Metodološki zvezki, Vol. 6, No. 2, 2009, 135-153

Clustering of Attribute and/or Relational Data

Anuška Ferligoj and Luka Kronegger[1]

Abstract

A large class of clustering problems can be formulated as an optimizational problem in which the best clustering is searched for among all feasible clustering according to a selected criterion function. This clustering approach can be applied to a variety of very interesting clustering problems, as it is possible to adapt it to a concrete clustering problem by an appropriate specification of the criterion function and/or by the definition of the set of feasible clusterings. Both, the blockmodeling problem (clustering of the relational data) and the clustering with relational constraint problem (clustering of the attribute and relational data) can be very successfully treated by this approach. It also opens many new developments in these areas. The paired clustering approaches are applied to the Slovenian scientific collaboration data.

1 Introduction

Grouping units into clusters so that those within a cluster are as similar to each other as possible, while units in different clusters as dissimilar as possible, is a very old problem. Although the clustering problem is intuitively simple and understandable, providing solution(s) remains a very exciting activity. The field of cluster analysis has its one society, the *International Federation of Classification Societies* which was formed in 1985 from several national classification societies. The society organizes every second year its conference and publishes two journals: the *Journal of Classification*, which was established in 1984 and the journal *Advances in Data Analysis and Classification* established in 2007.

Clustering of relational data is one of the clustering topics that was mostly developed in the field of social network analysis. There, for the clustering of relational data the term blockmodeling is used. On the other side, the clustering with relational constraint where a solution is searched according to the attribute and the relational data, was mostly developed in the field of cluster analysis. A unified approach is presented here.

2 Clustering problem

Cluster analysis (known also as classification and taxonomy) deals mainly with the following general problem: given a set of units, $\mathcal{U} = \{x_1, x_2, ..., x_n\}$, determine subsets,

[1] Faculty of Social Sciences, University of Ljubljana, Slovenia; Anuska.Ferligoj@fdv.uni-lj.si, Luka.Kronegger@fdv.uni-lj.si

called clusters, C, which are homogeneous and/or well separated according to the measured variables. The set of clusters forms a clustering. This problem can be formulated as an optimization problem:

Determine the clustering $\mathbf{C}^*$ for which

$$P(\mathbf{C}^*) = \min_{\mathbf{C} \in \Phi} P(\mathbf{C})$$

where $\mathbf{C}$ is a clustering of a given *set of units*, $\mathcal{U}$, Φ is the set of all feasible clusterings and $P : \Phi \rightarrow R$ is a *criterion function*.

There are several types of clusterings, e.g., partition, hierarchy, pyramid, fuzzy clustering, clustering with overlapping clusters. The most frequently used clusterings are partitions and hierarchies. A clustering $\mathbf{C} = \{C_1, C_2, ..., C_k\}$ is a *partition* of the set of units $\mathcal{U}$ if

$$\bigcup_i C_i = \mathcal{U}$$

$$i \neq j \Rightarrow C_i \cap C_j = \emptyset$$

A clustering $\mathbf{H} = \{C_1, C_2, ...C_k\}$ is a *hierarchy* if for each pair of clusters C_i and C_j from $\mathbf{H}$ it holds

$$C_i \cap C_j \in \{C_i, C_j, \emptyset\}$$

and it is a complete hierarchy if for each unit x it holds $\{x\} \in \mathbf{H}$, and $\mathcal{U} \in \mathbf{H}$.

Clustering criterion functions can be constructed *indirectly*, e.g., as a function of a suitable (dis)similarity measure between pairs of units (e.g., euclidean distance) or *directly*. In most cases, the criterion function is defined indirectly. For partitions into k clusters, the Ward criterion function (Ward, 1963)

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} \sum_{x \in C} d(x, t_C)$$

is usually used, where t_C is the center of the cluster C and is defined as

$$t_C = (\overline{u}_{1C}, \overline{u}_{2C}, ..., \overline{u}_{mC})$$

where $\overline{u}_{iC}$ is the average of the variable U_i, $i = 1, ...m$, for the units from the cluster C. d is the squared euclidean distance.

As the set of feasible clusterings is finite a solution of the clustering problem always exists. Since this set is usually very large it is not easy to find an optimal solution. In general, most of the clustering problems are NP-hard. For this reason, different efficient *heuristic* algorithms are used. There are many such algorithms and approaches. Among these, the agglomerative (hierarchical) and the relocation approach are most often used (see, e.g., Doreian et al., 2005).

2.1 Agglomerative approach

Agglomerative clustering approach usually assumes that all relevant information on the relationships between the n units from the set $\mathcal{U}$ is summarized by a symmetric pairwise dissimilarity matrix $D = [d_{ij}]$. The scheme of the agglomerative approach is:

Each unit is a cluster: $C_i = \{x_i\}$, $x_i \in \mathcal{U}$, $i = 1, 2, \ldots, n$;
repeat while there exist at least two clusters:
determine the nearest pair of clusters C_p and C_q:
$d(C_p, C_q) = \min_{u,v} d(C_u, C_v)$;
fuse the clusters C_p and C_q to form a new cluster $C_r = C_p \cup C_q$;
replace C_p and C_q by the cluster C_r;
determine the dissimilarities between the
cluster C_r and other clusters.

The result is a hierarchy that is usually presented by a dendrogram.

2.2 Relocation approach

This approach assumes that the user can specify the number of clusters in the partition.
The scheme of the relocation approach is:

Determine the initial clustering $\mathbf{C}$;
while
there exists $\mathbf{C}'$ such that $P(\mathbf{C}') \leq P(\mathbf{C})$,
where $\mathbf{C}'$ is obtained by moving a unit x_i from cluster
C_p to cluster C_q, or by interchanging units x_i and x_j
between two clusters in the clustering $\mathbf{C}$;
repeat:
substitute $\mathbf{C}'$ for $\mathbf{C}$.

While different criterion functions can be used in this approach, the Ward criterion function is usually used.

2.3 Benefits from the optimizational approach

The optimizational approach to clustering problem offers two possibilities to adapt to a concrete clustering problem: the definition of the criterion function P and the specification of the set of feasible clusterings Φ. The usual clustering problem seeks for a clustering according to the selected variables or attributes. Blockmodeling is searching for a clustering according to the relational data only. The solution can be obtained by an appropriately defined criterion function, described in Section 3. For clustering with relational constraint (attribute and relational data) an appropriately defined set of feasible clusterings is used (see Section 4).

3 Blockmodeling

3.1 Some definitions

Let $\mathcal{U}$ be a finite set of units and let the units be related by a binary relation $R \subseteq \mathcal{U} \times \mathcal{U}$ which determines a network $\mathbf{N} = (\mathcal{U}, R)$. R can be described by a corresponding binary

matrix $\mathbf{R} = [r_{ij}]_{n\times n}$ where

$$r_{ij} = \begin{cases} 1 & x_i R x_j \\ 0 & \text{otherwise} \end{cases}$$

In some applications r_{ij} can be a nonnegative real number expressing the strength of the relation R between units x_i and x_j.

One of the main procedural goals of social network analysis is to identify, in a given network, clusters of units that share structural characteristics defined in terms of the relation R. The units within a cluster have the same or similar connection patterns to the units of other clusters. A clustering $\mathbf{C}$ partitions also the relation R into *blocks* $R(C_i, C_j) = R \cap C_i \times C_j$. Each block is defined in terms of units belonging to clusters C_i and C_j and consists of all arcs from units in cluster C_i to units in cluster C_j. If $i = j$, the block $R(C_i, C_i)$ is called a *diagonal* block.

A *blockmodel* consists of structures obtained by shrinking all units from the same cluster of the clustering $\mathbf{C}$. For an exact definition of a blockmodel we must be precise about which blocks produce an arc in the *reduced graph* and which do not. The reduced graph can be presented also by a relational matrix, called an *image matrix*.

The partition is constructed by using structural information contained in R only, and units in the same cluster are equivalent to each other in terms of R alone. These units share a common structural position within the network.

Blockmodeling, as a set of empirical procedures, is based on the idea that units in a network can be grouped according to the extent to which they are equivalent, in terms of some *meaningful* definition of equivalence. In general different definitions of equivalence usually lead to distinct partitions.

3.2 Equivalences

Lorrain and White (1971) provided a definition of *structural equivalence*: Units are equivalent if they are connected to the rest of the network in *identical* ways. x and y are structurally equivalent if and only if:

s1.	$xRy \Leftrightarrow yRx$	s3.	$\forall z \in \mathcal{U} \setminus \{x, y\} : (xRz \Leftrightarrow yRz)$
s2.	$xRx \Leftrightarrow yRy$	s4.	$\forall z \in \mathcal{U} \setminus \{x, y\} : (zRx \Leftrightarrow zRy)$

From this definition it follows that only four possible ideal blocks can appear (Batagelj et al., 1992, Doreian et al., 2005)

Type 0.	$b_{ij} = 0$	Type 2.	$b_{ij} = 1 - \delta_{ij}$
Type 1.	$b_{ij} = \delta_{ij}$	Type 3.	$b_{ij} = 1$

where δ_{ij} is the Kronecker delta function and $i, j \in C$. The blocks of types 0 and 1 are called the *null* blocks and the blocks of types 2 and 3 the *complete* blocks. For the nondiagonal blocks $R(C_u, C_v), u \neq v$, only blocks of type 0 and type 3 are admissible.

Attempts to generalize the structural equivalence date back at least to Sailer (1978) and have taken various forms. Integral to all formulations is the idea that units are equivalent if they link in equivalent ways to other units that are also equivalent. Regular equivalence, as defined by White and Reitz (1983), is one such generalization.

The equivalence relation $\approx$ on $\mathcal{U}$ is a *regular equivalence* on network $\mathbf{N} = (\mathcal{U}, R)$ if and only if for all $x, y, z \in \mathcal{U}$, $x \approx y$ implies both

R1. $xRz \Rightarrow \exists w \in \mathcal{U} : (yRw \wedge w \approx z)$ R2. $zRx \Rightarrow \exists w \in \mathcal{U} : (wRy \wedge w \approx z)$

As was the case with structural equivalence, regular equivalence implies the existance of ideal blocks. The nature of these ideal blocks follows from the following theorem (Batagelj et al., 1992):

Theorem 1 *Let* $\mathbf{C} = \{C_i\}$ *be a partition corresponding to a regular equivalence* $\approx$ *on the network* $\mathbf{N} = (\mathcal{U}, R)$*. Then each block* $R(C_u, C_v)$ *is either null or it has the property that there is at least one 1 in each of its rows and in each of its columns. Conversely, if for a given clustering* $\mathbf{C}$*, each block has this property then the corresponding equivalence relation is a regular equivalence.*

Until now, a definition of equivalence was assumed for the *entire* network and the network was analyzed in terms of the permitted ideal blocks. Doreian, Batagelj and Ferligoj (2005) generalized the idea of a blockmodel to one where the blocks can conform to more types beyond the three mentioned above, and one where there is no single a priori definition of 'equivalence' for the entire network.

3.3 Blockmodeling as clustering problem

The problem of establishing a partition of units in a network, in terms of a considered equivalence, is a special case of the clustering problem – such that the criterion function reflects the considered equivalence. Such criterion functions can be constructed *indirectly* as a function of a compatible (dis)similarity measure between pairs of units. A dissimilarity d is *compatible* with the equivalence $\equiv$ if

$$x_i \equiv x_j \Leftrightarrow d(x_i, x_j) = 0$$

Not many dissimilarities are compatible with the equivalences mentioned above. The dissimilarity compatible with the structural equivalence is Corrected euclidean-like dissimilarity (Burt and Minor, 1983):

$$d(x_i, x_j) = \sqrt{(r_{ii} - r_{jj})^2 + (r_{ij} - r_{ji})^2 + \sum_{\substack{s=1 \\ s \neq i,j}}^{n} ((r_{is} - r_{js})^2 + (r_{si} - r_{sj})^2)}$$

After calculating appropriate dissimilarities one of the clustering approaches (e.g., hierarchical or relocation approach) can be used to obtain a clustering solution.

The other possible way of constructing the criterion function is the direct way that directly reflects the considered equivalence. Such a criterion function measures the fit of a clustering to an ideal one with perfect relations within each cluster and between clusters according to the selected type of equivalence. The criterion function $P(\mathbf{C})$ defined should be sensitive to the considered equivalence:

$$P(\mathbf{C}) = 0 \Leftrightarrow \mathbf{C} \text{ determines the equivalence.}$$

Given a clustering $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$, let $\mathcal{B}(C_u, C_v)$ denote the set of all ideal blocks corresponding to block $R(C_u, C_v)$. Then the global error of clustering $\mathbf{C}$ can be expressed as

$$P(\mathbf{C}) = \sum_{C_u, C_v \in \mathbf{C}} \min_{B \in \mathcal{B}(C_u, C_v)} d(R(C_u, C_v), B)$$

where the term $d(R(C_u, C_v), B)$ measures the difference (error or inconsistency) between the block $R(C_u, C_v)$ and the ideal block B.

E.g., for structural equivalence the term $d(R(C_u, C_v), B)$ can be expressed as

$$d(R(C_u, C_v), B) = \sum_{x \in C_u, y \in C_v} |r_{xy} - b_{xy}|$$

where r_{xy} is the observed tie and b_{xy} is the corresponding value in an ideal block. It is easy to verify that this criterion function is sensitive to structural equivalence.

A similar criterion function can be defined also for regular equivalence (See Doreian et al., 2005).

In the case of the direct clustering approach, where an appropriate criterion function that captures the selected equivalence is constructed, relocation approach can be used to solve the given blockmodeling problem (Batagelj et al., 1992).

3.4 Pre-specified blockmodeling

The *inductive* approaches for establishing blockmodels for a set of social relations defined over a set of units were discussed above. Some form of equivalence is specified and clusterings are sought that are consistent with a specified equivalence. Another view of blockmodeling is *deductive* in the sense of starting with a blockmodel that is specified in terms of substance prior to an analysis (e.g., cohesive model, core-periphery model, hierarchical model). In this case given a network, set of types of ideal blocks, and a family of reduced models, a clustering can be determined which minimizes the criterion function. For details see (Batagelj et al., 1998, Doreian et al. 2005).

3.5 Blockmodeling of multi-way network

It is also possible to formulate a generalized blockmodeling problem where the network is defined by several sets of units and ties between them. Therefore, several partitions – for each set of units a partition has to be determined. The generalized blockmodeling approach was adapted for 2-way networks (Doreian et al., 2004), and only for structural equivalence and the indirect approach for 3-way networks (Batagelj et al., 2007).

3.6 Blockmodeling of valued networks

Until now only binary networks were treated. Another interesting approach is the development of generalized blockmodeling of valued networks. Žiberna (2007) proposed several approaches to generalized blockmodeling of valued networks, where values of the ties are assumed to be measured on at least interval scale. The first approach is a straightforward generalization of the generalized blockmodeling of binary networks (Batagelj and

Ferligoj, 2000) to valued blockmodeling. The second approach is homogeneity blockmodeling. The basic idea of homogeneity blockmodeling is that the inconsistency of an empirical block with its ideal block can be measured by within block variability of appropriate values.

4 Clustering with relational constraints

The constrained clustering problem can be expressed as clustering problem where the constraints are considered in the definition of the set of the feasible clusterings. In the case of clustering with the relational constraint, the problem is to find clusterings as similar as possible according to attribute data and also considering the ties from a relation R. The clustering with constraints problem seeks to determine the clustering $\mathbf{C}^*$ for which the criterion function P has the minimal value among all clusterings from the set of feasible clusterings $\mathbf{C} \in \Phi$, where Φ *is determined by the relational constraints*. Generally, such a set of feasible clusterings can be defined as:

$$\Phi(R) = \{\mathbf{C} : \mathbf{C} \text{ is a partition of } \mathcal{U} \text{ and each cluster } C \in \mathbf{C} \text{ is a subgraph } (C\ ,\ R \cap C \times C) \text{ in the graph } (\mathcal{U}, R) \text{ with the required type of connectedness}\}$$

We can define different types of sets of feasible clusterings for the same relation R (Ferligoj and Batagelj, 1983). Some examples of clusterings with relational constraint $\Phi^i(R)$ are

type of clusterings	type of connectedness
$\Phi^1(R)$	weakly connected units
$\Phi^2(R)$	weakly connected units that contain at most one center
$\Phi^3(R)$	strongly connected units
$\Phi^4(R)$	clique
$\Phi^5(R)$	the existence of a trail containing all the units of the cluster

In the clustering type $\Phi^2(R)$ a center of a cluster C is the set of units $L \subseteq C$ iff the subgraph induced by L is strongly connected and $R(L) \cap (C \setminus L) = 0$ where $R(L) = \{y : \exists x \in L : xRy\}$.

In the case of a symmetric relation each cluster determines a connected subnetwork. If the relational matrix is permuted in such a way that first the units of the first cluster are given by rows and columns than the units of the second cluster and so on, and if we cut the relational matrix by clusters we obtain blocks of the relational matrix. In the blockmodeling terminology in the case of clustering with symmetric relational constraint the obtained diagonal blocks have to be at least regular. The nondiagonal blocks can be zero blocks or something else.

Standard clustering algorithms can be adapted for solving relational constrained clustering problems (e.g., the agglomerative hierarchical and the relocation approach) (Ferligoj and Batagelj, 1992, 1983). In the case of the agglomerative approach when searching for the nearest pair of clusters according to the attribute data two clusters can be fused only if the fused cluster satisfies the required type of the relational constraint (e.g., strong connectivity). In the last step of each iteration of the algorithm we have also to determine the

tie between the fused cluster and other clusters. These strategies are given in Ferligoj and Batagelj (1982, 1983).

Recently Batagelj, Ferligoj and Mrvar (2009) adapted the clustering with relational constraint approach for very large sets of units. To obtain an efficient algorithm for large networks they compute the dissimilarities between units according to the attribute data only for those ones that are connected according to the relational data. They determine the dissimilarities and ties between the fused clusters and the other ones considering only the dissimilarities of those pairs of units that are connected according to the relation R.

5 Software

All described clustering of attribute and relational data procedures are implemented in Pajek – program for analysis and visualization of large networks (Batagelj and Mrvar, 1998). It is freely available, for noncommercial use, at: http://pajek.imfm.si.

6 Application: Clustering of Slovenian sociologists

The described clustering approaches are applied to the collaboration network of Slovenian researchers and their publication performance. The dataset was obtained from the Current Research Information System (SICRIS) which includes the information of all active researchers registered at the Slovenian Research Agency and at the co-operative On-Line Bibliographic System & Services (COBISS) which officially maintains database of all publications available in Slovenian libraries.

In this study the units are researchers who were in September 2008 in SICRIS registered to work in the field of sociology in Slovenia. The collaboration between sociologists is operationalized by coauthorship of publications. A tie between two researchers is measured by coauthorship of an original article, a chapter (independent scientific component part) of a monograph, or a scientific monograph in the years from 1996 to 2007.

Publication performance is measured by the number of publications by type (articles in the journals with an impact factor, other original scientific articles, chapters in scientific monographs, and scientific monographs) and language (English, Slovenian, or other languages).

The network consists of 95 units and 224 ties. Practically the whole network is one component which means that all except 6 units are connected. These 6 units are in three dyads and were excluded from further analysis. All analyses presented here are for the 89 units in the single large component.

6.1 Publication performance of the Slovenian sociologists

The publication performance was measured by ten variables:

- number of articles in journals with an impact factor,
- number of articles in English language scientific journals,

- number of chapters in English language scientific monographs,
- number of scientific English language monographs,
- number of articles in Slovene scientific journals,
- number of chapters in Slovene scientific monographs,
- number scientific Slovene monographs,
- number of articles in scientific journals in other languages,
- number of chapters in scientific monographs in other languages,
- number of scientific monographs in other languages,

Dissimilarity between researchers was measured by the euclidean distance. The Ward dendrogram (see Figure 1) was obtained for clustering of 89 sociologists considering the standardized variables.

The dendrogram shows two clusters: at the top with the researchers with lower scientific performance and at the bottom with higher performance (see Table 1). Table 1 presents averages of the publication performance variables. From the dendrogram in Figure 1 and Table 1 we can see that the first cluster is quite homogenous but it splits into two subclusters:

- Subcluster 1: researchers with the lowest performance (most of the researchers in this subcluster are young researchers) and
- Subcluster 2: sociologists with still below average performance.

The second cluster is quite heterogenous with seven subclusters with typical scientific performance:

- Subcluster 3: they mostly publish Slovene monographs and publications in other languages,
- Subcluster 4: they typically publish English chapters and monographs,
- Subcluster 5: they mostly publish articles in Slovene journals,
- Subcluster 6 (singleton): (s)he mostly publishes chapters and scientific monographs in all languages and in below average articles in journals,
- Subcluster 7 (singleton): (s)he typically publishes articles in journals with an impact factor and monographs in English language,
- Subcluster 8: these two sociologists mostly publish English and Slovene chapters and monographs,
- Subcluster 9: they typically publish articles in English and Slovene journals.

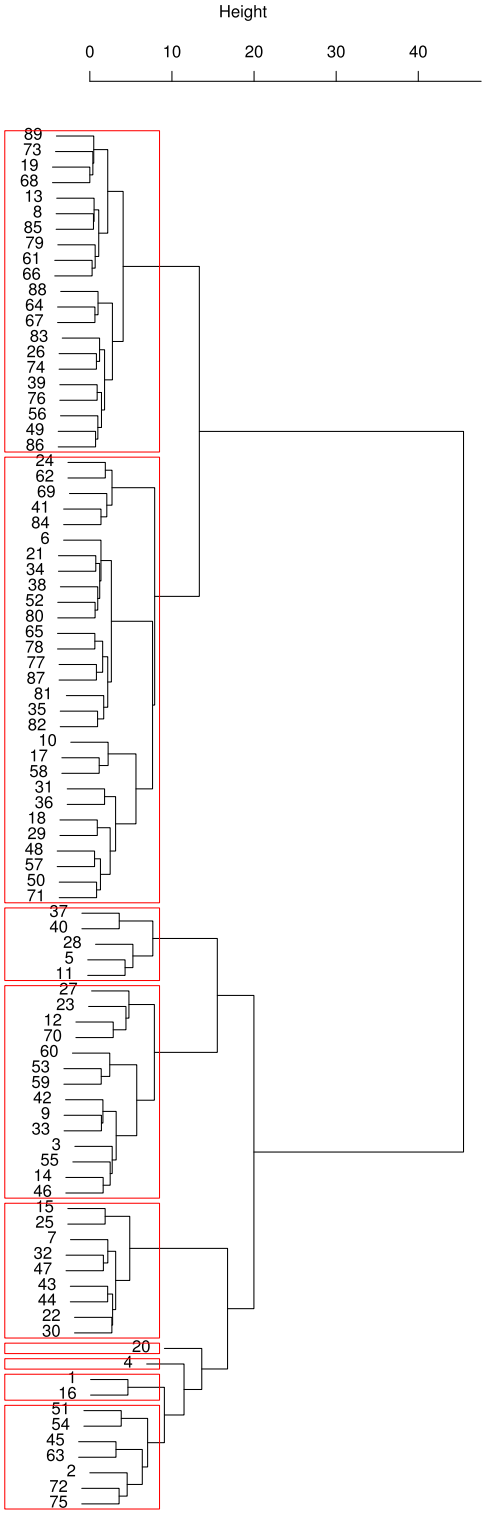


Figure 1: Hierarchical clustering of sociologists according to their publication performance.

Table 1: Average publication performance of obtained clusters.

cluster	N	journals with IF	in English language			in Slovenian language			in other languages		
			journals	chapters	books	journals	chapters	books	journals	chapters	books
1	21	0,43	0,48	0,71	0	1,71	0,62	0,67	0	0	0
2	29	0,41	1,21	2,14	0,1	3,52	3,1	1,86	0,14	0,07	0
3	5	0,2	1,4	3,2	0,2	3,8	3,2	4,4	2,8	2,8	0,2
4	14	0,71	0,93	4,43	1,43	3,5	2,86	3,36	0,43	0,71	0,07
5	9	0,56	0,89	1,11	0,44	10,89	5,78	2	0,11	0,11	0
6	1	0	3	9	2	3	11	12	0	2	3
7	1	14	9	4	3	3	5	0	2	1	0
8	2	1	2,5	8	1	7,5	12,5	9	0,5	0	0
9	7	2,86	7,14	4,14	0,14	9,43	4,86	2,71	0,57	0,14	0,29
together	89	0,82	1,57	2,51	0,4	4,39	3,21	2,29	0,36	0,35	0,08

6.2 Blockmodeling of the coauthorship network

Researcher are collaborating in many ways. Usually this collaboration results with a joint publication. Here we are interested to obtain clusters of Slovenian sociologists that publish together an article, a chapter in a scientific monograph, or a scientific monograph. We have to empasize that the measured coauthorship network measures coauthorship only between Slovenian researchers inside the field of sociology. (Nevertheless we know that they publish together also with researchers from the other fields in Slovenia and also with the researchers outside Slovenia.) To obtain such a clustering we used indirect and direct blockmodeling.

First the indirect approach for structural equivalence for coauthorship network was performed. The Ward dendrogram where the corrected euclidean-like dissimilarity was used is presented in Figure 2. According to the structure of the obtained dendrogram, there is a suggestion of clusterings into two or eight clusters.

In the case of the coauthorship network we can assume a core-periphery structure: one or several clusters of strongly connected scientists that strongly collaborate among themselves. The last cluster consists of scientists that do not collaborate among themselves and also do not with the scientists of the other clusters. They publish by themselves or with the researchers from the other scientific disciplines or with researchers outside of Slovenia. We performed the core-periphery pre-specified blockmodels (structural equivalence) into two to ten clusters. The highest drop of the criterion function was obtained for the blockmodels into three, five, and eight clusters. As eight clusters have been shown also by the dendrogram this clustering is further analyzed.

If we permute rows and columns in the relational matrix in such a way that first the units of the first cluster are given, than the units of the second cluster and so on, and if we cut the relational matrix by clusters we obtain Table 2. (Here a black square is drawn if the two researchers have at least one joint publication or a white square if they have not publish any joint publication.) In Table 2 the obtained blockmodel into eight clusters is presented (seven core clusters and one peripherical one). The seven core clusters are rather small ones and the peripherical large one (50 sociologists out of 89). Only cluster 4 is connected also to the first and second clusters, all other core clusters show cohesivness inside the cluster and nonconnectivity outside the cluster. The first two diagonal blocks are complete without any inconsistency compared to the ideal complete blocks. All white squares in complete blocks and black squares in zero blocks are inconsistencies. The number of all inconsistencies is the value of the criterion function. For the blockmodel

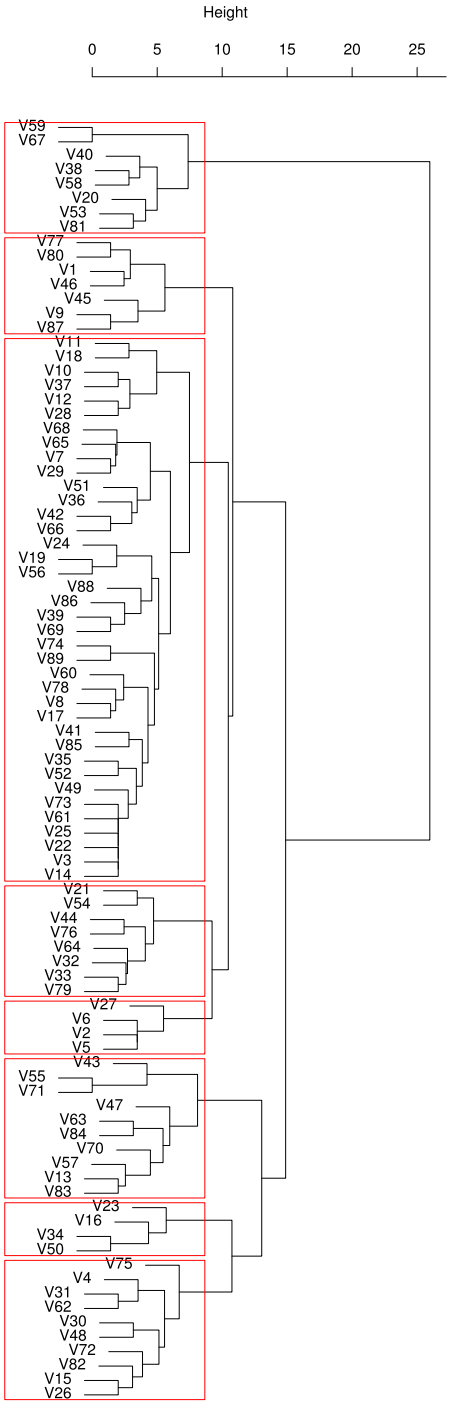
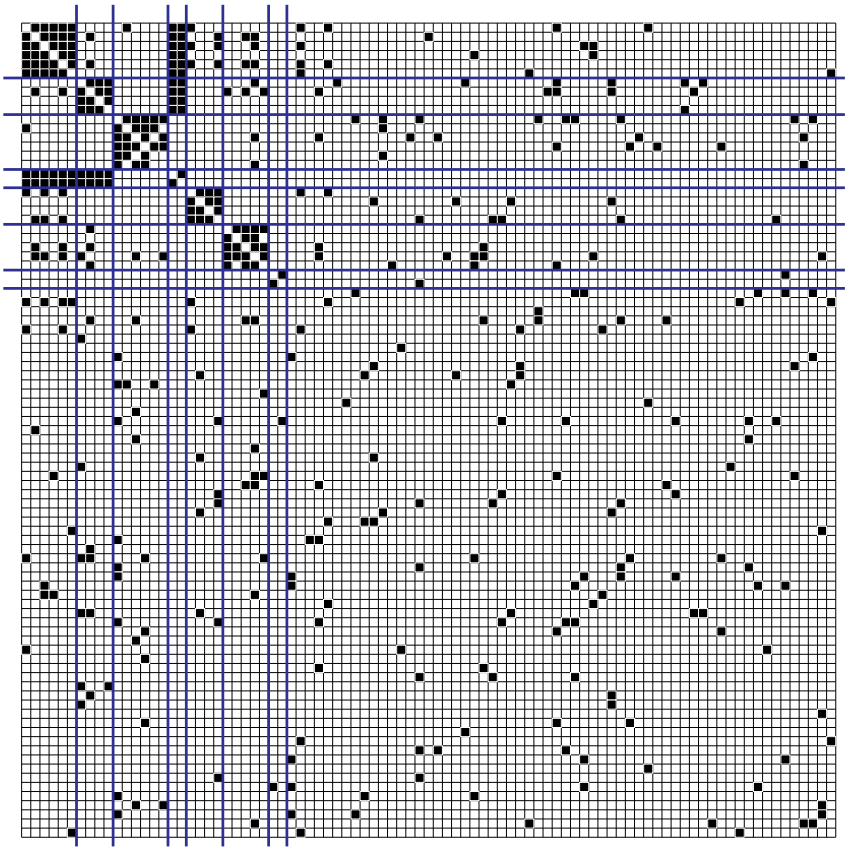


Figure 2: Hierarchical clustering of the sociologists according to their coauthorship network.

Table 2: Core-periphery structure of the coauthorship network.



presented in Table 2 the value of the criterion function is 310.

6.3 Coauthorship and publication performance of the Slovenian sociologists

Another view to the publication performance and coauthorship network of Slovenian sociologists is to jointly consider both type of data. We can search for clusters of researchers that jointly publish and are as similar as possible according to their publication performance. This can be done by clustering with relational constraint where clustering is done according to ten publication performance variables and relation is measured by coauthorship. We considered standardized variables, euclidean distance among researchers, and maximum agglomerative method. The obtained dendrogram is presented in Figure 3. From the dendrogram 14 typical clusters can be seen.

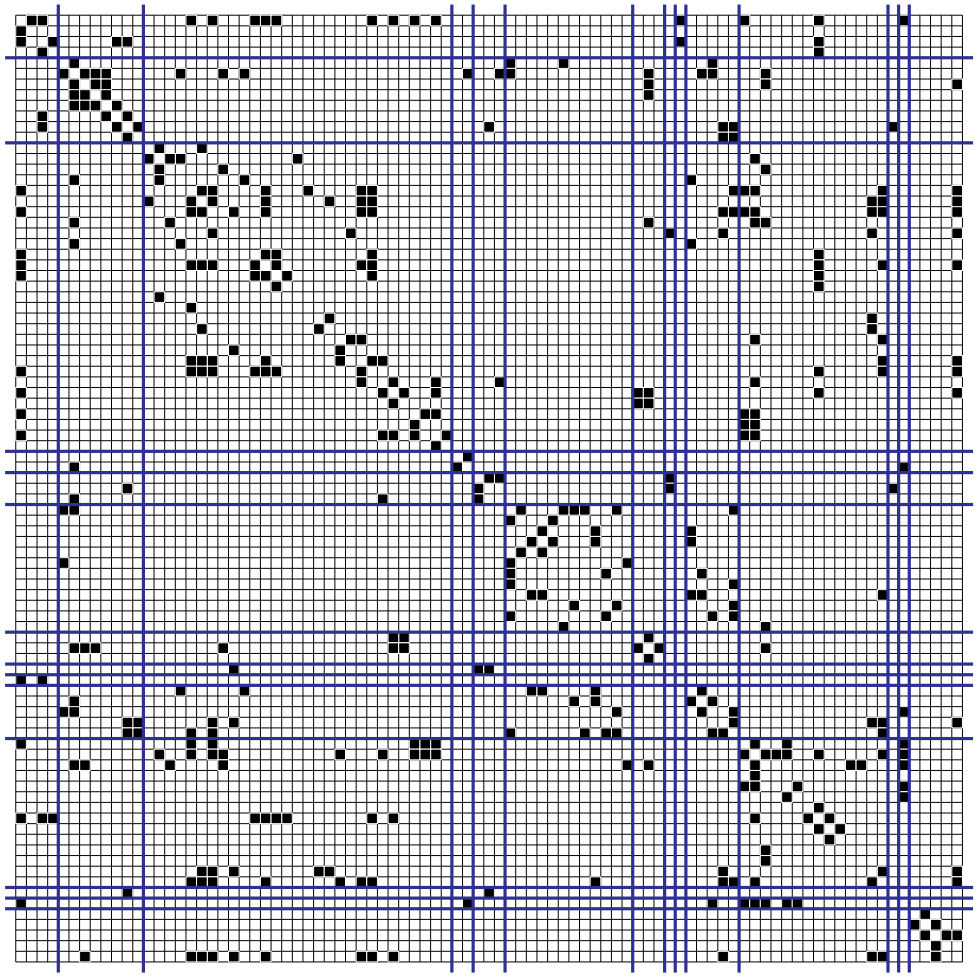
As we mentioned in Section 4, the clustering with relational constraint provides us clusters of units that are connected at least to another one inside the cluster - the diagonal blocks are at least regular. In Section 6.2 we have seen that there is a core-periphery structure in the coauthorship network. Therefore, the requirement that the diagonal blocks are at least regular is a quite stringent one. The clustering result is that it has 4 singletons. The obtained result can be nicely seen in Table 4 from the relational matrix where reasearchers

are permuted according to the obtained clustering. We can notice more inconsistencies in nondiagonal blocks (the diagonal ones are regular without inconsistencies). This is not a surprise as clustering with the relational constraint approach is not a blockmodeling approach. It searches for connected clusters according to the coauthorship relation in such a way that the researchers inside the clusters are as similar as possible according to the publication performance variables. The approach is imposing a cohesive structure.

In Table 4 the averages of publication performance variables for each obtained cluster are given. The comparison of Table 4 with Table 1 shows that the averages of the ten publication performance variables are in general lower in Table 4. This is not surprising as the sociologists inside a cluster obtained by the clustering with relational constraint have to be as similar as possible according to the publication style and each of them has to be a coauthor at least of another member of the cluster. But a similar interpretation can be done as in the case of the clustering of the researchers according to the publication performance variables only:

- Cluster 1: they have low publication performance, above average monographs,
- Cluster 2: they have below average performance, above average only articles in Slovene journals,
- Cluster 3: researchers with low publication performance,
- Cluster 4: they have above average only English publications,
- Cluster 5: they have low publication performance, above average only English and Slovene monographs,
- Cluster 6: below average performance,
- Cluster 7: they typically publish English monographs and articles in Slovene journals,
- Cluster 8 (singleton): (s)he has above average performance, mostly articles in all languages,
- Cluster 9 (singleton): (s)he has very low publication performance,
- Cluster 10: they have good publication performance, especially articles and chapters in all languages,
- Cluster 11: they publish mostly articles in journals in all languages,
- Cluster 12 (singleton): (s)he publishes in other languages,
- Cluster 13 (singleton): (s)he typically publishes articles in journals with an impact factor and monographs in English language,
- Cluster 14: they mostly publish chapters in English monographs and Slovene monographs.

Table 3: Clustering with coauthorship constraint structure.



Only the cluster 13 is equal to one of the clusters obtained according to the publication performance only (subcluster 7). Some clusters from Table 1 split into two or several clusters because the researchers are not coauthors of some publications. Therefore, some clusters have similar publication style but are not connected, e.g., clusters 1 and 5, or clusters 3 and 6.

6.4 Discussion of the obtained clusterings

Each of the clustering approaches reveals a different features of the collaboration and publication performance of the Slovenian sociologists. The clustering of the sociologists according to the publication performance variables only shows that they publish their research results in very specific ways. E.g., some of them publish mostly (only) in the Slovenian language, some of them just chapters in the scientific monographs, some of them typically in English journals. The results clearly show that there is no a typical common culture of the publishing performance in the field of sociology in Slovenia.

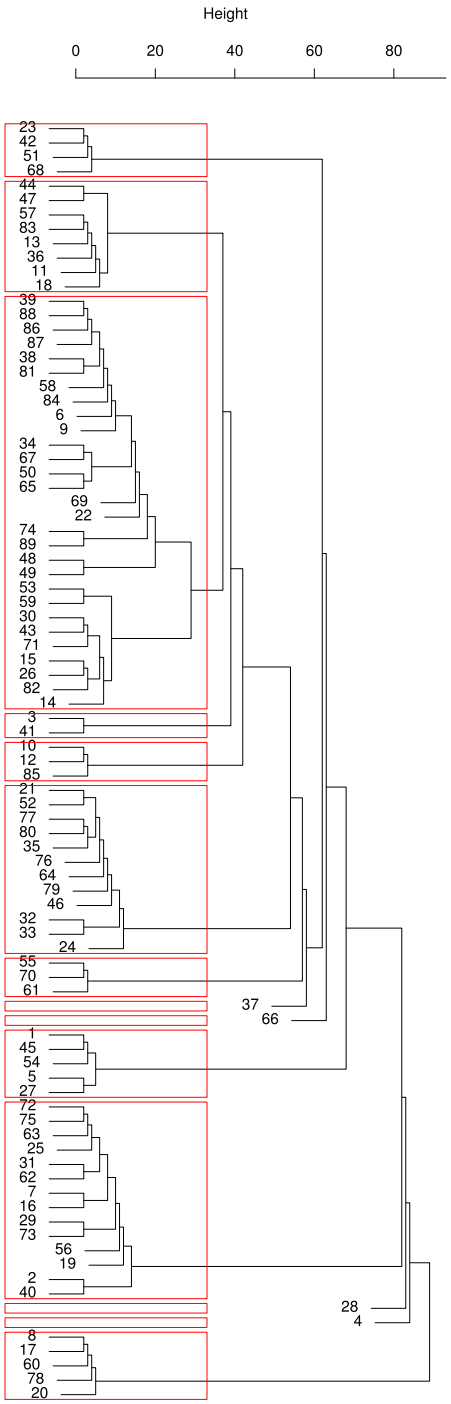


Figure 3: Hierarchical clustering of the sociologists according to their publication performance and coauthorship network.

Table 4: Average publication performance of obtained clustering with relational constraint.

		journals	in English language			in Slovenian language			in other languages		
cluster	N	with IF	journals	chapters	books	journals	chapters	books	journals	chapters	books
1	4	0,25	0,75	1,5	0,75	5	1,5	4,75	0	0,25	0,5
2	8	0,25	0,88	0,5	0	5,75	2,63	2,38	0,38	0,38	0
3	29	0,52	0,9	2,35	0,24	3,86	3,1	1,62	0,17	0,07	0
4	2	1	3	3	0,5	1,5	1	0	0	0,5	0
5	3	0	1,33	1	1	2,67	1,67	4,67	0,33	0,33	0
6	12	0,42	0,92	2,67	0,33	3,25	3,17	1,33	0	0	0
7	3	0,33	0,33	2,33	1,67	5	2	1,67	0,67	0,33	0
8	1	0	4	4	1	6	3	6	4	1	0
9	1	0	0	0	0	3	1	0	0	0	0
10	5	1	4,6	5,8	0,6	4,8	6,6	3,6	0	1,6	0,4
11	14	1,93	2,79	2,5	0,21	6,64	3,43	2,43	0,79	0,29	0
12	1	0	1	1	0	4	4	2	4	5	0
13	1	14	9	4	3	3	5	0	2	1	0
14	5	0,2	1,2	4,8	0,6	3	4,8	4,8	0	0,6	0,6
together	89	0,82	1,57	2,51	0,4	4,39	3,21	2,29	0,36	0,35	0,08

On the other side the clustering of the coauthorship network by the indirect approach and by direct blockmodeling approach clearly shows a core-periphery coauthorship structure. There are several small core groups of the Slovenian sociologists that publish among themselves and much less or not at all with the members of the other groups. The core groups overlap with the organizational structure: sociologists of each of these core groups are members of a research center. The periphery group composed by the sociologists that are mostly not coauthors with the other Slovenian sociologists is very large (more than 56 %). Most of them publish only as a single author. Probably some of them publish with the other Slovenian nonsociologists or with the researchers outside of Slovenia. More detailed analysis of the periphery group should be further studied. The coauthorship network is quite a sparse one which tells us, that the preferred publication culture by Slovenian sociologists is to publish as a single author.

The most challenging result is the clustering of the sociologists according to their publication performance considering also the coauthorship relation. Here we are looking for groups of sociologists that have as similar publication style as possible and are at the same time publishing together. For example, to search for well established sociologists with the best publication performance and that publish together, or a group of young sociologists that publish together and have not a strong publication record. Usually is the case that professors are publishing together with their students. These two groups have usually very different publication performance. Therefore, nevertheless they publish together they would not appear in the same cluster. The publication structure of the obtained clusters by the clustering with the relational constraint is similar to the one obtained by the clustering according to the publication performance only. The difference is that because of the coauthorship constraint some nonconnected clusters split to several connected ones, some others are completely rearranged. It is not surprising that four sociologists do not fit to any of the obtained clusters (having a similar performance style and at the same time publishing with the members of the cluster) and form a single unit cluster (singleton). As the coauthorship network is a sparse one, the obtained result is not a surprising one.

Of course, which clustering approach to use depends strongly on the research problem that we study.

7 Conclusion

The optimizational approach to the clustering problem can be applied to a variety of very interesting clustering problems, as it allows possible adaptations of a concrete clustering problem by an appropriate specification of the criterion function and by the definition of the set of feasible clusterings. Both the blockmodeling problem and the clustering with relational constraint problem are such cases. Possible applications of these quite different clustering approaches were presented by the analyses of the publication performance and coauthorship network of the Slovenian sociologists at the last ten years.

There are several possible further developments in blockmodeling, e.g., efficient direct approach for 3-way blockmodeling, blockmodeling for large networks, dynamic blockmodels.

References

[1] Batagelj, V., Doreian, P., and Ferligoj, A. (1992a): An optimizational approach to regular equivalence, *Social Networks*, 14, 121-135.

[2] Batagelj, V. and Ferligoj, A. (2000): Clustering relational data. In Gaul, W., Opitz, O., and Schader, M. (Eds.): *Data Analysis: Scientific Modeling and Practical Applications*, , Berlin: Springer, 3-15.

[3] Batagelj, V., Ferligoj, A., and Doreian, P. (1992b): Direct and indirect methods for structural equivalence, *Social Networks*, 14, 63-90.

[4] Batagelj, V., Ferligoj, A., and Doreian, P. (1998): Fitting to pre-specified blockmodels. In Hayashi, N., Jajima, K., Bock, H.H., and Baba, Y. (Eds.): *Classification and Related Methods*, Berlin: Springer, 199-206.

[5] Batagelj, V., Ferligoj, A., and Doreian, P. (2007): Indirect blockmodeling of 3-way networks. In Brito, P., Bertrand, P., Cucumel, G., and de Carvalho, F. (Eds.): *Selected Contributions in Data Analysis and Classification*. Berlin: Springer, 151-159.

[6] Batagelj, V., Ferligoj, A., and Mrvar A. (2009): Hierarchical clustering in large networks, submitted.

[7] Batagelj, V. and Mrvar, A. (1998): Pajek - a program for large network analysis, *Connections*, 21, 47-57.

[8] Doreian, P., Batagelj, V., and Ferligoj, A. (2004): Generalized blockmodeling of two-mode network data, *Social Networks*, 26, 29-53.

[9] Doreian, P., Batagelj, V., and Ferligoj, A. (2005): *Generalized Blockmodeling*, Cambridge: Cambrige University Press.

[10] Ferligoj, A. and Batagelj, V. (1982): Clustering with relational constraint, *Psychometrika*, 47, 413-426.

[11] Ferligoj, A. and Batagelj, V. (1983): Some types of clustering with relational constraints, *Psychometrika*, 48, 541-552.

[12] Burt, R.S. and Minor, M.J. (1983): *Applied Network Analysis*, Bevery Hills, CA: Sage.

[13] Lorrain, F. and White, H. (1971): Structural equivalence of individuals in social networks, *Journal of Mathematical Sociology*, 1, 49-80.

[14] Sailer, L. (1978): Structural equivalence: Meaning and definition, computation and aplication, *Social Networks*, 1, 73-90.

[15] Žiberna, A. (2007): Generalized blockmodeling of valued networks, *Social Networks*, 29, 105-126.

[16] Ward, J. (1963): Hierarchical grouping to optimize an objective function, *Journal of American Statistical Association*, 58, 236-244.

[17] White, D. and Reitz, K. (1983): Graph and semigroup homomorphisms on networks of relations, *Social Networks*, 5, 193-234.