necessary conditions and then apply our result to study maximum scattered linear sets of $\mathrm{PG}(1, q^5)$.

Keywords: Linearized polynomial, linear set, direction.

Math. Subj. Class.: 11T06, 51E20

[*] The research was supported by the Italian National Group for Algebraic and Geometric Structures and their Applications (GNSAGA - INdAM).

[†] The first author is supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. The first author acknowledges the support of OTKA Grant No. K 124950.

E-mail addresses: csajbokb@cs.elte.hu (Bence Csajbók), giuseppe.marino@unicampania.it, giuseppe.marino@unina.it (Giuseppe Marino), olga.polverino@unicampania.it (Olga Polverino)

1 Introduction

Let $\mathbb{F}_{q^n}$ denote the finite field of q^n elements where $q = p^h$ for some prime p. For $n > 1$ and $s \mid n$ the trace and norm over $\mathbb{F}_{q^s}$ of elements of $\mathbb{F}_{q^n}$ are defined as $\mathrm{Tr}_{q^n/q^s}(x) = x + x^{q^s} + \cdots + x^{q^{n-s}}$ and $\mathrm{N}_{q^n/q^s}(x) = x^{1+q^s+\cdots+q^{n-s}}$, respectively. When $s = 1$ then we will simply write $\mathrm{Tr}(x)$ and $\mathrm{N}(x)$. Every function $f\colon \mathbb{F}_{q^n} \to \mathbb{F}_{q^n}$ can be given uniquely as a polynomial with coefficients in $\mathbb{F}_{q^n}$ and of degree at most $q^n - 1$. The function f is $\mathbb{F}_q$-linear if and only if it is represented by a *q-polynomial*, that is,

$$f(x) = \sum_{i=0}^{n-1} a_i x^{q^i} \tag{1.1}$$

with coefficients in $\mathbb{F}_{q^n}$. Such polynomials are also called *linearized*. If f is given as in (1.1), then its adjoint (w.r.t. the symmetric non-degenerate bilinear form defined by $\langle x, y\rangle = \mathrm{Tr}(xy)$) is

$$\hat{f}(x) := \sum_{i=0}^{n-1} a_i^{q^{n-i}} x^{q^{n-i}},$$

i.e. $\mathrm{Tr}(xf(y)) = \mathrm{Tr}(y\hat{f}(x))$ for any $x, y \in \mathbb{F}_{q^n}$.

The aim of this paper is to study what can be said about two q-polynomials f and g over $\mathbb{F}_{q^n}$ if they satisfy

$$\mathrm{Im}\left(\frac{f(x)}{x}\right) = \mathrm{Im}\left(\frac{g(x)}{x}\right), \tag{1.2}$$

where by $\mathrm{Im}(f(x)/x)$ we mean the image of the rational function $f(x)/x$, i.e. $\{f(x)/x : x \in \mathbb{F}_{q^n}^*\}$.

For a given q-polynomial f, the equality (1.2) clearly holds with $g(x) = f(\lambda x)/\lambda$ for each $\lambda \in \mathbb{F}_{q^n}^*$. It is less obvious that (1.2) holds also for $g(x) = \hat{f}(\lambda x)/\lambda$, cf. [2, Lemma 2.6] and the first part of [8, Section 3], see also the proof of [18, Theorem 3.3.9].

When one of the functions in (1.2) is a monomial then the answer to the question posed above follows from McConnel's generalization [25, Theorem 1] of a result due to Carlitz [7] (see also Bruen and Levinger [6]).

Theorem 1.1 ([25, Theorem 1]). *Let p denote a prime, $q = p^h$, and $1 < d$ a divisor of $q - 1$. Also, let $F\colon \mathbb{F}_q \to \mathbb{F}_q$ be a function such that $F(0) = 0$ and $F(1) = 1$. Then*

$$(F(x) - F(y))^{\frac{q-1}{d}} = (x - y)^{\frac{q-1}{d}}$$

for all $x, y \in \mathbb{F}_q$ if and only if $F(x) = x^{p^j}$ for some $0 \le j < h$ and $d \mid p^j - 1$.

Indeed, when the function F of Theorem 1.1 is $\mathbb{F}_q$-linear, we easily get the following corollary (see Section 2 for the proof, or [16, Corollary 1.4] for the case when q is an odd prime).

Corollary 1.2. *Let $g(x)$ and $f(x) = \alpha x^{q^k}$, $q = p^h$, be q-polynomials over $\mathbb{F}_{q^n}$ satisfying Condition (1.2). Denote $\gcd(k, n)$ by t. Then $g(x) = \beta x^{q^s}$ with $\gcd(s, n) = t$ for some β with $\mathrm{N}_{q^n/q^t}(\alpha) = \mathrm{N}_{q^n/q^t}(\beta)$.*

Another case for which we know a complete answer to our problem is when $f(x) = \mathrm{Tr}(x)$.

Theorem 1.3 ([8, Theorem 3.7]). *Let $f(x) = \mathrm{Tr}(x)$ and let $g(x)$ be a q-polynomial over $\mathbb{F}_{q^n}$ such that*

$$\mathrm{Im}(f(x)/x) = \mathrm{Im}(g(x)/x).$$

Then $g(x) = \mathrm{Tr}(\lambda x)/\lambda$ for some $\lambda \in \mathbb{F}^*_{q^n}$.

Note that in Theorem 1.3 we have $\hat{f}(x) = f(x)$ and the only solutions for g are $g(x) = f(\lambda x)/\lambda$, while in Corollary 1.2 we have (up to scalars) $\varphi(n)$ different solutions for g, where φ is the Euler's totient function.

The problem posed in (1.2) is also related to the study of the directions determined by an additive function. Indeed, when f is additive, then

$$\mathrm{Im}(f(x)/x) = \left\{ \frac{f(x) - f(y)}{x - y} : x \neq y,\ x, y \in \mathbb{F}_{q^n} \right\},$$

is the *set of directions* determined by the graph of f, i.e. by the point set $\mathcal{G}_f := \{(x, f(x)) : x \in \mathbb{F}_{q^n}\} \subset \mathrm{AG}(2, q^n)$. Hence, in this setting, the problem posed in (1.2) corresponds to finding the $\mathbb{F}_q$-linear functions whose graph determines the same set of directions. The size of $\mathrm{Im}(f(x)/x)$ (for any f, not necessarily additive) was studied extensively. When f is $\mathbb{F}_q$-linear the following result holds.

Theorem 1.4 ([1, 3]). *Let f be a q-polynomial over $\mathbb{F}_{q^n}$, with maximum field of linearity $\mathbb{F}_q$. Then*

$$q^{n-1} + 1 \leq |\,\mathrm{Im}(f(x)/x)| \leq \frac{q^n - 1}{q - 1}.$$

The classical examples which show the sharpness of these bounds are the monomial functions x^{q^s}, with $\gcd(s, n) = 1$, and the $\mathrm{Tr}(x)$ function. However, these bounds are also achieved by other polynomials which are not "equivalent" to these examples (see Section 2 for more details).

Two $\mathbb{F}_q$-linear polynomials $f(x)$ and $h(x)$ of $\mathbb{F}_{q^n}[x]$ are *equivalent* if the two graphs $\mathcal{G}_f$ and $\mathcal{G}_h$ are equivalent under the action of the group $\Gamma\mathrm{L}(2, q^n)$, i.e. if there exists an element $\varphi \in \Gamma\mathrm{L}(2, q^n)$ such that $\mathcal{G}_f^{\varphi} = \mathcal{G}_h$. In such a case, we say that f and h are *equivalent* (via φ) and we write $h = f_\varphi$. It is easy to see that in this way we defined an equivalence relation on the set of q-polynomials over $\mathbb{F}_{q^n}$. If f and g are two q-polynomials such that $\mathrm{Im}(f(x)/x) = \mathrm{Im}(g(x)/x)$, then $\mathrm{Im}(f_\varphi(x)/x) = \mathrm{Im}(g_\varphi(x)/x)$ for any admissible $\varphi \in \Gamma\mathrm{L}(2, q^n)$ (see Proposition 2.6). This means that the problem posed in (1.2) can be investigated up to equivalence.

For $n \leq 4$, the only solutions for g in problem (1.2) are the trivial ones, i.e. either $g(x) = f(\lambda x)/x$ or $g(x) = \hat{f}(\lambda x)/x$ (cf. Theorem 2.8).

For the case $n = 5$, in Section 4, we prove the following main result.

Theorem 1.5. *Let $f(x)$ and $g(x)$ be two q-polynomials over $\mathbb{F}_{q^5}$, with maximum field of linearity $\mathbb{F}_q$, such that $\mathrm{Im}(f(x)/x) = \mathrm{Im}(g(x)/x)$. Then either there exists $\varphi \in \Gamma\mathrm{L}(2, q^5)$ such that $f_\varphi(x) = \alpha x^{q^i}$ and $g_\varphi(x) = \beta x^{q^j}$ with $\mathrm{N}(\alpha) = \mathrm{N}(\beta)$ for some $i, j \in \{1, 2, 3, 4\}$, or there exists $\lambda \in \mathbb{F}^*_{q^5}$ such that $g(x) = f(\lambda x)/\lambda$ or $g(x) = \hat{f}(\lambda x)/\lambda$.*

Finally, the relation between $\mathrm{Im}(f(x)/x)$ and the linear sets of rank n of the projective line $\mathrm{PG}(1, q^n)$ will be pointed out in Section 5. As an application of Theorem 1.5 we get a criterium of $\mathrm{P}\Gamma\mathrm{L}(2, q^5)$-equivalence for linear sets in $\mathrm{PG}(1, q^5)$ and this allows us to prove that the family of (maximum scattered) linear sets of rank n and of size $(q^n-1)/(q-1)$ in $\mathrm{PG}(1,q^n)$ found by Sheekey in [27] contains members which are not-equivalent to the previously known linear sets of this size.

2 Background and preliminary results

Let us start this section by the following immediate corollary of Theorem 1.4.

Proposition 2.1. *If* $\mathrm{Im}(f(x)/x)=\mathrm{Im}(g(x)/x)$ *for two q-polynomials f and g over* $\mathbb{F}_{q^n}$*, then their maximum fields of linearity coincide.*

Proof. Let $\mathbb{F}_{q^m}$ and $\mathbb{F}_{q^k}$ be the maximum fields of linearity of f and g, respectively. Suppose to the contrary $m<k$. Then $|\mathrm{Im}(g(x)/x)|\leq(q^n-1)/(q^k-1)<q^{n-k+1}+1\leq q^{n-m}+1\leq|\mathrm{Im}(f(x)/x)|$, a contradiction. □

Now we are able to prove Corollary 1.2.

Proof. The maximum field of linearity of $f(x)$ is $\mathbb{F}_{q^t}$, thus, by Proposition 2.1, $g(x)$ has to be a q^t-polynomial as well. Then for $t>1$ the result follows from the $t=1$ case (after substituting q for q^t and n/t for n) and hence we can assume that $f(x)$ and $g(x)$ are strictly $\mathbb{F}_q$-linear. By (1.2), we note that $g(1)=\alpha z_0^{q^k-1}$, for some $z_0\in\mathbb{F}_{q^n}^*$. Let $F(x):=g(x)/g(1)$, then F is a q-polynomial over $\mathbb{F}_{q^n}$, with $F(0)=0$ and $F(1)=1$. Also, from (1.2), for each $x\in\mathbb{F}_{q^n}^*$ there exists $z\in\mathbb{F}_{q^n}^*$ such that

$$\frac{F(x)}{x}=\left(\frac{z}{z_0}\right)^{q^k-1}.$$

This means that for each $x\in\mathbb{F}_{q^n}^*$ we get $\mathrm{N}\left(\frac{F(x)}{x}\right)=1$. By Theorem 1.1 (applied to the q-polynomial F with $d=q-1\mid q^n-1$ and using the fact that F is additive) it follows that $F(x)=x^{p^j}$ for some $0\leq j<nh$. Then Proposition 2.1 yields $p^j=q^s$ with $\gcd(s,n)=1$. We get the first part of the statement by putting $\beta=g(1)$. Then from the assumption (1.2) it is easy to deduce $\mathrm{N}(\alpha)=\mathrm{N}(\beta)$. □

We will use the following definition.

Definition 2.2. Let f and g be two equivalent q-polynomials over $\mathbb{F}_{q^n}$ via the element $\varphi\in\Gamma\mathrm{L}(2,q^n)$ represented by the invertible matrix

$$\begin{pmatrix} a & b\\ c & d\end{pmatrix}$$

and with companion automorphism σ of $\mathbb{F}_{q^n}$. Then

$$\left\{\begin{pmatrix} x\\ g(x)\end{pmatrix}:x\in\mathbb{F}_{q^n}\right\}=\left\{\begin{pmatrix} a & b\\ c & d\end{pmatrix}\begin{pmatrix} x^\sigma\\ f(x)^\sigma\end{pmatrix}:x\in\mathbb{F}_{q^n}\right\}. \tag{2.1}$$

Let

$$K_f^\varphi(x)=ax^\sigma+bf(x)^\sigma$$

and

$$H_f^\varphi(x)=cx^\sigma+df(x)^\sigma.$$

Proposition 2.3. *Let f and g be q-polynomials over $\mathbb{F}_{q^n}$ such that $g = f_\varphi$ for some $\varphi \in \Gamma\mathrm{L}(2, q^n)$. Then K_f^φ is invertible and $g(x) = H_f^\varphi((K_f^\varphi)^{-1}(x))$.*

Proof. It easily follows from (2.1). □

From (2.1) it is also clear that

$$\operatorname{Im}\left(\frac{f_\varphi(x)}{x}\right) = \left\{\frac{c + dz^\sigma}{a + bz^\sigma} : z \in \operatorname{Im}\left(\frac{f(x)}{x}\right)\right\} \tag{2.2}$$

and hence

$$|\operatorname{Im}(f_\varphi(x)/x)| = |\operatorname{Im}(f(x)/x)|. \tag{2.3}$$

From Equation (2.3) and Theorem 1.4 the next result easily follows.

Proposition 2.4. *If two q-polynomials over $\mathbb{F}_{q^n}$ are equivalent, then their maximum fields of linearity coincide.*

Note that $|\operatorname{Im}(g(x)/x)| = |\operatorname{Im}(f(x)/x)|$ does not imply the equivalence of f and g. In fact, in the last section we will list the known examples of q-polynomials f which are not equivalent to monomials but the size of $\operatorname{Im}(f(x)/x)$ is maximal. To find such functions was also proposed in [16] and, as it was observed by Sheekey, they determine certain MRD-codes [27].

Let us give the following definition.

Definition 2.5. An element $\varphi \in \Gamma\mathrm{L}(2, q^n)$ represented by the invertible matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

and with companion automorphism σ of $\mathbb{F}_{q^n}$ is said to be *admissible* w.r.t. a given q-polynomial f over $\mathbb{F}_{q^n}$ if either $b = 0$ or $-(a/b)^{\sigma^{-1}} \notin \operatorname{Im}(f(x)/x)$.

The following results will be useful later in the paper.

Proposition 2.6. *If* $\operatorname{Im}(f(x)/x) = \operatorname{Im}(g(x)/x)$ *for some q-polynomials over $\mathbb{F}_{q^n}$, then* $\operatorname{Im}(f_\varphi(x)/x) = \operatorname{Im}(g_\varphi(x)/x)$ *holds for each admissible $\varphi \in \Gamma\mathrm{L}(2, q^n)$.*

Proof. From $\operatorname{Im}(f(x)/x) = \operatorname{Im}(g(x)/x)$ it follows that any $\varphi \in \Gamma\mathrm{L}(2, q^n)$ admissible w.r.t. f is admissible w.r.t. g as well. Hence K_f^φ and K_g^φ are both invertible and we may construct f_φ and g_φ as indicated in Proposition 2.3. The statement now follows from Equation (2.2). □

Proposition 2.7. *Let f and g be q-polynomials over $\mathbb{F}_{q^n}$ and take some $\varphi \in \Gamma\mathrm{L}(2, q^n)$ with companion automorphism σ. Then $g_\varphi(x) = f_\varphi(\lambda^\sigma x)/\lambda^\sigma$ for some $\lambda \in \mathbb{F}_{q^n}^*$ if and only if $g(x) = f(\lambda x)/\lambda$.*

Proof. First we prove the "if" part. Since $g(x) = f(\lambda x)/\lambda = (\omega_{1/\lambda} \circ f \circ \omega_\lambda)(x)$, where ω_α denotes the scalar map $x \in \mathbb{F}_{q^n} \mapsto \alpha x \in \mathbb{F}_{q^n}$, direct computations show that $H_g^\varphi = \omega_{1/\lambda^\sigma} \circ H_f^\varphi \circ \omega_\lambda$ and $K_g^\varphi = \omega_{1/\lambda^\sigma} \circ K_f^\varphi \circ \omega_\lambda$. Then $g_\varphi = \omega_{1/\lambda^\sigma} \circ f_\varphi \circ \omega_{\lambda^\sigma}$ and the first part of the statement follows. The "only if" part follows from the "if" part applied to $g_\varphi(x) = f_\varphi(\lambda^\sigma x)/\lambda^\sigma$ and φ^{-1}; and from $(f_\varphi)_{\varphi^{-1}} = f$ and $(g_\varphi)_{\varphi^{-1}} = g$. □

Next we summarize what is known about problem (1.2) for $n \leq 4$.

Theorem 2.8. *Suppose* $\mathrm{Im}(f(x)/x) = \mathrm{Im}(g(x)/x)$ *for some q-polynomials over* $\mathbb{F}_{q^n}$*,* $n \leq 4$*, with maximum field of linearity* $\mathbb{F}_q$*. Then there exist* $\varphi \in \mathrm{GL}(2, q^n)$ *and* $\lambda \in \mathbb{F}_{q^n}^*$ *such that the following holds.*

- *If* $n = 2$ *then* $f_\varphi(x) = x^q$ *and* $g(x) = f(\lambda x)/\lambda$*.*
- *If* $n = 3$ *then either*
$$f_\varphi(x) = \mathrm{Tr}(x) \quad \textit{and} \quad g(x) = f(\lambda x)/\lambda$$
or
$$f_\varphi(x) = x^q \quad \textit{and} \quad g(x) = f(\lambda x)/\lambda \quad \textit{or} \quad g(x) = \hat{f}(\lambda x)/\lambda.$$
- *If* $n = 4$ *then* $g(x) = f(\lambda x)/\lambda$ *or* $g(x) = \hat{f}(\lambda x)/\lambda$*.*

Proof. In the $n = 2$ case $f(x) = ax + bx^q$, $b \neq 0$. Let φ be represented by the matrix
$$\begin{pmatrix} 1 & 0 \\ -a/b & 1/b \end{pmatrix}.$$
Then $\varphi \in \mathrm{GL}(2, q^2)$ maps $f(x)$ to x^q. Then Proposition 2.6 and Corollary 1.2 give $g_\varphi(x) = f_\varphi(\mu x)/\mu$ and hence Proposition 2.7 gives $g(x) = f(\lambda x)/\lambda$ for some $\lambda \in \mathbb{F}_{q^n}$. If $n = 3$ then according to [21, Theorem 5] and [8, Theorem 1.3] there exists $\varphi \in \mathrm{GL}(2, q^3)$ such that either $f_\varphi(x) = \mathrm{Tr}(x)$ or $f_\varphi(x) = x^q$. In the former case Proposition 2.6 and Theorem 1.3 give $g_\varphi(x) = f_\varphi(\mu x)/\mu$ and the assertion follows from Proposition 2.7. In the latter case Proposition 2.6 and Corollary 1.2 give $g_\varphi(x) = \alpha x^{q^i}$ where $i \in \{1, 2\}$ and $\mathrm{N}(\alpha) = 1$. If $i = 1$, then $g_\varphi(x) = f_\varphi(\mu x)/\mu$ where $\mu^{q-1} = \alpha$ and the assertion follows from Proposition 2.7. Let now $i = 2$ and denote by
$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$
the matrix of φ^{-1}. Also, let Δ denote the determinant of this matrix and recall that $f_\varphi(x) = x^q$, with $\varphi \in \mathrm{GL}(2, q^3)$. Then by Proposition 2.3
$$K_{f_\varphi}^{\varphi^{-1}}(x) = Ax + Bx^q$$
is invertible and its inverse is the map
$$\psi(x) := \frac{A^{q+q^2}x - A^{q^2}Bx^q + B^{1+q}x^{q^2}}{\mathrm{N}(A) + \mathrm{N}(B)}.$$
Also, by Proposition 2.3 we have
$$(f_\varphi)_{\varphi^{-1}}(x) = C\psi(x) + D\psi(x)^q,$$
which gives $f(x) = (f_\varphi)_{\varphi^{-1}}(x)$.

Using similar arguments, since $\mathrm{N}(\alpha) = 1$, direct computations show
$$g(x) = (g_\varphi)_{\varphi^{-1}}(x) = \frac{(A^{q+q^2}C + B^{q+q^2}D)x - B^{q^2}\Delta\alpha^{q^2+1}x^q + A^q\Delta\alpha x^{q^2}}{\mathrm{N}(A) + \mathrm{N}(B)},$$
and hence $g(x) = \hat{f}(\lambda x)/\lambda$ for each $\lambda \in \mathbb{F}_{q^3}^*$ with $\lambda^{q-1} = \Delta^{1-q}/\alpha^q$.

The case $n = 4$ is [8, Proposition 4.2]. □

Remark 2.9. Theorem 2.8 yields that there is a unique equivalence class of q-polynomials, with maximum field of linearity $\mathbb{F}_q$, when $n = 2$. For $n = 3$ there are two non-equivalent classes and they correspond to the classical examples: $\mathrm{Tr}(x)$ and x^q. Whereas, for $n = 4$, from [8, Sec. 5.3] and [5, Table on p. 54], there exist at least eight non-equivalent classes. The possible sizes for the sets of directions determined by these strictly $\mathbb{F}_q$-linear functions are q^3+1, q^3+q^2-q+1, q^3+q^2+1 and q^3+q^2+q+1 and each of them is determined by at least two non-equivalent q-polynomials. Also, by [13, Theorem 3.4], if f is a q-polynomial over $\mathbb{F}_{q^4}$ for which the set of directions is of maximum size then f is equivalent either to x^q or to $\delta x^q + x^{q^3}$, for some $\delta \in \mathbb{F}_{q^4}^*$ with $N(\delta) \neq 1$ (see [23]).

3 Preliminary results about $\mathrm{Tr}(x)$ and the monomial q-polynomials over $\mathbb{F}_{q^5}$

Let q be a power of a prime p. We will need the following results.

Proposition 3.1. *Let $f(x) = \sum_{i=0}^{4} a_i x^{q^i}$ and $g(x) = \mathrm{Tr}(x)$ be q-polynomials over $\mathbb{F}_{q^5}$. Then there is an element $\varphi \in \Gamma\mathrm{L}(2, q^5)$ such that $\mathrm{Im}(f_\varphi(x)/x) = \mathrm{Im}(g(x)/x)$ if and only if $a_1a_2a_3a_4 \neq 0$, $(a_1/a_2)^q = a_2/a_3$, $(a_2/a_3)^q = a_3/a_4$ and $\mathrm{N}(a_1) = \mathrm{N}(a_2)$.*

Proof. Let $\varphi \in \Gamma\mathrm{L}(2, q^5)$ such that $\mathrm{Im}(f_\varphi(x)/x) = \mathrm{Im}(g(x)/x)$. By Proposition 2.4, the maximum field of linearity of f is $\mathbb{F}_q$ and by Theorem 1.3 there exists $\lambda \in \mathbb{F}_{q^5}^*$ such that $f_\varphi(x) = \mathrm{Tr}(\lambda x)/\lambda$. This is equivalent to the existence of a, b, c, d, $ad - bc \neq 0$ and $\sigma\colon x \mapsto x^{p^h}$ such that

$$\left\{ \begin{pmatrix} y \\ \mathrm{Tr}(y) \end{pmatrix} : y \in \mathbb{F}_{q^5} \right\} = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x^\sigma \\ f(x)^\sigma \end{pmatrix} : x \in \mathbb{F}_{q^5} \right\}.$$

Then $cx^\sigma + df(x)^\sigma \in \mathbb{F}_q$ for each $x \in \mathbb{F}_{q^5}$. Let $z = x^\sigma$. Then

$$cz + d\sum_{i=0}^{4} a_i^\sigma z^{q^i} = c^q z^q + d^q \sum_{i=0}^{4} a_i^{\sigma q} z^{q^{i+1}},$$

for each z. As polynomials of z, the left and right-hand sides of the above equation coincide modulo $z^{q^5} - z$ and hence comparing coefficients yield

$$\begin{aligned} c + da_0^\sigma &= d^q a_4^{\sigma q}, \\ da_1^\sigma &= c^q + d^q a_0^{\sigma q}, \\ da_{k+1}^\sigma &= d^q a_k^{\sigma q}, \end{aligned}$$

for $k = 1, 2, 3$. If $d = 0$, then $c = 0$, a contradiction. Since $d \neq 0$, if one of a_1, a_2, a_3, a_4 is zero, then all of them are zero and hence f is $\mathbb{F}_{q^5}$-linear. This is not the case, so we have $a_1a_2a_3a_4 \neq 0$. Then the last three equations yield

$$\begin{aligned} \left(\frac{a_1}{a_2}\right)^q &= \frac{a_2}{a_3}, \\ \left(\frac{a_2}{a_3}\right)^q &= \frac{a_3}{a_4}, \end{aligned}$$

and by taking the norm of both sides in $da_2^{\sigma} = d^q a_1^{\sigma q}$ we get $\mathrm{N}(a_1) = \mathrm{N}(a_2)$.

Now assume that the conditions of the assertion hold. It follows that $a_3 = a_2^{q+1}/a_1^q$ and $a_4 = a_3^{q+1}/a_2^q = a_2^{q^2+q+1}/a_1^{q^2+q}$. Let $\alpha_i = a_i/a_1$ for $i = 0, 1, 2, 3, 4$. Then $\alpha_1 = 1$, $\mathrm{N}(\alpha_2) = 1$, $\alpha_3 = \alpha_2^{q+1}$ and $\alpha_4 = \alpha_2^{1+q+q^2}$. We have $\alpha_2 = \lambda^{q-1}$ for some $\lambda \in \mathbb{F}_{q^5}^*$. If

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 - \lambda^{1-q^4} a_0/a_1 & \lambda^{1-q^4}/a_1 \end{pmatrix},$$

then

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ f(x) \end{pmatrix} = \begin{pmatrix} x \\ x + \lambda^{1-q^4} x^q + \lambda^{q-q^4} x^{q^2} + \lambda^{q^2-q^4} x^{q^3} + \lambda^{q^3-q^4} x^{q^4} \end{pmatrix} = \begin{pmatrix} x \\ \mathrm{Tr}(x\lambda^{q^4})/\lambda^{q^4} \end{pmatrix},$$

i.e. $f_\varphi(x) = \mathrm{Tr}(\lambda^{q^4} x)/\lambda^{q^4}$, where φ is defined by the matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$ □

Proposition 3.2. *Let* $f(x) = \sum_{i=0}^{4} a_i x^{q^i}$, *with* $a_1 a_2 a_3 a_4 \neq 0$. *Then there is an element* $\varphi \in \Gamma\mathrm{L}(2, q^5)$ *such that* $\mathrm{Im}(f_\varphi(x)/x) = \mathrm{Im}(x^q/x)$ *if and only if one of the following holds:*

1. $(a_1/a_2)^q = a_2/a_3$, $(a_2/a_3)^q = a_3/a_4$ *and* $\mathrm{N}(a_1) \neq \mathrm{N}(a_2)$, *or*
2. $(a_4/a_1)^{q^2} = a_1/a_3$, $(a_1/a_2)^{q^2} = a_3/a_4$ *and* $\mathrm{N}(a_1) \neq \mathrm{N}(a_3)$.

In both cases, if the condition on the norms does not hold, then

$$\mathrm{Im}(f_\varphi(x)/x) = \mathrm{Im}(\mathrm{Tr}(x)/x).$$

Proof. We first note that the monomials x^{q^i} and $x^{q^{5-i}}$ are equivalent via the map

$$\psi := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Hence, by Corollary 1.2, the statement holds if and only if there exist a, b, c, d, $ad - bc \neq 0$, $\sigma\colon x \mapsto x^{p^h}$ and $i \in \{1, 2\}$ such that

$$\left\{ \begin{pmatrix} y \\ y^{q^i} \end{pmatrix} : y \in \mathbb{F}_{q^5} \right\} = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x^\sigma \\ f(x)^\sigma \end{pmatrix} : x \in \mathbb{F}_{q^5} \right\}. \tag{3.1}$$

If Condition 1 holds then let $\alpha_j = a_j/a_1$ for $j = 0, 1, 2, 3, 4$. So $\alpha_1 = 1$, $\mathrm{N}(\alpha_2) \neq 1$, $\alpha_3 = \alpha_2^{q+1}$, $\alpha_4 = \alpha_2^{1+q+q^2}$ and it turns out that

$$\begin{pmatrix} 1 & \alpha_2^{q^4} \\ \alpha_2^{1+q+q^2+q^3} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\alpha_0 & 1/a_1 \end{pmatrix} \begin{pmatrix} x \\ f(x) \end{pmatrix} = \begin{pmatrix} 1 & \alpha_2^{q^4} \\ \alpha_2^{1+q+q^2+q^3} & 1 \end{pmatrix} \begin{pmatrix} x \\ x^q + \alpha_2 x^{q^2} + \alpha_3 x^{q^3} + \alpha_4 x^{q^4} \end{pmatrix}.$$

Hence (3.1) is satisfied with $i = 1$, $\sigma\colon x \mapsto x$ and

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & \alpha_2^{q^4} \\ \alpha_2^{1+q+q^2+q^3} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\alpha_0 & 1/a_1 \end{pmatrix}.$$

If condition (1.2) holds then let $\alpha_j = a_j/a_3$ for $j = 0, 1, 2, 3, 4$. So $\alpha_3 = 1$, $\mathrm{N}(\alpha_1) \neq 1$, $\alpha_2 = \alpha_1^{1+q+q^3}$, $\alpha_4 = \alpha_1^{1+q^3}$ and (3.1) is satisfied with $i = 2$, $\sigma\colon x \mapsto x$ and

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \alpha_1^{1+q+q^3+q^4} & 1 \\ 1 & \alpha_1^{q^2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\alpha_0 & 1/a_3 \end{pmatrix}.$$

Suppose now that (3.1) holds and put $z = x^\sigma$. Then

$$\Big(za + b\sum_{j=0}^{4} a_j^\sigma z^{q^j}\Big)^{q^i} = cz + d\sum_{j=0}^{4} a_j^\sigma z^{q^j}$$

for each $z \in \mathbb{F}_{q^5}$ and hence, as polynomials in z, the left-hand side and right-hand side of the above equation coincide modulo $z^{q^5} - z$. The coefficients of z, z^{q^i} and z^{q^k} with $i \in \{1, 2\}$ and $k \in \{1, 2, 3, 4\} \setminus \{i\}$ give

$$\begin{aligned} b^{q^i} a_{-i}^{\sigma q^i} &= c + d a_0^\sigma, \\ a^{q^i} + b^{q^i} a_0^{\sigma q^i} &= d a_i^\sigma, \\ b^{q^i} a_{k-i}^{\sigma q^i} &= d a_k^\sigma, \end{aligned}$$

respectively, where the indices are considered modulo 5. Note that $db \neq 0$ since otherwise also $a = c = 0$ and hence $ad - bc = 0$. With $\{r, s, t\} = \{1, 2, 3, 4\} \setminus \{i\}$, the last three equations yield:

$$\begin{aligned} \left(\frac{a_{r-i}}{a_{s-i}}\right)^{q^i} &= \frac{a_r}{a_s}, \\ \left(\frac{a_{s-i}}{a_{t-i}}\right)^{q^i} &= \frac{a_s}{a_t}. \end{aligned}$$

First assume $i = 1$. Then we have

$$\left(\frac{a_1}{a_2}\right)^q = \frac{a_2}{a_3} \quad \text{and} \quad \left(\frac{a_2}{a_3}\right)^q = \frac{a_3}{a_4}.$$

If $\mathrm{N}(a_1) = \mathrm{N}(a_2)$, from Proposition 3.1 and Equation (2.3) it follows that $|\operatorname{Im}(x^q/x)| = |\operatorname{Im}(\operatorname{Tr}(x)/x)|$. Since $|\operatorname{Im}(x^q/x)| = (q^n - 1)/(q - 1)$ and $|\operatorname{Im}(\operatorname{Tr}(x)/x)| = q^{n-1} + 1$, we get a contradiction.

Now assume $i = 2$. Then we have $(a_4/a_1)^{q^2} = a_1/a_3$ and

$$\left(\frac{a_1}{a_2}\right)^{q^2} = \frac{a_3}{a_4}. \tag{3.2}$$

Multiplying these two equations yields $a_4^{q^2+1} = a_1 a_2^{q^2}$ and hence

$$a_2 = a_1^{1+q+q^3} / a_3^{q^3+q}. \tag{3.3}$$

By (3.2) this implies

$$a_4 = a_1^{q^3+1} / a_3^{q^3}. \tag{3.4}$$

If $\mathrm{N}(a_1) = \mathrm{N}(a_3)$, then also $\mathrm{N}(a_1) = \mathrm{N}(a_2) = \mathrm{N}(a_3) = \mathrm{N}(a_4)$. We show that in this case $\mathrm{Im}(f_\varphi(x)/x) = \mathrm{Im}(\mathrm{Tr}(x)/x)$, so we must have $\mathrm{N}(a_1) \neq \mathrm{N}(a_3)$. According to Proposition 3.1 it is enough to show $(a_1/a_2)^q = a_2/a_3$ and $(a_2/a_3)^q = a_3/a_4$. By (3.2) we have $(a_1/a_2)^q = (a_3/a_4)^{q^4}$, which equals a_2/a_3 if and only if $(a_2/a_3)^q = a_3/a_4$, i.e. $a_3^{1+q} = a_4 a_2^q$. Taking into account (3.3) and (3.4), this equality follows from $\mathrm{N}(a_1) = \mathrm{N}(a_3)$. □

4 Proof of the main theorem

In this section we prove Theorem 1.5. In order to do this, we use the following two results and the technique developed in [8].

Lemma 4.1 ([8, Lemma 3.4]). *Let f and g be two linearized polynomials over $\mathbb{F}_{q^n}$. If $\mathrm{Im}(f(x)/x) = \mathrm{Im}(g(x)/x)$, then for each positive integer d the following holds*

$$\sum_{x \in \mathbb{F}_{q^n}^*} \left(\frac{f(x)}{x} \right)^d = \sum_{x \in \mathbb{F}_{q^n}^*} \left(\frac{g(x)}{x} \right)^d.$$

Lemma 4.2 (See for example [8, Lemma 3.5]). *For any prime power q and integer d we have $\sum_{x \in \mathbb{F}_q^*} x^d = -1$ if $q - 1 \mid d$ and $\sum_{x \in \mathbb{F}_q^*} x^d = 0$ otherwise.*

Proposition 4.3. *Let $f(x) = \sum_{i=0}^4 a_i x^{q^i}$ and $g(x) = \sum_{i=0}^4 b_i x^{q^i}$ be two q-polynomials over $\mathbb{F}_{q^5}$ such that $\mathrm{Im}(f(x)/x) = \mathrm{Im}(g(x)/x)$. Then the following relations hold between the coefficients of f and g:*

$$a_0 = b_0, \tag{4.1}$$

$$a_1 a_4^q = b_1 b_4^q, \tag{4.2}$$

$$a_2 a_3^{q^2} = b_2 b_3^{q^2}, \tag{4.3}$$

$$a_1^{q+1} a_3^{q^2} + a_2 a_4^{q+q^2} = b_1^{q+1} b_3^{q^2} + b_2 b_4^{q+q^2}, \tag{4.4}$$

$$a_1 a_2^{q+q^3} + a_3^{1+q^3} a_4^q = b_1 b_2^{q+q^3} + b_3^{1+q^3} b_4^q, \tag{4.5}$$

$$\begin{gathered} a_1^{1+q+q^2} a_2^{q^3} + a_2^{1+q} a_3^{q^2+q^3} + a_1^q a_3^{1+q^2+q^3} + a_1^{q^2} a_2 a_3^{q^3} a_4^q + a_2^{1+q+q^3} a_4^{q^2} + \\ a_1^q a_2^{q^3} a_3 a_4^{q^2} + a_1 a_2^q a_3^{q^2} a_4^{q^3} + a_1^{1+q^2} a_4^{q+q^3} + a_3 a_4^{q+q^2+q^3} = \\ b_1^{1+q+q^2} b_2^{q^3} + b_2^{1+q} b_3^{q^2+q^3} + b_1^q b_3^{1+q^2+q^3} + b_1^{q^2} b_2 b_3^{q^3} b_4^q + b_2^{1+q+q^3} b_4^{q^2} + \\ b_1^q b_2^{q^3} b_3 b_4^{q^2} + b_1 b_2^q b_3^{q^2} b_4^{q^3} + b_1^{1+q^2} b_4^{q+q^3} + b_3 b_4^{q+q^2+q^3}, \end{gathered} \tag{4.6}$$

$$\begin{gathered}
\mathrm{N}(a_1)+\mathrm{N}(a_2)+\mathrm{N}(a_3)+\mathrm{N}(a_4)+\mathrm{Tr}(a_1^q a_2^{q^2+q^3+q^4}a_3+a_1^{q+q^3}a_2^{q^4}a_3^{1+q^2}+\\
a_1^{q+q^2}a_2^{q^3+q^4}a_4+a_1^{q+q^2+q^4}a_3^{q^3}a_4+a_2^q a_3^{q^2+q^3+q^4}a_4+a_1^{q^2}a_3^{q^3+q^4}a_4^{1+q}+\\
a_2^{q+q^3}a_3^{q^4}a_4^{1+q^2}+a_1^{q^2}a_2^{q^4}a_4^{1+q+q^3})=\\
\mathrm{N}(b_1)+\mathrm{N}(b_2)+\mathrm{N}(b_3)+\mathrm{N}(b_4)+\mathrm{Tr}(b_1^q b_2^{q^2+q^3+q^4}b_3+b_1^{q+q^3}b_2^{q^4}b_3^{1+q^2}+\\
b_1^{q+q^2}b_2^{q^3+q^4}b_4+b_1^{q+q^2+q^4}b_3^{q^3}b_4+b_2^q b_3^{q^2+q^3+q^4}b_4+b_1^{q^2}b_3^{q^3+q^4}b_4^{1+q}+\\
b_2^{q+q^3}b_3^{q^4}b_4^{1+q^2}+b_1^{q^2}b_2^{q^4}b_4^{1+q+q^3}).
\end{gathered}\tag{4.7}$$

Proof. Equations (4.1) – (4.5) follow from [8, Lemma 3.6]. To prove (4.6) we will use Lemma 4.1 with $d=q^3+q^2+q+1$. This gives us

$$\sum_{1\le i,j,m,n\le 4} a_i a_j^q a_m^{q^2} a_n^{q^3} \sum_{x\in\mathbb{F}_{q^5}^*} x^{q^i-1+q^{j+1}-q+q^{m+2}-q^2+q^{n+3}-q^3} =$$
$$\sum_{1\le i,j,m\le 4} b_i b_j^q b_m^{q^2} b_n^{q^3} \sum_{x\in\mathbb{F}_{q^5}^*} x^{q^i-1+q^{j+1}-q+q^{m+2}-q^2+q^{n+3}-q^3}.$$

By Lemma 4.2 we have $\sum_{x\in\mathbb{F}_{q^5}^*} x^{q^i-1+q^{j+1}-q+q^{m+2}-q^2+q^{n+3}-q^3}=-1$ if and only if

$$q^i+q^{j+1}+q^{m+2}+q^{n+3}\equiv 1+q+q^2+q^3 \pmod{q^5-1},\tag{4.8}$$

and zero otherwise. Suppose that the former case holds. The right-hand side of (4.8) is smaller than the left-hand side, thus

$$q^i+q^{j+1}+q^{m+2}+q^{n+3}=1+q+q^2+q^3+k(q^5-1),$$

for some positive integer k. We have $q^i+q^{j+1}+q^{m+2}+q^{n+3}\le q^4+q^5+q^6+q^7<1+q+q^2+q^3+(q^2+q+2)(q^5-1)$ and hence $k\le q^2+q+1$. If $i=1$, then $q^2\mid 1-k$ and hence $k=1$, $j=m=1$ and $n=2$, or $k=q^2+1$, $n=4$ and either $j=2$ and $m=3$, or $j=4$ and $m=1$. If $i>1$, then q^2 divides $q+1-k$ and hence $k=q+1$, or $k=q^2+q+1$. In the former case $i=j=n=2$ and $m=4$, or $i=j=2$ and $n=m=3$, or $i=3$, $j=1$, $m=4$ and $n=2$, or $i=3$, $j=1$ and $m=n=3$, or $m=1$, $i=2$, $j=4$ and $n=3$. In the latter case $i=3$ and $n=m=j=4$. Then (4.6) follows.

To prove (4.7) we follow the previous approach with $d=q^4+q^3+q^2+q+1$. We obtain

$$\sum a_i a_j^q a_m^{q^2} a_n^{q^3} a_r^{q^4}=\sum b_i b_j^q b_m^{q^2} b_n^{q^3} b_r^{q^4},$$

where the summation is on the quintuples (i,j,m,n,r) with elements taken from $\{1,2,3,4\}$ such that $L_{i,j,m,n,r}:=(q^i-1)+(q^{j+1}-q)+(q^{m+2}-q^2)+(q^{n+3}-q^3)+(q^{r+4}-q^4)$ is divisible by q^5-1. Then

$$L_{i,j,m,n,r}\equiv K_{i,j',m',n',r'} \pmod{q^5-1},$$

where

$$K_{i,j',m',n',r'}=(q^i-1)+(q^{j'}-q)+(q^{m'}-q^2)+(q^{n'}-q^3)+(q^{r'}-q^4),$$

such that $j' \equiv j+1$, $m' \equiv m+2$, $n' \equiv n+3$, $r' \equiv r+4 \pmod 5$ with

$$j' \in \{0,2,3,4\}, \quad m' \in \{0,1,3,4\}, \quad n' \in \{0,1,2,4\}, \quad r' \in \{0,1,2,3\}. \tag{4.9}$$

For $q=2$ and $q=3$ we can determine by computer those quintuples (i,j',m',n',r') for which $K_{i,j',m',n',r'}$ is divisible by q^5-1 and hence (4.7) follows. So we may assume $q>3$. Then

$$\begin{gathered} 3-q^2-q^3-q^4 = (q-1)+(1-q)+(1-q^2)+(1-q^3)+(1-q^4) \leq \\ K_{i,j',m',n',r'} \leq \\ (q^4-1)+(q^4-q)+(q^4-q^2)+(q^4-q^3)+(q^3-q^4) = 3q^4-1-q-q^2, \end{gathered}$$

and hence $L_{i,j,m,n,r}$ is divisible by q^5-1 if and only if $K_{i,j',m',n',r'}=0$. It follows that

$$q^i + q^{j'} + q^{m'} + q^{n'} + q^{r'} = 1+q+q^2+q^3+q^4. \tag{4.10}$$

For $h \in \{0,1,2,3,4\}$ let c_h denote the number of elements in the multiset $\{i,j',m',n',r'\}$ which equals h. So

$$\sum_{h=0}^{4} c_h q^h = 1+q+q^2+q^3+q^4$$

for some $0 \leq c_h \leq 5$ with $\sum_{h=0}^{4} c_h = 5$. We cannot have $c_0 = 5$ since $q>1$. If $c_i = 5$ for some $1 \leq i \leq 4$ then the left hand side of (4.10) is not congruent to 1 modulo q, a contradiction. It follows that $c_h \leq 4$. Thus for $q>3$ (4.10) holds if and only if $c_h = 1$ for $h=0,1,2,3,4$ and we have to find those quintuples (i,j',m',n',r') for which $i \in \{1,2,3,4\}$, $\{i,j',m',n',r'\} = \{0,1,2,3,4\}$ and (4.9) are satisfied. This can be done by computer and the 44 solutions yield (4.7). □

4.1 Proof of Theorem 1.5

Proof. Since f has maximum field of linearity $\mathbb{F}_q$, we cannot have $a_1=a_2=a_3=a_4=0$. If three of $\{a_1,a_2,a_3,a_4\}$ are zeros, then $f(x) = a_0x + a_ix^{q^i}$, for some $i \in \{1,2,3,4\}$. Hence with φ represented by

$$\begin{pmatrix} 1 & 0 \\ -a_0/a_i & 1/a_i \end{pmatrix}$$

we have $f_\varphi(x) = x^{q^i}$. Then Proposition 2.6 and Corollary 1.2 give $g_\varphi(x) = \beta x^{q^j}$ where $\mathrm{N}(\beta)=1$ and $j \in \{1,2,3,4\}$. Now, we distinguish three main cases according to the number of zeros among $\{a_1,a_2,a_3,a_4\}$.

Two zeros among $\{a_1,a_2,a_3,a_4\}$

Applying Proposition 4.3 we obtain $a_0=b_0$. The two non-zero coefficients can be chosen in six different ways, however the cases $a_1a_2 \neq 0$ and $a_1a_3 \neq 0$ correspond to $a_3a_4 \neq 0$ and $a_2a_4 \neq 0$, respectively, since $\mathrm{Im}(f(x)/x) = \mathrm{Im}(\hat{f}(x)/x)$. Thus, after possibly interchanging f with $\hat{f}$, we may consider only four cases.

First let

$$f(x) = a_0x + a_1x^q + a_4x^{q^4}, \quad a_1a_4 \neq 0.$$

Applying Proposition 4.3 we obtain $0 = b_2 b_3^{q^2}$. Since $b_1 b_4 \neq 0$, from (4.4) we get $b_2 = b_3 = 0$ and hence (4.7) gives

$$\mathrm{N}(a_1) + \mathrm{N}(a_4) = \mathrm{N}(b_1) + \mathrm{N}(b_4).$$

Also, from (4.2) we have $\mathrm{N}(a_1)\,\mathrm{N}(a_4) = \mathrm{N}(b_1)\,\mathrm{N}(b_4)$. It follows that either $\mathrm{N}(a_1) = \mathrm{N}(b_1)$ and $\mathrm{N}(a_4) = \mathrm{N}(b_4)$, or $\mathrm{N}(a_1) = \mathrm{N}(b_4)$ and $\mathrm{N}(a_4) = \mathrm{N}(b_1)$. In the first case $b_1 = a_1 \lambda^{q-1}$ for some $\lambda \in \mathbb{F}_{q^5}^*$ and by (4.2) we get $g(x) = f(\lambda x)/\lambda$. In the latter case $b_1 = a_4^q \lambda^{q-1}$ for some $\lambda \in \mathbb{F}_{q^5}^*$ and by (4.2) we get $g(x) = \hat{f}(\lambda x)/\lambda$.

Now consider

$$f(x) = a_1 x^q + a_3 x^{q^3}, \quad a_1 a_3 \neq 0.$$

Applying Proposition 4.3 and arguing as above we have either $b_2 = b_4 = 0$ or $b_1 = b_3 = 0$. In the first case from (4.6) we obtain

$$a_1^q a_3^{1+q^2+q^3} = b_1^q b_3^{1+q^2+q^3}$$

and together with (4.4) this yields $\mathrm{N}(a_1) = \mathrm{N}(b_1)$ and $\mathrm{N}(a_3) = \mathrm{N}(b_3)$. In this case $g(x) = f(\lambda x)/\lambda$ for some $\lambda \in \mathbb{F}_{q^5}^*$. If $b_1 = b_3 = 0$, then in $\hat{g}(x)$ the coefficients of x^{q^2} and x^{q^4} are zeros thus applying the result obtained in the former case we get $\lambda \hat{g}(x) = f(\lambda x)$ and hence after substituting $y = \lambda x$ and taking the adjoints of both sides we obtain $g(y) = \hat{f}(\mu y)/\mu$, where $\mu = \lambda^{-1}$.

The cases

$$f(x) = a_1 x^q + a_2 x^{q^2} \quad \text{and} \quad f(x) = a_2 x^{q^2} + a_3 x^{q^3}$$

can be handled in a similar way, applying Equations (4.2) – (4.7) of Proposition 4.3.

One zero among $\{a_1, a_2, a_3, a_4\}$

Since $\mathrm{Im}(f(x)/x) = \mathrm{Im}(\hat{f}(x)/x)$, we may assume $a_1 = 0$ or $a_2 = 0$.

First suppose $a_1 = 0$. Then by (4.2) either $b_1 = 0$ or $b_4 = 0$. In the former case putting together Equations (4.3), (4.4), (4.5) we get $\mathrm{N}(a_i) = \mathrm{N}(b_i)$ for $i \in \{2, 3, 4\}$ and hence there exists $\lambda \in \mathbb{F}_{q^5}^*$ such that $g(x) = f(\lambda x)/\lambda$. If $a_1 = b_4 = 0$, then in $\hat{g}(x)$ the coefficient of x^q is zero thus applying the previous result we get $g(x) = \hat{f}(\mu x)/\mu$, where $\mu = \lambda^{-1}$.

Now suppose $a_2 = 0$. Then by (4.3) either $b_2 = 0$ or $b_3 = 0$. Using the same approach but applying (4.2), (4.4) and (4.5) we obtain $g(x) = f(\lambda x)/\lambda$ or $g(x) = \hat{f}(\lambda x)/\lambda$.

Case $a_1 a_2 a_3 a_4 \neq 0$

We will apply (4.1) – (4.6) of Proposition 4.3. Note that Equations (4.2) and (4.3) yield $a_1 a_2 a_3 a_4 \neq 0 \Leftrightarrow b_1 b_2 b_3 b_4 \neq 0$. Multiplying (4.4) by a_2 and applying (4.3) yield

$$a_2^2 a_4^{q+q^2} - a_2 (b_1^{q+1} b_3^{q^2} + b_2 b_4^{q+q^2}) + a_1^{q+1} b_3^{q^2} b_2 = 0.$$

Taking (4.2) into account, this is equivalent to

$$(a_2 a_4^{q+q^2} - b_1^{q+1} b_3^{q^2})(a_2 a_4^{q+q^2} - b_2 b_4^{q+q^2}) = 0.$$

Multiplying (4.5) by a_1 and applying (4.2) yield

$$a_1^2a_2^{q+q^3} - a_1(b_1b_2^{q+q^3} + b_3^{1+q^3}b_4^q) + a_3^{1+q^3}b_4^qb_1 = 0.$$

Taking (4.3) into account, this is equivalent to

$$(a_1a_2^{q+q^3} - b_1b_2^{q+q^3})(a_1a_2^{q+q^3} - b_3^{1+q^3}b_4^q) = 0.$$

We distinguish four cases:

Case 1. $a_2a_4^{q+q^2} = b_1^{q+1}b_3^{q^2}$ and $a_1a_2^{q+q^3} = b_1b_2^{q+q^3}$,

Case 2. $a_2a_4^{q+q^2} = b_1^{q+1}b_3^{q^2}$ and $a_1a_2^{q+q^3} = b_3^{1+q^3}b_4^q$,

Case 3. $a_2a_4^{q+q^2} = b_2b_4^{q+q^2}$ and $a_1a_2^{q+q^3} = b_1b_2^{q+q^3}$,

Case 4. $a_2a_4^{q+q^2} = b_2b_4^{q+q^2}$ and $a_1a_2^{q+q^3} = b_3^{1+q^3}b_4^q$.

We show that these four cases produce the relations:

$$\mathrm{N}\left(\frac{b_1}{a_4}\right) = \frac{a_1a_2^{q+q^3}}{a_4^qa_3^{q^3+1}} = \frac{b_1b_2^{q+q^3}}{b_4^qb_3^{q^3+1}}, \tag{4.11}$$

$$\mathrm{N}\left(\frac{b_1}{a_4}\right) = 1, \tag{4.12}$$

$$\mathrm{N}\left(\frac{b_1}{a_1}\right) = 1, \tag{4.13}$$

$$\mathrm{N}\left(\frac{b_1}{a_1}\right) = \frac{a_3^{q^3+1}a_4^q}{a_1a_2^{q+q^3}} = \frac{b_1b_2^{q+q^3}}{b_3^{q^3+1}b_4^q}, \tag{4.14}$$

respectively.

To see (4.11) observe that from $a_2a_4^{q+q^2} = b_1^{q+1}b_3^{q^2}$ and (4.2) we get

$$\mathrm{N}\left(\frac{b_1}{a_4}\right) = \left(\frac{b_1^{q+1}}{a_4^{q+q^2}}\right)^{q^2+1}\frac{b_1^{q^4}}{a_4} = \left(\frac{a_2^{q^2+1}}{b_3^{q^2+q^4}}\right)\frac{a_1^{q^4}}{b_4} = \frac{a_1a_2^{q+q^3}}{b_4^qb_3^{q^3+1}}, \tag{4.15}$$

where the last equation follows from $\mathrm{N}(b_1/a_4)^q = \mathrm{N}(b_1/a_4)$. Hence by $a_1a_2^{q+q^3} = b_1b_2^{q+q^3}$ and (4.5) we get (4.11).

Equation (4.12) immediately follows from (4.15) taking $a_1a_2^{q+q^3} = b_3^{1+q^3}b_4^q$ into account.

Now we show (4.14). By (4.2), we get

$$\mathrm{N}\left(\frac{b_1}{a_1}\right) = \mathrm{N}\left(\frac{a_4}{b_4}\right) = \left(\frac{a_4}{b_4}\right)^{q+q^2}\left(\left(\frac{a_4}{b_4}\right)^{q+q^2}\right)^{q^2}\left(\frac{a_4}{b_4}\right). \tag{4.16}$$

Since $\mathrm{N}(b_1/a_1)^q = \mathrm{N}(b_1/a_1)$, by $a_2a_4^{q+q^2} = b_2b_4^{q+q^2}$, the previous equation becomes

$$\mathrm{N}\left(\frac{b_1}{a_1}\right) = \frac{b_2^{q^3+q}}{a_2^{q^3+q}}\frac{a_4^q}{b_4^q} \tag{4.17}$$

and taking $a_1 a_2^{q+q^3} = b_3^{1+q^3} b_4^q$ and (4.3) into account we get (4.14).

Equation (4.13) immediately follows from (4.17) taking $a_1 a_2^{q+q^3} = b_1 b_2^{q+q^3}$ and (4.2) into account.

- In Case 3 by (4.13) we get $b_1 = a_1 \lambda^{q-1}$ for some $\lambda \in \mathbb{F}_{q^5}^*$ and by (4.2) and (4.3) we have $g(x) = f(\lambda x)/\lambda$.
- Analogously, in Case 2 $g(x) = \hat{f}(\lambda x)/\lambda$.
- Case 4 is just Case 3 after replacing g by $\hat{g}$ since $\mathrm{Im}(g(x)/x) = \mathrm{Im}(\hat{g}(x)/x)$.

This allows us to restrict ourself to Case 1.

Taking (4.2) and (4.3) into account, it will be useful to express a_1, a_2, a_3 as follows:

$$a_1 = \frac{b_1 b_4^q}{a_4^q}, \quad a_2 = \frac{b_1^{q+1} b_3^{q^2}}{a_4^{q+q^2}}, \quad a_3 = \frac{b_2^{q^3} b_4^{1+q^4}}{a_1^{q^3+q^4}}. \tag{4.18}$$

We are going to simplify (4.6). Using Equations (4.18) and (4.2) it is easy to see that

$$\begin{aligned} &a_2^{1+q} a_3^{q^2+q^3} = b_2^{1+q} b_3^{q^2+q^3}, && a_1^{1+q^2} a_4^{q+q^3} = b_1^{1+q^2} b_4^{q+q^3}, \\ &a_1^{q^2} a_2 a_3^{q^3} a_4^q = b_1 b_2^q b_3^{q^2} b_4^{q^3}, && a_1^q a_2^{q^3} a_3 a_4^{q^2} = b_1^q b_2^{q^3} b_3 b_4^{q^2}, \\ &a_1 a_2^q a_3^{q^2} a_4^{q^3} = b_1^{q^2} b_2 b_3^{q^3} b_4^q \end{aligned}$$

and hence

$$\begin{aligned} a_1^{1+q+q^2} a_2^{q^3} + a_1^q a_3^{1+q^2+q^3} + a_2^{1+q+q^3} a_4^{q^2} + a_3 a_4^{q+q^2+q^3} = \\ b_1^{1+q+q^2} b_2^{q^3} + b_1^q b_3^{1+q^2+q^3} + b_2^{1+q+q^3} b_4^{q^2} + b_3 b_4^{q+q^2+q^3}. \end{aligned} \tag{4.19}$$

The following equations can be proved applying (4.2), (4.3) and (4.18):

$$\mathrm{N}\left(\frac{b_1}{a_4}\right) b_3 b_4^{q+q^2+q^3} = a_2^{q^3} a_1^{1+q+q^2}, \tag{4.20}$$

$$\mathrm{N}\left(\frac{a_4}{b_1}\right) b_4^{q^2} b_2^{1+q+q^3} = a_1^q a_3^{1+q^2+q^3}, \tag{4.21}$$

$$\mathrm{N}\left(\frac{b_1}{a_4}\right) b_1^q b_3^{1+q^2+q^3} = a_2^{1+q+q^3} a_4^{q^2}, \tag{4.22}$$

$$\mathrm{N}\left(\frac{a_4}{b_1}\right) b_2^{q^3} b_1^{1+q+q^2} = a_3 a_4^{q+q^2+q^3}. \tag{4.23}$$

Then (4.19) can be written as

$$\begin{aligned} (\mathrm{N}(b_1/a_4) - 1)\left(b_3 b_4^{q+q^2+q^3} + b_1^q b_3^{1+q^2+q^3}\right) = \\ \frac{\mathrm{N}(b_1/a_4) - 1}{\mathrm{N}(b_1/a_4)} \left(b_4^{q^2} b_2^{1+q+q^3} + b_2^{q^3} b_1^{1+q+q^2}\right). \end{aligned}$$

If $\mathrm{N}(b_1/a_4) = 1$, then (4.15) equals 1 and hence $a_1 a_2^{q+q^3} = b_4^q b_3^{q^3+1}$ which means that we are in Case 2. Then again $g(x) = \hat{f}(\lambda x)/\lambda$.

Otherwise dividing by $\mathrm{N}(b_1/a_4)-1$ and substituting $\mathrm{N}(b_1/a_4)=b_1b_2^{q+q^3}/b_4^qb_3^{q^3+1}$ we obtain

$$b_1b_2^{q+q^3}(b_3b_4^{q+q^2+q^3}+b_1^qb_3^{1+q^2+q^3})=b_4^qb_3^{q^3+1}(b_4^{q^2}b_2^{1+q+q^3}+b_2^{q^3}b_1^{1+q+q^2}).$$

Substituting $\mathrm{N}(b_1/a_4)b_4^qb_3^{q^3+1}/b_2^{q+q^3}$ for b_1 and using the fact that $\mathrm{N}(b_1/a_4)\in\mathbb{F}_q$ we obtain

$$\left(1-\mathrm{N}\left(\frac{b_1}{a_4}\right)^2\mathrm{N}\left(\frac{b_3}{b_2}\right)\right)\left(\mathrm{N}\left(\frac{b_1}{a_4}\right)b_4^{q+q^3}b_3-b_2^{1+q+q^3}\right)=0.$$

This gives us two possibilities:

$$\mathrm{N}\left(\frac{b_1}{a_4}\right)b_4^{q+q^3}b_3=b_2^{1+q+q^3}, \tag{4.24}$$

or

$$\mathrm{N}\left(\frac{b_2}{b_3}\right)=\mathrm{N}\left(\frac{b_1}{a_4}\right)^2. \tag{4.25}$$

First consider the case when (4.25) holds.

We show $\mathrm{N}(a_1)=\mathrm{N}(b_1)$, that is, (4.13). We have $a_2a_4^{q+q^2}=b_1^{q+1}b_3^{q^2}$ from (4.18) and hence $\mathrm{N}(a_2)\,\mathrm{N}(a_4)^2=\mathrm{N}(b_1)^2\,\mathrm{N}(b_3)$. It follows that

$$\mathrm{N}\left(\frac{b_1}{a_4}\right)^2=\mathrm{N}\left(\frac{a_2}{b_3}\right).$$

Combining this with (4.25) we obtain $\mathrm{N}(b_2)=\mathrm{N}(a_2)$. Then $\mathrm{N}(b_1)=\mathrm{N}(a_1)$ follows from $a_1a_2^{q+q^3}=b_1b_2^{q+q^3}$ since we are in Case 1.

From now on we can suppose that (4.24) holds.

Then (4.11) yields

$$\left(\frac{b_1}{b_2}\right)^{q^2}=\frac{b_3}{b_4}. \tag{4.26}$$

Multiplying both sides of (4.24) by $b_4^{q^2}$ and applying (4.20) gives

$$a_2^{q^3}a_1^{1+q+q^2}=b_2^{1+q+q^3}b_4^{q^2}. \tag{4.27}$$

Then multiplying (4.20) by (4.21) and taking (4.27) into account we obtain

$$a_1^qa_3^{1+q^2+q^3}=b_3b_4^{q+q^2+q^3}. \tag{4.28}$$

Multiplying (4.22) and (4.23) yield

$$(b_1^qb_3^{1+q^2+q^3})(b_2^{q^3}b_1^{1+q+q^2})=(a_2^{1+q+q^3}a_4^{q^2})(a_3a_4^{q+q^2+q^3}).$$

On the other hand, from (4.19), and taking (4.27) and (4.28) into account, it follows that

$$b_1^qb_3^{1+q^2+q^3}+b_2^{q^3}b_1^{1+q+q^2}=a_2^{1+q+q^3}a_4^{q^2}+a_3a_4^{q+q^2+q^3}.$$

Hence

$$b_1^qb_3^{1+q^2+q^3}=a_2^{1+q+q^3}a_4^{q^2}\quad\text{and}\quad b_2^{q^3}b_1^{1+q+q^2}=a_3a_4^{q+q^2+q^3},$$

or

$$b_1^q b_3^{1+q^2+q^3} = a_3 a_4^{q+q^2+q^3} \quad \text{and} \quad b_2^{q^3} b_1^{1+q+q^2} = a_2^{1+q+q^3} a_4^{q^2}.$$

In the former case (4.22) yields $\mathrm{N}(b_1/a_4) = 1$, which is (4.12). In the latter case (4.11) and (4.23) gives

$$\frac{b_4^q b_3^{q^3+1}}{b_1 b_2^{q+q^3}} b_2^{q^3} b_1^{1+q+q^2} = \mathrm{N}(a_4/b_1) b_2^{q^3} b_1^{1+q+q^2} = b_1^q b_3^{1+q^2+q^3},$$

and hence

$$\frac{b_4}{b_2} = \left(\frac{b_3}{b_1}\right)^q. \tag{4.29}$$

Equation (4.26) is equivalent to

$$b_4 b_1^{q^2} = b_3 b_2^{q^2}, \tag{4.30}$$

while (4.29) is equivalent to

$$b_4 b_1^q = b_3^q b_2.$$

Dividing these two equations by each other yield

$$b_2^{q^2-1} = b_3^{q-1} b_1^{q^2-q}.$$

It follows that there exists $\lambda \in \mathbb{F}_q^*$ such that

$$b_2^{q+1} = \lambda b_3 b_1^q, \tag{4.31}$$

thus

$$b_3 = b_2^{q+1}/(b_1^q \lambda) \tag{4.32}$$

and by (4.30)

$$b_4 = b_2^{1+q+q^2}/(b_1^{q+q^2} \lambda). \tag{4.33}$$

Then (4.11) can be written as

$$\mathrm{N}\left(\frac{b_1}{a_4}\right) = \frac{b_1 b_2^{q+q^3}}{b_4^q b_3^{q^3+1}} = \mathrm{N}\left(\frac{b_1}{b_2}\right) \lambda^3,$$

and hence

$$\mathrm{N}\left(\frac{b_2}{a_4}\right) = \lambda^3. \tag{4.34}$$

By (4.2), (4.34) and (4.33) we get

$$\mathrm{N}(a_1) = \mathrm{N}(b_2)^2/(\mathrm{N}(b_1)\lambda^2). \tag{4.35}$$

By (4.18), (4.32) and (4.34) we have

$$\mathrm{N}(a_2) = \mathrm{N}(b_1)\lambda. \tag{4.36}$$

By (4.18), (4.35) and (4.33) we get

$$\mathrm{N}(a_3) = \mathrm{N}(b_2)^3/(\mathrm{N}(b_1)^2\lambda^6), \tag{4.37}$$

and by (4.34) we have

$$\mathrm{N}(a_4) = \mathrm{N}(b_2)/\lambda^3. \tag{4.38}$$

Before we go further, we simplify (4.7) and prove

$$\mathrm{N}(a_1) + \mathrm{N}(a_2) + \mathrm{N}(a_3) + \mathrm{N}(a_4) = \mathrm{N}(b_1) + \mathrm{N}(b_2) + \mathrm{N}(b_3) + \mathrm{N}(b_4). \tag{4.39}$$

It is enough to show

$$\begin{gathered}
\mathrm{Tr}(\overbrace{a_1^q a_2^{q^2+q^3+q^4} a_3}^{A_1} + \overbrace{a_1^{q+q^3} a_2^{q^4} a_3^{1+q^2}}^{A_2} + \overbrace{a_1^{q+q^2} a_2^{q^3+q^4} a_4}^{A_3} + \overbrace{a_1^{q+q^2+q^4} a_3^{q^3} a_4}^{A_4} + \\
\overbrace{a_2^q a_3^{q^2+q^3+q^4} a_4}^{A_5} + \overbrace{a_1^{q^2} a_3^{q^3+q^4} a_4^{1+q}}^{A_6} + \overbrace{a_2^{q+q^3} a_3^{q^4} a_4^{1+q^2}}^{A_7} + \overbrace{a_1^{q^2} a_2^{q^4} a_4^{1+q+q^3}}^{A_8}) = \\
\mathrm{Tr}(\overbrace{b_1^q b_2^{q^2+q^3+q^4} b_3}^{B_1} + \overbrace{b_1^{q+q^3} b_2^{q^4} b_3^{1+q^2}}^{B_7} + \overbrace{b_1^{q+q^2} b_2^{q^3+q^4} b_4}^{B_3} + \overbrace{b_1^{q+q^2+q^4} b_3^{q^3} b_4}^{B_8} + \\
\overbrace{b_2^q b_3^{q^2+q^3+q^4} b_4}^{B_5} + \overbrace{b_1^{q^2} b_3^{q^3+q^4} b_4^{1+q}}^{B_6} + \overbrace{b_2^{q+q^3} b_3^{q^4} b_4^{1+q^2}}^{B_2} + \overbrace{b_1^{q^2} b_2^{q^4} b_4^{1+q+q^3}}^{B_4}),
\end{gathered}$$

which can be done by proving $\mathrm{Tr}(A_i) = \mathrm{Tr}(B_i)$ for $i = 1, 2, \dots, 8$. Expressing a_3 with a_4 in (4.18), and using (4.2) as well, we get $a_3 = b_2^{q^3} a_4^{q^4+1} / b_1^{q^3+q^4}$. Then a_1, a_2, a_3 can be eliminated in all of the A_i, $i \in \{1, 2 \dots, 8\}$. It turns out that this procedure eliminates also a_4 when $i \in \{2, 4, 7, 8\}$ and we obtain

$$A_2 = B_2^{q^2}, \quad A_4 = B_4^{q^2}, \quad A_7 = B_7^{q^3} \quad \text{and} \quad A_8 = B_8^{q^2}.$$

In each of the other cases what remains is $\mathrm{N}(a_4)$ times an expression in b_1, b_2, b_3, b_4. Then by using (4.11) we can also eliminate $\mathrm{N}(a_4)$ and hence A_i can be expressed in terms of b_1, b_2, b_3, b_4. This gives $A_1 = B_1$ and $A_5 = B_5$. Applying also (4.26) and (4.29) we obtain $A_3 = B_3^{q^2}$ and $A_6 = B_6$.

Let $x = \mathrm{N}(b_2/b_1)$. Multiplying both sides of (4.39) by $\lambda^6/\mathrm{N}(b_1)$, taking into account (4.35), (4.36), (4.37) and (4.38) for the left hand side and (4.32) and (4.33) for the right hand side we get the following equation

$$x^2\lambda^4 + \lambda^7 + x^3 + x\lambda^3 = \lambda^6 + x\lambda^6 + x^2\lambda + \lambda x^3.$$

After rearranging we get:

$$(1 - \lambda)(x - \lambda)(x - \lambda^2)(x - \lambda^3) = 0.$$

First suppose $\lambda \neq 1$. Then we have three possibilities.

1. If

$$x = \lambda,$$

in which case $N(b_2) = N(a_2)$ follows from (4.36). Since $\gcd(q - 1, q^5 - 1) = \gcd(q^2-1, q^5-1)$, in $\mathbb{F}_{q^5}^*$ the set of $(q-1)$-th powers is the same as the set of (q^2-1)-th powers and hence there exists and element $\nu \in \mathbb{F}_{q^5}^*$ such that $b_2 = \nu^{q^2-1} a_2$. Therefore, since we are in Case 1, from $a_1 a_2^{q+q^3} = b_1 b_2^{q+q^3}$ we obtain $b_1 = \nu^{q-1} a_1$. Equations (4.2) and (4.3) give $g(x) = f(\nu x)/\nu$.

2. If
$$x = \lambda^3,$$
then $N(a_4) = N(b_1)$ follows from (4.38). Then (4.15) equals 1 and hence $a_1 a_2^{q+q^3} = b_4^q b_3^{q^3+1}$ which means that we are in Case 2, thus $g(x) = \hat{f}(\mu x)/\mu$.
3. If
$$x = \lambda^2,$$
then we show that there exists $\varphi \in \Gamma\mathrm{L}(2, q^5)$ such that either
$$\mathrm{Im}(g_\varphi(x)/x) = \mathrm{Im}(x^q/x) \quad \text{or} \quad \mathrm{Im}(g_\varphi(x)/x) = \mathrm{Im}(\mathrm{Tr}(x)/x).$$
In the former case by Proposition 2.6 and Corollary 1.2 we get $f_\varphi(x) = \alpha x^{q^i}$ and $g_\varphi(x) = \beta x^{q^j}$ for some $i, j \in \{1, 2, 3, 4\}$, with $\mathrm{N}(\alpha) = \mathrm{N}(\beta) = 1$. In the latter case, by Theorem 1.3 and by Propositions 2.6 and 2.7, there exists $\mu \in \mathbb{F}_{q^5}^*$ such that $g(x) = f(\mu x)/\mu$.

According to part 2 of Proposition 3.2, it is enough to show
$$(b_4/b_1)^{q^2} = b_1/b_3, \quad (b_1/b_2)^{q^2} = b_3/b_4.$$
(Note that there is no need to confirm $\mathrm{N}(b_1) \neq \mathrm{N}(b_3)$ since otherwise the result follows from the last part of Proposition 3.2 and from Theorem 1.3.) The second equation is just (4.26), thus it is enough to prove the first one.

First we show
$$b_2 b_3^{q+q^3} = b_1^{1+q+q^3}. \tag{4.40}$$
From (4.31) we have
$$\mathrm{N}\left(\frac{b_2}{b_1}\right) = \lambda^2 = \left(\frac{b_2^{q+1}}{b_3 b_1^q}\right)^2,$$
and hence after rearranging
$$\frac{b_2^{q^2+q^3+q^4} b_3}{b_1^{1+q^2+q^3+q^4}} = \frac{b_2^{q+1}}{b_3 b_1^q}.$$
On the right-hand side we have λ, which is in $\mathbb{F}_q$, thus, after taking q-th powers on the left and q^3-th powers on the right, the following also holds
$$\frac{b_2^{q^3+q^4+1} b_3^q}{b_1^{q+q^3+q^4+1}} = \frac{b_2^{q^3+q^4}}{b_3^{q^3} b_1^{q^4}}.$$
After rearranging we obtain (4.40).

Now we show that $(b_4/b_1)^{q^2} = b_1/b_3$ is equivalent to (4.40). Expressing b_4 from (4.26) we get
$$(b_4/b_1)^{q^2} = b_1/b_3 \iff b_3^{1+q^2} b_2^{q^4} = b_1^{1+q^2+q^4},$$
where the equation on the right-hand side is just the q^4-th power of (4.40).

Finally, consider the case $\lambda = 1$. Then
$$b_3 = b_2^{q+1}/b_1^q, \quad b_4 = b_2^{1+q+q^2}/b_1^{q+q^2}$$

and it follows from Proposition 3.2 that there exists $\varphi \in \Gamma\mathrm{L}(2,q^5)$ such that either

$$\mathrm{Im}(g_\varphi(x)/x) = \mathrm{Im}(x^q/x) \quad \text{or} \quad \mathrm{Im}(g_\varphi(x)/x) = \mathrm{Im}(\mathrm{Tr}(x)/x).$$

As above, the assertion follows either from Proposition 2.6 and Corollary 1.2 or from Theorem 1.3 and by Propositions 2.6 and 2.7.

This finishes the proof when $\prod_{i=1}^{4} a_i b_i \neq 0$. □

5 New maximum scattered linear sets of $\mathrm{PG}(1, q^5)$

A point set L of a line $\Lambda = \mathrm{PG}(W, \mathbb{F}_{q^n}) = \mathrm{PG}(1,q^n)$ is said to be an $\mathbb{F}_q$-*linear set* of Λ of rank n if it is defined by the non-zero vectors of an n-dimensional $\mathbb{F}_q$-vector subspace U of the two-dimensional $\mathbb{F}_{q^n}$-vector space W, i.e.

$$L = L_U := \{\langle \mathbf{u} \rangle_{\mathbb{F}_{q^n}} : \mathbf{u} \in U \setminus \{\mathbf{0}\}\}.$$

One of the most natural questions about linear sets is their equivalence. Two linear sets L_U and L_V of $\mathrm{PG}(1,q^n)$ are said to be $\mathrm{P\Gamma L}$-*equivalent* (or simply *equivalent*) if there is an element in $\mathrm{P\Gamma L}(2,q^n)$ mapping L_U to L_V. In the applications it is crucial to have methods to decide whether two linear sets are equivalent or not. This can be a difficult problem and some results in this direction can be found in [8, 12]. If L_U and L_V are two equivalent $\mathbb{F}_q$-linear sets of rank n in $\mathrm{PG}(1,q^n)$ and φ is an element of $\Gamma\mathrm{L}(2,q^n)$ which induces a collineation mapping L_U to L_V, then $L_{U^\varphi} = L_V$. Hence the first step to face with the equivalence problem for linear sets is to determine which $\mathbb{F}_q$-subspaces can define the same linear set.

For any q-polynomial $f(x) = \sum_{i=0}^{n-1} a_i x^{q^i}$ over $\mathbb{F}_{q^n}$, the graph

$$\mathcal{G}_f = \{(x, f(x)) : x \in \mathbb{F}_{q^n}\}$$

is an $\mathbb{F}_q$-vector subspace of the 2-dimensional vector space $V = \mathbb{F}_{q^n} \times \mathbb{F}_{q^n}$ and the point set

$$L_f := L_{\mathcal{G}_f} = \{\langle (x, f(x)) \rangle_{\mathbb{F}_{q^n}} : x \in \mathbb{F}_{q^n}^*\}$$

is an $\mathbb{F}_q$-linear set of rank n of $\mathrm{PG}(1,q^n)$. In this context, the problem posed in (1.2) corresponds to find all $\mathbb{F}_q$-subspaces of V of rank n (cf. [8, Proposition 2.3]) defining the linear set L_f. The maximum field of linearity of f is the maximum field of linearity of L_f, and it is well-defined (cf. Proposition 2.1 and [8, Proposition 2.3]). Also, by the Introduction (Section 1), for any q-polynomial f over $\mathbb{F}_{q^n}$, the linear sets L_f, L_{f_λ} (with $f_\lambda(x) := f(\lambda x)/\lambda$ for each $\lambda \in \mathbb{F}_{q^n}^*$) and $L_{\hat{f}}$ coincide (cf. [2, Lemma 2.6] and the first part of [8, Section 3]). If f and g are two equivalent q-polynomials over $\mathbb{F}_{q^n}$, i.e. $\mathcal{G}_f$ and $\mathcal{G}_g$ are equivalent w.r.t. the action of the group $\Gamma\mathrm{L}(2,q^n)$, then the corresponding $\mathbb{F}_q$-linear sets L_f and L_g of $\mathrm{PG}(1,q^n)$ are $\mathrm{P\Gamma L}(2,q^n)$-equivalent. The converse does not hold (see [12] and [8] for further details).

The relation between the problem posed in (1.2) and the equivalence problem of linear sets of the projective line is summarized in the following result.

Proposition 5.1. *Let L_f and L_g be two $\mathbb{F}_q$-linear sets of rank n of $\mathrm{PG}(1,q^n)$. Then L_f and L_g are $\mathrm{P\Gamma L}(2,q^n)$-equivalent if and only if there exists an element $\varphi \in \Gamma\mathrm{L}(2,q^n)$ such that* $\mathrm{Im}(f_\varphi(x)/x) = \mathrm{Im}(g(x)/x)$. □

Linear sets of rank n of $\mathrm{PG}(1,q^n)$ have size at most $(q^n-1)/(q-1)$. A linear set L_U of rank n whose size achieves this bound is called *maximum scattered.* For applications of these objects we refer to [26] and [19].

Definition 5.2 ([15, 22])**.** A maximum scattered $\mathbb{F}_q$-linear set L_U of rank n in $\mathrm{PG}(1,q^n)$ is of *pseudoregulus type* if it is $\mathrm{P\Gamma L}(2,q^n)$-equivalent to L_f with $f(x)=x^q$ or, equivalently, if there exists an element $\varphi\in\mathrm{GL}(2,q^n)$ such that

$$L_{U^\varphi}=\{\langle(x,x^q)\rangle_{\mathbb{F}_{q^n}} : x\in\mathbb{F}^*_{q^n}\}.$$

By Proposition 5.1 and Corollary 1.2, it follows

Proposition 5.3. *An $\mathbb{F}_q$-linear set L_f of rank n of* $\mathrm{PG}(1,q^n)$ *is of pseudoregulus type if and only if $f(x)$ is equivalent to x^{q^i} for some i with* $\gcd(i,n)=1$. □

For the proof of the previous result see also [20].

The known pairwise non-equivalent families of q-polynomials over $\mathbb{F}_{q^n}$ which define maximum scattered linear sets of rank n in $\mathrm{PG}(1,q^n)$ are

1. $f_s(x)=x^{q^s}$, $1\le s\le n-1$, $\gcd(s,n)=1$ ([4, 11]),
2. $g_{s,\delta}(x)=\delta x^{q^s}+x^{q^{n-s}}$, $n\ge 4$, $\mathrm{N}_{q^n/q}(\delta)\notin\{0,1\}$[1], $\gcd(s,n)=1$ ([23] for $s=1$, [24, 27] for $s\ne 1$),
3. $h_{s,\delta}(x)=\delta x^{q^s}+x^{q^{s+n/2}}$, $n\in\{6,8\}$, $\gcd(s,n/2)=1$, $\mathrm{N}_{q^n/q^{n/2}}(\delta)\notin\{0,1\}$, for the precise conditions on δ and q see [9, Theorems 7.1 and 7.2][2],
4. $k_b(x)=x^q+x^{q^3}+bx^{q^5}$, $n=6$, with $b^2+b=1$, $q\equiv 0,\pm1 \pmod 5$ ([10]).

Remark 5.4. All the previous polynomials in cases 2, 3, and 4 above are examples of functions which are not equivalent to monomials but the set of directions determined by their graph has size $(q^n-1)/(q-1)$, i.e. the corresponding linear sets are maximum scattered. The existence of such linearized polynomials is briefly discussed also in [16, p. 132].

For $n=2$ the maximum scattered $\mathbb{F}_q$-linear sets coincide with the Baer sublines. For $n=3$ the maximum scattered linear sets are all of pseudoregulus type and the corresponding q-polynomials are all $\mathrm{GL}(2,q^3)$-equivalent to x^q (cf. [21]). For $n=4$ there are two families of maximum scattered linear sets. More precisely, if L_f is a maximum scattered linear set of rank 4 of $\mathrm{PG}(1,q^4)$, with maximum field of linearity $\mathbb{F}_q$, then there exists $\varphi\in\mathrm{GL}(2,q^4)$ such that either $f_\varphi(x)=x^q$ or $f_\varphi(x)=\delta x^q+x^{q^3}$, for some $\delta\in\mathbb{F}^*_{q^4}$ with $\mathrm{N}_{q^4/q}(\delta)\notin\{0,1\}$ (cf. [13]). It is easy to see that $L_{f_1}=L_{f_s}$ for any s with $\gcd(s,n)=1$, and f_i is equivalent to f_j if and only if $j\in\{i,n-i\}$. Also, the graph of $g_{s,\delta}$ is $\mathrm{GL}(2,q^n)$-equivalent to the graph of $g_{n-s,\delta^{-1}}$.

In [23, Theorem 3] Lunardon and Polverino proved that $L_{g_{1,\delta}}$ and L_{f_1} are not $\mathrm{P\Gamma L}(2,q^n)$-equivalent when $q>3$, $n\ge 4$. This was extended also for $q=3$ [10, Theorem 3.4]. Also in [10], it has been proven that for $n=6,8$ the linear sets L_{f_1}, $L_{g_{s,\delta}}$, $L_{h_{s',\delta'}}$ and L_{k_b} are pairwise non-equivalent for any choice of s,s',δ,δ',b.

In this section we prove that one can find for each $q>2$ a suitable δ such that $L_{g_{2,\delta}}$ of $\mathrm{PG}(1,q^5)$ is not equivalent to the linear sets $L_{g_{1,\mu}}$ of $\mathrm{PG}(1,q^5)$ for each $\mu\in\mathbb{F}^*_{q^5}$, with $\mathrm{N}_{q^5/q}(\mu)\notin\{0,1\}$. In order to do this, we first reformulate Theorem 1.5 as follows.

[1] This condition implies $q\ne 2$.

[2] Also here $q>2$, otherwise the linear set defined by $h_{s,\delta}$ is never scattered.

Theorem 5.5 (Theorem 1.5)**.** *Let $f(x)$ and $g(x)$ be two q-polynomials over $\mathbb{F}_{q^5}$ such that $L_f = L_g$. Then either $L_f = L_g$ is of pseudoregulus type or there exists some $\lambda \in \mathbb{F}^*_{q^5}$ such that $g(x) = f(\lambda x)/\lambda$ or $g(x) = \hat{f}(\lambda x)/\lambda$ holds.*

From [27, Theorem 8] and [24, Theorem 4.4] it follows that the family of $\mathbb{F}_q$-subspaces $U_{g_{s,\delta}}$, $s \notin \{1, n-1\}$, $\gcd(s, n) = 1$, contains members which are not ΓL-equivalent to the previously known $\mathbb{F}_q$-subspaces defining maximum scattered linear sets of $\mathrm{PG}(1, q^n)$. Our next result shows that the corresponding family $L_{g_{s,\delta}}$ of linear sets contains (at least for $n = 5$) examples which are not PΓL-equivalent to the previously known maximum scattered linear sets.

Theorem 5.6. *Let $g_{2,\delta}(x) = \delta x^{q^2} + x^{q^3}$ for some $\delta \in \mathbb{F}^*_{q^5}$ with $\mathrm{N}(\delta)^5 \neq 1$. Then $L_{g_{2,\delta}}$ is not $\mathrm{P\Gamma L}(2, q^5)$-equivalent to any linear set $L_{g_{1,\mu}}$ and hence it is a new maximum scattered linear set.*

Proof. Suppose, contrary to our claim, that $L_{g_{2,\delta}}$ is $\mathrm{P\Gamma L}(2, q^5)$-equivalent to a linear set $L_{g_{1,\mu}}$. From Proposition 5.1 and Theorem 5.5, taking into account that $L_{g_{1,\mu}}$ is not of pseudoregulus type, it follows that there exist $\varphi \in \Gamma\mathrm{L}(2, q^5)$ and $\lambda \in \mathbb{F}^*_{q^5}$ such that either $(g_{2,\delta})_\varphi(x) = g_{1,\mu}(\lambda x)/\lambda$ or $(g_{2,\delta})_\varphi(x) = \hat{g}_{1,\mu}(\lambda x)/\lambda$. This is equivalent to say that there exist $\alpha, \beta, A, B, C, D \in \mathbb{F}_{q^5}$ with $AD - BC \neq 0$ and a field automorphism τ of $\mathbb{F}_{q^5}$ such that

$$\left\{ \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x^\tau \\ g_{2,\delta}(x)^\tau \end{pmatrix} : x \in \mathbb{F}_{q^5} \right\} = \left\{ \begin{pmatrix} z \\ \alpha z^q + \beta z^{q^4} \end{pmatrix} : z \in \mathbb{F}_{q^5} \right\},$$

where $N(\alpha) \neq N(\beta)$ and $\alpha\beta \neq 0$. We may substitute x^τ by y, then

$$\alpha(Ay + B\delta^\tau y^{q^2} + By^{q^3})^q + \beta(Ay + B\delta^\tau y^{q^2} + By^{q^3})^{q^4} = Cy + D\delta^\tau y^{q^2} + Dy^{q^3}$$

for each $y \in \mathbb{F}_{q^5}$. Comparing coefficients yields $C = 0$ and

$$\alpha A^q + \beta B^{q^4} \delta^{q^4\tau} = 0, \tag{5.1}$$

$$\beta B^{q^4} = D\delta^\tau, \tag{5.2}$$

$$\alpha B^q \delta^{q\tau} = D, \tag{5.3}$$

$$\alpha B^q + \beta A^{q^4} = 0. \tag{5.4}$$

Conditions (5.2) and (5.3) give

$$B^{q^4-q} = \delta^{(q+1)\tau} \alpha/\beta. \tag{5.5}$$

On the other hand from (5.4) we get $A^q = -B^{q^3}\alpha^{q^2}/\beta^{q^2}$ and substituting this into (5.1) we have

$$B^{q^3-q^4} = \delta^{q^4\tau} \beta^{q^2+1}/\alpha^{q^2+1}. \tag{5.6}$$

Equations (5.5) and (5.6) give $N(\beta/\alpha) = N(\delta)^{2\tau}$ and $N(\alpha/\beta)^2 = N(\delta)^\tau$, respectively. It follows that $N(\delta)^{5\tau} = 1$ and hence $N(\delta)^5 = 1$, a contradiction. □

6 Open problems

We conclude the paper by the following open problems.

1. Is it true also for $n > 5$ that for any pair of q-polynomials $f(x)$ and $g(x)$ of $\mathbb{F}_{q^n}[x]$, with maximum field of linearity $\mathbb{F}_q$, if $\mathrm{Im}(f(x)/x) = \mathrm{Im}(g(x)/x)$ then either there exists $\varphi \in \Gamma\mathrm{L}(2, q^n)$ such that $f_\varphi(x) = \alpha x^{q^i}$ and $g_\varphi(x) = \beta x^{q^j}$ with $\mathrm{N}(\alpha) = \mathrm{N}(\beta)$ and $\gcd(i, n) = \gcd(j, n) = 1$, or there exists $\lambda \in \mathbb{F}^*_{q^n}$ such that $g(x) = f(\lambda x)/\lambda$ or $g(x) = \hat{f}(\lambda x)/\lambda$?
2. Is it possible, at least for small values of $n > 4$, to classify, up to equivalence, the q-polynomials $f(x) \in \mathbb{F}_{q^n}[x]$ such that $|\,\mathrm{Im}(f(x)/x)| = (q^n - 1)/(q - 1)$? Find new examples!
3. Is it possible, at least for small values of n, to classify, up to equivalence, the q-polynomials $f(x) \in \mathbb{F}_{q^n}[x]$ such that $|\,\mathrm{Im}(f(x)/x)| = q^{n-1} + 1$? Find new examples!
4. Is it possible, at least for small values of n, to classify, up to equivalence, the q-polynomials $f(x) \in \mathbb{F}_{q^n}[x]$ such that in the multiset $\{f(x)/x : x \in \mathbb{F}^*_{q^n}\}$ there is a unique element which is represented more than $q - 1$ times? In this case the linear set L_f is an *i-club* of rank n and when $q = 2$, then such linear sets correspond to translation KM-arcs cf. [14] (a KM-arc, or $(q + t, t)$-arc of type $(0, 2, t)$, is a set of $q+t$ points of $\mathrm{PG}(2, 2^n)$, such that each line meets the point set in 0, 2 or in t points, cf. [17]). Find new examples!
5. Determine the equivalence classes of the set of q-polynomials in $\mathbb{F}_{q^4}[x]$.
6. Determine, at least for small values of n, all the possible sizes of $\mathrm{Im}(f(x)/x)$ where $f(x) \in \mathbb{F}_{q^n}[x]$ is a q-polynomial.

References

[1] S. Ball, The number of directions determined by a function over a finite field, *J. Comb. Theory Ser. A* **104** (2003), 341–350, doi:10.1016/j.jcta.2003.09.006.

[2] D. Bartoli, M. Giulietti, G. Marino and O. Polverino, Maximum scattered linear sets and complete caps in Galois spaces, *Combinatorica* **38** (2018), 255–278, doi:10.1007/s00493-016-3531-6.

[3] A. Blokhuis, S. Ball, A. E. Brouwer, L. Storme and T. Szőnyi, On the number of slopes of the graph of a function defined on a finite field, *J. Comb. Theory Ser. A* **86** (1999), 187–196, doi:10.1006/jcta.1998.2915.

[4] A. Blokhuis and M. Lavrauw, Scattered spaces with respect to a spread in $\mathrm{PG}(n, q)$, *Geom. Dedicata* **81** (2000), 231–243, doi:10.1023/a:1005283806897.

[5] G. Bonoli and O. Polverino, $\mathbb{F}_q$-linear blocking sets in $\mathrm{PG}(2, q^4)$, *Innov. Incidence Geom.* **2** (2005), 35–56, http://www.iig.ugent.be/online/2/volume-2-article-2-online.pdf.

[6] A. Bruen and B. Levinger, A theorem on permutations of a finite field, *Canad. J. Math.* **25** (1973), 1060–1065, doi:10.4153/cjm-1973-113-4.

[7] L. Carlitz, A theorem on permutations in a finite field, *Proc. Amer. Math. Soc.* **11** (1960), 456–459, doi:10.2307/2034798.

[8] B. Csajbók, G. Marino and O. Polverino, Classes and equivalence of linear sets in $\mathrm{PG}(1, q^n)$, *J. Comb. Theory Ser. A* **157** (2018), 402–426, doi:10.1016/j.jcta.2018.03.007.

[9] B. Csajbók, G. Marino, O. Polverino and C. Zanella, A new family of MRD-codes, *Linear Algebra Appl.* **548** (2018), 203–220, doi:10.1016/j.laa.2018.02.027.

[10] B. Csajbók, G. Marino and F. Zullo, New maximum scattered linear sets of the projective line, *Finite Fields Appl.* **54** (2018), 133–150, doi:10.1016/j.ffa.2018.08.001.

[11] B. Csajbók and C. Zanella, On scattered linear sets of pseudoregulus type in $\mathrm{PG}(1, q^t)$, *Finite Fields Appl.* **41** (2016), 34–54, doi:10.1016/j.ffa.2016.04.006.

[12] B. Csajbók and C. Zanella, On the equivalence of linear sets, *Des. Codes Cryptogr.* **81** (2016), 269–281, doi:10.1007/s10623-015-0141-z.

[13] B. Csajbók and C. Zanella, Maximum scattered $\mathbb{F}_q$-linear sets of $\mathrm{PG}(1, q^4)$, *Discrete Math.* **341** (2018), 74–80, doi:10.1016/j.disc.2017.07.001.

[14] M. De Boeck and G. Van de Voorde, A linear set view on KM-arcs, *J. Algebraic Combin.* **44** (2016), 131–164, doi:10.1007/s10801-015-0661-7.

[15] G. Donati and N. Durante, Scattered linear sets generated by collineations between pencils of lines, *J. Algebraic Combin.* **40** (2014), 1121–1134, doi:10.1007/s10801-014-0521-x.

[16] F. Göloğlu and G. McGuire, On theorems of Carlitz and Payne on permutation polynomials over finite fields with an application to $x^{-1} + L(x)$, *Finite Fields Appl.* **27** (2014), 130–142, doi:10.1016/j.ffa.2014.01.004.

[17] G. Korchmáros and F. Mazzocca, On $(q + t)$-arcs of type $(0, 2, t)$ in a Desarguesian plane of order q, *Math. Proc. Cambridge Philos. Soc.* **108** (1990), 445–459, doi:10.1017/s0305004100069346.

[18] M. Lavrauw, *Scattered Spaces With Respect to Spreads, and Eggs in Finite Projective Spaces*, Ph.D. thesis, Eindhoven University of Technology, Eindhoven, 2001, `https://search.proquest.com/docview/304739268`.

[19] M. Lavrauw, Scattered spaces in Galois geometry, in: A. Canteaut, G. Effinger, S. Huczynska, D. Panario and L. Storme (eds.), *Contemporary Developments in Finite Fields and Applications*, World Scientific Publishing, Hackensack, NJ, pp. 195–216, 2016, doi:10.1142/9762, papers based on the 12th International Conference on Finite Fields and Their Applications (Fq12) held at Skidmore College, Saratoga Springs, NY, July 2015.

[20] M. Lavrauw, J. Sheekey and C. Zanella, On embeddings of minimum dimension of $\mathrm{PG}(n, q) \times \mathrm{PG}(n, q)$, *Des. Codes Cryptogr.* **74** (2015), 427–440, doi:10.1007/s10623-013-9866-8.

[21] M. Lavrauw and G. Van de Voorde, On linear sets on a projective line, *Des. Codes Cryptogr.* **56** (2010), 89–104, doi:10.1007/s10623-010-9393-9.

[22] G. Lunardon, G. Marino, O. Polverino and R. Trombetti, Maximum scattered linear sets of pseudoregulus type and the Segre variety $\mathcal{S}_{n,n}$, *J. Algebraic Combin.* **39** (2014), 807–831, doi:10.1007/s10801-013-0468-3.

[23] G. Lunardon and O. Polverino, Blocking sets and derivable partial spreads, *J. Algebraic Combin.* **14** (2001), 49–56, doi:10.1023/a:1011265919847.

[24] G. Lunardon, R. Trombetti and Y. Zhou, Generalized twisted Gabidulin codes, *J. Comb. Theory Ser. A* **159** (2018), 79–106, doi:10.1016/j.jcta.2018.05.004.

[25] R. McConnel, Pseudo-ordered polynomials over a finite field, *Acta Arith.* **8** (1963), 127–151, doi:10.4064/aa-8-2-127-151.

[26] O. Polverino, Linear sets in finite projective spaces, *Discrete Math.* **310** (2010), 3096–3107, doi:10.1016/j.disc.2009.04.007.

[27] J. Sheekey, A new family of linear maximum rank distance codes, *Adv. Math. Commun.* **10** (2016), 475–488, doi:10.3934/amc.2016019.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 609–623
https://doi.org/10.26493/1855-3974.1681.4ec
(Also available at http://amc-journal.eu)

Embedding of orthogonal Buekenhout-Metz unitals in the Desarguesian plane of order q^2

Gábor Korchmáros , Alessandro Siciliano

Dipartimento di Matematica, Informatica ed Economia,
Università degli Studi della Basilicata,
Viale dell'Ateneo Lucano 10 - 85100 Potenza (Italy)

Received 19 April 2018, accepted 18 February 2019, published online 5 June 2019

Abstract

A unital, that is a 2-$(q^3+1, q+1, 1)$ block-design, is embedded in a projective plane π of order q^2 if its points are points of π and its blocks are subsets of lines of π, the point-block incidences being the same as in π. Regarding unitals $\mathcal{U}$ which are isomorphic, as a block-design, to the classical unital, T. Szőnyi and the authors recently proved that the natural embedding is the unique embedding of $\mathcal{U}$ into the Desarguesian plane of order q^2. In this paper we extend this uniqueness result to all unitals which are isomorphic, as block-designs, to orthogonal Buekenhout-Metz unitals.

Keywords: Unital, embedding, finite Desarguesian plane.

Math. Subj. Class.: 51E05, 51E20

1 Introduction

A *unital* is a set of $q^3 + 1$ points equipped with a family of subsets, each of size $q + 1$, such that every pair of distinct points are contained in exactly one subset of the family. In Design Theory, such subsets are usually called *blocks* so that unitals are 2-$(q^3+1, q+1, 1)$ block-designs. A unital $\mathcal{U}$ is *embedded* in a projective plane π of order q^2, if its points are points of π, its blocks are subsets of lines of π and the point-block incidences being the same as in π.

Sufficient conditions for a unital to be embeddable in a projective plane are given in [21]. Computer aided searches suggest that there should be plenty of unitals, especially for small values of q, but those embeddable in a projective plane are quite rare, see [3, 6, 27]. Very recently, the GAP package UnitalSz was released [25]. This package contains methods for the embeddings of unitals in the finite projective plane.

E-mail addresses: gabor.korchmaros@unibas.it (Gábor Korchmáros), alessandro.siciliano@unibas.it (Alessandro Siciliano)

In the finite Desarguesian projective plane of order q^2, a unital arises from a unitary polarity: the points of the unital are the absolute points, and the blocks are the non-absolute lines of the polarity. This unital is called *classical unital*. The following result comes from [23].

Theorem 1.1. *Let $\mathcal{U}$ be a unital embedded in* $\mathrm{PG}(2, q^2)$ *which is isomorphic, as a block-design, to a classical unital. Then $\mathcal{U}$ is the classical unital of* $\mathrm{PG}(2, q^2)$.

Buekenhout [11] constructed unitals in any translation planes with dimension at most two over their kernel by using the Andrè/Bruck-Bose representation. Buekenhout's work was completed by Metz [24] who was able to prove by a counting argument that when the plane is Desarguesian then Buekenhout's construction provides not only the classical unital but also non-classical unitals in $\mathrm{PG}(2, q^2)$ for all $q > 2$. These unitals are called *Buekenhout-Metz unitals*, and they are the only known unitals in $\mathrm{PG}(2, q^2)$. With the terminology in [5], an *orthogonal Buekenhout-Metz unital* is a Buekenhout-Metz unital arising from an elliptic quadric in Buekehout's construction.

In this paper, we prove the following result:

Main Theorem. Let $\mathcal{U}$ be a unital embedded in $\mathrm{PG}(2, q^2)$ which is isomorphic, as block-design, to an orthogonal Buekenhout-Metz unital. Then $\mathcal{U}$ is an orthogonal Buekenhout-Metz unital.

Our approach is different from that adopted in [23]. Our idea is to exploit two different models of $\mathrm{PG}(2, q^2)$ in $\mathrm{PG}(5, q)$, one of them is a variant of the so-called $\mathrm{GF}(q)$-linear representation. We start off with a representation of a non-classical Buekenhout-Metz unital given in one of these models of $\mathrm{PG}(2, q^2)$, then we exhibit a linear collineation of $\mathrm{PG}(5, q)$ that takes this representation to a representation of a classical unital in the other model of $\mathrm{PG}(2, q^2)$. At this point to finish the proof we only need some arguments from the proof of Theorem 1.1 together with the characterization of the orthogonal Buekenhout-Metz unitals due to Casse, O'Keefe, Penttila and Quinn [12, 29].

2 Preliminary results

The study of unitals in finite projective planes has been greatly aided by the use of the Andrè/Bruck-Bose representation of these planes [1, 9, 10]. Let $\mathrm{PG}(4, q)$ denote the projective 4-dimensional space over the finite field $\mathrm{GF}(q)$, and let Σ be some fixed hyperplane of $\mathrm{PG}(4, q)$. Let $\mathcal{N}$ be a line spread of Σ, that is a collection of $q^2 + 1$ mutually skew lines of Σ. We consider the following incidence structure: the *points* are the points of $\mathrm{PG}(4, q)$ not in Σ, the *lines* are the planes of $\mathrm{PG}(4, q)$ which meet Σ in a line of $\mathcal{N}$ and *incidence* is defined by inclusion. This incidence structure is an affine translation plane of order q^2 which is at most two-dimensional over its kernel. It can be completed to a projective plane $\pi(\mathcal{N})$ by the addition of an ideal line L_∞ whose points are the elements of the spread $\mathcal{N}$. Conversely, any translation plane of order q^2 with $\mathrm{GF}(q)$ in its kernel can be modeled this way [9]. Moreover, it is well known that the resulting plane is Desarguesian if and only if $\mathcal{N}$ is a Desarguesian spread [10].

Our first step is to outline the usual representation of $\mathrm{PG}(2, q^2)$ in $\mathrm{PG}(5, q)$ due to Segre [30] and Bose [7]. While such representation is usually thought of in a projective setting, algebraic dimensions are more amenable to an introductory discussion of it, so we will mainly take a vector space approach along all this section.

Look at $\mathrm{GF}(q^2)$ as the two-dimensional vector space over $\mathrm{GF}(q)$ with basis $\{1, \epsilon\}$, so that every $x \in \mathrm{GF}(q^2)$ is uniquely written as $x = x_0 + x_1\epsilon$, for $x_0, x_1 \in \mathrm{GF}(q)$. Then the vectors (x, y, z) of $V(3, q^2)$ are viewed as the vectors $(x_1, x_2, y_1, y_2, z_1, z_2)$ of $V(6, q)$ where

$$\begin{aligned} x &= x_0 + x_1\epsilon, \\ y &= y_0 + \epsilon y_1 \text{ and} \\ z &= z_0 + \epsilon z_1. \end{aligned}$$

Therefore the points of $\mathrm{PG}(2, q^2)$ are two-dimensional subspaces in $V(6, q)$, and hence lines of $\mathrm{PG}(5, q)$, the five-dimensional projective space arising from $V(6, q)$. Such lines are the members of a Desarguesian line-spread $\mathcal{S}$ of $\mathrm{PG}(5, q)$ which gives rise to a point-line incidence structure $\Pi(\mathcal{S})$ where points are the elements of $\mathcal{S}$, and lines are the three-dimensional subspaces of $\mathrm{PG}(5, q)$ spanned by two elements of $\mathcal{S}$, incidence being inclusion. Obviously, $\Pi(\mathcal{S}) \simeq \mathrm{PG}(2, q^2)$, and $\Pi(\mathcal{S})$ is the $\mathrm{GF}(q)$-*linear representation of* $\mathrm{PG}(2, q^2)$ in $\mathrm{PG}(5, q)$. Since $\mathrm{PG}(5, q)$ is naturally embedded in $\mathrm{PG}(5, q^2)$, we also have an embedding of $\mathrm{PG}(2, q^2)$ in $\mathrm{PG}(5, q^2)$ via $\Pi(\mathcal{S})$.

Actually, we will use a different embedding of $\mathrm{PG}(2, q^2)$ in $\mathrm{PG}(5, q^2)$ which is more suitable for computation.

In $V(6, q^2)$, let $\widehat{V}$ be the set of all vectors (x, x^q, y, y^q, z, z^q) with $x, y, z \in \mathrm{GF}(q^2)$. With the usual sum and multiplication by scalars from $\mathrm{GF}(q)$, $\widehat{V}$ is a six-dimensional vector space over $\mathrm{GF}(q)$. On the other hand, $V(6, q)$ is naturally embedded in $V(6, q^2)$. Therefore, the question arises whether there exists an invertible endomorphism of $V(6, q^2)$ that takes $\widehat{V}$ to $V(6, q)$. The affirmative answer is given by the following proposition.

Proposition 2.1. *$\widehat{V}$ is linearly equivalent to $V(6, q)$ in $V(6, q^2)$.*

Proof. Write $V(6, q)$ as the direct sum $W^{(1)} \oplus W^{(2)} \oplus W^{(3)}$, with

$$\begin{aligned} W^{(1)} &= \{(a, b, 0, 0, 0, 0) : a, b \in \mathrm{GF}(q)\} \\ W^{(2)} &= \{(0, 0, a, b, 0, 0) : a, b \in \mathrm{GF}(q)\} \\ W^{(3)} &= \{(0, 0, 0, 0, a, b) : a, b \in \mathrm{GF}(q)\}. \end{aligned}$$

Clearly, each $W^{(i)}$ is isomorphic to $V(2, q) = \{(a, b) : a, b \in \mathrm{GF}(q)\}$. Take a basis $\{u_1, u_2\}$ of $V(2, q)$ together with a Singer cycle σ of $V(2, q)$. Since σ has two distinct eigenvalues, both in $\mathrm{GF}(q^2) \setminus \mathrm{GF}(q)$, we find two linearly independent eigenvectors v_1, v_2 that form a basis for $V(2, q^2)$. Such a basis $\{v_1, v_2\}$ is called a *Singer basis* with respect to $V(2, q)$ [15]. In this context, $V(2, q) = \{xv_1 + x^q v_2 : x \in \mathrm{GF}(q^2)\}$ [14].

Applying this argument to $W^{(i)}$ with $i = 1, 2, 3$, gives a Singer basis $\{v_1^{(i)}, v_2^{(i)}\}$ of $W^{(i)}$ such that $W^{(i)} = \{xv_1^{(i)} + x^q v_2^{(i)} : x \in \mathrm{GF}(q^2)\}$. In this basis we have

$$V(6, q) = \{xv_1^{(1)} + x^q v_2^{(1)} + yv_1^{(2)} + y^q v_2^{(2)} + zv_1^{(3)} + z^q v_2^{(3)} : x, y, z \in \mathrm{GF}(q^2)\}. \quad (2.1)$$

Now, the result follows from the fact that the change from any basis of $V(6, q^2)$ to the basis $\{v_1^{(i)}, v_2^{(i)} : i = 1, 2, 3\}$ is carried out by an invertible endomorphism over $\mathrm{GF}(q^2)$. □

We call the vector space $\widehat{V}$ the *cyclic representation of* $V(6, q)$ *over* $\mathrm{GF}(q^2)$.

To state Proposition 2.1 in terms of projective geometry, let $\mathrm{PG}(5,q)$ denote the projective space arising from $V(6,q)$. Also, let $\mathrm{PG}(\widehat{V}) = \{\langle v\rangle_q : v \in \widehat{V}\}$ be the five-dimensional projective space whose points are the one-dimensional $\mathrm{GF}(q)$-subspaces spanned by vectors in $\widehat{V}$.

Corollary 2.2. $\mathrm{PG}(\widehat{V})$ *is projectively equivalent to* $\mathrm{PG}(5,q)$ *in* $\mathrm{PG}(5,q^2)$.

We call the the projective space $\mathrm{PG}(\widehat{V})$ the *cyclic representation of* $\mathrm{PG}(5,q)$ *over* $\mathrm{GF}(q^2)$.

Recall that a 2×2 *q-circulant* (or *Dickson*) *matrix* over $\mathrm{GF}(q^2)$ is a matrix of the form

$$D = \begin{pmatrix} d_1 & d_2 \\ d_2^q & d_1^q \end{pmatrix}$$

with $d_1, d_2 \in \mathrm{GF}(q^2)$.

Let $\mathcal{B}$ denote the basis $\{v_1^{(i)}, v_2^{(i)} : i = 1,2,3\}$ of $\widehat{V}$.

Proposition 2.3. *In the basis $\mathcal{B}$, the matrix associated to any endomorphism of $\widehat{V}$ is of the form*

$$\begin{pmatrix} D_{11} & D_{12} & D_{13} \\ D_{21} & D_{22} & D_{23} \\ D_{31} & D_{32} & D_{33} \end{pmatrix}, \tag{2.2}$$

where D_{ij} is a 2×2 q-circulant matrix over $\mathrm{GF}(q^2)$.

Proof. It is easily seen that any matrix of type (2.2) is associated to an endomorphism of $\widehat{V}$.

Conversely, take an endomorphism τ of $V(6,q^2)$ and let $T = (t_{ij})$, $t_{ij} \in \mathrm{GF}(q^2)$, be the matrix of τ in the basis $\mathcal{B}$. For a generic array $\mathbf{x} = (x, x^q, y, y^q, z, z^q) \in \widehat{V}$,

$$T\mathbf{x}^t = \begin{pmatrix} \vdots \\ t_{k,1}x + t_{k,2}x^q + t_{k,3}y + t_{k,4}y^q + t_{k,5}z + t_{k,6}z^q \\ \vdots \end{pmatrix}, \text{for } k = 1,\dots,6.$$

If $y = z = 0$, a necessary condition for $T\mathbf{x}^t \in \widehat{V}$ is

$$(t_{k,1}x + t_{k,2}x^q)^q = t_{k+1,1}x + t_{k+1,2}x^q\ ,$$

for $k = 1,3,5$, that is,

$$(t_{k,2}^q - t_{k+1,1})x + (t_{k,1}^q - t_{k+1,2})x^q = 0,$$

for $k = 1,3,5$ and for all $x \in \mathrm{GF}(q^2)$. This shows that the polynomial in x of degree q on the left hand side of the last equation has at least q^2 roots. Therefore, it must be the zero polynomial. Hence $t_{k+1,1} = t_{k,2}^q$ and $t_{k+1,2} = t_{k,1}^q$, for $k = 1,3,5$. To end the proof, it is enough to repeat the above argument for $x = z = 0$ and then for $x = y = 0$. □

Next we exhibit quadratic forms on $V(6,q^2)$ which induce quadratic forms on $\widehat{V}$.

The vector space $V(2n,q)$ has precisely two (nondegenerate) quadratic forms, and they differ by their Witt-index, that is the dimension of their maximal totally singular subspaces;

see [22, 32]. These dimensions are $n-1$ and n, and the quadratic form is *elliptic* or *hyperbolic*, respectively. In terms of the associated projective space $\mathrm{PG}(2n-1,q)$, the elliptic (resp. hyperbolic) quadratic form defines an *elliptic* (resp. *hyperbolic*) quadric of $\mathrm{PG}(2n-1,q)$.

Fix a basis $\{1,\epsilon\}$ for $\mathrm{GF}(q^2)$ over $\mathrm{GF}(q)$, and write $x = x_0 + \epsilon x_1$, for $x \in \mathrm{GF}(q^2)$ with $x_0, x_1 \in \mathrm{GF}(q)$. Here, ϵ is taken such that $\epsilon^2 = \xi$ with ξ a nonsquare in $\mathrm{GF}(q)$ for q odd, and that $\epsilon^2 + \epsilon = s$ with $s \in C_1$ and $s \neq 1$ for q even, where C_1 stands for the set of elements in $\mathrm{GF}(q)$ with absolute trace 1. Furthermore, Tr denotes the trace map $x \in \mathrm{GF}(q^2) \to x + x^q \in \mathrm{GF}(q)$.

Proposition 2.4. *Let $\alpha, \beta \in \mathrm{GF}(q^2)$ satisfy the following conditions:*

$$\begin{cases} 4\alpha^{q+1} + (\beta^q - \beta)^2 \textit{ is nonsquare in } \mathrm{GF}(q), \textit{for q odd}, \\ \alpha^{q+1}/(\beta^q + \beta)^2 \in C_0 \textit{ with } \beta \in \mathrm{GF}(q^2) \setminus \mathrm{GF}(q), \textit{for q even,} \end{cases}$$

where C_0 stands for the set of elements in $\mathrm{GF}(q)$, q even, with absolute trace 0. Let $Q_{\alpha,\beta}$ be the quadratic form on $V(6,q^2)$ given by

$$\begin{aligned} &Q_{\alpha,\beta}(X_1, X_2, Y_1, Y_2, Z_1, Z_2) = \\ &\quad \delta^q X_1 Z_2 + \delta X_2 Z_1 + \alpha\delta Y_1^2 + \alpha^q \delta^q Y_2^2 + \mathrm{Tr}(\delta\beta) Y_1 Y_2, \end{aligned} \tag{2.3}$$

with $\delta = \epsilon$ or $\delta = 1$ according as q is odd or even. then the restriction $\widehat{Q}_{\alpha,\beta}$ of $Q_{\alpha,\beta}$ on $\widehat{V}$ defines an elliptic quadratic form on $\widehat{V}$.

Proof. Two cases are treated separately according as q is odd or even.

If q is odd, let $b_{\alpha,\beta}$ denote the symmetric bilinear form on $V(6,q^2)$ associated to $Q_{\alpha,\beta}$. The matrix of $b_{\alpha,\beta}$ in the canonical basis is

$$B_{\alpha,\beta} = \begin{pmatrix} O_2 & O_2 & E \\ O_2 & A_{\alpha,\beta} & O_2 \\ \overline{E} & O_2 & O_2 \end{pmatrix},$$

with

$$E = \begin{pmatrix} 0 & \epsilon^q \\ \epsilon & 0 \end{pmatrix}, \quad \overline{E} = \begin{pmatrix} 0 & \epsilon \\ \epsilon^q & 0 \end{pmatrix} \quad \text{and} \quad A_{\alpha,\beta} = \begin{pmatrix} 2\alpha\epsilon & \mathrm{Tr}(\epsilon\beta) \\ \mathrm{Tr}(\epsilon\beta) & 2\alpha^q\epsilon^q \end{pmatrix}.$$

A straightforward computation shows that $B_{\alpha,\beta}$ induces a symmetric bilinear form on $\widehat{V}$. Let $\widehat{Q}_{\alpha,\beta}$ denote the resulting quadratic form on $\widehat{V}$.

Since $\det A_{\alpha,\beta} = 4\alpha^{q+1} + (\beta^q - \beta)^2$ is nonsquare in $\mathrm{GF}(q)$, it follows that $Q_{\alpha,\beta}$ is nondegenerate. Hence $\widehat{Q}_{\alpha,\beta}$ is nondegenerate, as well. Let H be the four-dimensional subspace $\{(x, x^q, 0, 0, z, z^q) : x, z \in GF(q^2)\}$ of $\widehat{V}$. Then the restriction of $\widehat{Q}_{\alpha,\beta}$ on H is a hyperbolic quadratic form, as $L_1 = \{(x, x^q, 0, 0, 0, 0) : x \in \mathrm{GF}(q^2)\}$ and $L_2 = \{(0,0,0,0,z,z^q) : z \in GF(q^2)\}$ are totally isotropic subspaces with trivial intersection. The orthogonal space of H with respect to $b_{\alpha,\beta}$ is $L = \{(0,0,y,y^q,0,0) : y \in \mathrm{GF}(q^2)\}$. By [22, Proposition 2.5.11], $\widehat{Q}_{\alpha,\beta}$ is elliptic if and only if the restriction of $\widehat{Q}_{\alpha,\beta}$ on L is elliptic, that is,

$$\mathrm{Tr}(\alpha\epsilon y^2 + \epsilon\beta y^{q+1}) = 0 \tag{2.4}$$

has no solution $y \in \mathrm{GF}(q^2)$ other than 0.

Write $y = y_0 + \epsilon y_1$, $\alpha = a_0 + \epsilon a_1$ and $\beta = b_0 + \epsilon b_1$ with $y_0, y_1, a_0, a_1, b_0, b_1 \in \mathrm{GF}(q)$. As $\epsilon^q = -\epsilon$ and $\epsilon^2 = \xi$, we have

$$\begin{aligned} y^q &= y_0 - \epsilon y_1 \\ y^{q+1} &= y_0^2 - \xi y_1^2 \\ y^2 &= y_0^2 + \xi y_1^2 + 2\epsilon y_0 y_1 \\ y^{2q} &= y_0^2 + \xi y_1^2 - 2\epsilon y_0 y_1 \\ \alpha\epsilon y^2 &= \xi(2a_0 y_0 y_1 + a_1(y_0^2 + \xi y_1^2)) + \epsilon(a_0(y_0^2 + \xi y_1^2) + 2\xi a_1 y_0 y_1) \\ \alpha^q\epsilon^q y^{2q} &= \xi(2a_0 y_0 y_1 + a_1(y_0^2 + \xi y_1^2)) - \epsilon(a_0(y_0^2 + \xi y_1^2) + 2\xi a_1 y_0 y_1), \end{aligned}$$

whence

$$\mathrm{Tr}(\alpha\epsilon y^2) = 2\xi(2a_0 y_0 y_1 + a_1(y_0^2 + \xi y_1^2)).$$

Moreover,

$$\mathrm{Tr}(\epsilon\beta y^{q+1}) = 2\xi b_1(y_0^2 - \xi y_1^2).$$

Then Equation (2.4) has a nontrivial solution $y \in \mathrm{GF}(q^2)$ if and only if $(y_0, y_1) \neq (0,0)$ with $y_0, y_1 \in \mathrm{GF}(q)$ is a solution of

$$(a_1 + b_1)y_0^2 + 2a_0 y_0 y_1 + \xi(a_1 - b_1)y_1^2 = 0. \tag{2.5}$$

By a straightforward computation, (2.5) occurs if and only if $4\alpha^{q+1} + (\beta^q - \beta)^2 = u^2$ for some $u \in \mathrm{GF}(q)$. But the latter equation contradicts our hypothesis. Therefore, Equation (2.4) has no nontrivial solution in $\mathrm{GF}(q^2)$ and hence $\widehat{Q}_{\alpha,\beta}$ is elliptic.

For q even, the above approach still works up to some differences due to the fact that the well known formula solving equations of degree 2 fails in even characteristic. For completeness, we give all details.

If q is even, the restriction of $Q_{\alpha,\beta}$ on $\widehat{V}$ is a quadratic form $\widehat{Q}_{\alpha,\beta}$ on $\widehat{V}$, and the matrix of the associated bilinear form b_β is

$$B_\beta = \begin{pmatrix} O_2 & O_2 & E \\ O_2 & A_\beta & O_2 \\ E & O_2 & O_2 \end{pmatrix},$$

where

$$E = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad A_\beta = \begin{pmatrix} 0 & \mathrm{Tr}(\beta) \\ \mathrm{Tr}(\beta) & 0 \end{pmatrix}.$$

Since $\beta \notin \mathrm{GF}(q)$, a straightforward computation shows that the radical of b_β is trivial, which gives $\widehat{Q}_{\alpha,\beta}$ is nonsingular. As for the odd q case, the orthogonal space of H with respect to b_β is L. Therefore, $\widehat{Q}_{\alpha,\beta}$ is elliptic if and only if

$$\mathrm{Tr}(\alpha y^2 + \beta y^{q+1}) = 0 \tag{2.6}$$

has no nontrivial solution $y \in \mathrm{GF}(q^2)$.

As before, let $y = y_0 + \epsilon y_1, \alpha = a_0 + \epsilon a_1$ and $\beta = b_0 + \epsilon b_1$ with $y_0, y_1, a_0, a_1, b_0, b_1 \in \mathrm{GF}(q)$. As $\epsilon^q = \epsilon + 1$ and $\epsilon^2 = \epsilon + s$, with $s \in C_1$, we have

$$\begin{aligned}
y^q &= y_0 + y_1 + \epsilon y_1\\
y^{q+1} &= y_0^2 + y_0y_1 + sy_1^2\\
y^2 &= y_0^2 + sy_1^2 + \epsilon y_1^2\\
y^{2q} &= y_0^2 + (s+1)y_1^2 + \epsilon y_1^2\\
\alpha y^2 &= a_0y_0^2 + s(a_0 + a_1)y_1^2 + \epsilon(a_0y_1^2 + a_1y_0^2 + (s+1)a_1y_1^2)\\
\alpha^q y^{2q} &= a_0y_0^2 + s(a_0 + a_1)y_1^2 + (a_0y_1^2 + a_1y_0^2 + (s+1)a_1y_1^2)\\
&\quad + \epsilon(a_0y_1^2 + a_1y_0^2 + (s+1)a_1y_1^2),
\end{aligned}$$

whence

$$\mathrm{Tr}(\alpha y^2) = a_0y_1^2 + a_1y_0^2 + (s+1)a_1y_1^2,$$

and

$$\mathrm{Tr}(\beta y^{q+1}) = b_1(y_0^2 + y_0y_1 + sy_1^2).$$

Therefore, Equation (2.6) has a nontrivial solution in $\mathrm{GF}(q^2)$ if and only if

$$(a_1 + b_1)y_0^2 + b_1y_0y_1 + (a_0 + a_1 + sa_1 + sb_1)y_1^2 = 0.$$

Assume $y = y_0 \in \mathrm{GF}(q)$ is a nontrivial solution of (2.6). Then $a_1 = b_1$. This gives

$$\frac{\alpha^{q+1}}{(\beta^q + \beta)^2} = \frac{a_0^2}{a_1^2} + \frac{a_0}{a_1} + s \in C_1,$$

a contradiction since

$$\frac{a_0^2}{a_1^2} + \frac{a_0}{a_1} \in C_0.$$

Assume that $y = y_0 + \epsilon y_1 \in \mathrm{GF}(q^2)$, with $y_1 \neq 0$, is a solution of (2.6). Then $y_0y_1^{-1}$ is a solution of

$$(a_1 + b_1)X^2 + b_1X + a_0 + a_1 + s(a_1 + b_1) = 0, \tag{2.7}$$

where $b_1 \neq 0$.

Let $Y = (a_1 + b_1)b_1^{-1}X$. Replacing X by Y in (2.7) gives $Y^2 + Y + d = 0$ where

$$d = \frac{a_0^2 + a_1a_0 + sa_1^2}{b_0^2} + \frac{a_0^2 + a_1^2}{b_0^2} + \frac{a_0 + a_1}{b_0} + s.$$

Here, $d \in C_1$ by

$$\frac{a_0^2 + a_1a_0 + sa_1^2}{b_0^2} = \frac{\alpha^{q+1}}{(\beta^q + \beta)^2} \in C_0.$$

This shows that Equation (2.7) has no nontrivial solution in $\mathrm{GF}(q)$. Hence Equation (2.6) has no nontrivial solution in $\mathrm{GF}(q^2)$, as well. Therefore $\widehat{Q}_{\alpha,\beta}$ is elliptic. □

Let $\widehat{\mathcal{Q}}_{\alpha,\beta}$ stand for the elliptic quadric in $\mathrm{PG}(\widehat{V})$ defined by the quadratic form $\widehat{Q}_{\alpha,\beta}$ on $\widehat{V}$. Then the coordinates of the points of $\mathrm{PG}(\widehat{V})$ that lie on $\widehat{\mathcal{Q}}_{\alpha,\beta}$ satisfy the equation

$$\delta^q XZ^q + \delta X^q Z + \alpha\delta Y^2 + \alpha^q\delta^q Y^{2q} + \mathrm{Tr}(\delta\beta)Y^{q+1} = 0, \tag{2.8}$$

with $\delta = \epsilon$ or $\delta = 1$ according as q is odd or even.

3 The GF(q)-linear representation of Buekenhout-Metz unitals

In the light of Proposition 2.1, we introduce another incidence structure $\Pi(\widehat{\mathcal{S}})$.

Let $\widehat{\phi}$ be the bijective map defined by

$$\widehat{\phi}\colon \quad \begin{array}{ccc} V(3,q^2) & \longrightarrow & \widehat{V} \\ (x,y,z) & \longmapsto & (x,x^q,y,y^q,z,z^q) \end{array} .$$

By Proposition 2.1, $\widehat{\phi}$ is the field reduction of $V(3,q^2)$ over $\mathrm{GF}(q)$ in the basis $\{v_1^{(i)}, v_2^{(i)}, i=1,2,3\}$ of $V(6,q^2)$.

The points of $\mathrm{PG}(2,q^2)$ are mapped by $\widehat{\phi}$ to the two-dimensional $\mathrm{GF}(q)$-subspaces of $\widehat{V}$ of the form

$$\{(\lambda x, \lambda^q x^q, \lambda y, \lambda^q y^q, \lambda z, \lambda^q z^q) : \lambda \in \mathrm{GF}(q^2)\}, \text{ for } x,y,z \in \mathrm{GF}(q^2),$$

and hence lines of $\mathrm{PG}(\widehat{V})$. Such lines form a line-spread $\widehat{\mathcal{S}}$ of $\mathrm{PG}(\widehat{V})$. By Proposition 2.1 and Corollary 2.2, $\widehat{\mathcal{S}}$ is projectively equivalent to $\mathcal{S}$ in $\mathrm{PG}(5,q^2)$. Hence, $\widehat{\mathcal{S}}$ is also a Desarguesian line-spread of $\mathrm{PG}(\widehat{V})$. Therefore, in $\mathrm{PG}(5,q^2)$ $\Pi(\widehat{\mathcal{S}})$ is projectively equivalent to the $\mathrm{GF}(q)$-linear representation $\Pi(\mathcal{S})$ of $\mathrm{PG}(2,q^2)$.

The following lemma goes back to Singer, see [31].

Lemma 3.1. *Let ω be a primitive element of $\mathrm{GF}(q^2)$ over $\mathrm{GF}(q)$ with minimal polynomial $f(T) = T^2 - p_1T - p_0$. then the multiplication by ω in $\mathrm{GF}(q^2)$ defines a Singer cycle of $V(2,q) = \{(a,b) : a,b \in \mathrm{GF}(q)\}$ whose matrix is the companion matrix of $f(T)$.*

Proposition 3.2. *Any endomorphism of $V(3,q^2)$ with matrix $A = (a_{ij})$ defines the endomorphism of $\widehat{V}$ with matrix*

$$\begin{pmatrix} D_{11} & D_{12} & D_{13} \\ D_{21} & D_{22} & D_{23} \\ D_{31} & D_{32} & D_{33} \end{pmatrix},$$

where $D_{ij} = \mathrm{diag}(a_{ij}, a_{ij}^q)$.

The Frobenius transformation $\psi\colon (x,y,z) \mapsto (x^q,y^q,z^q)$ of $V(3,q^2)$ defines the endomorphism of $\widehat{V}$ with matrix

$$\begin{pmatrix} \widehat{F} & 0 & 0 \\ 0 & \widehat{F} & 0 \\ 0 & 0 & \widehat{F} \end{pmatrix},$$

where

$$\widehat{F} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Proof. The Singer cycle defined by a primitive element ω of $\mathrm{GF}(q^2)$ over $\mathrm{GF}(q)$ acts on the $\mathrm{GF}(q)$-vector space $\{(x,x^q) : x \in \mathrm{GF}(q^2)\}$ by the matrix $D = \mathrm{diag}(\omega, \omega^q)$. For every entry a_{ij} of A, write $a_{ij} = \omega^{e(i,j)}$, $0 \le e(i,j) \le q^2 - 2$. From Lemma 3.1, the multiplication by a_{ij} in $\mathrm{GF}(q^2)$ defines the endomorphism with matrix $D^{e(i,j)} = \mathrm{diag}(a_{ij}, a_{ij}^q)$. From this the first part of the proposition follows. The second part comes from Cooperstein's paper [14]. □

Remark 3.3. From a result due to Dye [16], the stabilizer of the Desarguesian partition $\mathcal{K}$ in $\mathrm{GL}(6, q)$ is the semidirect product of the field extension subgroup $\mathrm{GL}(3, q^2)$ by the cyclic subgroup $\langle\psi\rangle$ generated by the Frobenius transformation. In terms of projective geometry, the stabilizer of the Desarguesian spread $\mathcal{S}$ in $\mathrm{PGL}(6, q)$ is $(\mathrm{GL}(3, q^2) \rtimes \langle\psi\rangle)/\,\mathrm{GF}(q)^*$ [16]. It should be noted that the center of $\mathrm{GL}(\widehat{V})$ is the subgroup $\{cI : c \in \mathrm{GF}(q)^*\}$. Proposition 3.2 provides the representation in $\mathrm{GL}(\widehat{V})$ and $\mathrm{PGL}(\widehat{V})$ of these stabilizers.

In [2] and [17] the orthogonal Buekenhout-Metz unitals are coordinatized in $\mathrm{PG}(2, q^2)$. Let L_∞ be the line of $\mathrm{PG}(2, q^2)$ with equation $Z = 0$ and $P_\infty = \langle(1, 0, 0)\rangle_{q^2}$.

Theorem 3.4. *Let $\alpha, \beta \in \mathrm{GF}(q^2)$ such that*

$$\begin{cases} 4\alpha^{q+1} + (\beta^q - \beta)^2 \text{ is nonsquare in } \mathrm{GF}(q), \text{for } q \text{ odd}, \\ \alpha^{q+1}/(\beta^q + \beta)^2 \in C_0 \text{ with } \beta \in \mathrm{GF}(q^2) \setminus \mathrm{GF}(q), \text{for } q \text{ even}. \end{cases}$$

Then

$$U_{\alpha,\beta} = \{\langle(\alpha y^2 + \beta y^{q+1} + r, y, 1)\rangle_{q^2} : y \in \mathrm{GF}(q^2), r \in \mathrm{GF}(q)\} \cup \{P_\infty\}$$

is an orthogonal Buekenhout-Metz unital. $U_{\alpha,\beta}$ is classical if and only if $\alpha = 0$.

Conversely, every orthogonal Buekenhout-Metz unital can be expressed as $U_{\alpha,\beta}$ for some $\alpha, \beta \in \mathrm{GF}(q^2)$ which satisfy the above conditions.

We go back to the projective equivalence of $\Pi(\mathcal{S})$ and $\Pi(\widehat{\mathcal{S}})$ arising from the bijective map $\widehat{\phi}$. The line set $\widehat{\phi}(U_{\alpha,\beta}) = \{\widehat{\phi}(P) : P \in U_{\alpha,\beta}\}$ can be regarded as the restriction on $U_{\alpha,\beta}$ of the $\mathrm{GF}(q)$-linear representation of $\mathrm{PG}(2, q^2)$ in $\mathrm{PG}(\widehat{V})$.

Remark 3.5. Thas [33] showed that the $\mathrm{GF}(q)$-linear representation of the classical unital is a partition of an elliptic quadric in $\mathrm{PG}(5, q)$. Thas's result is obtained here when the representation $\widehat{\phi}(U_{0,\beta})$ is used. Let $\delta = \epsilon$ for odd q, and $\delta = 1$ for even q. For any $\beta \in \mathrm{GF}(q^2)$ satisfying the conditions of Theorem 3.4, $U_{0,\beta}$ is the set of absolute points of the unitary polarity associated to the Hermitian form h_β of $V(3, q^2)$ with matrix

$$H_\beta = \begin{pmatrix} 0 & 0 & \delta^q \\ 0 & \mathrm{Tr}(\delta\beta) & 0 \\ \delta & 0 & 0 \end{pmatrix}.$$

Hence $U_{0,\beta}$ has equation

$$\delta X^q Z + \delta^q X Z^q + \mathrm{Tr}(\delta\beta) Y^{q+1} = 0.$$

Let Tr denote the trace map of $\mathrm{GF}(q^2)$ over $\mathrm{GF}(q)$. For any $v, v' \in V(3, q^2)$,

$$\mathrm{Tr}(h_\beta(v, v')) = \begin{cases} b_{0,\beta}(\widehat{\phi}(v), \widehat{\phi}(v')), & \text{for } q \text{ odd} \\ b_\beta(\widehat{\phi}(v), \widehat{\phi}(v')), & \text{for } q \text{ even.} \end{cases}$$

This shows that the points in $\widehat{\phi}(U_{0,\beta})$ belong to $\widehat{\mathcal{Q}}_{0,\beta}$. In particular, the line set $\widehat{\phi}(U_{\alpha,\beta})$ is a partition of $\widehat{\mathcal{Q}}_{0,\beta}$.

We now put in evidence the relation between the elliptic quadric $\widehat{\mathcal{Q}}_{\alpha,\beta}$ and the Buekenhout representation of $U_{\alpha,\beta}$ in the André/Bruck-Bose model of $\mathrm{PG}(2,q^2)$.

The subspace $\Lambda = \{\langle(x,x^q,y,y^q,c,c)\rangle_q : c \in \mathrm{GF}(q), x,y \in \mathrm{GF}(q^2)\}$ is an hyperplane of $\mathrm{PG}(\widehat{V})$ containing the 3-dimensional subspace $\Sigma = \{\langle(x,x^q,y,y^q,0,0)\rangle_q : x,y \in \mathrm{GF}(q^2)\}$. The line set $\mathcal{N} = \{\widehat{\phi}(P) : P \in L_\infty\}$ is a Desarguesian line spread of Σ. Hence, $\mathcal{N}$ defines the André/Bruck-Bose model of $\mathrm{PG}(2,q^2)$ in Λ: the points are the lines of $\mathcal{N}$ and the points of Λ not in Σ, the *lines* are the planes of Λ not in Σ which meet Σ in a line of $\mathcal{N}$ and $\mathcal{N}$ itself, *incidence* is defined by inclusion. We denote by $\pi(\mathcal{N})$ this model of $\mathrm{PG}(2,q^2)$. The set $\overline{U}_{\alpha,\beta} = \bigcup_{P\in U_{\alpha,\beta}}(\widehat{\phi}(P)\cap\Lambda)$ is the Buekenhout representation of $U_{\alpha,\beta}$ in $\pi(\mathcal{N})$.

The hyperplane Λ is the orthogonal space of the point $R = \langle(1,1,0,0,0,0)\rangle_q$ with respect the polarity associated with the quadric $\widehat{\mathcal{Q}}_{\alpha,\beta}$. Since $R \in \widehat{\mathcal{Q}}_{\alpha,\beta}$, the intersection between Λ and $\widehat{\mathcal{Q}}_{\alpha,\beta}$ is a cone $\Gamma_{\alpha,\beta}$ projecting an elliptic quadric from R and containing the spread element $\widehat{\phi}(P_\infty) = \{\langle(x,x^q,0,0,0,0)\rangle_q : x \in \mathrm{GF}(q^2)\}$ as a generator.

Proposition 3.6. *The cone $\Gamma_{\alpha,\beta}$ coincides with the Buekenhout representation $\overline{U}_{\alpha,\beta}$ of $U_{\alpha,\beta}$ in $\pi(\mathcal{N})$, that is,*

$$\bigcup_{P\in U_{\alpha,\beta}}(\widehat{\phi}(P)\cap\Lambda) = \Gamma_{\alpha,\beta}.$$

Proof. We have $\widehat{\phi}(P_\infty) = \widehat{\mathcal{Q}}_{\alpha,\beta}\cap\Sigma$. For any $P = \langle(ay^2+\beta y^{q+1},y,1)\rangle_{q^2} \in U_{\alpha,\beta}$,

$$\widehat{\phi}(P) = \{\langle(\lambda(ay^2+\beta y^{q+1}),\lambda^q(a^qy^{2^q}+\beta^qy^{q+1}),\lambda y,\lambda^qy^q,\lambda,\lambda^q)\rangle_q : \lambda\in\mathrm{GF}(q^2)\}.$$

Then $\widehat{\phi}(P)\cap\Lambda = \langle(\alpha y^2+\beta y^{q+1}+r,\alpha^qy^{2q}+\beta^qy^{q+1}+r,y,y^q,1,1)\rangle_q$. From a straightforward calculation involving Equation (2.8) of $\widehat{\mathcal{Q}}_{\alpha,\beta}$ it follows that $\widehat{\phi}(P)\cap\Lambda \in \Gamma_{\alpha,\beta}$. Since the size of $\bigcup_{P\in U_{\alpha,\beta}\setminus\{P_\infty\}}(\widehat{\phi}(P)\cap\Lambda)$ equals the size of $\Gamma_{\alpha,\beta}\setminus\widehat{\phi}(P_\infty)$ the result follows. □

Remark 3.7. The affine points of $\Gamma_{\alpha,\beta}$ satisfy the equation

$$\delta^qX+\delta X^q+\alpha\delta Y^2+\alpha^q\delta^qY^{2q}+\mathrm{Tr}(\delta\beta)Y^{q+1}=0, \tag{3.1}$$

with $\delta=\epsilon$ or $\delta=1$ according as q is odd or even. It may be observed that Equation (3.1) is the equation of the affine points of $U_{\alpha,\beta}$ [13, 20]. Equation (3.1) in homogeneous form is

$$\delta^qXZ^{2q-1}+\delta X^qZ^q+\alpha\delta Y^2Z^{2q-2}+\alpha^q\delta^qY^{2q}+\mathrm{Tr}(\delta\beta)Y^{q+1}Z^{q-1}=0,$$

which is satisfied by the points of the $\mathrm{GF}(q)$-linear representation $\widehat{\phi}(U_{\alpha,\beta})$ of $U_{\alpha,\beta}$.

In [28], Polverino proved that the $\mathrm{GF}(q)$-linear representation of an orthogonal Buekenhout-Metz unital cover the $\mathrm{GF}(q)$-points of an algebraic hypersurface of degree four minus the complements of a line in a three-dimensional subspace. She also showed that the hypersurface is reducible if and only if the unital is classical. Polverino's result is obtained here when the representation $\widehat{\phi}(U_{0,\beta})$ is used. Let $\mathcal{F}$ be the hypersurface of $\mathrm{PG}(5,q^2)$ with equation

$$\mathcal{F}\colon \delta^qX_1Z_1Z_2^2+\delta X_2Z_1^2Z_2+\alpha\delta Y_1^2Z_2^2+\alpha^q\delta^qY_2^2Z_1^2+\mathrm{Tr}(\delta\beta^q)Y_1Y_2Z_1Z_2=0.$$

The intersection $\widehat{\mathcal{F}}$ of $\mathcal{F}$ with $\mathrm{PG}(\widehat{V})$ consists of all points of $\mathrm{PG}(\widehat{V})$ satisfying the equation

$$\delta^q XZ^{2q+1} + \delta X^q Z^{q+2} + \alpha\delta Y^2 Z^{2q} + \alpha^q \delta^q Y^{2q} Z^2 + \mathrm{Tr}(\delta\beta^q) Y^{q+1} Z^{q+1} = 0. \quad (3.2)$$

Clearly, $\widehat{\mathcal{F}}$ contains the three-dimensional subspace Σ. By the above arguments, the $\mathrm{GF}(q)$-linear representation $\widehat{\phi}(U_{\alpha,\beta})$ covers the points in $\widehat{\mathcal{F}}$ minus the complements of $\widehat{\phi}(L_\infty)$ in Σ. Furthermore, Equation (3.2) defines an algebraic hypersurface of degree four of $\mathrm{PG}(5, q)$. A straightforward, though tedious, calculation shows that Equation (3.2) is precisely the algebraic hypersurface provided by Polverino in [28].

As elliptic quadrics in $\mathrm{PG}(\widehat{V})$ are projectively equivalent, some linear collineation τ_α of $\mathrm{PG}(\widehat{V})$ takes $\widehat{\mathcal{Q}}_{0,\beta}$ to $\widehat{\mathcal{Q}}_{\alpha,\beta}$. Actually we need such a linear collineation τ_α with some extra-property.

Proposition 3.8. *In* $\mathrm{PG}(\widehat{V})$ *there exists a linear collineation* τ_α *which takes* $\widehat{\mathcal{Q}}_{0,\beta}$ *to* $\widehat{\mathcal{Q}}_{\alpha,\beta}$*, preserves the subspaces* Λ*,* Σ*, and fixes* $\widehat{\phi}(P_\infty)$ *pointwise. Therefore it maps the cone* $\Gamma_{0,\beta}$ *into* $\Gamma_{\alpha,\beta}$*.*

Proof. The restriction $\widehat{Q}_{\alpha,\beta}|_L$ on the subspace $L = \{(0, 0, y, y^q, 0, 0) : y \in \mathrm{GF}(q^2)\}$ of $\widehat{Q}_{\alpha,\beta}$ given by (2.3) is the quadratic form defined by

$$\widehat{Q}_{\alpha,\beta}|_L(y, y^q) = \alpha\delta y^2 + \alpha^q \delta^q y^{2q} + \mathrm{Tr}(\delta\beta) y^{q+1} \in \mathrm{GF}(q)$$

which is of elliptic type by the proof of Proposition 2.4. As two such forms are equivalent, some endomorphism of L maps $\widehat{Q}_{0,\beta}|_L$ to $\widehat{Q}_{\alpha,\beta}|_L$. In a natural way, as in the proof of Proposition 2.3, we may identify any endomorphism of L with a 2 × 2 q-circulant matrix. Doing so, the endomorphism with matrix

$$D = \begin{pmatrix} d_1 & d_2 \\ d_2^q & d_1^q \end{pmatrix},$$

where

$$\begin{aligned} d_1^{q+1} + d_2^{q+1} &= 1 \\ d_1 d_2^q &= \alpha\delta\, \mathrm{Tr}(\delta\beta)^{-1}, \end{aligned}$$

maps $\widehat{Q}_{0,\beta}|_L$ to $\widehat{Q}_{\alpha,\beta}|_L$. Let τ_α be the linear collineation of $\mathrm{PG}(\widehat{V})$ defined by the matrix

$$D_\alpha = \begin{pmatrix} I_2 & O_2 & O_2 \\ O_2 & D & O_2 \\ O_2 & O_2 & I_2 \end{pmatrix}.$$

It is easily seen that τ_α preserves the subspaces Λ, Σ, and fixes $\widehat{\phi}(P_\infty)$ pointwise, and that it maps the cone $\Gamma_{0,\beta}$ into $\Gamma_{\alpha,\beta}$. □

Remark 3.9. Bearing in mind Remark 3.3, one can ask whether τ_α is an incidence preserving map of $\Pi(\widehat{\mathcal{S}})$. The answer is negative by $d_1 d_2 \neq 0$ and Proposition 3.2. This implies that $\Gamma_{0,\beta}$ and $\Gamma_{\alpha,\beta}$ are Buekenhout representations of unitals of $\mathrm{PG}(2, q^2)$ and that they are not projectively equivalent. In particular, this provides a new proof for the existence of non-classical unitals embedded in $\mathrm{PG}(2, q^2)$.

It is clear that the image $\widehat{\mathcal{S}}^{\tau_\alpha}$ of the Desarguesian line-spread $\widehat{\mathcal{S}}$ under the linear collineation τ_α is a Desarguesian line-spread and it defines the $\mathrm{GF}(q)$-linear representation $\Pi(\widehat{\mathcal{S}}^{\tau_\alpha})$ of $\mathrm{PG}(2,q^2)$.

4 The proof of the Main Theorem

In our proof the models of $\mathrm{PG}(2,q^2)$ treated in Section 3 play a role. Two of them arose from Desarguesian line-spreads of $\mathrm{PG}(\widehat{V})$ denoted by $\widehat{\mathcal{S}}$ and $\widehat{\mathcal{S}}^{\tau_\alpha}$ respectively, the third was the André/Bruck-Bose model $\pi(\mathcal{N})$ in the 4-dimensional subspace Λ.

In $\mathrm{PG}(2,q^2)$ consider a unital $\mathcal{U}$ isomorphic, as a block-design, to an orthogonal Buekenhout-Metz unital $U_{\alpha,\beta}$ with $\alpha \neq 0$. It is known [2, 17] that $U_{\alpha,\beta}$ has a special point which is the unique fixed point of the automorphism group of $U_{\alpha,\beta}$. Hence the automorphism group of $\mathcal{U}$ fixes a unique point of $\mathcal{U}$. Up to a change of the homogeneous coordinate system in $\mathrm{PG}(2,q^2)$, the special point of $U_{\alpha,\beta}$ is $P_\infty = \langle(1,0,0)\rangle_{q^2}$ and the tangent line of $U_{\alpha,\beta}$ at P_∞ is $L_\infty\colon Z=0$. Up to a linear collineation, $P_\infty \in \mathcal{U}$ is the fixed point of the automorphism group of $\mathcal{U}$ and L_∞ is the tangent to $\mathcal{U}$ at P_∞. Therefore, $\mathcal{U}$ and $U_{\alpha,\beta}$ share P_∞ and L_∞.

We interpret the isomorphism between $\mathcal{U}$ and $U_{\alpha,\beta}$ in each of the above three models of $\mathrm{PG}(2,q^2)$. The representation $\widehat{\mathcal{U}} = \{\widehat{\phi}(P) : P \in \mathcal{U}\}$ of $\mathcal{U}$ in $\Pi(\widehat{\mathcal{S}})$ is isomorphic, as a block-design, to $\widehat{U}_{\alpha,\beta} = \{\widehat{\phi}(P) : P \in U_{\alpha,\beta}\}$. The Buekenhout representation $\overline{\mathcal{U}} = \bigcup_{P\in\mathcal{U}} (\widehat{\phi}(P)\cap\Lambda)$ of $\mathcal{U}$ in $\pi(\mathcal{N})$ is isomorphic, as a block-design, to $\overline{U}_{\alpha,\beta} = \bigcup_{P\in U_{\alpha,\beta}} (\widehat{\phi}(P)\cap\Lambda)$. Here, by Proposition 3.6, $\overline{U}_{\alpha,\beta}$ is the cone $\Gamma_{\alpha,\beta}$. This gives that the representation $\widetilde{\mathcal{U}} = \{L \in \widehat{\mathcal{S}}^{\tau_\alpha} : L\cap\Lambda \subset \overline{U}\}$ of $\mathcal{U}$ in $\Pi(\widehat{\mathcal{S}}^{\tau_\alpha})$ is isomorphic, as a block-design, to $\widetilde{U}_{\alpha,\beta} = \{L \in \widehat{\mathcal{S}}^{\tau_\alpha} : L\cap\Lambda \subset \Gamma_{\alpha,\beta}\}$.

From Proposition 3.8, the lines which are the points of $\widetilde{U}_{\alpha,\beta}$ partition the elliptic quadric $\widehat{\mathcal{Q}}_{\alpha,\beta} = \widehat{\mathcal{Q}}_{0,\beta}^{\tau_\alpha}$. On the other hand, from Remark 3.5, $\widehat{\mathcal{Q}}_{0,\beta}$ is partitioned by lines which are the points of the classical unital $\widehat{U}_{0,\beta}$ in $\Pi(\widehat{\mathcal{S}})$. This yields that $\widetilde{U}_{\alpha,\beta}$ coincides with $\widehat{U}_{0,\beta}^{\tau_\alpha}$. It turns out that $\widetilde{U}_{\alpha,\beta}$ is a classical unital in $\Pi(\widehat{\mathcal{S}}^{\tau_\alpha})$, and hence $\widetilde{\mathcal{U}}$ is isomorphic, as a block-design, to the classical unital.

Now we quote the following result from [23] which was the keystone in the proof of Theorem 1.1.

Lemma 4.1. *Let $\mathcal{U}$ be a unital embedded in a Desarguesian finite projective plane π and isomorphic, as a block-design, to the classical unital. For any block B of $\mathcal{U}$, let ℓ be the line of π containing B. Then B is an orbit of a cyclic subgroup of order $q+1$ contained in the projectivity group of ℓ. This implies that B is a Baer subline of ℓ.*

We emphasize that the proof of Lemma 4.1 only uses arguments involving point-block incidences of $\mathcal{U}$ viewed as a block-design embedded in π.

Therefore, Lemma 4.1 applies to $\widetilde{\mathcal{U}}$. Thus, every block of $\widetilde{\mathcal{U}}$ is a Baer subline of $\Pi(\widehat{\mathcal{S}}^{\tau_\alpha})$, that is, a regulus of $\mathrm{PG}(\widehat{V})$. From this, each block of $\overline{\mathcal{U}}$ is the intersection of these reguli with Λ. In particular, each block of $\overline{\mathcal{U}}$ through $\widehat{\phi}(P_\infty)$ is the union of $\widehat{\phi}(P_\infty)$ with q collinear affine points, and this implies that each block of $\widehat{\mathcal{U}}$ through $\widehat{\phi}(P_\infty)$ is a regulus of $\mathrm{PG}(\widehat{V})$ whose lines are in $\widehat{\mathcal{S}}$. Under $\widehat{\phi}$, these reguli correspond to Baer sublines of $\mathrm{PG}(2,q^2)$ through P_∞. This yields that the points of $\mathcal{U}$ on each of the q^2 secant lines to $\mathcal{U}$ form a Baer subline through P_∞. By the characterization of such unitals of $\mathrm{PG}(2,q^2)$

given in [12, 29], we may conclude that $\mathcal{U}$ is a Buekenhout-Metz unital. By definition, the Buekenhout representation $\overline{\mathcal{U}}$ of $\mathcal{U}$ is a cone that project an ovoid $\mathcal{O}$ from a point of $\widehat{\phi}(P_\infty)$ not in $\mathcal{O}$. Here an *ovoid* is a set of q^2+1 points in a 3-dimensional subspace of Λ no three of which are collinear.

To conclude the proof we only need to prove that $\mathcal{O}$ is an elliptic quadric. Since the ovoids in $\mathrm{PG}(3,q)$ with odd q are elliptic quadrics, see [4, 26], we assume $q = 2^h$. In $\mathrm{PG}(3,2^h)$, there are known two ovoids, up to projectivities, namely the elliptic quadric which exist for $h \geq 1$, and the Tits ovoid which exists for odd $h \geq 3$; see [18, Chapter 10]. Let Ω be the 3-dimensional subspace of Λ containing $\mathcal{O}$. Note that $\mathcal{O} = \Omega \cap \overline{\mathcal{U}}$. Set α_∞ to be the plane $\Omega \cap \Sigma$. Then α_∞ meets $\mathcal{O}$ exactly in the point $\mathcal{O} \cap \widehat{\phi}(P_\infty)$, and it is a simple matter to show that α_∞ contains only one line $\widehat{\phi}(P)$ of $\mathcal{N}$. Also, $\widehat{\phi}(P)$ is distinct from $\widehat{\phi}(P_\infty)$. Let $\alpha_1, \dots, \alpha_q$ denote the further planes of Ω through $\widehat{\phi}(P)$. As these planes are lines of $\pi(\mathcal{N})$ through the point $\widehat{\phi}(P)$, each of them meets $\overline{\mathcal{U}}$ in 1 or $q+1$ points. This holds true for $\mathcal{O}$.

It is well known [19, Section 12.3] that in a finite Desarguesian projective plane through any point off a unital there are exactly $q+1$ tangent lines, that is, lines of the plane that intersects the unital in exactly one point. In terms of the unital $\overline{\mathcal{U}}$ this property states that there is only one plane among $\alpha_1, \dots, \alpha_q$ that meets $\mathcal{O}$ in exactly one point. Let α_1 denote this plane. Then the block $\alpha_i \cap \mathcal{O}$ of $\overline{\mathcal{U}}$, for $i = 2, \dots, q$, is the intersection of α_i with a regulus in $\mathrm{PG}(\widehat{V})$. Since that regulus does not contain $\widehat{\phi}(P)$, the block $\alpha_i \cap \mathcal{O}$ is a conic C_i of α_i, for $i = 2, \dots, q$. Thus the blocks $\alpha_i \cap \mathcal{O}$, for $i = 2, \dots, q$, are $q-1$ conics that partition all but two points of $\mathcal{O}$. By [8, Theorem 5] $\mathcal{O}$ is an elliptic quadric.

References

[1] J. André, Über nicht-Desarguessche Ebenen mit transitiver Translationsgruppe, *Math. Z.* **60** (1954), 156–186, doi:10.1007/bf01187370.

[2] R. D. Baker and G. L. Ebert, On Buekenhout-Metz unitals of odd order, *J. Comb. Theory Ser. A* **60** (1992), 67–84, doi:10.1016/0097-3165(92)90038-v.

[3] J. Bamberg, A. Betten, C. E. Praeger and A. Wassermann, Unitals in the Desarguesian projective plane of order 16, *J. Statist. Plann. Inference* **144** (2014), 110–122, doi:10.1016/j.jspi.2012.10.006.

[4] A. Barlotti, Un'estensione del teorema di Segre-Kustaanheimo, *Boll. Un. Mat. Ital. Serie 3* **10** (1955), 498–506, http://www.bdim.eu/item?id=BUMI_1955_3_10_4_498_0.

[5] S. Barwick and G. Ebert, *Unitals in Projective Planes*, Springer Monographs in Mathematics, Springer, New York, 2008, doi:10.1007/978-0-387-76366-8.

[6] A. Betten, D. Betten and V. D. Tonchev, Unitals and codes, *Discrete Math.* **267** (2003), 23–33, doi:10.1016/s0012-365x(02)00600-3.

[7] R. C. Bose, On a representation of the Baer subplanes of the Desarguesian plane $\mathrm{PG}(2,q^2)$ in a projective five dimensional space $\mathrm{PG}(5,q)$, in: *Colloquio Internazionale sulle Teorie Combinatorie, Tomo I*, Accademia Nazionale dei Lincei, Rome, 1976 pp. 381–391, proceedings of a conference held in Rome, September 3 – 15, 1973.

[8] M. R. Brown, C. M. O'Keefe and T. Penttila, Triads, flocks of conics and $Q^-(5,q)$, *Des. Codes Cryptogr.* **18** (1999), 63–70, doi:10.1023/a:1008376900914.

[9] R. H. Bruck and R. C. Bose, The construction of translation planes from projective spaces, *J. Algebra* **1** (1964), 85–102, doi:10.1016/0021-8693(64)90010-9.

[10] R. H. Bruck and R. C. Bose, Linear representations of projective planes in projective spaces, *J. Algebra* **4** (1966), 117–172, doi:10.1016/0021-8693(66)90054-8.

[11] F. Buekenhout, Existence of unitals in finite translation planes of order q^2 with a kernel of order q, *Geom. Dedicata* **5** (1976), 189–194, doi:10.1007/bf00145956.

[12] L. R. A. Casse, C. M. O'Keefe and T. Penttila, Characterizations of Buekenhout-Metz unitals, *Geom. Dedicata* **59** (1996), 29–42, doi:10.1007/bf00181524.

[13] D. B. Chandler, The sizes of the intersections of two unitals in $\mathrm{PG}(2, q^2)$, *Finite Fields Appl.* **25** (2014), 255–269, doi:10.1016/j.ffa.2013.10.001.

[14] B. N. Cooperstein, External flats to varieties in $\mathbb{PG}(M_{n,n}(\mathrm{GF}(q)))$, *Linear Algebra Appl.* **267** (1997), 175–186, doi:10.1016/s0024-3795(97)80049-3.

[15] N. Durante and A. Siciliano, Non-linear maximum rank distance codes in the cyclic model for the field reduction of finite geometries, *Electron. J. Combin.* **24** (2017), P2.33, https://www.combinatorics.org/ojs/index.php/eljc/article/view/v24i2p33.

[16] R. H. Dye, Spreads and classes of maximal subgroups of $\mathrm{GL}_n(q)$, $\mathrm{SL}_n(q)$, $\mathrm{PGL}_n(q)$ and $\mathrm{PSL}_n(q)$, *Ann. Mat. Pura Appl.* **158** (1991), 33–50, doi:10.1007/bf01759298.

[17] G. L. Ebert, On Buekenhout-Metz unitals of even order, *European J. Combin.* **13** (1992), 109–117, doi:10.1016/0195-6698(92)90042-x.

[18] J. W. P. Hirschfeld, *Finite Projective Spaces of Three Dimensions*, Oxford Mathematical Monographs, Oxford University Press, New York, 1985.

[19] J. W. P. Hirschfeld, *Projective Geometries over Finite Fields*, Oxford Mathematical Monographs, Oxford University Press, New York, 2nd edition, 1998.

[20] J. W. P. Hirschfeld and G. Korchmáros, Arcs and curves over a finite field, *Finite Fields Appl.* **5** (1999), 393–408, doi:10.1006/ffta.1999.0260.

[21] A. M. W. Hui and P. P. W. Wong, On embedding a unitary block design as a polar unital and an intrinsic characterization of the classical unital, *J. Comb. Theory Ser. A* **122** (2014), 39–52, doi:10.1016/j.jcta.2013.09.007.

[22] P. Kleidman and M. Liebeck, *The Subgroup Structure of the Finite Classical Groups*, volume 129 of *London Mathematical Society Lecture Note Series*, Cambridge University Press, Cambridge, 1990, doi:10.1017/cbo9780511629235.

[23] G. Korchmáros, A. Siciliano and T. Szőnyi, Embedding of classical polar unitals in $\mathrm{PG}(2, q^2)$, *J. Comb. Theory Ser. A* **153** (2018), 67–75, doi:10.1016/j.jcta.2017.08.002.

[24] R. Metz, On a class of unitals, *Geom. Dedicata* **8** (1979), 125–126, doi:10.1007/bf00147935.

[25] G. P. Nagy and D. Mezőfi, UnitalSZ – a GAP package, Version 0.5, 23 March 2018, https://nagygp.github.io/UnitalSZ/.

[26] G. Panella, Caratterizzazione delle quadriche di uno spazio (tridimensionale) lineare sopra un corpo finito, *Boll. Un. Mat. Ital. Serie 3* **10** (1955), 507–513, http://www.bdim.eu/item?id=BUMI_1955_3_10_4_507_0.

[27] T. Penttila and G. F. Royle, Sets of type (m, n) in the affine and projective planes of order nine, *Des. Codes Cryptogr.* **6** (1995), 229–245, doi:10.1007/bf01388477.

[28] O. Polverino, Linear representation of Buekenhout-Metz unitals, *Discrete Math.* **267** (2003), 247–252, doi:10.1016/s0012-365x(02)00618-0.

[29] C. T. Quinn and R. Casse, Concerning a characterisation of Buekenhout-Metz unitals, *J. Geom.* **52** (1995), 159–167, doi:10.1007/bf01406836.

[30] B. Segre, Teoria di Galois, fibrazioni proiettive e geometrie non desarguesiane, *Ann. Mat. Pura Appl.* **64** (1964), 1–76, doi:10.1007/bf02410047.

[31] J. Singer, A theorem in finite projective geometry and some applications to number theory, *Trans. Amer. Math. Soc.* **43** (1938), 377–385, doi:10.2307/1990067.

[32] D. E. Taylor, *The Geometry of the Classical Groups*, volume 9 of *Sigma Series in Pure Mathematics*, Heldermann Verlag, Berlin, 1992.

[33] J. A. Thas, Semipartial geometries and spreads of classical polar spaces, *J. Comb. Theory Ser. A* **35** (1983), 58–66, doi:10.1016/0097-3165(83)90026-2.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 625–639
https://doi.org/10.26493/1855-3974.1605.43f
(Also available at http://amc-journal.eu)

The dimension of the negative cycle vectors of signed graphs

Alex Schaefer [*], Thomas Zaslavsky

Binghamton University (SUNY), Department of Mathematical Sciences,
Binghamton, NY 13902-6000, U.S.A.

Received 16 February 2018, accepted 7 March 2019, published online 6 June 2019

Abstract

A *signed graph* is a graph Γ with edges labeled "$+$" and "$-$". The sign of a cycle is the product of its edge signs. Let $\mathrm{SpecC}(\Gamma)$ denote the list of lengths of cycles in Γ. We equip each signed graph with a vector whose entries are the numbers of negative k-cycles for $k \in \mathrm{SpecC}(\Gamma)$. These vectors generate a subspace of $\mathbb{R}^{|\,\mathrm{SpecC}(\Gamma)|}$. Using matchings with a strong permutability property, we provide lower bounds on the dimension of this space; in particular, we show for complete graphs, complete bipartite graphs, and a few other graphs that this space is all of $\mathbb{R}^{|\,\mathrm{SpecC}(\Gamma)|}$.

Keywords: Signed graph, negative cycle vector, permutable matching.

Math. Subj. Class.: 05C22, 05C38

1 Introduction

A *signed graph* Σ is a graph Γ whose edges have sign labels, either "$+$" or "$-$". The sign of a cycle in the graph is the product of the signs of its edges. Write $c_l^-(\Sigma)$ for the number of negative cycles of length l in Σ and collect these numbers in the *negative cycle vector* $c^-(\Sigma) = (c_3^-, c_4^-, \dots, c_n^-) \in \mathbb{R}^{n-2}$, where n is the order of Σ. We are interested in the structure of the collection $\mathrm{NCV}(\Gamma)$ of all negative cycle vectors of signings of a fixed underlying simple graph Γ.

The negative cycle numbers are of interest for several reasons. Ours is that, while the structure of a signed graph is more complex than that of an unsigned graph, much of that complexity is traceable to the distribution of negative cycles. We think negative cycle vectors are a step towards better understanding of those cycles. Beyond this, negative cycle numbers have been an object of interest since the first days of signed graph theory. When

[*] Present address: University of Kansas, Department of Mathematics, Lawrence, KS 66045-7594, U.S.A.
E-mail addresses: alex.scha4@ku.edu (Alex Schaefer), zaslav@math.binghamton.edu (Thomas Zaslavsky)

signed graphs were introduced by Harary [2] to be applied to a problem in social psychology by Cartwright and Harary [1], one of their concerns was to measure how unbalanced a signed graph is. One measure they proposed was the proportion of negative cycles, i.e., $\big[\sum_l c_l^-(\Sigma)\big] / \big[\sum_l c_l(\Gamma)\big]$, where c_l denotes the total number of l-cycles in the graph. This proportion is hard to calculate even for signed complete graphs, since the number of cycles can be exponential in the order n and the negative cycle numbers are also complicated.

There are at least three natural questions raised by the existence of the collections $\operatorname{NCV}(\Gamma)$. Most simply, since any set of points in $\mathbb{R}^{n-2}$ lies in a smallest subspace, what subspace do they span? That is the question we address here. The *cycle spectrum* $\operatorname{SpecC}(\Gamma)$ is the list of lengths of cycles in Γ. The finite set $\operatorname{NCV}(\Gamma)$ generates a subspace of $\mathbb{R}^{n-2}$ that is contained in the subspace $\mathbb{R}_{\operatorname{NCV}(\Gamma)}$ consisting of all vectors that are 0 in the coordinates that correspond to cycle lengths not in the cycle spectrum of Γ. We develop a general approach to the dimension question in terms of "permutable matchings" (see Section 2.3) that allows us to prove that, for $\Gamma = K_n$, $K_{m,n}$, and the Petersen graph, $\operatorname{NCV}(\Gamma)$ spans $\mathbb{R}_{\operatorname{SpecC}(\Gamma)}$; it also gives us a lower bound on dimension for the Heawood graph and one other graph family. We also solve a few examples with an *ad hoc* method.

Knowing the span of the negative cycle vectors, what is their convex hull? In [5] and [8] Popescu and Tomescu gave inequalities bounding the numbers of negative cycles in a signed complete graph, which may be a step towards the answer for K_n (see Section 5). A related question: Do the facets of the convex cone generated by $\operatorname{NCV}(\Gamma)$ have combinatorial meaning?

The ultimate question: Which vectors in the convex hull are actually the vectors of signed graphs? Kittipassorn and Mészáros [3], inspired by the theory of two-graphs from finite group theory and geometry (see [7]) gave strong restrictions on the number of negative triangles in a signed K_n. This is a step towards the answer for K_n.

We discuss these questions further in Section 5.

Our work was originally focused on complete graphs and complete bipartite graphs. Those cases and others led the first author to the following conjecture, to which we do not know any counterexample.

Conjecture 1.1 (Schaefer, 2015)**.** *For any graph* Γ, $\dim \operatorname{NCV}(\Gamma) = |\operatorname{SpecC}(\Gamma)|$, *the number of different lengths of cycles in* Γ.

2 Background

2.1 Graphs

A *graph* is a pair $\Gamma = (V, E)$, where $V = \{v_1, \ldots, v_n\}$ is a finite set of *vertices* and E is a finite set of unordered pairs of vertices, called *edges*. Our graphs are all unlabeled, simple, and undirected. Thus, all cycle lengths are between 3 and n.

The cycle spectrum $\operatorname{SpecC}(\Gamma)$ is the set of cycle lengths that appear in Γ. The number of cycles of length l in Γ is $c_l = c_l(\Gamma)$. The *cycle vector* of Γ is $c(\Gamma) = (c_3, c_4, \ldots, c_n)$; sometimes we omit the components that correspond to lengths l not in the cycle spectrum. The number of cycle lengths in Γ, $|\operatorname{SpecC}(\Gamma)|$, is clearly fundamental since $\dim \operatorname{NCV}(\Gamma) \leq |\operatorname{SpecC}(\Gamma)|$.

2.2 Signed graphs

A *signed graph* is a triple $\Sigma = (V, E, \sigma)$ where $\Gamma = (V, E)$ is a graph, called the *underlying graph* of Σ, and $\sigma\colon E \to \{+, -\}$ is the *sign function*. Two signed graphs are *isomorphic* if there is an isomorphism of underlying graphs that preserves edge signs. The *sign of a cycle* is the product of the signs of its edges; a signed graph in which every cycle is positive is called *balanced*. The *negative edge set* E^- is the set of negative edges of Σ and the *negative subgraph* is $\Sigma^- = (V, E^-)$, the spanning subgraph of negative edges. We sometimes write Γ_N for Γ signed so that N is its set of negative edges.

Switching Σ means choosing a vertex subset $X \subseteq V$ and negating all the edges between X and its complement. Switching yields an equivalence relation on the set of all signings of a fixed underlying graph. If Σ_2 is isomorphic to a switching of Σ_1, we say that Σ_1 and Σ_2 are *switching isomorphic*. This relation is an equivalence relation on signed graphs; we denote the equivalence class of Σ by $[\Sigma]$. A signed graph is balanced if and only if it is switching isomorphic to the all-positive graph. Signed graphs that are switching isomorphic, like those in Figure 1, have the same negative cycle vector.

The negative cycle vector of Σ is $c^-(\Sigma) = (c_3^-(\Sigma), c_4^-(\Sigma), \ldots, c_n^-(\Sigma))$, where $c_l^- = c_l^-(\Sigma)$ is the number of negative cycles of length l. As with $c(\Gamma)$, we may omit the components of $c^-(\Sigma)$ that correspond to lengths l not in the cycle spectrum. Also, we may write either $c^-(\Sigma)$ or $c^-(\sigma)$, the latter when only the signature σ is varying.

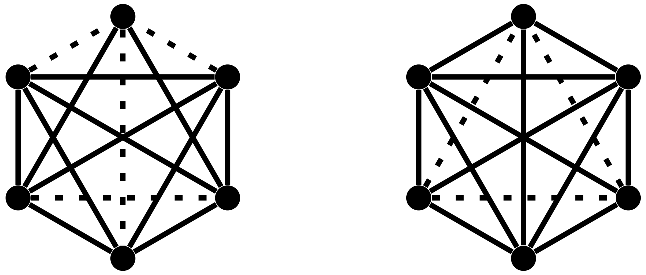


Figure 1: Two switching equivalent signings of K_6, with the same negative cycle vector $(10, 18, 36, 36)$. Solid lines are positive, dashed lines are negative.

The *negation* of Σ is $-\Sigma = (V, E, -\sigma)$, in which the sign of every edge is negated. Sometimes Σ and $-\Sigma$ are switching isomorphic, e.g., when Σ is bipartite or when it is a signed complete graph whose negative subgraph is self-complementary.

2.3 Permutable matchings

A *matching* in Γ is a set M of pairwise nonadjacent edges; it is *perfect* if $V(M) = V$. A matching M—or any other edge set—is *permutable* if some subgroup of the automorphism group of Γ acts on the edges of M as the symmetric group $S_{|M|}$. We base our results largely on permutable matchings, having noticed their utility for complete and complete bipartite graphs. The advantage of permutability is that, in counting negative cycles using a permutable matching, any two equicardinal subsets belong to the same number of negative cycles of each length. That makes it feasible to calculate the numbers in the vectors we use to estimate the dimension of $\mathrm{NCV}(\Gamma)$.

Our introduction of permutable matchings led to the question: Which graphs have permutable matchings? That has been investigated by Schaefer and Swartz in [6]; they find large families of examples. On the other hand, there are only a few kinds of graph with permutable perfect matchings; Schaefer and Swartz determine them all.

3 Rank and dimension

The dimension of $\mathrm{NCV}(\Gamma)$ is the rank of the matrix whose rows are the negative cycle vectors of all signatures of Γ. The columns of this matrix that correspond to lengths $k \in \{3, 4, \ldots, n\} \setminus \mathrm{SpecC}(\Gamma)$ are all zero; thus, we may ignore them. Since the rank cannot be greater than $|\mathrm{SpecC}(\Gamma)|$, if we produce a submatrix of that rank we have proved that $\dim \mathrm{NCV}(\Gamma) = |\mathrm{SpecC}(\Gamma)|$. That is what we now endeavor to do with the aid of a permutable matching.

Even if permutable matchings fail to reach the spectral upper bound, they imply a lower bound. However, we are happy to say that in our three main examples, permutable matchings solve the dimension problem.

The rank of a matrix A is written $\mathrm{rk}(A)$.

3.1 Any negative edge set

We begin with the most general calculation. Given a signed graph Γ_N with an arbitrary negative edge set $N \subseteq E$, how many negative cycles are there of each length? For $X \subseteq N$ let $f_l(X) =$ the number of l-cycles that intersect N precisely in X. We get a formula for f_l by Möbius inversion from $g_l(X) =$ the number of l-cycles that contain X, since

$$g_l(X) = \sum_{X \subseteq Y \subseteq N} f_l(Y),$$

which implies that

$$f_l(X) = \sum_{X \subseteq Y \subseteq N} (-1)^{|Y|-|X|} g_l(Y).$$

The number of negative l-cycles is the number of l-cycles that intersect N in an odd number of edges; therefore,

$$\begin{aligned} c_l^-(\Gamma_N) = \sum_{X \subseteq N,\, |X| \text{ odd}} f_l(X) &= \sum\sum_{X \subseteq Y \subseteq N,\, |X| \text{ odd}} (-1)^{|Y|-|X|} g_l(Y) \\ &= \sum_{Y \subseteq N} g_l(Y) \sum_{X \subseteq Y,\, |X| \text{ odd}} (-1)^{|Y|-|X|} \\ &= \sum_{\emptyset \neq Y \subseteq N} (-2)^{|Y|-1} g_l(Y). \end{aligned} \tag{3.1}$$

This applies to every underlying graph Γ.

3.2 A matrix calculation

Now assume we have a graph Γ of order n together with m unbalanced sign functions $\sigma_1, \ldots, \sigma_m$ in addition to the all-positive function $\sigma_0 \equiv +$. To avoid redundancy we want the associated signed graphs to be switching nonisomorphic. For instance, choosing more than half the edges at a vertex to be negative is switching equivalent to choosing fewer than half, so we would not want the negative edge set to contain more than $\frac{1}{2}\deg(v)$ of the edges incident with any vertex v.

We construct the matrix of the negative cycle vectors of all signings σ_s and their negatives, with columns segregated by parity. The rows are one for $+\Gamma$ (i.e., $\sigma_0 \equiv +$), then m rows for the unbalanced signatures σ_s, $0 < s \leq m$, then $-\Gamma$ (the signature $-\sigma_0 \equiv -$), then the m negations $-\sigma_s$. The relationship between the upper and lower halves is that

$$c_l^-(-\sigma_s) = \begin{cases} c_l - c_l^-(\sigma_s) & \text{if } l \text{ is odd,} \\ c_l^-(\sigma_s) & \text{if } l \text{ is even.} \end{cases}$$

The resulting matrix is

$$\left(\begin{array}{ccc|ccc} 0 & 0 & \cdots & 0 & 0 & \cdots \\ c_3^-(\sigma_1) & c_5^-(\sigma_1) & \cdots & c_4^-(\sigma_1) & c_6^-(\sigma_1) & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ c_3^-(\sigma_m) & c_5^-(\sigma_m) & \cdots & c_4^-(\sigma_m) & c_6^-(\sigma_m) & \cdots \\ c_3 & c_5 & \cdots & 0 & 0 & \cdots \\ c_3 - c_3^-(\sigma_1) & c_5 - c_5^-(\sigma_1) & \cdots & c_4^-(\sigma_1) & c_6^-(\sigma_1) & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ c_3 - c_3^-(\sigma_m) & c_5 - c_5^-(\sigma_m) & \cdots & c_4^-(\sigma_m) & c_6^-(\sigma_m) & \cdots \end{array}\right). \tag{3.2}$$

The last column in the left half is that of $n-1$ or n depending on whether n is even or odd; in the right half it is that of n or $n-1$, respectively. Row operations reduce this matrix to

$$\left(\begin{array}{ccc|ccc} 0 & 0 & \cdots & 0 & 0 & \cdots \\ c_3^-(\sigma_1) & c_5^-(\sigma_1) & \cdots & 0 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ c_3^-(\sigma_m) & c_5^-(\sigma_m) & \cdots & 0 & 0 & \cdots \\ c_3 & c_5 & \cdots & 0 & 0 & \cdots \\ 0 & 0 & \cdots & c_4^-(\sigma_1) & c_6^-(\sigma_1) & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ 0 & 0 & \cdots & c_4^-(\sigma_m) & c_6^-(\sigma_m) & \cdots \end{array}\right). \tag{3.3}$$

Ignoring the first row of zeros, this is a block matrix

$$A = \begin{pmatrix} U & O \\ c_{\text{odd}}(\Gamma) & \mathbf{0} \\ O & R \end{pmatrix}.$$

The middle row $c_{\text{odd}}(\Gamma)$, consisting of the odd-cycle numbers of Γ, corresponds to $-\Gamma$. The upper left block U is the matrix of negative odd-cycle vectors of the unbalanced signatures σ_s, and the lower right block R is the matrix of negative even-cycle vectors of the same signatures. We infer the fundamental fact that:

Lemma 3.1. *The rank of the negative cycle matrix* (3.2) *equals the sum of the ranks of* $\begin{pmatrix} U \\ c_{\text{odd}}(\Gamma) \end{pmatrix}$ *and* R.

For a bipartite graph $U = O$ and $c_{\text{odd}} = \mathbf{0}$, so only R needs to be considered.

3.3 Permutable negative matchings

Henceforth we assume we have chosen a fixed permutable matching M_m of m edges in Γ. For each $s = 1, 2, \dots, m$ we choose a submatching $M_s \subseteq M_m$ of s edges and we define the signature σ_s as that of the signed graph Γ_{M_s}. (It does not matter which M_s we use, because M_m is permutable.) This generates a matrix of negative cycle vectors as in (3.2).

Permutability implies that $g_l(Y)$ depends only on $|Y|$ so we may define $G_l(k) = g_l(Y)$ for any one k-edge subset $Y \subseteq M_m$. Then (3.1) becomes

$$c_l^-(\Gamma_{M_s}) = \sum_{k=1}^{s}(-2)^{k-1}\binom{s}{k}G_l(k) = \sum_{k=1}^{n}(-2)^{k-1}\frac{G_l(k)}{k!}(s)_k, \tag{3.4}$$

where $(x)_k$ denotes the falling factorial, $(x)_k = x(x-1)\cdots(x-[k-1])$. We may let k run up to n in the second summation because if $k > s$, the falling factorial equals 0. Formula (3.4) gives $c_l^-(\Gamma_{M_s})$ as a polynomial function $p_l(s)$ without constant term, of degree d_l where d_l is the largest integer k for which $G_l(k) > 0$; that is, d_l is the largest size of a submatching of M_m that is contained in some cycle of length l. We leave d_l undefined if no l-cycle intersects M_m. Clearly, $d_l \leq m$.

(This method works equally well for subsets of any permutable edge set N in any graph. It is easy to see that there are only three possible kinds of permutable set: a matching, a subset of the edges incident to a vertex, and the three edges of a triangle. In K_n a permutable set of edges at a vertex is useless since then the entire matrix (3.2) has rank 1. We have not seen a graph where a triangle's edges might help find the dimension.)

A column of U or R is not all zero if and only if it corresponds to a cycle length l for which there exists an l-cycle in Γ that intersects M_m. Such a column contains m values of the polynomial $p_l(s)$. Since p_l has degree at most m and no constant term, these values determine p_l completely.

Now a nonzero column in U or R for cycle length l looks like this:

$$\begin{pmatrix} p_l(1) \\ p_l(2) \\ \vdots \\ p_l(m) \end{pmatrix} = \begin{pmatrix} \alpha_l 1^{d_l} + \cdots \\ \alpha_l 2^{d_l} + \cdots \\ \vdots \\ \alpha_l m^{d_l} + \cdots \end{pmatrix}, \tag{3.5}$$

since p_l is a polynomial of degree d_l; here $\alpha_l = (-2)^{d_l-1}G_l(d_l)/d_l!$. Moreover, $d_l = \mu(l) > 0$ for a nonzero column, where we define

$$\mu(l) = \max_{C_l} |C_l \cap M_m|, \tag{3.6}$$

maximized over all l-cycles C_l.

Define δ_{odd} to be the number of distinct degrees d_l for odd lengths l whose column in U is not zero, and let δ_{even} be the number of distinct degrees d_l for even lengths whose column in R is not zero. If some values of d_l for, e.g., odd lengths l happen to be equal, they are counted only once. Thus, δ_{odd} may be less than the number of nonzero columns. The number of distinct polynomial degrees represented in the columns of U is δ_{odd}, and similarly for R the number is δ_{even}. Let Δ_{odd} be the set of distinct degrees d_l counted by δ_{odd}, and similarly for Δ_{even}.

Lemma 3.2. *The rank of U is at least δ_{odd} and that of R is at least δ_{even}.*

The rank of $\begin{pmatrix} U \\ c_{\text{odd}} \end{pmatrix}$ *is* $\operatorname{rk}(U) + 1$ *if there is an odd length l such that an l-cycle exists in Γ but no l-cycle intersects M_m.*

Proof. In U choose one column of each different degree d_l. Divide by the leading coefficient α_l, which is necessarily nonzero; this does not affect the rank. Now add columns of the form $\left(s^d\right)_{s=1}^{m}$ for every $d = 1, 2, \ldots, m$ that is not in Δ_{odd}. Column operations allow us to eliminate the lower-degree terms of the column (3.5), leaving a Vandermonde matrix with 1^d in the top row and m^d in the bottom row of column d for each $d = 1, 2, \ldots, m$, the rank of which is m. Now reverse the column operations; the rank remains the same, so the columns of U must have full column rank.

The same reasoning applies to R.

The extra 1 in the rank of $\begin{pmatrix} U \\ c_{\text{odd}} \end{pmatrix}$ arises from the fact that, under the assumption, it has a column that is zero in U but is nonzero in c_{odd}. □

3.4 Theorems

Lemma 3.2 yields our principal general theorem. Given a matching M_m and a cycle length $l \in \operatorname{SpecC}(\Gamma)$, define $\mu(l)$ by Equation (3.6).

Theorem 3.3. *Let M_m be a permutable m-matching in Γ. Then*

$$\begin{aligned} &|\{\mu(l) : \textit{odd } l \in \operatorname{SpecC}(\Gamma)\}| + |\{\mu(l) > 0 : \textit{even } l \in \operatorname{SpecC}(\Gamma)\}| \\ &\qquad\qquad \leq \dim \operatorname{NCV}(\Gamma) \leq |\operatorname{SpecC}(\Gamma)|. \end{aligned} \tag{3.7}$$

Suppose that all values $\mu(l)$ for even lengths $l \in \operatorname{SpecC}(\Gamma)$ are distinct and positive, and all values $\mu(l)$ for odd lengths $l \in \operatorname{SpecC}(\Gamma)$ are distinct. Then $\operatorname{NCV}(\Gamma)$ spans $\mathbb{R}_{\operatorname{SpecC}(\Gamma)}$.

Proof. The first part follows directly from Lemma 3.2 since

$$\dim \operatorname{NCV}(\Gamma) \geq \operatorname{rk}(A) = \operatorname{rk}\begin{pmatrix} U \\ c_{\text{odd}} \end{pmatrix} + \operatorname{rk}(R) \geq \delta_{\text{odd}} + \delta_{\text{even}}.$$

Moreover, if there is an odd length l such that $\mu(l) = 0$, then $\operatorname{rk}\begin{pmatrix} U \\ c_{\text{odd}} \end{pmatrix} = \operatorname{rk}(U) + 1 \geq \delta_{\text{even}} + 1$; that explains why we do not exclude $\mu(l) = 0$ from being counted in the odd-length part of (3.7).

In the second part, $\delta_{\text{even}} =$ the number of even cycle lengths in Γ and δ_{odd} or (if some odd $l \in \operatorname{SpecC}(\Gamma)$ has $\mu(l) = 0$) $\delta_{\text{odd}} + 1 =$ the number of odd cycle lengths in Γ. Then the left-hand side of Formula (3.7) equals $|\operatorname{SpecC}(\Gamma)|$. □

There is a simpler statement that applies to graphs with a permutable matching that is sufficiently omnipresent, i.e., meeting the condition of Theorem 3.4. Given m, define $\nu_{\text{odd}}(m) =$ the number of odd lengths $l < 2m$ in $\operatorname{SpecC}(\Gamma)$, $+1$ if there is an odd cycle length $l \geq 2m$, and define $\nu_{\text{even}}(m) =$ the number of even lengths $l < 2m$ in $\operatorname{SpecC}(\Gamma)$, $+1$ if there is an even cycle length $l \geq 2m$.

Theorem 3.4. *Suppose M_m is a permutable m-matching in Γ and for every length $l \in \operatorname{SpecC}(\Gamma)$ there exists a cycle C_l such that $|C_l \cap M_m| = \min(m, \lfloor l/2 \rfloor)$. Then*

$$\dim \operatorname{NCV}(\Gamma) \geq \nu_{\text{odd}}(m) + \nu_{\text{even}}(m).$$

The hypothesis can be lessened since, if there is any cycle length $l \geq 2m$, it suffices to have one length $l \geq 2m$ for which there is a C_l with $|C_l \cap M_m| = m$.

Proof. The hypotheses imply that

$$d_l = \begin{cases} \lfloor l/2 \rfloor & \text{if } l \leq 2m, \\ m & \text{if } l \geq 2m. \end{cases}$$

We count the number of distinct values d_l for odd and even cycle lengths. For odd l we get $(l-1)/2$ if $l \in \operatorname{SpecC}(\Gamma)$ and $l < 2m$, and we get m if and only if there exists a cycle length $l \geq 2m$. The total is $\nu_{\text{odd}}(m)$. The computation of $\nu_{\text{even}}(m)$ is similar.

The values of $\mu(l)$ in Theorem 3.3 are the same as those of d_l unless there is a cycle length for which no l-cycle intersects M_m; but that is ruled out by our hypotheses. Theorem 3.4 follows. □

A connected graph is *bipancyclic* if it is bipartite with vertex classes of size p and q and has a cycle of every even length from 4 to $2\min(p,q)$. (This extends the usual definition, which assumes $p = q$.) This is the bipartite analog of pancyclicity, in which the graph has a cycle of every length from 3 to n, the order of the graph.

Corollary 3.5. *Assume Γ is pancyclic and has a permutable m-matching M_m, and for every l with $3 \leq l \leq n$ there is an l-cycle C_l with $|C_l \cap M_m| = \min(m, \lfloor l/2 \rfloor)$. Then*

$$\begin{aligned} \dim \operatorname{NCV}(\Gamma) &= n-2 && \textit{if } 2m \geq n-1, \\ n-2 \geq \dim \operatorname{NCV}(\Gamma) &\geq 2m-1 && \textit{if } 2m \leq n-2. \end{aligned}$$

Assume Γ is bipancyclic and has vertex class sizes p, q with $p \leq q$, and it has a permutable m-matching M_m such that for every k with $2 \leq k \leq p$ there is a $2k$-cycle C_{2k} with $|C_{2k} \cap M_m| = \min(m, k)$. Then

$$\begin{aligned} \dim \operatorname{NCV}(\Gamma) &= p-1 && \textit{if } m = p, \\ p-1 \geq \dim \operatorname{NCV}(\Gamma) &\geq m-1 && \textit{if } m \leq p-1. \end{aligned}$$

The hypotheses can be lessened in the same way as those of Theorem 3.4.

Proof. If Γ is pancyclic, ν_{odd} counts all the numbers $3, 5, \ldots, 2m-1$ plus 1 for $2m+1$ if $n > 2m$, and ν_{even} counts the numbers $4, 6, \ldots, 2m-2$ plus 1 for $2m$ since $n \geq 2m$. Thus

$$\nu_{\text{odd}} + \nu_{\text{even}} = \begin{cases} (m) + (m-1) = 2m-1 & \text{if } n > 2m, \\ (m-1) + (m-1) = 2m-2 & \text{if } n = 2m. \end{cases}$$

The conclusion follows easily.

If Γ is bipancyclic, then $\nu_{\text{even}} = m-1$ and the conclusion follows easily. □

The two most complete graphs are easy consequences of any of the preceding results, but especially of Corollary 3.5.

Corollary 3.6. *For a complete graph K_n with $n \geq 3$,*

$$\dim \mathrm{NCV}(K_n) = n - 2.$$

For a complete bipartite graph $K_{p,q}$ with $p, q \geq 2$,

$$\dim \mathrm{NCV}(K_{p,q}) = \min(p, q) - 1.$$

4 Examples

4.1 The complete graph

Our original example was K_n. The biggest permutable edge set is a perfect or near-perfect matching. This turns out to be "perfect" for our purposes. But first, let us see the negative cycle vectors of all signings of small complete graphs.

The vectors for K_3 are

$$(0),\ (1)$$

(from the balanced and unbalanced triangle). The vectors for K_4 are

$$(0,0),\ (2,2),\ (4,0)$$

(the all-positive graph, one negative edge, and two nonadjacent negative edges). Here are the vectors for K_5:

$$(0,0,0),\ (3,6,6),\ (4,8,8),\ (5,10,6),\ (6,8,4),\ (7,6,6),\ (10,0,12);$$

and for K_6:

$$\begin{array}{llll}
(0,0,0,0), & (4,12,24,24), & (6,18,36,36), & (8,20,32,24), \\
(10,18,36,36), & (8,24,40,32), & (10,22,36,28), & (12,24,24,32), \\
(10,26,36,28), & (8,24,48,32), & (14,18,36,36), & (12,24,32,32), \\
(12,20,40,24), & (10,30,36,20), & (16,12,48,24), & (20,0,72,0).
\end{array}$$

The number of switching isomorphism classes of complete graphs grows super-exponentially [4]. Since two signed graphs which yield different vectors must belong to different classes, one naturally wonders about the converse property, that the vector uniquely identifies a switching class. This is true up through K_7 but false for K_8: see Figure **??** below (found by Gary Greaves, whose assistance we greatly appreciate). Thus when $n = 8$ there are fewer vectors than classes; for $n > 8$ see Question 5.5.

Now we compute the function G_l of Section 3.3. Consider the signed K_n's whose negative edges are s nonadjacent edges, for $0 \leq s \leq \lfloor n/2 \rfloor$. It is straightforward to compute g_l. For a fixed $k \geq 1$ and set Y with $|Y| = k$, we need to form an l-cycle using Y and $l - k$ other edges. (Since Y is a matching, we know that $l \geq 2k$.) So we choose $l - 2k$ of the remaining $n - 2k$ vertices, and then create our cycle as follows: imagine contracting the edges in Y; the resultant vertices, together with the other $l - 2k$ vertices, will form an $l - k$-cycle in the contracted graph (which will eventually give an l-cycle in K_n). Cyclically order these $l - k$ "vertices"; this orders the vertices in our actual cycle while ensuring the

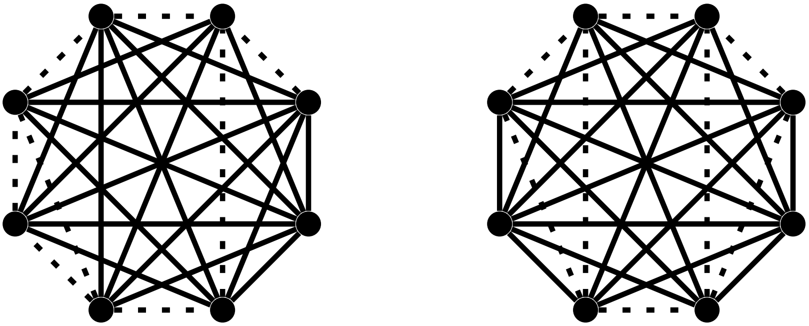


Figure 2: Two switching inequivalent signings of K_8 with the same negative cycle vector $(28, 108, 336, 848, 1440, 1248)$.

edges from Y remain. There are $\frac{(l-k-1)!}{2}$ ways to do this. Then, we expand the contracted edges to regain them; there are 2 ways to do this for each edge. So we have

$$g_l(Y) = \binom{n-2k}{l-2k}(l-k-1)! \cdot 2^{k-1},$$

whence

$$G_l(k) = \binom{n-2k}{l-2k}(l-k-1)! \cdot 2^{k-1}.$$

By Equation (3.4), $c_l^-(s)$ is a polynomial in s of degree $d_l = \lfloor l/2 \rfloor$ and the general formula is

$$c_l^-(s) = \sum_{k=1}^{n} \binom{s}{k}(-4)^{k-1}\binom{n-2k}{l-2k}(l-k-1)!,$$

For example, $c_3^-(s) = s(n-2)$ and $c_4^-(s) = s(n^2+5n+8) - 2s^2$. This formula for $c_l^-(s)$ demonstrates that the degrees d_l of the odd polynomials are all distinct, and the same for the even polynomials; consequently our main Theorem 3.3 itself implies that the matrix of negative cycle vectors $c^-(s)$ has full rank $n-2$.

Alternatively, in K_n with a maximum matching, $\Delta_{\text{odd}} = \{3, 5, \dots\}$ (odd numbers up to n) and $\Delta_{\text{even}} = \{4, 6, \dots\}$ (even numbers up to n). So, by Lemma 3.2, for K_n the ranks of U and R are $\lceil n/2 \rceil - 1$ and $\lfloor n/2 \rfloor - 1$, respectively, which sum to $n-2$.

4.2 Complete bipartite graphs

We now examine $K_{p,q}$, which always has $p \leq q$. We use a maximum matching M_p, i.e., we set $m = p$.

To get $c_{2l}^-(K_{p,q})$ we compute g_{2l}, where the subscript is now $2l$ because all cycles have even length. Call the two independent vertex sets $A = \{a_1, \dots, a_p\}$ and $B = \{b_1, \dots, b_q\}$. For a fixed k-edge set $Y = \{a_{i_1}b_{j_1}, \dots, a_{i_k}b_{j_k}\} \subseteq M_p$, where $k \leq l$, we need to form a $2l$-cycle using Y and $2l-2k$ other vertices. Fix one edge $y_1 \in Y$, say $y_1 = a_{i_1}b_{j_1}$. Choose $l-k$ of the remaining $p-k$ vertices from A, in order, in one of $(p-k)_{l-k}$ ways; $l-k$ of the remaining $q-k$ vertices from B, also in order, in one of $(q-k)_{l-k}$ ways; and shuffle the sequences together as $(a_{i_{k+1}}, b_{j_{k+1}}, \dots, a_{i_l}, b_{j_l})$. Insert Y into this $2(l-k)$-sequence by inserting y_1 before $a_{i_{k+1}}$, which we may do because each Y edge must be between an A vertex and a B vertex; treating the resulting sequence as cyclically ordered, which can be done in only one way since the A neighbor of y_1 appears after y_1; then ordering $Y \setminus \{y_1\}$ in one of $(k-1)!$ ways as $(y_2, \dots, y_k)$; and finally inserting $y_2, \dots, y_k$ anywhere into the cycle in one of

$$\binom{[2(l-k)+1]+[k-1]-1}{[2(l-k)+1]-1} = \binom{2l-k-1}{k-1}$$

ways. Note that when those edges are inserted into the cycle, there is only one way to orient each edge. The net result is that

$$G_{2l}(k) = g_{2l}(Y) = (p-k)_{l-k}(q-k)_{l-k} \cdot (k-1)! \binom{2l-k-1}{k-1}.$$

Then by Equation (3.4), for $2 \le l \le p$,

$$c^-_{2l}(s) = \sum_{k=1}^{p} (s)_k \frac{(-2)^{k-1}}{k} (p-k)_{l-k}(q-k)_{l-k} \binom{2l-k-1}{k-1}.$$

This explicit formula for the negative cycle vectors $c^-(s)$, with Theorem 3.3, implies that $\dim \mathrm{NCV}(K_{p,q}) = p = \min(p,q)$.

4.3 The Petersen graph

Next we consider the Petersen graph P, which has four cycle lengths, 5, 6, 8, and 9, so $\dim \mathrm{NCV}(P) \le 4$. It lacks a permutable 4-matching. In fact:

Theorem 4.1. *A 3-regular graph that is arc transitive cannot have a permutable 4-matching.*

Proof. By [6, Theorem 1.1] an arc-transitive graph with a permutable m-matching, where $m \ge 4$, must have degree at least m. □

The Petersen graph does have a permutable 3-matching, in fact, two kinds.

The first kind consists of alternate edges of a C_6. In the language of Theorem 3.3, we must compute $\mu(l) = |\max\{C_l \cap M_3\}|$ for each cycle length. We find with little difficulty that $\mu(5) = 2$, $\mu(6) = 3$, $\mu(8) = 2$, and $\mu(9) = 3$. Therefore $|\Delta_{\text{odd}}| = 2$ and $|\Delta_{\text{even}}| = 2$, whence, despite only having a 3-matching, we can deduce that $\dim \mathrm{NCV}(P) = 4$. We even know the negative cycle vectors corresponding to negative 0-, 1-, 2-, and 3-submatchings and the negated signatures; they are (in order of matching size)

$$\begin{array}{llll} (0,0,0,0), & (4,4,8,12), & (6,6,8,10), & (6,10,0,10) \\ (12,0,0,20), & (8,4,8,8), & (6,8,8,10), & (6,10,0,10). \end{array}$$

The bottom vector in each column corresponds to the negated signing.

The second kind of permutable 3-matching consists of three edges at distance 3. The first matching type also is three equally spaced edges in a C_9, but not every such subset of a C_9 is also a set of alternating edges of a C_6; the other such subsets are 3-matchings of the second kind. This second kind generates negative cycle vectors from negated submatchings and the corresponding negated sign functions whose dimension is only 3, not 4, since with this matching the negated signatures are switching isomorphic to unnegated signatures. This shows that not all permutable m-matchings in a graph are equally useful.

4.4 The Heawood graph

The Heawood graph H is bipartite and has five cycle lengths, 6, 8, 10, 12, and 14, so $\dim \mathrm{NCV}(H) \leq 5$. It has a permutable 3-matching, indeed three different kinds, for instance alternate edges of a 6-cycle. Using that 3-matching we find that $\mu(6) = 3$, $\mu(8) = 2$, $\mu(10) = 3$, $\mu(12) = 3$, and $\mu(14) = 3$. These are two different values, thus $\dim \mathrm{NCV}(H) \geq 2$. The results for the other two kinds of permutable 3-matching are the same except that $\mu(6) = 2$. In every case μ has two values.

Our matching method, in principle, cannot prove more because H has no permutable 4-matching (see Theorem 4.1). Nonetheless we suspect the dimension equals $|\operatorname{SpecC}(H)|$.

4.5 Other graphs with permutable perfect matchings, and the cube

Schaefer and Swartz found all graphs that have a permutable perfect matching. Besides K_n and $K_{p,p}$ they are the hexagon C_6, the octahedron graph O_3, and three general examples: the join $K_p \vee \overline{K}_p$ of a complete graph with its complement, the *matching join* $K_p \veebar K_p$ obtained from two copies of K_p by inserting a perfect matching between the two copies, and the matching join $K_p \veebar \overline{K}_p$, obtained by hanging a pendant edge from each vertex of K_p.

Our treatment of them leads us to one other family, the cyclic prisms $C_p \,\Box\, K_2$.

4.5.1 The simple four

Trivially,

$$\dim \mathrm{NCV}(C_6) = 1 = |\operatorname{SpecC}(C_6)|.$$

It is easy to verify by hand that O_3 satisfies the conditions of Corollary 3.5, so

$$\dim \mathrm{NCV}(O_3) = |\operatorname{SpecC}(O_3)| = 4.$$

As for $K_p \veebar \overline{K}_p$, since the pendant edges contribute nothing to cycles, $\operatorname{SpecC}(K_p \veebar \overline{K}_p) = \operatorname{SpecC}(K_p)$ and $\mathrm{NCV}(K_p \veebar \overline{K}_p) = \mathrm{NCV}(K_p)$; thence

$$\dim \mathrm{NCV}(K_p \veebar \overline{K}_p) = |\operatorname{SpecC}(K_p \veebar \overline{K}_p)| = p.$$

It is also easy to show that $K_p \vee \overline{K}_p$ satisfies the conditions of Corollary 3.5. Thus,

$$\dim \mathrm{NCV}(K_p \vee \overline{K}_p) = |\operatorname{SpecC}(K_p \vee \overline{K}_p)| = 2p.$$

4.5.2 The matching join of two complete graphs

This graph, $K_p \veebar K_p$, is pancyclic, but its permutable matchings are peculiar. One kind is any matching in a K_p. A maximum matching $M_{\lfloor p/2 \rfloor}$ in K_p has $\mu(l) = \min(p, \lfloor l/2 \rfloor)$, hence $\dim \mathrm{NCV}(K_p \veebar K_p) \geq p$ by reasoning similar to that for K_p. The matching $M_p^{\vee}$ that joins the copies of K_p prevents a permutable matching from having edges in both copies. The only other permutable matchings are subsets of $M_p^{\vee}$. This matching only generates $\lfloor p/2 \rfloor$ switching nonisomorphic signatures since negating a subset of $M_p^{\vee}$ switches to negating the complementary subset. By itself, therefore, choosing our grand matching M_m to be $M_p^{\vee}$ does not give a better lower bound than p. Nonetheless we feel the dimension is likely to be $n - 2 = 2p - 2$.

The smallest case, $K_3 \veebar K_3$, is the triangular prism. There are four cycle lengths; the cycle count vector is $(c_3, c_4, c_5, c_6) = (2, 3, 6, 3)$. The required dimension can be found directly. There are four unbalanced signatures; see Figure 3. The negative cycle vectors are linearly independent so $\dim \mathrm{NCV}(K_3 \veebar K_3) = |\operatorname{SpecC}(v)|$, in agreement with Conjecture 1.1.

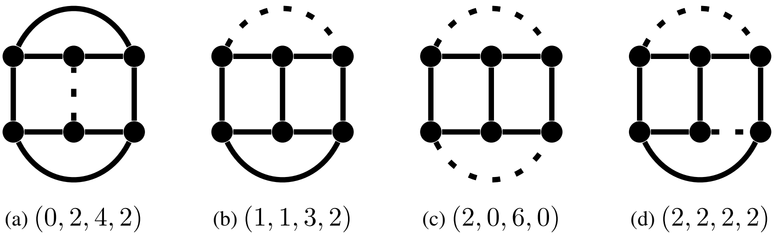

(a) $(0, 2, 4, 2)$ (b) $(1, 1, 3, 2)$ (c) $(2, 0, 6, 0)$ (d) $(2, 2, 2, 2)$

Figure 3: The four unbalanced switching classes of the prism $K_3 \veebar K_3$ and their negative cycle vectors.

As for permutable matchings in the triangular prism, $M_3^{\veebar}$ gives $\mu(3) = 0$, $\mu(4) = \mu(5) = \mu(6) = 2$, thus $\dim \mathrm{NCV}(K_3 \veebar K_3) \geq 3$, less than the true value. A strange permutable matching gives the right dimension. Choose M_2 to consist of one edge from each triangle, not both in a C_4. Then $\mu(3) = \mu(4) = 1$ and $\mu(5) = \mu(6) = 2$, so by Theorem 3.3, $\dim \mathrm{NCV}(K_3 \veebar K_3) = 4$, the exact value. This example and the Petersen graph demonstrate that useful permutable matchings need not be perfect matchings.

4.5.3 Prisms, with cube

The triangular prism lends support to our belief that $\dim \mathrm{NCV}(K_p \veebar K_p) = 2p - 2$. However, it is atypical since it is also a prism, the Cartesian product $C_p \,\Box\, K_2$ with $p = 3$.

Prisms with $p > 3$ do not have permutable perfect matchings but they make good examples, especially the next case, the cube $Q_3 = C_4 \,\Box\, K_2$. It is bipartite and has only three cycle lengths: 4, 6, and 8. Three unbalanced signatures whose negative cycle vectors are linearly independent are

σ_1, with one negative edge, e. It has $c^-(\sigma_1) = (2, 8, 4)$;

σ_2, with a second negative edge, parallel to e and sharing a quadrilateral with it. It has $c^-(\sigma_2) = (2, 12, 4)$;

σ_3, with a second negative edge, also parallel to e but not in a common quadrilateral. It has $c^-(\sigma_3) = (2, 4, 2)$.

Thus, $\dim \mathrm{NCV}(Q_3) = |\operatorname{SpecC}(Q_3)|$, again agreeing with Conjecture 1.1.

5 Questions

Here are what we consider the principal open questions concerning negative cycle numbers and vectors. The purpose is to find connections between the structure of Γ and the signed cycle structure of signatures of Γ. We list them in order of increasing refinement. Complete graphs seems to be the simplest example with interesting properties so we recommend them as the first object of study, except of course in Question 5.1.

5.1 Dimension

Resolve Conjecture 1.1. If it is false, can $\dim \mathrm{NCV}(\Gamma)$ be determined in terms of structural properties of Γ?

5.2 Cone

The zero vector is the most obvious negative cycle vector of every graph. That suggests looking at the convex cone generated by $\mathrm{NCV}(\Gamma)$. In particular, we wonder whether the facets or edges of that cone have combinatorial meaning.

5.3 Polytope

The convex hull $\operatorname{conv} \mathrm{NCV}(\Gamma)$ is a natural object of interest, and in particular its facets, which represent the complete set of inequalities satisfied by all negative cycle vectors. Almost nothing is known about these inequalities even for K_n. We looked at complete graphs of orders up to 6 but they were too small to suggest a conjecture.

If Σ is a signed K_n with frustration index $m = l(\Sigma) \leq n/2$, the negative cycle numbers for lengths $l < n/2$ (approximately) must satisfy bounds found by Popescu and Tomescu [5, Corollary 1]; the lower bounds occur when E^- is an m-edge star and the upper bounds when E^- is an m-edge matching. Since the bounds depend on the frustration index, they do not appear to constrain $\operatorname{conv} \mathrm{NCV}(\Gamma)$, but perhaps something relevant can be made of them.

5.4 Characterization

The negative cycle numbers of a signed K_n, Σ, must satisfy divisibility conditions found by Popescu and Tomescu [5, Section 4]. Aside from that and the work of Kittpassorn and Mészáros [3] on $c_3^-(\Sigma)$—that is, sizes of n-vertex two-graphs—it is not known which integral vectors in $\operatorname{conv} \mathrm{NCV}(K_n)$ can be negative cycle vectors. Surely, a characterization will be difficult if not impossible.

We know of no partial results for other graphs.

5.5 Collapsing pairs

Concerning Gary Greaves' counterexample mentioned in Section 4.1, we propose:

Conjecture 5.1. *For every $n \geq 8$ there are pairs of switching nonisomorphic signed K_n's that have the same negative cycle vector.*

In a related question, we ask whether the number I_n of switching isomorphism types of signed complete graphs [4] is asymptotic to the number $|NCV(K_n)|$ of negative cycle vectors of those graphs; that is, whether $|NCV(K_n)|/I_n \to 1$. If not, does it approach 0?

5.6 Conclusion

Evidently, there is much to discover before we can say the negative cycles in signed graphs are well understood.

References

[1] D. Cartwright and F. Harary, Structural balance: a generalization of Heider's theory, *Psychol. Rev.* **63** (1956), 277–293, doi:10.1037/h0046049.

[2] F. Harary, On the notion of balance of a signed graph, *Michigan Math. J.* **2** (1953–54), 143–146, doi:10.1307/mmj/1028989917.

[3] T. Kittipassorn and G. Mészáros, Frustrated triangles, *Discrete Math.* **338** (2015), 2363–2373, doi:10.1016/j.disc.2015.06.006.

[4] C. L. Mallows and N. J. A. Sloane, Two-graphs, switching classes and Euler graphs are equal in number, *SIAM J. Appl. Math.* **28** (1975), 876–880, doi:10.1137/0128070.

[5] D. R. Popescu and I. Tomescu, Negative cycles in complete signed graphs, *Discrete Appl. Math.* **68** (1996), 145–152, doi:10.1016/0166-218x(95)00010-o.

[6] A. Schaefer and E. Swartz, Graphs that contain multiply transitive matchings, submitted, arXiv:1706.08964 [math.CO].

[7] J. J. Seidel, A survey of two-graphs, in: *Colloquio Internazionale sulle Teorie Combinatorie, Tomo I*, Accademia Nazionale dei Lincei, Rome, pp. 481–511, 1976, proceedings of the Convegni Lincei, No. 17, held in Rome, September 3 – 15, 1973.

[8] I. Tomescu, Sur le nombre des cycles négatifs d'un graphe complet signé, *Math. Sci. Humaines* **53** (1976), 63–67, http://www.numdam.org/item/MSH_1976__53__63_0.

ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 16 (2019) 641
https://doi.org/10.26493/1855-3974.1842.e16
(Also available at http://amc-journal.eu)

Corrigendum to: On zero sum-partition of Abelian groups into three sets and group distance magic labeling

Sylwia Cichacz

*Faculty of Applied Mathematics, AGH University of Science and Technology,
Al. Mickiewicza 30, 30-059 Kraków, Poland*

Received 31 October 2018, accepted 25 January 2019, published online 7 June 2019

In the paper [1] by me, Theorem 4.4 is stated incorrectly and contradicts Theorem 4.5. Therefore, Theorem 4.4 should have been stated as follows:

Theorem 4.4. *Let $G = K_{n_1,n_2,n_3}$ be a complete tripartite graph such that $1 \leq n_1 \leq n_2 \leq n_3$ and $n = n_1 + n_2 + n_3$. The graph G is a group distance magic graph if and only if ($n_2 > 1$ and $n_1 + n_2 + n_3 \neq 2^p$ for some positive integer p) or ($n_1 \neq 2$ and $n_2 > 2$).*

In the previous version in the proof for $n_1 + n_2 + n_3 = 2^p$ the case $n_1 \neq 2$ and $n_2 > 2$ is not considered. The statement follows directly from Theorem 4.5.

Let $n_1 + n_2 + \cdots + n_t = n$ and $G = K_{n_1,n_2,\ldots,n_t}$. In the introduction is stated that G is Γ-distance magic if and only if Γ has the $\mathrm{CSP}(t)$-property. It is not true. It should be stated that G is Γ-distance magic if and only if for the partition $n = n_1 + n_2 + \cdots + n_t$ of n there is a partition of Γ into pairwise disjoint subsets $A_1, A_2, \ldots, A_t$, such that $|A_i| = n_i$ and for some $\nu \in \Gamma$, $\sum_{a \in A_i} a = \nu$ for $1 \leq i \leq t$.

References

[1] S. Cichacz, On zero sum-partition of Abelian groups into three sets and group distance magic labeling, *Ars Math. Contemp.* **13** (2017), 417–425, doi:10.26493/1855-3974.1054.fcd.

E-mail address: cichacz@agh.edu.pl (Sylwia Cichacz)

ARS MATHEMATICA
CONTEMPORANEA

Author Guidelines

Before submission

Papers should be written in English, prepared in LaTeX, and must be submitted as a PDF file.

The title page of the submissions must contain:

- *Title*. The title must be concise and informative.
- *Author names and affiliations*. For each author add his/her affiliation which should include the full postal address and the country name. If avilable, specify the e-mail address of each author. Clearly indicate who is the corresponding author of the paper.
- *Abstract*. A concise abstract is required. The abstract should state the problem studied and the principal results proven.
- *Keywords*. Please specify 2 to 6 keywords separated by commas.
- *Mathematics Subject Classification*. Include one or more Math. Subj. Class. codes – see http://www.ams.org/msc.

After acceptance

Articles which are accepted for publication must be prepared in LaTeX using class file amcjoucc.cls and the bst file amcjoucc.bst (if you use BibTeX). If you don't use BibTeX, please make sure that all your references are carefully formatted following the examples provided in the sample file.

All files can be found on-line at:

https://amc-journal.eu/index.php/amc/about/submissions/#authorGuidelines

Abstracts: Be concise. As much as possible, please use plain text in your abstract and avoid complicated formulas. Do not include citations in your abstract. All abstracts will be posted on the website in fairly basic HTML, and HTML can't handle complicated formulas. It can barely handle subscripts and greek letters.

Cross-referencing: All numbering of theorems, sections, figures etc. that are referenced later in the paper should be generated using standard LaTeX \label{...} and \ref{...} commands. See the sample file for examples.

Theorems and proofs: The class file has pre-defined environments for theorem-like statements; please use them rather than coding your own. Please use the standard \begin{proof} ... \end{proof} environment for your proofs.

Spacing and page formatting: Please do not modify the page formatting and do not use \medbreak, \bigbreak, \pagebreak etc. commands to force spacing. In general, please let LaTeX do all of the space formatting via the class file. The layout editors will modify the formatting and spacing as needed for publication.

Figures: Any illustrations included in the paper must be provided in PDF format, or via LaTeX packages which produce embedded graphics, such as TikZ, that compile with PdfLaTeX. (Note, however, that PSTricks is problematic.) Make sure that you use uniform lettering and sizing of the text. If you use other methods to generate your graphics, please provide .pdf versions of the images (or negotiate with the layout editor assigned to your article).

Subscription

Yearly subscription:	150 EUR

Any author or editor that subscribes to the printed edition will receive a complimentary copy of *Ars Mathematica Contemporanea*.

Subscription Order Form

Name: ....
E-mail: ....
Postal Address: ....
....
....
....

I would like to subscribe to receive ...... copies of each issue of
Ars Mathematica Contemporanea in the year 2019.

I want to renew the order for each subsequent year if not cancelled by e-mail:
□ Yes □ No

Signature: ....

Please send the order by mail, by fax or by e-mail.

By mail: Ars Mathematica Contemporanea
UP FAMNIT
Glagoljaška 8
SI-6000 Koper
Slovenia

By fax: +386 5 611 75 71

By e-mail: info@famnit.upr.si

Printed in Slovenia by *IME_TISKARNE*