Tie Persistence in Academic Social Networks

Djamila Mohdeb, Adelhak Boubetra and Mourad Charikhi
Department of Computer Science, University of Bordj Bou Arreridj, Bordj Bou Arreridj, Algeria
E-mail: djamila.mhb@gmail.com, {boubetraabd, mcharikhi}@yahoo.fr

Keywords: link persistence, link prediction, link strength, social network evolution, academic social network

Received: May 12, 2016

This paper attempts to shed light on the importance of some social academic-related factors in determining the strength of links in academic social networks. Our purpose is to assess the extent to which the frequency of the tie, the academic closeness between its actors, and the scientific contributions of the actors in the tie can affect the scientific collaboration relationship between them. We propose a model that relies on this three link strength indicators in order to predict the tie persistence in academic social networks. We experimented the model on a social network extracted from the DBLP computer science bibliographic network. We compared the output of the model with that of the link prediction baseline methods. The results show better performance of the proposed model.

Povzetek: Prispevek analizira vpliv socialnih povezav v omrežjih na akademski uspeh s pomočjo DBLP.

1 Introduction

The investigation of academic networks is increasingly an important topic in the area of social networks mining. Comprehending these complex networks is important to understand the trends of knowledge production through the world. A typical academic network contains a set of multi-typed entities (scientists, papers, journals, institutions…etc.) linked by a set of multi-typed associations (Figure 1-a). The collaboration network is the mainly used social projection of the scientific academic network. It consists of a set of nodes representing scientists, and a set of links representing collaboration relations between nodes. Frequently, researchers use co-authorship relations to construct collaboration networks as they denote formal cooperation between scientific actors. A collaboration network is composed by connecting every set of authors who share the same publications (Figure 1-b). This type of networks exhibits in general the same characteristics as social networks. They are of "small world" type, where the clustering coefficient, which describes the transitivity in the network, is high. As a result, the average distance between any two scientists in the network is short, and it does not usually exceed five or six degrees [33]. They are also scale free following a power law in several node properties and their structures are affected by the preferential attachment phenomenon [18, 37].

Studying the evolution and the dynamics of collaboration networks remains a continuing concern in social networks mining since the advances of science depend crucially on this type of interactions between scientists [23]. Studies in this field focus on the analysis of the observed changes in the network structure caused by both the links and the nodes. Among link analysis tasks, the link prediction problem [28] is one of most studied subjects in link mining literature. A link prediction model attempts to predict the appearance, the persistence, and the disappearance of a social network links relying on some of its given snapshots in the past. However, in this paper, we do not address the entire link prediction problem but only the sub-question that concerns the driving factors behind the persistence of the ties in academic social networks. The tie persistence seems to be an occasionally studied problem despite its importance. This importance is related mainly to the existence of a minority of nodes and links that persists always in spite of the rapid dynamicity of the network overtime. Identifying the driving factors behind the structure persistence is as important as identifying the driving factors behind the structure evolution. Thus, this work attempts to resolve the link persistence problem using a link strength based technique that can measure the collaborative importance of the existent collaboration relationships in the network. This technique relies on three strength indicators that have been proposed in the social psychology literature [8, 34]: the frequency of interactions between the actors, their contributions in the relation, and the social closeness between them. Furthermore, the possible validity of the important relation is verified according to two relevant academic-related attributes that mostly must be taken into account in the context of scientific collaborations: the scientific productivity of the relation and the professional rank (status) of the scientists involved within. Our proposed tie persistence prediction model combines these link strength indicators to assess the strength of the scientific collaboration relationships between researchers in order to identify the persistent ties in a dynamic and time-varying academic social milieu.

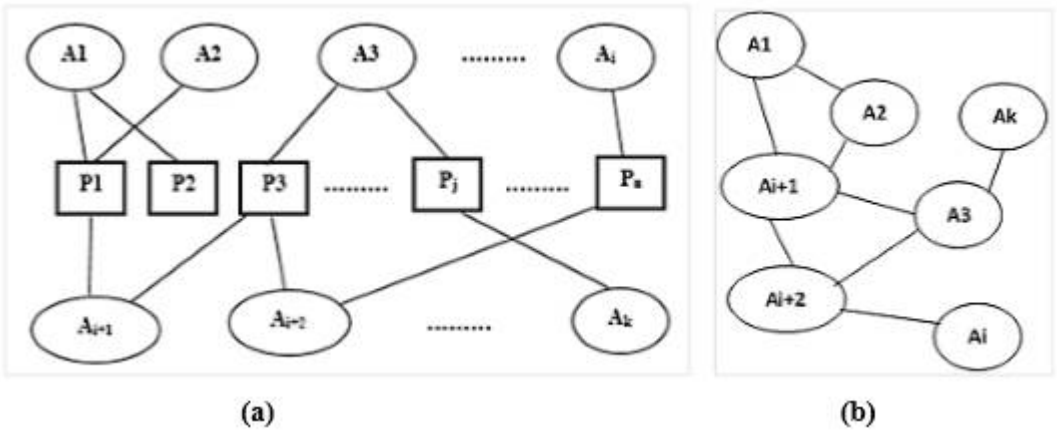

Figure 1: An example of an academic social network (b) extracted from a bibliographic network (a) (Authors (A) who share the same papers (P) in the bibliographic bipartite graph (a) are connected with co-authorship links in the academic social network (b)).

The remaining part of the paper is organized as follows: we begin with a brief overview of the previous research in the area. Then we explain our methodology for predicting the persistence of collaboration relations. We continue by presenting the performed experiments to validate the proposed model. Next, we report the findings of the research and discuss their implications. Then, we investigate the influence of the parameters of the model on its performance. Finally, we conclude with a brief summary of the findings and some suggestions for further research.

2 Related work

Social networks evolution problem addresses the question on how a social network evolves over time. Consequently, several sub-questions rise from this problematic. The most important are about the laws that govern the evolution and the factors that influence it. In this context, the study of the evolution of academic social networks has been a popular research topic in these recent few years. In the literature, the evolution of academic social networks may be analyzed on two levels: the macro level (the entire network) and the micro level (the simplest components of the network) [3]. Our work focuses on understanding the micro-level changes at the actor level. Specifically, it aims to predict the persistence of a tie between two nodes. This issue is a sub-question of the well-known link prediction problem, which addresses predicting the new links that join a social network in a given future time. Naturally, the link persistence issue is not independent from link prediction. This is for the reason that an actor's future links (which a link prediction model tries to predict) may incorporate also the old links that continue to be present in the future. The earliest studies on the link prediction are that proposed by Adamic and Adar [1] and by Liben-Nowell and Kleinberg [28] for social networks. They proposed unsupervised models basing on computing similarity scores between the network nodes using graph-based similarity measures that rely mainly on the topological structure of the network or on the node attributes. Later, the work of Hasan et al. [20] has argued for the effectiveness of the supervised models rather than the unsupervised ones. In addition to these two classical approaches, researchers have developed several models following various paradigms that include for example similarity-based models, feature-based models, probabilistic models, relational models, graphical models, linear algebraic models; and random walks based models [4, 16]. An in-depth survey of these approaches may be found in [21]. The link prediction models can obviously predict the link persistence between two nodes by restricting the model application to only direct (1-hop) neighbors. However, since the main concern of the link prediction problem is to predict the new relations and not the repeated ones, so predicting persistent links using a link prediction model may run the risk of providing modest results. Therefore, it will be interesting to develop independent models in which the only goal is to identify factors that drive the persistence of the tie between two nodes.

In this regards, social psychology literature provided important evidence about the various factors that influence the link persistence and decay. These factors are mostly: structural embeddedness (common acquaintances) [11, 14, 32], homophily [11, 32], social support [38], frequent contact (interaction) [38], social closeness [38], distance [32], status, and experience [11]. Moreover, there is some evidence on the "liability of newness", which means that newly formed ties tend to decay more quickly than old-timer ties [11]. On the other hand, there are few models that tried to treat the link persistence prediction problem. For instance, Hidalgo et al. [22] used a rule-based technique to predict the tie persistence in mobile phone social networks relying on their observations about the correlation between network topological variables (degree, clustering, reciprocity and topological overlap) with the tie persistence. Akoglu and Dalvi [2] proposed a logistic regression-based model for tie persistence prediction in large phone and SMS networks, which takes into account the fact that node and link attributes like neighborhood overlap, reciprocity, clustering coefficient, and node degree affect the link persistence between the actors.

Using decision tree and logistic regression based models, Raeder et al. [35] demonstrated that persistent ties in a cell-phone network are those characterized by high-levels of interaction frequency coupled with relatively constant re-activations of the tie overtime. On the contrary, ties that are candidates to decay, are characterized by relatively low levels of interaction and non-reciprocity. Apart from mobile phone social networks, Kirvan-Swaine et al. [26] studied the tie decay in the online social network of Twitter. Their findings revealed that reciprocity, embeddedness, power, and status influence significantly the tie breaking between follower-followee links in the online social network.

Differently from these studies, this paper proposes an unsupervised model for the tie persistence prediction in academic social networks. The proposed approach combines academic-related tie and node attributes to estimate the strength of relations between scientists. The model then reckons on its expectation about the possible scientists' collaboration preferences to validate or invalidate the collaborative importance of relations and their probable continuity in the future.

Algorithm 1. Tie persistence prediction model

Input s : source author

Output $Imp(t)$ collaborative importance score of the target author t

For *each author* s *in the network* **do**
  Extract bylines of all publications of s
  Calculate the collaborative importance of each byline
  For *each co-author* t *belongs to the publications bylines of* s **do**
    Calculate the collaborative productivity of the relation between s and t
    If *"Productive collaboration"* **then**
      Assign t the collaborative importance of the publication byline he belongs to
    Else
      If *the professional rank (academic age) of* s *equals to* "$Senior$" **then**
        collaborative importance of t $Imp(t) = 0$ // s has no need to t
      End if
    End if
  End for
End For

3 Tie persistence prediction model

The strength of social links may be estimated using various strength indicators that can reflect the different dimensions of a given relationship. In the case of academic links, the strength may be better captured using indicators that are related to the knowledge production since the knowledge production represents the ultimate goal of academic collaboration relations [23].

To predict the persistence or the dissolution of an existent collaboration relationship between two scientists, our proposed model follows two steps (see Algorithm 1). First, it measures the collaborative importance of the existing relations of a given scientist using three strength indicators: the frequency of the relation, the contribution of the concerned actor within, and the social-academic closeness between its actors. Second, the model decides to retain or to terminate the existing relation depending on its expectation about the behavior of the concerned author toward this relation given its collaborative importance to him. Formally saying for each author a, we collect the publication bylines from his papers. The publication byline is the list of n authors who have co-written a given paper p. For each set of authors in each paper, the model measures the collaborative importance of the relation according to the author a. Given the collaborative importance of the relation, the model then verifies its collaborative productivity and checks the professional rank of the author in order to decide finally whether it is better to keep or to terminate the existing relation.

Therefore, if a given relation passes the two steps successfully, every co-author who belongs to its related publication byline will take the collaborative importance value of the publication byline; otherwise, the model will suppose the inutility of a future collaboration relation between the concerned co-authors. Below, we give a detailed explanation of the model.

3.1 Computing collaborative importance

$$Imp(pbl, a) = (freq(pbl)/nbrPub(a)) * (contrib(a, pbl)/freq(pbl)) * (cl(pbl)/freq(pbl)) \quad (1)$$

Where:

- a is the author, pbl is the publication byline to which an author a belongs.
- $freq(pbl)$ is the number of times the author a has published papers having the byline pbl.
- $nbrPub(a)$ is the total number of publications of the author a.
- $contrib(a, pbl)$ is the contribution of the author a in the paper in comparison with his co-authors in the publication byline pbl.
- $cl(pbl)$ is the social-academic closeness factor of the publication byline pbl.

3.1.1 Frequency

The frequency [8, 34] is an intuitive indicator of the link strength. It represents the number of times a set of authors have participated to the publication of the same papers. A high value of frequency of a publication byline indicates some trust between its members.

3.1.2 Social-academic closeness factor

The closeness [8, 34] encompasses a wide variety of meanings characterizing the social proximity between actors in social networks. To estimate this proximity, relationship scholars have conceptualized multiple measures such as RCI (Relationship Closeness Inventory) [7], IOS (Inclusion-of-Other-in-Self Scale) [3], and URCS (Unidimensional Relationship Closeness Scale) [13]. These measures are not deterministic models but scoring systems relying on questionnaires attempting to capture the various dimensions of the relationship.

In [3], Aron et .al (the developers of IOS measure) postulated that in close relationship "people are motivated to include another in the self in order to include that other's resources". These resources may be anything that can "facilitate the achievement of goals". Obviously, in academic social networks, the knowledge is that valuable resource a scientist hope others will share with him/her. Co-authored publications characterize scientific relations, but the type of publications may reveal the social-academic closeness between the actors of these relations. The concept of closeness in our model is oriented to estimate mainly the familiarity between the collaborators. Therefore, we suppose that a book type publication is more important than a journal paper type, and a journal paper type is more important than a conference paper type. This is based on the relevance of the "book" type as the most valuable publication and on previous observations [17, 18] that have shown that authors sharing journal papers are professionally and socially closer than authors sharing common conference papers. This is for the reason that journal papers have a much higher impact than conference papers as they receive more citations [17]. In addition, a relevant work requires more time to be produced and the relative length of the time spent in the publication production may multiply the chance of familiarity between the paper's co-authors.

Formally, the social-academic closeness factor for a given publication having a byline pbl is expected to respect the constraint:

$$cl(pbl, type\_pub = \text{“}book\text{”}) > cl(pbl, type\_pub = \text{“}journal\ paper\text{”}) > cl(pbl, type\_pub = \text{“}conference\ paper\text{”})$$

- **Estimating the social-academic closeness from the type of publication**

We use in our model a scoring system bit similar to psychological measures described above in order to assess the social-academic closeness between publication co-authors. First, we construct an ordered list arranging publications types according to their relevance $L = \{1: book, 2: journal\ paper, 3: conference\ paper\}$

Then, we penalize a given type of publication by discounting from its initial default value V a portion equals to θ (θ is a model parameter) multiplied by the order of the publication type in the arrangement list of publication types L.

$$cl = V - (k-1) * \theta \qquad (k >= 1,\ (k-1) * \theta \leq V) \tag{2}$$

- V is the default value of publication. It is estimated to be $V = 1$.
- θ a regular portion the value of publication V loses by the degradation from a publication type to another.
- k is the order of the type of publication in the arrangement list.

3.1.3 Author contribution in the relation

The investment in the relation is another relevant strength indicator proposed in [8, 34]. Contribution is a domain-specific concept that can take different meanings according to the context it is used in. In academic social networks, the contribution of a scientist in a collaboration relation can be reflected in the credit that he deserves in the related publication in comparison with his co-authors. The proposed model estimates this credit using the Network-Based Allocation (NBA) model of co-authorship credit proposed by Kim and Diesner [24]. The NBA model uses the order of the author in the publication byline in addition to the length of the author list involved in to calculate his final credit. Noting that in many research fields, the order reveals reliable information about the contribution of the author in the publication with the exception of some disciplines such as Mathematics, Economics or High Energy Physics, which follow in their publications alphabetical order of authors [12, 24].

The NBA model is flexible in partitioning the credit between the co-authors of a given paper. It is based on the idea that each author belonging to a publication byline of length N and having an initial co-authorship credit equals to v, distributes a portion of his credit (equals to v_t) in equal amounts to his preceding authors on the byline. We can calculate final credits for each coauthor as follows:

$$\begin{aligned} v &= V/N && (1 \leq V, 2 \leq N) \\ v_t &= d * v && (0 \leq d \leq 1) \\ &\begin{cases} v_r^N = v + v_t \sum_{n=1}^{N-r} 1/N - n & (r = 1, 2 \leq N) \\ v_r^N = (v - v_t) + v_t \sum_{n=1}^{N-r} 1/N - n & (1 < r < N, 2 \leq N) \\ v_r^N = v - v_t & (r = N,\ 2 \leq N) \end{cases} \end{aligned} \tag{3}$$

Where:

- v is the initial co-authorship credit given to each author.
- V is the value of the paper (assumed equals to 1), N is the number of the authors on a paper.
- v_t is the transferable credit, calculated by assigning a distribution factor $d \in [0,1]$ to the initial co-authorship credit v. The distribution factor d is the ratio of initial credit that should be distributed by each coauthor.
- r is the order of authors

Co-writers deserve equal co-authorship credits in a given publication if $d = 0$. If $d = 1$ this means that, the first author have the higher possible value of contribution in the publication and the role of non-first authors is negligible.

3.2 Predicting scientist's collaboration preferences

A collaboration relationship between two scientific actors becomes subject to some academic reckonings to be continued or terminated even if it seems strong. These reckonings are mainly related to the academic attributes of the author and the effect of the output of this relation on his scientific career. Our approach assumes that the persistence of an important collaboration relation between an author and his coauthor depends on two relevant academic-related factors: the professional rank of the author (status) and the collaborative productivity of the relationship. Relying on early observations [15, 30], our model assumes that a newcomer or a junior researcher needs to conserve his important relations despite its unproductivity for the reason that his rank as a beginner obligates him to develop his coauthors network by exploiting these important relations for his benefit. By contrast, an experienced researcher has always the possibility to terminate any unproductive relation that cannot offer him a scientific advantage.

3.2.1 Author professional rank

The professional rank or the status of an author is related to his scientific and professional career. It naturally influences the collaboration preferences [5, 9] since the collaboration choices of an experienced scientist differs widely from the collaboration choices of a novice scientist. The reason behind this is the relatively large or small scientific network that a senior or a beginner scientist have respectively.

3.2.2 Collaborative productivity

The productivity rate is an essential factor to validate the importance of a collaboration relation. Nevertheless, considering this factor differs according to the academic professional experience of the scientist [5, 9]. The proposed model considers a co-authorship relation as a "productive collaboration" if the number of collaborations between the author and the coauthor equals to at least the median of the duration of their relationship. Noting that the duration is a useful tie strength indicator to estimate the strength of links between actors in social networks [8, 34].

$$R\ (a,c)\ \ is\ a\ "Productive\ collaboration" \Rightarrow nbrCollab\ (a,c) \geq (d+1)/2 \quad (d = t_l - t_f) \tag{4}$$

Where:

- $nbrCollab\ (a,c)$ is the number of collaborations between the author a and his coauthor c.
- d is the duration of the relationship between a and c.
- t_l is the time (in year) of the last collaboration between the author and the coauthor.
- t_f is the time (in year) of the first collaboration between the author and the coauthor.

4 Experiments

4.1 Dataset

To demonstrate the performance of our approach, data are extracted from the well-known DBLP Computer Science Bibliography database, a huge digital library from the University of Trier, which covers publications in various computer science fields.

We selected randomly from the « DBLP » database a set of 2250 authors from different research areas in computer science who appeared between year 1993 and 2008. After that, we equally divided this subset into three author sets basing on the academic age of the authors in the DBLP bibliographic network. We measured the academic age of a scientist as the number of years since his first publication. The academic age obtained from the DBLP do not exactly reflect the professional rank of the author but can offer a hint about his experience (except for cases where an author's publications are not indexed by DBLP).

Therefore, we had the following sets:

- The Newcomers set: authors with an academic age less than six years.
- The Juniors set: authors with an academic age between six and ten years.
- The Seniors set: authors with an academic age greater than ten years.

Table 1 summarizes dataset statistics.

Table 1: Dataset statistics.

	N	**E**	**S**	**Avg. C**
Seniors	18 834	62 550	750	26.54
Juniors	17 699	57 655	750	22.68
Newcomers	17 076	56 417	750	21.78
	53 609	176 622	2250	23.66

N, E: number of nodes and edges in the full network, S: number of source authors, Avg. C: average number of co-authors per source

We draw the reader's attention to the fact that the quality of our data can be affected by the performance of the method used by DBLP to resolve the author name disambiguation problem [27]. Relying on the recent study of Kim and Diesner [25], the value ranges of the findings may vary if we use a different method for name disambiguation. Fortunately, as the latter study confirmed, this may not have a distortive effect on the general trend of the network evolution on which our findings depend.

4.2 Experimental setup

First, for learning the link persistence prediction model, we formed a general network combining all the co-authorship networks from 2003 to 2014 that correspond to the scientists belonging to the three author subsets. Let an author pair be (s, t), we call s the source author, and t the target author. The source authors are those who belong to the three author sets mentioned above (newcomers, juniors, and seniors). The target authors are the direct neighbors (i.e. 1-hop neighbors) of the source authors. Then, we chose the sub-network data between 2003 and 2008 as training set, and the sub-network data between 2009 and 2014 as testing set.
Second, the general parameters that we used for framing the link persistence model are the following:

- The professional rank of an author was calculated according to his academic age.
- As in the default setting of NBA co-authorship credit model [24], we assumed $d = 0.5$ as the ratio of the initial credit that should be distributed by each author belonging to the publication byline of a given paper (Eq. 3). The advantage of this setting is maximizing the contribution of first authors as well as avoiding neglecting the contribution of non-first authors.
- For the social-academic closeness factor in Eq. 2, we assumed $\theta = 0.25$ the ratio of the initial value of the publication an author loses according to the type of the publication. As such, the social-academic closeness factor is $1{,}3/4{,}1/2$ for the publication types: book, journal paper, and conference paper, respectively. We recall that our estimation of the social-academic closeness factor is based on a simple intuitive scoring system because we cannot exactly measure this value due to its psychological complex nature.

4.3 Evaluation framework

In order to show the effectiveness of our tie persistence prediction method for social academic networks, we compared its performance with the baseline methods used for the link prediction problem since they can also measure link persistence. The baseline methods considered are Common Neighbor (CN), Jaccard's Coefficient (JC), Adamic/Adar (AA), Preferential Attachment (PA), and Page Rank (PR). Formal descriptions of these methods are illustrated in Table 2.

Common Neighbor [19] is a simple metric that counts the number of shared neighbors (i.e. the number of paths of length 2) between two nodes. The Jaccard's coefficient [36] divides the common neighbors of a pair of nodes by the size of the union of their neighbors. The Adamic/Adar measure [1] weighs the rarer common features more heavily. These three metrics are related to the positive impact of the common acquaintances (structural embeddedness) on the tie formation and persistence between social actors. From the perspective of the tie persistence problem, they provide information about the social and scientific circles where a scientist moves. It is then reasonable to assume that, if two related scientists deal with the same scientific entourage, it is likely that their relation persists. The Preferential attachment [6] of two nodes is the product of their degrees. In our context, it is used to assume that a scientist tend to keep relations with highly connected scientists who have a better status [9, 12]. The PageRank algorithm [10] ranks the node proportionally to the probability that it will be attained through a random walk on the network.

Table 2: Link prediction baseline metrics.

Metrics	Description
Common Neighbor (CN)	$CN\ (x,y) = \lvert\Gamma(x) \cap \Gamma(y)\rvert$
Jaccard's coefficient (JC)	$JC\ (x,y) = \frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x) \cup \Gamma(y)\rvert}$
Adamic/Adar (AA)	$AA\ (x,y) = \sum_{z \in \Gamma(x)\cap\Gamma(y)} 1 / \log\lvert\Gamma(z)\rvert$
Preferential attachment (PA)	$PA\ (x,y) = \lvert\Gamma(x) * \Gamma(y)\rvert$
PageRank (named as Rooted PageRank in [28])	Similarity score between x and y is measured as the stationary distribution of y under the following random walk: • With probability β, return to x. • With probability $1-\beta$, move to a random neighbor.

x and y denote two given nodes in the social network. $\Gamma(x), \Gamma(y)$ represent the set of neighbors of x and y respectively.

For evaluating the methods used in this study, we employed a threshold curve metric: AUCPR (Area under the Precision Recall Curve) and two fixed threshold metrics: Precision and Recall. Precision is the probability that a randomly selected positive prediction by the classifier is correct. Recall is the probability that a randomly selected positive instance is detected by the classifier. A Precision-Recall (PR) curve plots precision vs. recall. AUCPR is thought to give a more reliable informative view of an algorithm's performance in comparison with the other common performance evaluation measures especially for the link prediction

Table 3: PRECISION performance results.

	LPP	**AA**	**CN**	**JC**	**PA**	**PR**
Newcomers	0.4597	0.4429	**0.4615**	0.4613	0.4535	0.3198
Juniors	**0.3621**	0.3494	0.3542	0.3542	0.3504	0.2770
Seniors	**0.3846**	0.3601	0.3731	0.3731	0.3644	0.2790
Avg. Precision	**0.4021**	0.3841	0.3963	0.3962	0.3894	0.2919

LPP: Link Persistence Prediction method, AA: Adamic-Adar, CN: Common Neighbor, JC: Jaccard's Coefficient, PA: Preferential Attachment, PR: Page Rank

Table 4: RECALL performance results.

	LPP	**AA**	**CN**	**JC**	**PA**	**PR**
Newcomers	**0.6161**	0.5571	0.5003	0.5002	0.5306	0.3918
Juniors	**0.5704**	0.5080	0.4561	0.4561	0.5060	0.3583
Seniors	**0.5424**	0.4970	0.4394	0.4394	0.4973	0.3423
Avg. Recall	**0.5763**	0.5207	0.4653	0.4653	0.5113	0.3642

Table 5: AUCPR performance results.

	LPP	**AA**	**CN**	**JC**	**PA**	**PR**
Newcomers	**0.6264**	0.5748	0.5557	0.5551	0.5558	0.4071
Juniors	**0.4823**	0.4549	0.4281	0.4281	0.4283	0.3533
Seniors	**0.4902**	0.4461	0.4254	0.4254	0.4253	0.3546
Avg. AUCPR	**0.5330**	0.4920	0.4697	0.4695	0.4698	0.3717

task [29]. This is mainly related to its fairness and efficiency in overcoming the class imbalance, which is not much present in the tie persistence prediction problem but very frequent in the link prediction problem. A high area under the curve characterizes both high recall and high precision.

$$Precision = TP/(TP + FP)$$
$$Recall = TP/(TP + FN)$$

5 Results and discussion

Performance results measured in Precision, Recall, and AUCPR for all baseline methods and link persistence prediction method (LPP) are presented in Tables 3, 4, and 5. The values in bold face indicate the best overall prediction performance for the corresponding dataset.

It is evident that Precision, Recall, and AUCPR agree about the best method and show in general the same performance trend. Interestingly, we note well performance of link persistence prediction method (LPP): about 57% of persisting ties are correctly classified by the model (recall) and about 40% of the ties that the model predicts to persist do in fact persist.

The proposed link persistence prediction method reveals the best performance in all the three datasets (except for its precision value in the Newcomers set) and provides significant enhancement over the link prediction baseline methods. AUCPR show notable performance of the proposed model in comparison with link prediction baselines with approximately 8% as relative improvement. The gain is remarkable from the perspective of Recall in which it reaches 10.5% but somewhat minimal from the perspective of Precision with only 1% as relative improvement. Apart from our proposed method, we note insufficient performance of the path-based method PageRank, which is explicable due to the fact that the target authors are direct neighbors. It is clear that a low distance such as 1-hop distance may have a negative impact on the effectiveness of a random walk based predictor. In contrast, we note good performance of the neighbor-based link prediction methods Common Neighbor, Adamic/Adar, Jaccard's coefficient, and Preferential attachment. As well, the results show that the performance of the neighbor-based methods is comparable in terms of Precision but the Adamic/Adar beats the three other metrics in terms of Recall and AUCPR. While the existing works in link prediction reported the performance of Adamic/Adar in co-authorship networks [28], the results of this predictor and the other neighbor-based metrics are consistent also with previous studies that signed the positive effect of the structural embeddedness and the actors' status on the tie persistence in social networks [2, 11, 14, 26, 32, 35].

Turning to the academic context of our model, these findings may be justified from different angles. An important thing to notice is that academic social networks are extremely dynamic networks. Rare are the nodes or the links that continue to accompany a scientist to a long period since the first collaboration even if they are academically strong. Moreover, the decay of a strong collaboration relation during a given time period sometimes can be confused by the tie inactivity. So, a possible explanation of our findings may be related to the social independence of the author as researcher. A scientist in an environment where there is no need to sentimental support as in real social networks seeks always a scientific support, which can be obtained from

different scientific entities or scientists who share with him the same ideas or the same research interests. Indeed, as he raises, his knowledge and expertise increase, his research interests develop, and his collaboration network evolves and gets larger. This is the reason behind the fact that the Recall in all the tested methods is higher than the Precision. All the methods can better expect the relevance of a collaboration relation but show less performance in expecting its probable continuity in the future.

This is well apparent when observing the impact of the author's professional experience on the persistence of his direct collaboration relations. In general, the results show that the higher the author's professional rank the lower the persistence prediction accuracy. This is reasonable since it is more difficult to expect the collaboration strategies of an experienced author for the reason that his choices are mainly independent, irregular, and do not follow clear trends. On the contrary, a scientist in his earlier career deals only with a few number of collaborators for a limited number of years. Mostly, his collaborators consist in his advisor and some colleagues who work with the same advisor. As the expertise of the scientist increases, his scientific network expands including a growing number of novice, junior, and senior researchers providing a large collaboration network with a high number of weak ties and a small number of strong ties, logically in a future step, many links will decay and few links will persist.

Certainly, the model needs to be refined and improved, but we can say that the present findings illustrate that the frequency, the contribution of the authors, and the type of the publication (that expresses the social-academic closeness between the co-authors) play a significant role in determining the persistence of a scientific collaboration relation. Furthermore, they encourage the consideration of the probable collaboration preferences that a scientist may pursue during his scientific career regarding some academic-related attributes such as academic experience and collaborative productivity in order to provide an additional gain in the prediction performance.

Table 6: LPP Performance results with different values of parameters d, θ, and different settings of publication types order.

d	Avg. PREC	Avg. REC	Avg. AUCPR
d = 0	**0.4220**	0.5465	**0.5447**
d=0.25	0.4043	0.5727	0.5385
d = 0.5	0.4021	**0.5763**	0.5330
d=0.75	0.3770	0.5730	0.5055
d = 1	0.3790	0.4755	0.4813
Order of Pub Types			
B-C-J	0.4008	0.5814	0.5374
B-J-C	**0.4021**	0.5763	0.5330
C-B-J	0.4012	**0.5825**	**0.5384**
C-J-B	0.3990	0.5809	0.5347
Theta (θ)			
T = 0.1	0.3947	0.5761	0.5273
T=0.25	**0.4021**	**0.5763**	**0.5330**
T = 0.3	0.3889	0.5714	0.5191
T = 0.4	0.3784	0.5638	0.5043

6 Influence of the model parameters

Next, we investigate the impact of co-authorship credit and social-academic closeness factor on the performance of the proposed tie persistent predictor. We depicts in Figure 2, 3, and 4 the plots of performance (Precision, Recall, and AUCPR) resulted from applying different values of parameters d (NBA co-authorship credit model), and θ (social-academic closeness factor), and from changing the relevance order of publication types (social-academic closeness factor). When the effect of a parameter is under experimentation, the other parameters are assigned with the default values that have been described previously in Section 4.2. Table 6 describes the overall Precision, Recall, and AUCPR on the complete dataset (that combines the three author sets Newcomers, Juniors, and Seniors) according to the various values of the aforementioned parameters.

6.1 The parameter d

Our results (Figure 2, Table 6) indicate that the performance of the proposed method varies inversely with d. The more the co-authorship credits get far from equality between the co-authors of the same paper ($d = 0$), the more the prediction accuracy of the proposed model gets lower. The only exception is for the Recall that peaks the best value when $d = 0.5$ (i.e. the first co-author has more than 50% as contribution in a given paper). It is a difficult conclusion to draw without careful investigations that computer scientists in their publications share equal credits with their co-authors because the first authors in a given publication have generally the greater contribution within [24]. Instead, we can assume that contributions may take other forms beyond the formal way (co-authoring the publication). This includes for example informal discussions between co-authors, supervision (advising), technical or academic assistance…etc. and all other practices that strengthen the academic relations between scientists but unfortunately, they are difficult to estimate formally.

6.2 The order of publication types and the parameter θ

The three versions of tie persistence predictor (LPP) that are relevant to the three orders B-C-J, C-B-J, and C-J-B respectively (Figure 3, Table 6), show comparable performance results in Precision, Recall and AUCPR. As for the B-J-C order (the default order of publication types), a lower Recall, a comparable AUCPR, and a slightly greater Precision are marked. Two observations

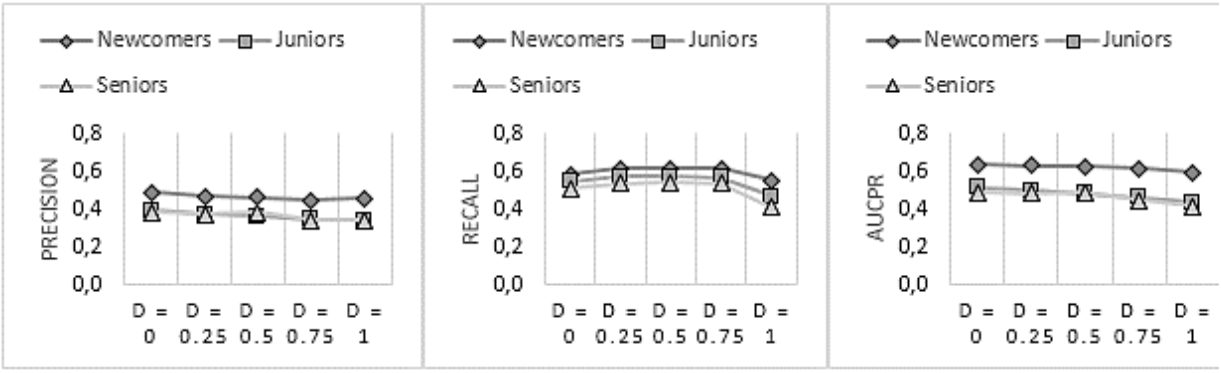

Figure 2: Performance of LPP with different values of parameter d.

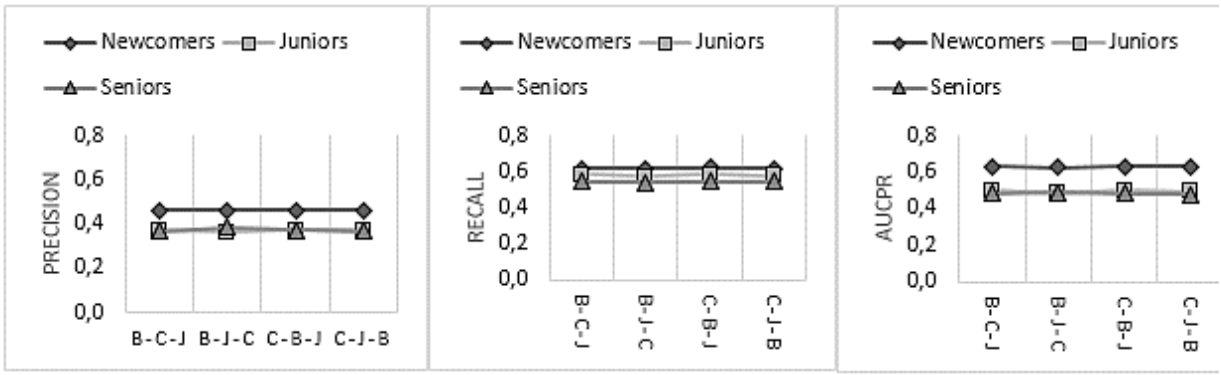

Figure 3: Performance of LPP with different settings of publication types order.

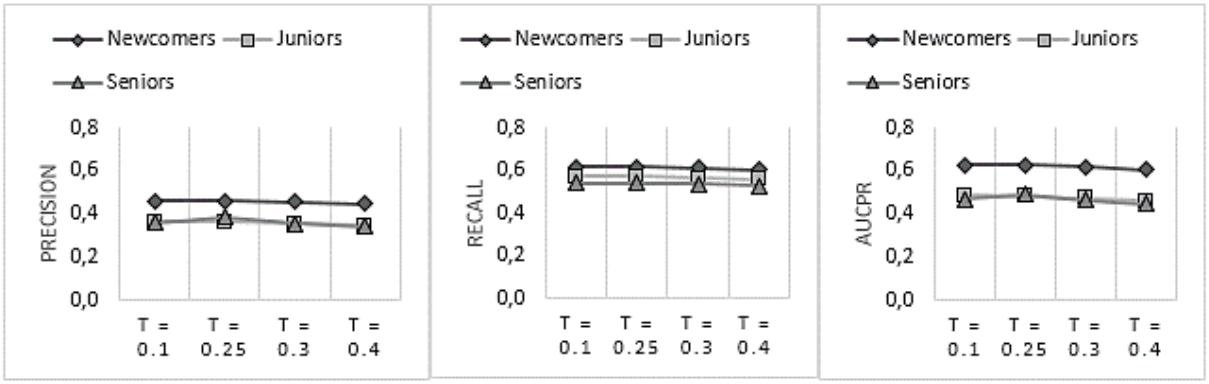

Figure 4: Performance of LPP with different values of parameter θ.

are worth noting here. First, the results reveal that "books" do not play a relevant role in determining the social-academic closeness between computer scientists. We think that this is not due to the unimportance of books but to the fact that this type of publication is not frequent in computer science field [18]. Second, if we ignore the "book" publication type, we observe that while using the ordinary order (B-J-C) seems yield to slightly more precision, the overall performance remains nearly comparable to the performance of the other order settings where conference papers are considered more relevant than journal papers (B-C-J, C-B-J, and C-J-B). In the DBLP bibliographic database, journal papers are less frequent than conference papers even though they have much higher impact [17]. Consequently, the outcome that we can assume from these conflicting findings is that the high frequency of conference papers in computer science collaboration networks reinforces positively the role of this type of publication in defining the social-academic closeness factor between computer scientists.

The parameter θ maintains the difference in relevance between the three publication types: book, journal paper, and conference paper. We tested θ with the default order B-J-C. The results (Figure 4, Table 6) show that the greater the difference between relevance values of publication types, the lower the prediction accuracy. This means that θ should be an appropriate value, which does not underestimate the role of publications, whatever their types, in maintaining the academic closeness between collaborators. $\theta = 0.25$ seems to be a suitable value since it gives convenient social-academic closeness values, which contribute to the well performance of the tie persistent predictor.

7 Conclusion

Studying the dynamics of a tie is a crucial step to comprehend the structure and the evolution overtime of social networks. In an academic social network, this is related to a number of factors that must be examined in order to better fix their actual effects on maintaining the connectivity of the network. We modeled a tie persistence prediction approach basing on estimating the tie strength using three factors: the frequency of collaborations, the social-academic closeness, and the scientific contributions of the scientists; and taking into account two other scientists' academic-related attributes: the collaborative productivity and the professional rank. Experimenting the model, we found significant impact of the aforementioned factors on the persistence or the dissolution of collaboration relations between scientists in academic social networks. Our findings also reported that a strong collaboration relation does not always persist due to other academic reckonings that are not easy to expect mostly for experienced scientists. It would be interesting then to develop useful techniques that have the ability to catch such unexpected collaboration choices. There is much room for improvement, particularly designing better metrics to estimate the contribution of the author in the relation and the social-academic closeness between the co-authors. As well, using the academic age to infer approximately the author's professional rank is a limited method. It would be better to develop efficient schemes that may provide realistic information about scientists' status in academic social networks. Further research might also investigate the impact of other academic-related attributes and other tie strength indicators, which have not been invested in this paper such as trust, reciprocity, and breadth of topics. Finally, applying the proposed model on larger academic networks and on academic networks from other academic fields may improve the performance of the tie persistence predictor and provide much understanding of collaboration trends in these disciplines.

References

[1] Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the Web. *Social Networks, 25*(3), 211–230.

[2] Akoglu, L., & Dalvi, B. (2010). Structure, tie persistence and event detection in large phone and SMS networks. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs - MLG '10* (pp. 10–17).

[3] Aron, A., Aron, E., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology, 63*, 596–612.

[4] Backstrom, L., & Leskovec, J. (2010). Supervised random walks: Predicting and recommending links in social networks. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 635–644.

[5] Bahr, A. H., & Zemon, M. (2000). Collaborative authorship in the journal literature: Perspectives for academic librarians who wish to publish. *College & Research Libraries, 61*(5), 410–419.

[6] Barabãsi, A. L. et al. (2002). Evolution of the social network of scientific collaborations. *Physica A 311*, 590–614.

[7] Berscheid, E., Snyder, M., & Omoto, A. M. (1989). The relationship closeness inventory: Assessing the closeness of interpersonal relationships. *Journal of Personality and Social Psychology, 57*, 792–807.

[8] Blumstein, P., & Kollock, P. (1988). Personal Relationships. *Annual Review of Sociology, 14*, 467–490.

[9] Bozeman, B., & Corley, E. (2004). Scientists' collaboration strategies: Implications for scientific and technical human capital. *Research Policy, 33*(4), 599–616.

[10] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1/7), 107–17.

[11] Burt, R. S. (2000). Decay functions. *Social Networks, 22*, 1–28.

[12] Costas, R., & Bordons, M. (2011). Do age and professional rank influence the order of authorship in scientific publications? Some evidence from a micro-level perspective. *Scientometrics, 88*(1), 145–161.

[13] Dibble, J. L., Levine, T. R., & Park, H. S. (2012). The Unidimensional Relationship Closeness Scale (URCS): Reliability and validity evidence for a new measure of relationship closeness. *Psychological Assessment, 24*(3), 565–572.

[14] Feld, S. L. (1997). Structural embeddedness and stability of interpersonal relations. *Social Networks. 19*, 91-95.

[15] Fonseca, L., Velloso, S., Wofchuk, S., & De Meis, L. (1998). The relationship between advisors and students. *Scientometrics, 41*(3), 299–312.

[16] Fouss, F., Pirotte, A., Renders, J. M., & Saerens, M. (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering, 19*(3), 355–369.

[17] Franceschet M. (2010). The role of conference publications in computer science: a bibliometric view. *Communications of the ACM, 53*(12), 129-132.

[18] Franceschet, M. (2011). Collaboration in computer science: A network science approach. *Journal of the American Society for Information Science and Technology, 62*(10), 1992–2012.

[19] Girvan, M. & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America, 99*(12), 7821–7826.

[20] Hasan, M. Al, Chaoji, V., Salem, S., & Zaki, M. (2006). Link prediction using supervised learning.

SDM'06: Workshop on Link analysis, Counter terrorism and Security.

[21] Hasan, M. Al, & Zaki, M. J. (2011). A Survey in link prediction in social networks. In *Social Network Data Analytics* (pp. 243–275).

[22] Hidalgo, C. A., & Rodriguez-Sickert, C. (2008). The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and Its Applications*, *387*(12), 3017–3024.

[23] Katz, J. S., & Martin, B. R. (1997). What is research collaboration?. *Research Policy*, *26*(1), 1–18.

[24] Kim, J., & Diesner, J. (2014). A network-based approach to coauthorship credit allocation. *Scientometrics*, 1–16.

[25] Kim, J., & Diesner, J. (2015). The effect of data pre-processing on understanding the evolution of collaboration networks. *Journal of Informetrics*, *9*(1), 226-236.

[26] Kivran-Swaine, F., Govindan, P., & Naaman, M. (2011). The impact of network structure on breaking ties in online social networks: Unfollowing on Twitter. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1101–1104.

[27] Ley, M. (2009). DBLP: some lessons learned. *Proceedings of VLDB Endow.*, *2*(2), 1493-1500.

[28] Liben-Nowell, D., & Kleinberg, J. (2003). The link prediction problem for social networks. *Proceedings of the Twelfth Annual ACM International Conference on Information and Knowledge Management (CIKM)*, 556–559.

[29] Lichtenwalter, R., & Chawla, N. V. (2012). Link prediction: Fair and effective evaluation. *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, *ASONAM 2012*, 376–383.

[30] Long, J. S., & McGinnis, R. (1985). The effects of the mentor on the academic career. *Scientometrics*, *7*(3-6), 255–280.

[31] Mali, F., Kronegger, L., Doreian, P., & Ferligoj, A. (2012). Dynamic scientific co-authorship networks. *Understanding Complex Systems*, 195–232.

[32] Martin, J. L., & Yeung, K. T. (2006). Persistence of close personal ties over a 12-year period. *Social Networks*, *28*(4), 331–362.

[33] Newman, M. E. J. (2001). Scientific collaboration networks: I. Network construction and fundamental results. *Physical Review E*, *64*(1), 1–8.

[34] Perlman, D., & Fehr, B. (1987). The development of intimate relationships. In *Intimate Relationships Development and Deterioration* (pp. 13–42).

[35] Raeder, T., Lizardo, O., Hachen, D., & Chawla, N. V. (2011). Predictors of short-term decay of cell phone contacts in a large-scale communication network. *Social Networks*, *33*(4), 245–257.

[36] Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. *Introduction to Modern Information Retrieval*.

[37] Velden, T., Haque, A. U., & Lagoze, C. (2010). A new approach to analyzing patterns of collaboration in co-authorship networks: Mesoscopic analysis and interpretation. *Scientometrics*, *85*(1), 219–242.

[38] Wellman, B., Wong, R. Y., Tindall, D., & Nazer, N. (1997). A decade of network change: Turnover, persistence and stability in personal communities. *Social Networks*, *19*, 27–50.