Research Paper

Predictive Characteristics of Co-authorship Networks: Comparing the Unweighted, Weighted, and Bipartite Cases

Raf Guns†

Centre for R&D Monitoring (ECOOM), University of Antwerp, Antwerp 2020, Belgium

Citation: Raf Guns (2016). Predictive Characteristics of Co-authorship Networks: Comparing the Unweighted, Weighted, and Bipartite Cases.

Received: May 17, 2016
Revised: Jun. 15, 2016
Accepted: Jun. 26, 2016

Abstract

Purpose: This study aims to answer the question to what extent different types of networks can be used to predict future co-authorship among authors.

Design/methodology/approach: We compare three types of networks: unweighted networks, in which a link represents a past collaboration; weighted networks, in which links are weighted by the number of joint publications; and bipartite author-publication networks. The analysis investigates their relation to positive stability, as well as their potential in predicting links in future versions of the co-authorship network. Several hypotheses are tested.

Findings: Among other results, we find that weighted networks do not automatically lead to better predictions. Bipartite networks, however, outperform unweighted networks in almost all cases.

Research limitations: Only two relatively small case studies are considered.

Practical implications: The study suggests that future link prediction studies on co-occurrence networks should consider using the bipartite network as a training network.

Originality/value: This is the first systematic comparison of unweighted, weighted, and bipartite training networks in link prediction.

Keywords Network evolution; Link prediction; Weighted networks; Bipartite networks; Two-mode networks

1 Introduction

Citations, collaborations, Web links, and other phenomena of interest to the field of informetrics can be studied from a network perspective (Otte & Rousseau, 2002). Co-authorship networks are among the most studied types of networks and can be considered an approximation of collaboration networks (Katz & Martin, 1997). During the latest decades, it has been shown that co-authorship networks share

† Corresponding author: Raf Guns (E-mail: raf.guns@uantwerpen.be).

JDIS
Journal of Data and Information Science
Vol. 1 No. 3, 2016
pp 59–78
DOI: 10.20309/jdis.201620
http://www.jdis.org

several typical characteristics with other kinds of complex networks. These include high clustering coefficients, low average shortest path lengths, and highly skewed degree distributions (Watts & Strogatz, 1998).

An important milestone in the study of networks was reached by the introduction of random network models (Erdős & Rényi, 1959). These models describe the class of networks in which n nodes are given that are connected by a given number of random links (or, alternatively, where random links occur with a given probability p). In a random network each link has an equal probability of occurrence, regardless of the network's structure. Despite their theoretical significance random networks do not exhibit all characteristics that are typical of many real-world networks.

Several more advanced models have been proposed that seek to characterize and explain how a network can obtain these characteristics. Let us take the preferential attachment model (equivalent to a success-breeds-success model; Barabási & Albert, 1999; Price, 1976) as a well-known example. In this model, new nodes are added to the network at each time step. Each new node connects to k existing nodes. The probability for a new node to connect to an existing node is proportional to the existing node's degree. In other words, contrary to the random network model, the probability for a given link to involve a specific node depends on the state of the network.

If one agrees that the state of a network influences a link's probability, it can be seen that some links are very likely to emerge, whereas others are extremely unlikely. Put another way, every non-random network has certain predictive characteristics that make some links more probable than others. The task of predicting which links are most likely to occur in a future state of the network is known as *link prediction* (Liben-Nowell & Kleinberg, 2007).

We may distinguish between two types of link prediction applications (Guns, 2014) that have sometimes been confounded in the literature:

1) Network evolution prediction, and
2) Network reconstruction.

Network evolution prediction (Liben-Nowell & Kleinberg, 2007) concerns the situation where one is given a temporal snapshot of an evolving network. The task is to predict a future state of the network. *Network reconstruction* (Guimerà & Sales-Pardo, 2009), on the other hand, concerns the case where a network is damaged by randomly deleting links and/or adding random spurious links[①]. The task is to

[①] We only consider the case where a network is damaged by random deletions and/or additions. The task becomes more challenging if the most important links (e.g. those with high edge betweenness centrality) are targeted first. In other words, we want to test the *error tolerance* rather than the *attack tolerance* (Jalili, 2011).

reconstruct the original network based on the damaged network. We will refer to the given network (i.e. the older snapshot or the damaged network) as the *training network* and to the network that should be obtained (i.e. the more recent snapshot or the undamaged network) as the *test network*.

In this paper, we study the predictive characteristics of co-authorship networks for network evolution prediction. Specifically, we want to compare three different 'formats' of the same network: unweighted, weighted and bipartite. Previous studies mainly considered the unweighted case, but translocating weighted networks has received more attention (Koren, North, & Volinsky, 2006; Lü & Zhou, 2010; Murata & Moriyasu, 2007; Lü & Zhou, 2010; Zhu & Xia, 2016). This is not surprising as (1) weighted networks are very well studied (e.g. Barrat et al., 2004; Newman, 2001b), making them obvious targets to consider for link prediction optimization, and (2) the unweighted networks are usually derived from the weighted networks, which in their turn are derived from the bipartite ones.

Figure 1 illustrates the difference between the three types. On the left, we have a two-mode or bipartite author-publication network. One can, for instance, see that publication A was written by authors Mark, Zoe, and Jane. This bipartite network corresponds to the weighted co-authorship network in the middle, where link weights denote the number of joint publications. Finally, omitting the link weights results in the unweighted network on the right.

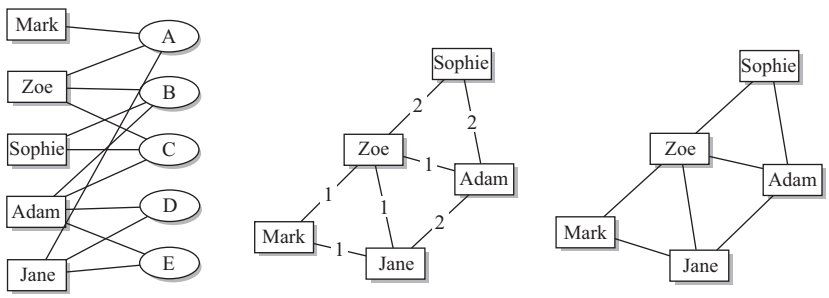

Figure 1. Example: Bipartite author-publication network and its corresponding weighted and unweighted co-authorship networks.

The weighted co-authorship network contains more information than the unweighted one since every unweighted network corresponds to infinitely many weighted ones. At the same time, a weighted co-authorship network typically corresponds to multiple bipartite networks. The non-trivial problem of determining the exact number of corresponding bipartite networks is beyond the scope of the present article. Bipartite networks carry more information than their weighted projections, which in turn carry more information than unweighted networks. We

will study and compare the performance of link predictors applied to unweighted, weighted and bipartite networks. Henceforth, we may refer to them as unweighted, weighted and bipartite predictors, respectively.

We formulate the following hypotheses. First, we have two hypotheses regarding *positive stability*, the simple re-occurrence of links in the test network:

H1a: Other things being equal, if two links occur in the weighted training network, the one with a higher link weight is more likely to re-occur in the test network than the other one.
H1b: Other things being equal, if the bipartite training network contains two publications with respectively n and m authors, and $n > m$, the latter authors are more likely to be linked in the test network than the former.

Hypothesis H1a simply states that two authors with more joint papers are more likely to write a joint paper in the future. Hypothesis H1b could be formulated slightly less formally as follows:

The intensity of collaboration (and hence the probability of renewed collaboration) decreases with the number of additional collaborators.

The rationale behind this hypothesis can perhaps best be explained by comparing cases of 'mega-authorship' (Kretschmer & Rousseau, 2001) or 'hyperauthorship' (Cronin, 2001) – papers written by hundreds or even thousands of authors – with papers authored by two or three researchers. It seems extremely unlikely that the intensity of close collaboration in the former case is as high as it is in the latter case.

Next, we have two hypotheses regarding prediction performance:

H2a: The performance of a weighted predictor is better than the performance of its unweighted counterpart.
H2b: The performance of a bipartite predictor is better than the performance of both its weighted and unweighted counterpart.

Hypothesis H2a stems from the fact that weighted networks carry more information. It therefore seems logical that they also allow for better predictions. Similar considerations lead, *mutatis mutandis*, to hypothesis H2b. Since there are many predictors in each family, hypotheses H2a and H2b need to be tested for each predictor separately. We see no *a priori* reason to believe that these hypotheses are valid for some predictors and not for others. Guns (2011) compared prediction performance for unweighted networks to bipartite networks but omitted weighted networks. In this paper, we attempt to systematically compare all three forms.

2 Data

We use two datasets of co-authorship between researchers. The first dataset will be referred to as the *UA* dataset; it consists of all co-authored publications at the University of Antwerp (Belgium) across all faculties and departments during the

period 2001–2006. This case study is based on the University of Antwerp's own publication database, which is virtually exhaustive for the time period considered. The dataset is divided into two time slices: for training we use the period 2001–2003 and the test period is 2004–2006.

The second dataset will be referred to as the *INF* dataset and consists of co-authored publications in the field of informetrics during the period 1990–2009. This dataset is based on data found in Thomson Reuters' Web of Science. Here, we use 1990–2004 as the training period and 2005–2009 as the test period. For more details on how the data were acquired and cleansed, we refer to (Guns, Liu, & Mahbuba, 2011).

Table 1 provides some basic statistics on the two datasets. In both cases, we only include those authors that occur at least once in both training and test dataset. This explains why the author numbers are the same for the training and test datasets. In other words, authors that stop publishing after the training period or that start publishing in the test period are excluded from the data, as is customary in link prediction studies (e.g. Liben-Nowell & Kleinberg, 2007). In rapidly expanding fields or fields with a high turnover, this would reduce the usefulness of link prediction. Note that the average number of papers per author is higher than the number of papers divided by the number of authors, because of multi-authored papers.

Table 1. Descriptive statistics of the two datasets.

		UA	INF
Training	Number of authors	1,102	397
	Number of papers	7,569	1,118
	Average number of papers per author	11.9	4.3
	Average number of authors per paper	1.7	1.5
Test	Number of authors	1,102	397
	Number of papers	7,939	1,069
	Average number of papers per author	12.8	4.4
	Average number of authors per paper	1.8	1.6

3 Methods

The primary goal of the current article is to compare the influence of the nature of the training network on prediction performance. Specifically, we are interested in the question to what extent bipartite, weighted and unweighted networks can impact the accuracy of the prediction results. We will mainly focus on a number of predictors that have been thoroughly tested in the literature. Here, we will briefly recapitulate their definitions for unweighted unipartite networks and outline if and how they can be applied to weighted and bipartite networks. All predictors were used as implemented in the *linkpred* software (Guns, 2014).

We use the following notation. Each predictor determines a likelihood score $s(u, v)$ that specifies the likelihood of a link occurring between nodes u and v in the test network. The set of neighbors that a node is connected to is called its *neighborhood*. The neighborhood of v is denoted by $\Gamma(v)$. We use $|\cdot|$ to denote set cardinality; hence, $|\Gamma(v)|$ is the *degree* of v, the number of adjacent nodes. Finally, $w(x, y)$ denotes the weight of the link between x and v.

3.1 Unweighted Predictors

The first predictor is *common neighbors*, which simply measures the number of common neighbors for two nodes:

$$s(u,v) = \left|\Gamma(u) \cap \Gamma(v)\right|. \tag{1}$$

Because it is so simple and it occurs in the seminal work by Liben-Nowell and Kleinberg (2007), this predictor is used (if only as a comparison point) in virtually every link prediction study. It is intuitively clear that, for instance, two unconnected people who have a large amount of friends in common are likely to eventually meet and befriend each other. This intuition was confirmed empirically in collaboration networks by Newman (2001a), who found that "[a] pair of scientists who have five mutual previous collaborators, for instance, are about twice as likely to collaborate as a pair with only two, and about 200 times as likely as a pair with none." In other words, many social networks exhibit a natural tendency towards forming triangles and hence, the common neighbors predictor is directly related to the clustering coefficient.

Several normalizations of the common neighbors predictor have been proposed in the literature. Because most of these normalizations yield very similar results (Guns, 2009), we choose one to represent this set of predictors, namely the *cosine* measure:

$$s(u,v) = \frac{\left|\Gamma(u) \cap \Gamma(v)\right|}{\sqrt{\left|\Gamma(u)\right|}\sqrt{\left|\Gamma(v)\right|}}. \tag{2}$$

Let $d(u, v)$ denote the length of the shortest path from u to v. The *graph distance* predictor then is defined as:

$$s(u,v) = \frac{1}{d(u,v)}. \tag{3}$$

Katz (1953) proposed a centrality indicator whose stated aim was to overcome the limitations of plain degree centrality. Interestingly, the same procedure can be used to obtain a measure of relatedness between two nodes. The latter has become known as the *Katz predictor*. Let $\mathbf{A}$ denote the adjacency matrix of the unweighted network G, such that $\mathbf{A}_{ij}$ is 1 if there is a link between nodes i and j, and 0 otherwise.

Then, each element $\mathbf{A}^k_{ij}$ of $\mathbf{A}^k$ (the k^{th} power of $\mathbf{A}$) has a value equal to the number of walks with length k from i to j (Wasserman & Faust, 1994). Generally, longer walks indicate a weaker association between the start and end node. Katz (1953) therefore introduces a parameter β ($0 < \beta < 1$), representing "the probability of effectiveness of a single link." Based on the assumption that more and shorter walks between two nodes indicate a higher likelihood of link formation, the Katz predictor is defined as:

$$s(i,j) = \sum_{k=1}^{\infty} \beta^k \mathbf{A}^k_{ij} . \tag{4}$$

The intuition behind *rooted PageRank* (Liben-Nowell & Kleinberg, 2007) can be understood from the perspective of a random walker. The random walker starts at a fixed node i. At each step, it moves to a random neighbor of the current node. Contrary to ordinary PageRank, rooted PageRank does not allow teleportation to a randomly chosen node with probability $(1–\alpha)/n$. Instead, there is a probability $(1–\alpha)$ that the walker will teleport back to i. Note that the literature mentions several measures that are closely related to rooted PageRank, such as cycle-free effective conductance (Koren et al., 2006) and escape probability (Song et al., 2009).

3.2 Weighted Predictors

We now turn to the question how the aforementioned predictors can be applied to weighted networks, such that they take link weight into account.

Murata and Moriyasu (2007) propose the following weighted variant of *common neighbors*:

$$s(u,v) = \sum_{Z \in \Gamma(u) \cap \Gamma(v)} \frac{w(u,z) + w(z,v)}{2} . \tag{5}$$

A potential downside of their approach is that this weighted variant lacks a theoretical basis and has been constructed *ad hoc*. Since the neighborhood-based predictors are based on similarity measures that have been used in information retrieval (Boyce, Meadow, & Kraft, 1994; Salton & McGill, 1983; van Rijsbergen, 1979), link weights can be taken into account by using the vector interpretation of the set-based similarity measures (Egghe & Michel, 2002; Egghe, 2009).

If element i is a member of the set, its corresponding vector element $x_i = 1$. Otherwise, $x_i = 0$. Assume there are two sets A and B, with corresponding vectors $\vec{X}$ and $\vec{Y}$. If

$$A = \{i \in 1,\ldots,n \mid x_i = 1\}$$
$$B = \{i \in 1,\ldots,n \mid y_i = 1\} ,$$

the step from sets to vectors can be taken using the following equation:

$$\begin{aligned} |A| &= \left\|\vec{X}\right\|^2 = \sum_{i=1}^{n} x_i^2 \\ |B| &= \left\|\vec{Y}\right\|^2 = \sum_{i=1}^{n} y_i^2 \\ |A \cap B| &= \vec{X} \cdot \vec{Y} = \sum_{i=1}^{n} x_i \cdot y_i \end{aligned} \quad . \tag{6}$$

The advantage of the vector-based measures over the set-theoretic ones is that they are not limited to the binary case, i.e. vector elements can have any numeric value.

To apply these in the context of link prediction, vectors $\vec{X}$ and $\vec{Y}$ are, respectively, the adjacency vectors of nodes *u* and *v*. We thus obtain the weighted interpretation of common neighbors:

$$s(u,v) = \sum_{i=1}^{n} x_i \cdot y_i \ . \tag{7}$$

Note that Equations (5) and (7) do not result in the same ranking. This is illustrated by Figure 2. According to (5), $s(u,v) = 3.5 > s(u',v') = 3$, whereas according to (7), $s(u,v) = 6 < s(u',v') = 9$. In our opinion, the ranking by (7) is more logical: both links need to have sufficient weight to indicate a 'strong' common neighbor and, hence, the situation on the right should have a higher likelihood score.

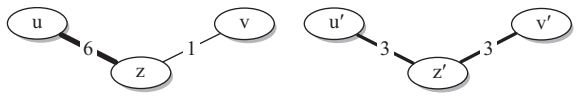

Figure 2. Comparison of two definitions of weighted common neighbors.

Likewise, *weighted cosine* becomes:

$$s(u,v) = \frac{\sum x_i \cdot y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \ . \tag{8}$$

Equation (8) indicates why this is called the cosine measure: it is the cosine of the angle between two vectors. In other words, this measure takes into account the direction but not the magnitude of the two vectors.

We now turn to the *weighted graph distance* predictor. Let $p(u, v)$ denote a path between *u* and *v* of length *t*. Let w_i (i = 1,..., t) denote the weight of the i^{th} link in the path. The weighted path length can be defined by taking the sum of the inverses of each weight (Brandes, 2001; Egghe & Rousseau, 2003; Newman, 2001b).

$$w_p(u,v) = \sum_{i=1}^{t} \frac{1}{w_i} \ . \tag{9}$$

One can then define the distance or dissimilarity between u and v as the length of the shortest weighted path:

$$d(u,v) = \min_{p} w_p(u,v). \tag{10}$$

Egghe and Rousseau (2003) remark that "A direct link is very important. Yet, if weights are high enough, it is possible that an indirect connection leads to a smaller dissimilarity than the direct link." Consider Figure 3. This network contains three paths between u and v. The shortest weighted path, using Equation (10), is the path u–y–v, with a length of 1/2 + 1/3 = 5/6. Interestingly, we note that the direct link and the path via x have the same weighted path length. This may be perfectly appropriate in some cases, especially if there is no real cost attached to the number of nodes one has to traverse (assuming that the link weights are high enough). However, it seems likely that a larger number of intermediary nodes may have a negative effect on at least some social networks. For instance, if the nodes in Figure 3 are authors and edge weights represent the number of joint publications, one could argue that the 'cost' of traversing the direct link is smaller than that of the path via x or even of the path via y.

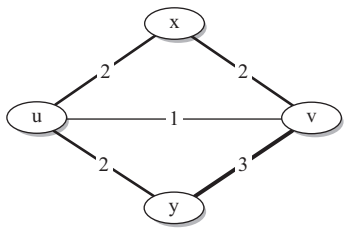

Figure 3. Example network with three paths between u and v.

On the basis of similar considerations, Opsahl, Agneessens, and Skvoretz (2010) propose a generalization of weighted path length in a proximity-based weighted network:

$$w_{p,\alpha}(u,v) = \sum_{i=1}^{t} \frac{1}{w_i^{\alpha}}, \tag{11}$$

where α is an extra tuning parameter. Here, we are only interested in those cases where $0 \leq \alpha \leq 1$. If $\alpha = 0$, Equation (11) reduces to the path length of the corresponding unweighted path—i.e. only the number of intermediary nodes is taken into account. If $\alpha = 1$, (10) reduces to (8)—i.e. only the edge weights are taken into account. Thus, setting $0 < \alpha < 1$ allows one to find a balance between both extremes. For instance, if $\alpha = 0.5$, the paths in Figure 3 get the following weights: direct link: 1; path via y: 1.28; path via x: 1.41. The weighted shortest path then is formed by the direct link.

We thus obtain the weighted graph distance predictor:

$$s(u,v) = \frac{1}{\min_p w_{p,\alpha}(u,v)}. \quad (12)$$

The *weighted Katz predictor* also uses Equation (4). The only difference is that in this case the adjacency matrix can contain any positive integer. As explained by Guns and Rousseau (2014), the value of an element $\mathbf{A}^k_{ij}$ in the weighted case is the number of walks with length k in the equivalent multigraph.

Finally, *weighted rooted PageRank* is very similar to *unweighted rooted PageRank*. The only difference is that link weights determine the transition probability to neighboring nodes. For instance, if we applied weighted rooted PageRank to the network in Figure 3 and the random walker was positioned at node v, node y (x) would be three (two) times more likely than u to be visited next.

3.3 Bipartite Predictors

The bipartite author-publication networks are unweighted. Hence, it is, at least mathematically speaking, possible to apply the predictors from the 'unweighted predictors' section to the bipartite network as well. Here, we consider some specific issues that may arise when doing so.

First, we emphasize that our current analysis is only concerned with predicting links between authors themselves, and not links between articles or between authors and articles. This can be achieved by simply distinguishing between 'eligible' (author) nodes and 'non-eligible' (article) nodes. We only retain predictions that involve two eligible nodes.

Let us now consider the neighbor-based predictors (*common neighbors* and *cosine*). If u and v are bottom nodes (authors), their neighbors are by definition top nodes (articles). In other words, common neighbors, cosine and variants can be used for bipartite networks as well. However, their interpretation is completely different. Most importantly, applying Equations (1) and (2) to a bipartite network never yields a *genuine* prediction (appearance or disappearance of a link), only an estimate of the likelihood a co-authorship link in the training network will *sustain* in the test network. While this is a valid application in itself, it is different from what is typically considered as link prediction.

Note that this limitation does not mean that the bipartite neighbor-based predictors are the same, since their ranking is different. The common neighbors predictor ranks in decreasing order of the number of joint publications. Cosine is similar, but normalizes with respect to the total number of publications of the authors.

The *bipartite graph distance predictor* will not be considered in our analysis. The reason is that it yields the same ranking as its unweighted unipartite counterpart: the shortest path length for any bottom node pair is simply doubled in the bipartite case.

Contrary to graph distance, the *bipartite Katz predictor* can result in a ranking that is different from the unipartite predictor. Moreover, compared to the neighbor-based predictors, Katz has the advantage that it is not restricted to neighboring nodes (or neighbors of neighbors). However, note that the bipartite variant does not incorporate hypothesis H1b. This can be illustrated through the example in Figure 4. Assume that the top nodes represent publications and the bottom nodes represent authors. In both cases, authors *a* and *b* have two collaborators in common and can be connected via two paths of length 4. Consequently, the Katz predictor ranks both cases equally high, even though our hypothesis implies that the left case should rank higher than the right one.

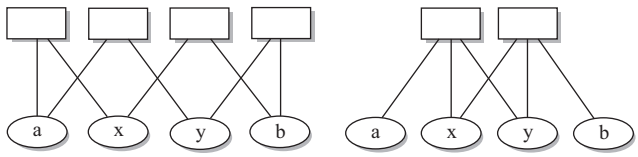

Figure 4. Hypothetical example: two cases of author–publication networks.

Finally, *bipartite rooted PageRank* has the same advantage as Katz when compared to neighbor-based predictors: it is not restricted to predicting co-authors that *u* and *v* have in common, let alone that it is – like bipartite common neighbors or cosine – restricted to authors that already collaborate in the training network. At each step, the random walker switches between top and bottom nodes (disregarding teleportation). If the random walker is at a given paper, each of the paper's authors has an equal chance of being the next hop. This implies that rooted PageRank values are lower in cases where papers have more authors: rooted PageRank ranks the right case of Figure 4 lower than the left one. In other words, if hypothesis H1b is correct, this should be reflected in the performance of bipartite rooted PageRank.

3.4 Evaluation

In the context of link prediction, the notions of recall and precision can be defined as follows. Precision *P* is the fraction of correctly predicted links compared to all predictions. Recall *R* is the fraction of correctly predicted links compared to all links in the test network. The *F*-score is the harmonic mean of precision and recall:

$$F = \frac{2PR}{P+R}. \tag{13}$$

If either *P* or *R* is low, *F* will be low as well. We rank predictions in decreasing order of their likelihood score. The values obtained for precision, recall and *F*-score depend on how many predictions are taken into account.

We use two measures for evaluation. Precision-at-20 (P@20) is the precision for the top 20 predictions and gauges a predictor's ability to make a small number of high-quality predictions. F_{max} is the highest F-score obtained by a predictor. As argued by Guns, Lioma, and Larsen (2012), this measure offers a good summary of the overall capacities of (in our case here) a predictor.

4 Results

4.1 Results for Positive Stability

Links in a network tend to be positively stable: we find that 61.7% (UA) and 62.7% (INF) of the links in the training network reoccur in the test network. Only 14.6% (UA) and 18.2% (INF) of the links in the test networks do not occur in the training network. All this suggests that plain reoccurrence is actually a fairly strong predictor.

What happens if we incorporate link weights? We test this on the UA dataset and find that the F-score is virtually unaffected but precision-at-20 increases from 0.48 to 1.00. Similar results are obtained for the INF dataset (Table 2). All in all, this lends strong support to hypothesis H1a.

To test the influence of the number of authors per paper, we determine for each author pair in the training network the median number of authors on their co-authored publications. This results in an increase of precision, but a decrease (UA) or status quo (INF) in F-score. The effect of this factor would likely be higher if tested on datasets where the overall number of authors per paper is higher. Indeed, the highest number of co-authors is 10 in the UA dataset and 6 in the INF dataset. Based on our current investigation, however, the evidence in favor of hypothesis H1b is limited.

Table 2. Results for positive stability.

	UA		INF	
	F_{max}	P@20	F_{max}	P@20
Reoccurrence	0.59	0.48	0.61	0.48
Reoccurrence with weights	0.59	1.00	0.61	0.81
Reoccurrence with author numbers	0.55	0.52	0.61	0.57

4.2 Results for Prediction Performance

The relative performance of different predictors has been studied quite extensively (e.g. Guimerà & Sales-Pardo, 2009; Guns, 2014; Liben-Nowell & Kleinberg, 2007; Song et al., 2009) and is not the main focus of the present paper. Rather, we are interested in comparing the influence of the network type on predictor performance. Hence, we will present results per predictor family.

Tables 3 and 4 contain the results for the neighborhood-based predictors for UA and INF, respectively. In all cases, we observe that the unweighted variants perform as well as or better than the weighted ones, both for *F*-score and precision. Although in most cases the difference between the two is negligible, this is quite surprising. We will discuss this remarkable result in the last section of the paper. For now, we just observe that it appears to counter hypothesis H2a.

The bipartite predictors perform much better (with the exception of precision for common neighbors applied to the UA dataset, which has a perfect score for all three types). This does not really corroborate H2b: though, as explained above, the bipartite neighborhood-based predictors only 'predict' the recurrence of already existing co-authorship relations. In other words, their good performance is related to hypothesis H1a.

Table 3. Results for neighborhood-based predictors (UA).

		Unweighted	Weighted	Bipartite
Common neighbors	F_{max}	0.3964	0.3928	0.5866
	P@20	1	1	1
Cosine	F_{max}	0.4025	0.397	0.5865
	P@20	0.1429	0.0952	0.7619

Table 4. Results for neighborhood-based predictors (INF).

		Unweighted	Weighted	Bipartite
Common neighbors	F_{max}	0.4091	0.3935	0.6109
	P@20	0.8571	0.5714	0.8095
Cosine	F_{max}	0.4332	0.4305	0.6217
	P@20	0.0476	0.0476	1

We have seen that Opsahl et al. (2010)'s refinement of graph distance introduces a tuning parameter that determines to what extent the measure takes link weights into account. Figure 5 shows the results for UA; the resulting picture for INF looks very similar. First, recall that $\alpha = 0$ corresponds to the unweighted case, whereas $\alpha = 1$ corresponds to the 'fully' weighted case. Neither of the two extremes turns out to be optimal. If weights are not used, precision suffers but as α increases, the maximum *F*-score decreases. The optimal situation for both datasets appears to lie around α values between 0.1 and 0.2. As has been explained above, we do not apply graph distance to a bipartite network, since the resultant ranking is the same as one would obtain for an unweighted network.

Figures 6–8 display the results for the Katz predictor for values of β between 0.1 and 0.9. For UA we find that, in general, the unweighted case has the lowest predictive power (Figure 6). In terms of precision (not shown), both the weighted

Research Paper

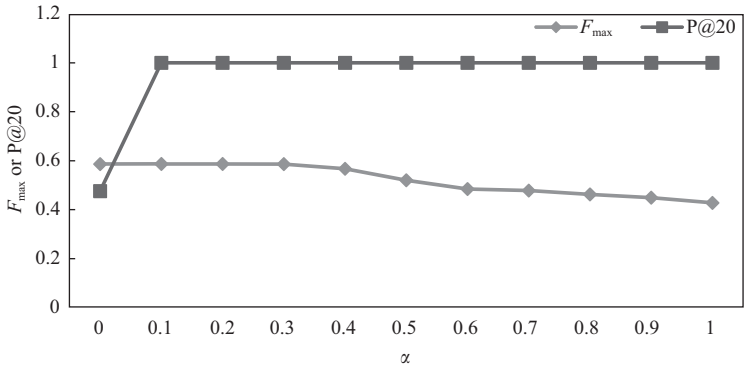

Figure 5. Influence of tuning parameter α on graph distance predictor (UA).

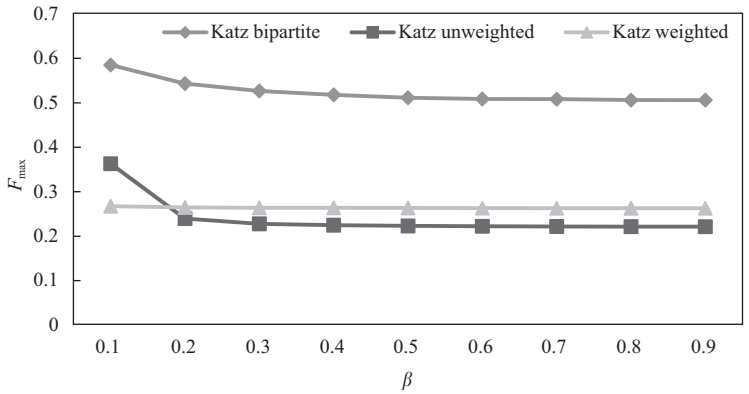

Figure 6. Comparison of Katz predictor on unweighted, weighted, and bipartite network (UA, F_{max}).

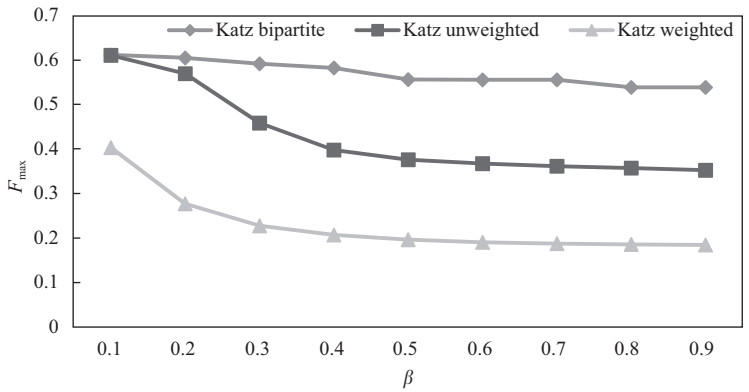

Figure 7. Comparison of Katz predictor on unweighted, weighted, and bipartite network (INF, F_{max}).

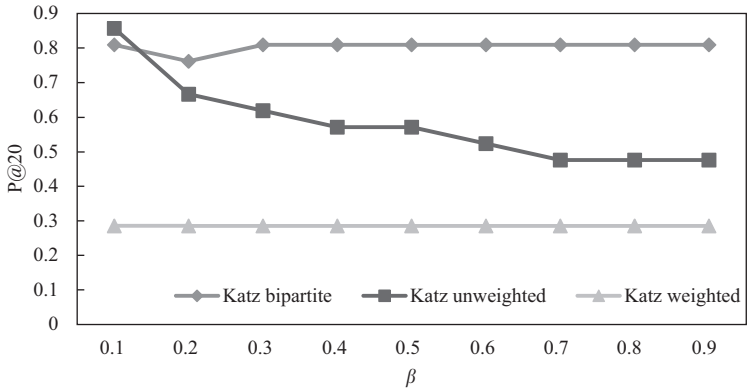

Figure 8. Comparison of Katz predictor on unweighted, weighted, and bipartite network (INF, P@20).

and bipartite variants obtain perfect scores for each value of β, while the unweighted variant yields values between 0.57 and 0.67. In terms of F-score, however, the advantage of the bipartite variant is quite clear, while the difference between the weighted and unweighted variants is fairly small. For INF the bipartite variant again scores best, followed by the unweighted variant and finally the weighted one. The difference between bipartite and unweighted increases with higher values of β.

Finally, we turn to the rooted PageRank predictor. Results are shown in Figures 9–12. For UA we find that the differences between the three network types yield rather small differences in terms of F-score (Figure 9). The differences are more pronounced for precision (Figure 10). The overall ranking by either measure, however, is similar: bipartite scores better than weighted, which in turn scores better than unweighted.

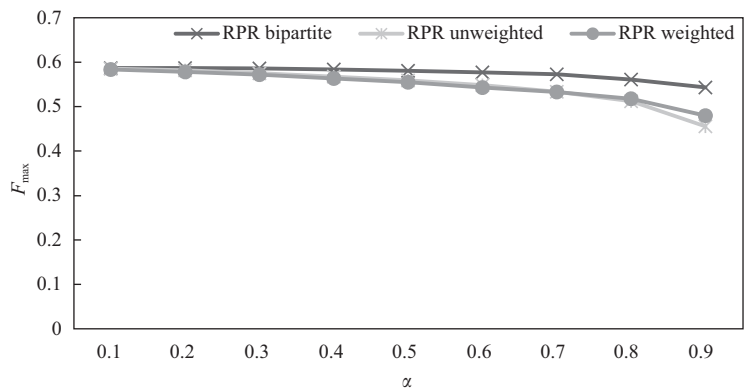

Figure 9. Comparison of rooted PageRank predictor on unweighted, weighted, and bipartite network (UA, F_{max}).

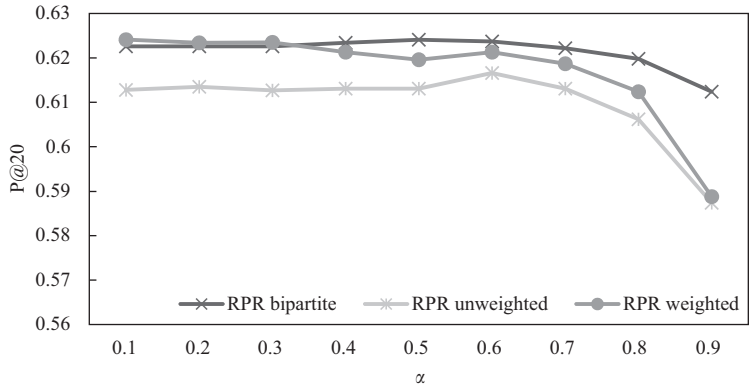

Figure 10. Comparison of rooted PageRank predictor on unweighted, weighted, and bipartite network (UA, P@20).

A similar conclusion holds for the INF dataset (Figures 11 and 12). Here, we see that unweighted and weighted obtain the same precision scores, but in terms of *F*-score the weighted variant is better. The bipartite variant, however, scores best overall.

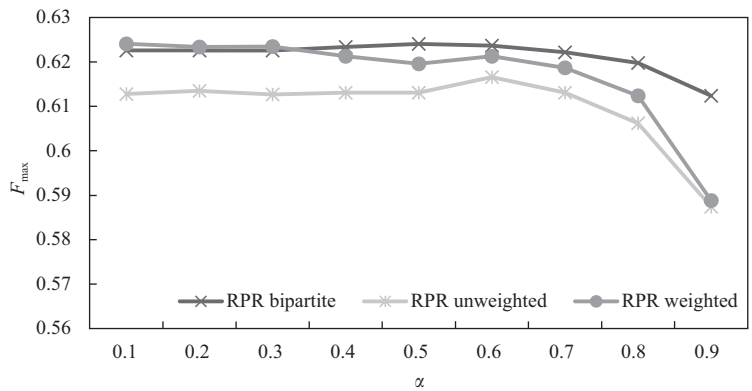

Figure 11. Comparison of rooted PageRank predictor on unweighted, weighted, and bipartite network (INF, F_{max}).

5 Discussion and Conclusion

In this paper we have compared predictive characteristics of three types of networks representing co-authorship: unweighted, weighted, and bipartite. The overall hypothesis of the paper is that more information will enhance a network's predictive potential. This general hypothesis was operationalized in four specific hypotheses.

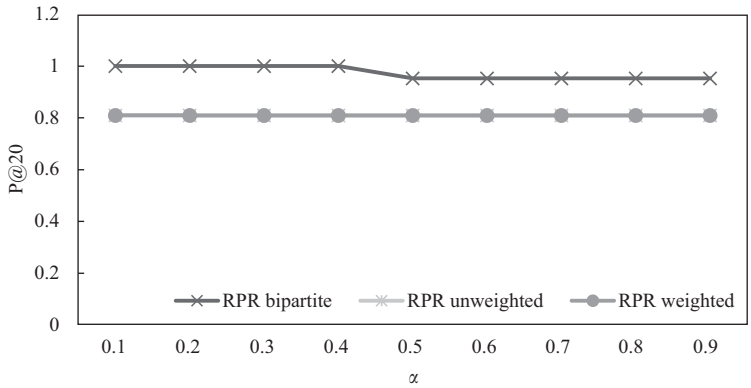

Figure 12. Comparison of rooted PageRank predictor on unweighted, weighted, and bipartite network (INF, P@20).

We have shown that collaborations tend to reoccur and that the more joint papers, the more likely the authors are to be co-authors in the future (hypothesis H1a). We can explain this using a barrier metaphor. In the context of collaboration and co-authorship, initiating new collaborations requires overcoming certain 'barriers' (geographic, cognitive, personal…); once this has happened and collaboration has been successful enough to result in at least one shared publication, future collaborations become easier. More co-authored papers imply that the barrier has been lowered even further.

We find only limited evidence for hypothesis H1b, however. A possible explanation could be that the numbers of authors per paper in our data are simply too small to have a clear effect. Nevertheless, our other results show that bipartite networks can lead to an improvement in prediction performance.

Hypothesis H2a states that the performance of a weighted predictor is better than the performance of its unweighted counterpart. This hypothesis is confirmed by the results for rooted PageRank (both UA and INF) and Katz (only UA), but contradicted by the other results. Our results for graph distance may help to clarify this unexpected finding. The best graph distance results were obtained when the tuning parameter α was around 0.1 to 0.2, corresponding to a limited influence of link weights on the result. This suggests that prediction with a weighted network *can* improve on prediction with an unweighted one, but that the importance of link weights should not be overestimated. Overall, hypothesis H2a is probably an overstatement and should be rejected in its current form. It is an interesting question for future research to find out to what extent it is possible to come up with variants of other predictors

that allow for tuning the influence of link weights. For instance, the generalization of degree centrality proposed by Opsahl et al. (2010) might form the basis of a more refined common neighbor's predictor which does this.

Hypothesis H2b states that the performance of a bipartite predictor is better than the performance of both its weighted and unweighted counterpart. We feel that the results for Katz and rooted PageRank confirm this hypothesis (we leave common neighbors, cosine and graph distance out of the discussion for reasons explained earlier). Only in a few specific cases and with specific parameter values did we observe better results for the weighted than the bipartite case. In most cases, however, the bipartite predictors clearly outperformed the weighted as well as the unweighted ones.

The results of the present study suggest that bipartite networks are underused in current link prediction studies. We have shown that the performance of some common predictors can be improved by applying them to a bipartite network. This finding also raises the question if new predictors that are attuned to bipartite networks can be created. More generally, the results illustrate that bipartite networks can yield results that cannot be obtained by unweighted or weighted networks. Bibliometric studies of co-authorship (but also co-word, co-citation, bibliographic coupling, etc.) might be enhanced by considering the bipartite form of the network.

It would be interesting to apply the same exercise to other networks that are derived from a bipartite network, either co-authorship networks or others. For instance, our case studies contain hardly any instances of hyperauthorship. While we expect that datasets that do include such cases would actually benefit from the bipartite approach, this has as of yet not been shown empirically. We leave this as a suggestion for further research.

As a general concluding comment, it is worth repeating that the current approach to link prediction does not include the appearance or disappearance of nodes. A more comprehensive prediction model that includes these aspects would probably need to incorporate more information than just network topology but also node or link attributes. Multi-relational networks or property graphs (Rodriguez & Neubauer, 2010) are one possible way of approaching this (Guns, 2012).

Acknowledgements

I thank Ronald Rousseau for useful comments on a previous draft and for his excellent supervision of my PhD dissertation. I also thank the anonymous reviewers for their comments.

References

Barabási, A.L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509.

Barrat, A., Barthélémy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. Proceedings of the National Academy of Sciences, 101(11), 3747–3752.

Boyce, B.R., Meadow, C.T., & Kraft, D.H. (1994). Measurement in information science. San Diego: Academic Press.

Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of Mathematical Sociology, 25, 163–177.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? Journal of the American Society for Information Science and Technology, 52(7), 558–569.

Egghe, L. (2009). New relations between similarity measures for vectors based on vector norms. Journal of the American Society for Information Science and Technology, 60(2), 232–239.

Egghe, L., & Michel, C. (2002). Strong similarity measures for ordered sets of documents in information retrieval. Information Processing & Management, 38(6), 823–848.

Egghe, L., & Rousseau, R. (2003). A measure for the cohesion of weighted networks. Journal of the American Society for Information Science and Technology, 54(3), 193–202.

Erdős, P., & Rényi, A. (1959). On random graphs, I. Publicationes Mathematicae (Debrecen), 6, 290–297.

Guimerà, R., & Sales-Pardo, M. (2009). Missing and spurious interactions and the reconstruction of complex networks. Proceedings of the National Academy of Sciences of the United States of America, 106(52), 22073–22078.

Guns, R. (2009). Generalizing link prediction: Collaboration at the University of Antwerp as a case study. Proceedings of the American Society for Information Science & Technology, 46(1), 1–15.

Guns, R. (2011). Bipartite networks for link prediction: Can they improve prediction performance? In Proceedings of ISSI 2011 – 13th International Conference of the International Society for Scientometrics and Informetrics (pp. 249–260). Leiden: Leiden University Press.

Guns, R. (2012). Missing links: Predicting interactions based on a multi-relational network structure with applications in informetrics. Antwerp. (University of Antwerp Ph.D dissertation)

Guns, R. (2014). Link prediction. In Ding, Y., Rousseau, R., & Wolfram, D. (Eds.), Measuring Scholarly Impact: Methods and Practice (pp. 35–55). Berlin: Springer.

Guns, R., Lioma, C., & Larsen, B. (2012). The tipping point: F-score as a function of the number of retrieved items. Information Processing & Management, 48(6), 1171–1180.

Guns, R., Liu, Y.X., & Mahbuba, D. (2011). Q-measures and betweenness centrality in a collaboration network: A case study of the field of informetrics. Scientometrics, 87(1), 133–147.

Guns, R., & Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. Scientometrics, 101(2), 1461–1473.

Jalili, M. (2011). Error and attack tolerance of small-worldness in complex networks. Journal of Informetrics, 5(3), 422–430.

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

Katz, L. (1953). A new status index derived from sociometric analysis. Psychometrika, 18(1), 39–43.

Koren, Y., North, S.C., & Volinsky, C. (2006). Measuring and extracting proximity in networks. In KDD2006: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (pp. 245–255). New York: ACM.

Kretschmer, H., & Rousseau, R. (2001). Author inflation leads to a breakdown of Lotka's law. Journal of the American Society for Information Science and Technology, 52(8), 610–614.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 58(7), 1019–1031.

Lü, L., & Zhou, T. (2010). Link prediction in weighted networks: The role of weak ties. EPL (Europhysics Letters), 89(1), 18001.

Murata, T., & Moriyasu, S. (2007). Link prediction of social networks based on weighted proximity measures. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, (pp. 85–88). Washington, DC: IEEE Computer Society.

Newman, M.E. (2001a). Clustering and preferential attachment in growing networks. Physical Review E, 64(2), 025102.

Newman, M.E. (2001b). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E, 64 (1), 016132.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks, 32(3), 245–251.

Otte, E., & Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information sciences. Journal of Information Science, 28(6), 441–453.

Price, D.J. de Solla. (1976). A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science, 27(5), 292–306.

Rodriguez, M.A., & Neubauer, P. (2010). Constructions from dots and lines. Bulletin of the American Society for Information Science and Technology, 36(6), 35–41.

Salton, G., & McGill, M.J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Song, H.H., Cho, T.W., Dave, V., Zhang, Y., & Qiu, L. (2009). Scalable proximity estimation and link prediction in online social networks. In IMC 2009: Proceedings of the 9th ACM Internet Measurement Conference (pp. 322–335). New York: ACM.

Van Rijsbergen, C.J. (1979). Information retrieval (Second ed.). Glasgow: Department of Computer Science, University of Glasgow.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications. Cambridge: University Press.

Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. Nature, 393(6684), 440–442.

Zhu, B.Y., & Xia, Y.X. (2016). Link prediction in weighted networks: A weighted mutual information model. PLoS ONE, 11(2), e0148265.