Bipartite networks for link prediction: Can they improve prediction performance?

Raf Guns[1]

[1] *raf.guns@ua.ac.be*
University of Antwerp, IBW, Venusstraat 35, B-2000 Antwerpen, Belgium

Abstract

Link prediction is concerned with the question to what extent it is possible to predict which links will exist in a future snapshot of an evolving network. Several measures, called predictors, have been proposed to achieve more accurate predictions. However, these are typically based on unipartite networks, even though many of the examined networks have an underlying bipartite structure. In this contribution, we examine whether predictions based on the original bipartite networks are able to improve prediction accuracy.

We present bipartite versions of several neighbor-based predictors, the Katz predictor and the Rooted PageRank predictor, and discuss their strong and weak points. Both the unipartite and bipartite predictors are then tested on two co-authorship networks. The results show that some bipartite predictors form a considerable improvement to their unipartite counterparts. We also note remarkable differences between the two case studies and discuss possible reasons for this disparity.

Introduction

Social network analysis (SNA) is the research field that tries to describe and explain social structures by means of network theory. SNA researchers have devised many techniques and measures, which have since been 'imported' into other fields of research. During the last decade, bibliometricians and informetricians have shown a growing interest in applying SNA techniques (e.g., Liu et al., 2005; Otte & Rousseau, 2002; Yan & Ding, 2009). This is quite natural; it is well-known that many interactions studied in informetrics can be represented as networks, such as citation networks, collaboration networks, web link networks etc. Mathematically, a network $G = (V, E)$ consists of a set of vertices or nodes V and a set of edges or links $E \subseteq V \times V$, which connect nodes. We will refer to connected nodes as neighbors. A node's neighborhood is the set of all nodes connected to it. In a weighted network, each link has a weight, indicating the link strength.

Contrary to random networks (Erdős & Renyi, 1959), social and informetric networks evolve according to certain regularities. The phenomenon of preferential attachment (Barabási & Albert, 1999) is also well-known: new nodes are more likely to link to existing nodes with many neighbors than to existing ones with few neighbors. Morris (2005) presents an interesting growth model of papers and their references, which is based on similar principles. In other words, there are features present in a network which make certain links more likely than others.

Link prediction

Liben-Nowell and Kleinberg (2003) were among the first to examine the following question: to what extent is it possible to predict which links will appear in the future? More specifically, given a snapshot of an evolving social network, how can one predict which new links will appear in some future snapshot of the same network? Liben-Nowell and Kleinberg (2003, 2007) refer to the question as the link-prediction problem. They call the current snapshot the training network and the future snapshot – if it is known – the test network.

Normally, one tries to capture the probability of a link between two given nodes using some measure or indicator, such as the number of neighbors they have in common. Such measures are referred to as *predictors*. Formally, a predictor p is a function p: $V \times V \rightarrow \mathbf{R}$, that maps a pair of nodes to a real-valued likelihood score β, which is usually rescaled to the interval

[0, 1]. The score β expresses the likelihood that a link between the two nodes will exist in the predicted network. This leads to a set of node pairs with a corresponding likelihood score. If one ranks these in decreasing order of β and cuts off the list at some threshold, one obtains a set of 'most likely' links, which together comprise the predicted network. The predictor thus indirectly specifies what the predicted network looks like. Since the predictor only assigns a score to node pairs from the training network, the problem of new nodes (and links to/from them) is outside the scope of link prediction.

In link prediction research, the test network is known; hence, one can compare the training network with the test network. Predictor performance can then be measured using measures such as precision and recall, the AUC measure (Bradley, 1997) etc.

Link prediction forms the basis of several applications, some of which have already been explored in the literature. A principal application is *recommendation*, which hinges on finding related but unlinked items. Prediction of links in an incomplete network leads the way to another application, viz. *detection of missing information*. It is a well-known problem that the data in citation databases like Web of Science is incomplete and sometimes erroneous, affecting the results of scientometric indicators (Jacsó, 2009). Link prediction can hence be used as a tool in detecting missing (or spurious) links in the database.

At a theoretical level, the main promise of link prediction is as a practical way of testing and *evaluating network formation and evolution models*. Predictors normally derive from an explicit or implicit hypothesis of how and why links arise in a network. The performance of a predictor therefore also may help to test the validity of the underlying hypothesis.

Bipartite networks

Bipartite networks are defined as follows. A bipartite network $G = (T, B, E)$ consists of two disjoint sets of nodes T and B, and $E \subseteq T \times B$ is the set of links. We call T the set of top nodes and B the set of bottom nodes. Bipartite networks are different from 'regular' (unipartite) networks in that links can only exist between top nodes and bottom nodes. Mathematically, they can be considered a subclass of networks in general (taking $V = T \cup B$). In practice, it is usually the case that top nodes and bottom nodes represent different kinds of entities (e.g., articles and keywords, authors and journals, authors and publications…).

Every bipartite network has a corresponding weighted unipartite network. Given the bipartite network $G = (T, B, E)$, its corresponding unipartite (weighted) network is $G' = (B, E', w)$. If $N(a)$ denotes the neighborhood of node a, then E' is the set of all bottom node pairs $\{u, v\}(u, v \in B)$ for which $N(u) \cap N(v) \neq \emptyset$. The weighting function $w : B \times B \to \mathbf{N}$ then is

$$w(u, v) = |N(u) \cap N(v)| \tag{1}$$

For every bipartite network, there is exactly one corresponding unipartite network using this procedure (but note that switching top and bottom nodes yields another with a different node set). This is not true the other way around: a weighted unipartite network may correspond to many bipartite networks. Consider the example in Figure 11, – the weighted unipartite network on the right corresponds to the two bipartite networks on the left.

In many aspects of informetrics, the distinction between asymmetrical occurrence matrices and symmetrical co-occurrence matrices is crucial (Leydesdorff & Vaughan, 2006; Schneider & Borlund, 2007). The relation between a bipartite network and its corresponding unipartite network is, in essence, the same as that between an occurrence matrix and its corresponding co-occurrence matrix. Indeed, the adjacency matrix of a bipartite network is of the form

$$A = \begin{pmatrix} [0] & X \\ X^T & [0] \end{pmatrix}$$

It consists of four submatrices, two of which are square matrices that only contain zeros (represented here as [0]). It is clear that all information is actually contained in the submatrix X (or its transpose), which can be interpreted as an occurrence matrix. The adjacency matrix of the corresponding unipartite network is the corresponding co-occurrence matrix. This implies that there is a direct relation between bipartite and unipartite networks on the one hand and occurrence and co-occurrence matrices on the other.

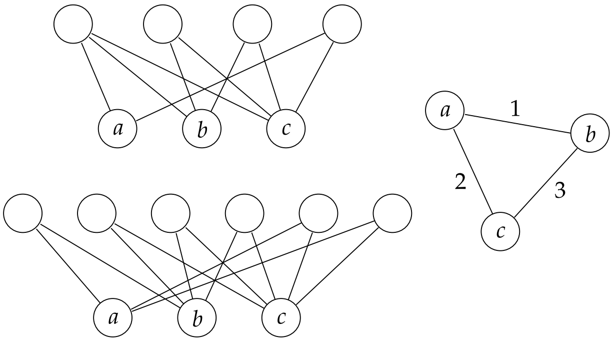

Figure 11. A unipartite network corresponds to many bipartite networks

An important kind of informetric network is the collaboration network, where nodes represent researchers, connected by collaboration links. Most often, the collaboration network is 'approximated' by a co-authorship network (e.g., Liu et al., 2005), – we will do the same and treat both terms as synonyms. Collaboration networks are very popular as case studies in link prediction research. There exist several good reasons for this: the results are of interest to a wide variety of researchers, it is relatively easy to gather enough data, and the networks are undirected. This is an advantage, since directed networks complicate things; in a directed network, one would also have to predict the link direction. Like many networks in informetrics, the collaboration network is in fact a 'derived' network, in that it corresponds to an underlying bipartite network. In this case, the bipartite network consist of publications as top nodes, authors as bottom nodes, and authorships as links. This bipartite network is unweighted.

Let us assume that Figure 11 represents a co-authorship network on the right and two corresponding author–publication networks on the left. In the upper bipartite network, *a* and *c* have co-authored two publications, one of which also has *b* as a co-author. In the lower bipartite network, *a* and *c* again have co-authored two publications, but this time, there are no co-authors. This may indicate that *a* and *c* have worked more closely together in the latter case. This is especially clear when one compares papers written by tens or even hundreds of authors – 'mega-authorship' (Kretschmer & Rousseau, 2001) or 'hyperauthorship' (Cronin, 2001) – with those written by two or three authors. It seems extremely unlikely that the degree of close collaboration in the former case is as high as in the latter case.

Since the bipartite network contains more information, it seems plausible that link prediction on the basis of a bipartite network can be more accurate. Indeed, typically link prediction tries to gauge how 'close' two potential collaborators are, – the extra information contained in the bipartite network may help to make better assessments in this regard. The research question we consider here is twofold: (a) can one make more accurate predictions on the basis of the bipartite network compared to the corresponding unipartite network, and (b) which indicators are best suited to exploit the specific features of the bipartite network? Note that we will only consider the case where the training network is a bipartite author–publication network and the test network and predicted network are unipartite author collaboration networks.

Predictors

We first present an overview of some well-known predictors used for unipartite networks (Clauset, Moore & Newman, 2008; Guns, 2009; Huang, Li & Chen, 2005; Liben-Nowell & Kleinberg, 2007). This overview is not intended to be exhaustive, but it does present members of the main 'families' of predictors. Some predictors with known bad performance, like the Preferential Attachment predictor (Guns, 2009; Liben-Nowell & Kleinberg, 2007), are also excluded. Subsequently, we discuss how these predictors may be applied to bipartite networks.

Social networks tend to have high clustering, in other words, two nodes that are connected to a third node have a high probability of being connected themselves. The Common Neighbors predictor is defined as the size of the intersection of their neighborhoods:

$$p(u, v) = |N(u) \cap N(v)| \tag{2}$$

Common Neighbors is however sensitive to size; if *u* and *v* have many neighbors, they are automatically more likely to have more neighbors in common. Therefore, several variants of Common Neighbors have been devised. Many of these variants have been thoroughly studied in information science, especially in Information Retrieval and science mapping. A well-known example is the Cosine measure (Salton & McGill, 1983):

$$\mathrm{Cos} = p(u, v) = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u) \cdot N(v)|}} \tag{3}$$

Other variants include the Jaccard index, the Overlap and Dice's measure. We note that these variants have both a set-theoretic and a vector-based interpretation (Egghe & Michel, 2002), both of which can be used for link prediction.

Some variants of Common Neighbors have originated and been studied mainly outside information science, such as the Adamic/Adar measure (Adamic & Adar, 2003):

$$p(u, v) = \sum_{z \in N(u) \cap N(v)} \frac{1}{\log |N(z)|} \tag{4}$$

or the Resource Allocation index (Zhou, Lü & Zhang, 2009):

$$p(u, v) = \sum_{z \in N(u) \cap N(v)} \frac{1}{|N(z)|} \tag{5}$$

Although their ranking is not the same, they are quite similar in form and 'function': rare (low-degree) neighbors of *u* and *v* have more weight than high-degree ones.

So far, we have not looked beyond the neighbors of two given nodes. A step further is taking into account the path(s) between them. Let *d*(*u*, *v*) denote the length of the shortest path from *u* to *v*. The Graph Distance predictor is then defined as:

$$p(u, v) = \frac{1}{d(u, v)} \tag{6}$$

A more complex measure that takes *all* paths up to a given length (not just shortest paths) between two nodes into account was proposed by Katz (1953). Although originally intended as an indicator of centrality, Katz's method can also be used to obtain a measure of relatedness between two nodes and hence as a predictor:

$$p(v_i, v_j) = \sum_{k=1}^{\infty} \lambda^k a_{ij}^{(k)} \tag{7}$$

where λ ($0 < \lambda < 1$) is a parameter that expresses the “probability of effectiveness of a single link” and $a_{ij}^{(k)}$ is the element corresponding to nodes v_i and v_j in the k-th power of the adjacency matrix, i.e. the number of paths with length k from v_i to v_j.

Finally, we mention the Rooted PageRank predictor, a variation of PageRank (Brin & Page, 1998; Langville & Meyer, 2005). The PageRank score of a node can be intuitively understood as the probability that a random walker visits a given node at a specific point in time. At each step, the random walker either moves to a neighbor of the current node (probability α) or ‘teleports’ to a random node in the network (probability $1 - \alpha$). PageRank is the largest eigenvector of the PageRank matrix $\mathbf{G}_{PR}$:

$$\mathbf{G}_{PR} = \alpha\mathbf{G} + (1 - \alpha)\mathbf{E} \tag{8}$$

where $\mathbf{G}$ is the stochastic matrix for the adjacency matrix of the network and $\mathbf{E}$ is a $n \times n$ matrix, with each element equal to $1 / n$. If $\mathbf{e}$ is a column vector of ones, then $\mathbf{E} = \mathbf{ee}^{\mathrm{T}} / n$.

Rooted PageRank simply uses the inherent possibility of personalization of PageRank (Brin & Page, 1998) and can also be explained from the perspective of a random walker. The random walker starts at a fixed root node v_k. At each step, it either moves to a neighbor of the current node (probability α) or teleports back to v_k (probability $1 - \alpha$). Rooted PageRank is then the largest eigenvector of the Rooted PageRank matrix $\mathbf{G}_{RPR}$:

$$\mathbf{G}_{RPR} = \alpha\mathbf{G} + (1 - \alpha)\mathbf{eg}^{T} \tag{9}$$

where

$$\mathbf{g}_m = \begin{cases} 1 & \text{if } m = k; \\ 0 & \text{otherwise.} \end{cases}$$

Typically, the node most related to v_i is v_i itself, followed by its neighbors, and so on.

How can these predictors be translated to bipartite networks? For the neighbor-based predictors, this is fairly straightforward. If u and v are bottom nodes (authors), then 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, all of them simply yield the training network as a result (if the threshold is low enough). In other word, applying equations (2) to (5) 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. This does not mean that these predictors are the same, since their ranking is different. The Common Neighbors predictor ranks in decreasing order of number of joint publications. Cosine is similar, but normalizes with respect to the total number of publications of the authors. The Adamic/Adar and Resource Allocation predictors favor joint publications with no or few other co-authors.

It would be possible to apply the Graph Distance predictor (6) to a bipartite network with bottom nodes u and v, but the resultant ranking would be the same as for the unipartite network: the shortest path length for any bottom node pair is simply doubled in the bipartite case. For this reason, we will not consider Graph Distance in the analysis.

Contrary to Graph Distance, the unipartite and bipartite Katz predictor can result in different rankings. Compared to the neighbor-based predictors, Katz has the advantage that it is not restricted to neighboring nodes (or neighbors of neighbors). However, the bipartite variant has a theoretical downside, which can be explained through the example in Figure 2. Assume that the top nodes represent publications and the bottom nodes represent authors. In the left case, authors a and b have both collaborated with a single third author, whereas in the right case, they were co-authors in a larger team. Nevertheless, because the left case contains one path of length 4 from *a* to *b* and the right one contains 4 paths of length 4, the Katz predictor would rank the right case higher.

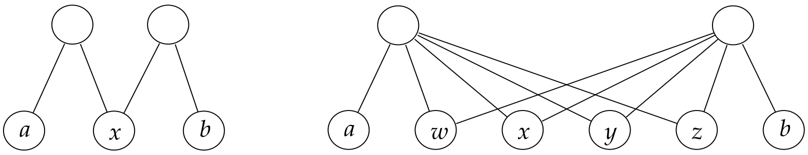

Figure 2. Hypothetical example: two cases of author–publication networks

Finally, 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 and its many variants – 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 would rank the right case of Figure 2 lower than the left one. As explained earlier, we consider this an attractive property.

Data and methodology

We have tested the predictors described in the previous section on two case studies of collaboration. The first case consists of co-authored publications at a single institution (University of Antwerp, Belgium) across all 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 training network is based on the period 2001–2003 and the test network is based on 2004–2006. The second case consists of co-authored publications in the field of informetrics during the period 1990–2009. This case study 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. We will refer to the first one as the 'UA' case study and to the second one as the 'Informetrics' case study.

In both cases, we have a bipartite training network (authors–publications) and a unipartite one (authors as nodes). The test network is a unipartite network. Table 1 contains basic data about both case studies. We restrict the networks to those nodes that are common to both training and test network; hence the number of authors in the training networks and the node number of the test network is the same.

Table 1. Descriptive statistics of the case studies

Network		*UA*	*Informetrics*
Unipartite training network	Number of nodes	1102	171
	Number of links	3212	178
	Average clustering coefficient	0.418	0.515
Test network	Number of nodes	1102	171
	Number of links	3550	188
	Average clustering coefficient	0.385	0.510
Bipartite training network	Number of authors	1102	171
	Number of papers	8671	862
	Number of links	13088	929
	Average number of papers per	11.877	5.433

	author		
	Average number of authors per paper	1.509	1.078

Each predictor makes a number of false and correct predictions, resulting in four groups: true positives (correct predictions), false positives (predictions which are not in the test network), false negatives (unpredicted links in the test network), and true negatives (no link in predicted or test network). We use recall and precision to evaluate a predictor's performance. Recall is defined as the number of correct predictions divided by the number of links in the test network. Precision is defined as the number of correct predictions divided by the number of links in the predicted network. Since each predicted link has a likelihood score, we can construct recall–precision charts (with predicted links in decreasing order of likelihood).

Of course, one can also choose a threshold and simply use the predictions above that threshold. In fact, this is probably more realistic: in most cases, one is mainly interested in a small group of high-quality predictions, rather than a broad swath of low-quality predictions. We therefore also look at the 'Precision at *k*' measure (Manning, Raghavan & Schütze, 2008); specifically, we record the precision of the first $k = 20$ predictions for Informetrics and the first $k = 60$ for UA. We take *k* larger for UA, because UA as a whole is a much larger data set (Table 1).

Results

In this section, we discuss the results of applying the unipartite and bipartite predictors, using recall–precision charts. Table 2 contains the Precision at *k* scores for all predictors for both case studies. Bipartite predictors in Table 2 are marked with the letters 'BP'.

First, we look at the results for the neighbor-based predictors. Figure 3 contains the recall–precision charts for (the unipartite and bipartite versions of) the Common Neighbors and Cosine predictors. The plots of other neighbor-based predictors like Jaccard and Overlap (not shown) are very similar to the Cosine plot. It is interesting to note that in both cases the simple Common Neighbors predictor outperforms the more sophisticated Cosine variant. This is consistent with the findings in older research (Liben-Nowell & Kleinberg, 2007). This is even clearer when considering Precision at *k*: in both cases, Cosine does not only worse than Common Neighbors but is in fact the least precise of all examined predictors (see Table 2).

Turning to the bipartite predictors, one can see some remarkable differences between both case studies. In the UA case study, both bipartite predictors achieve much higher accuracy than the unipartite ones (but keep in mind that the bipartite predictors only 'predict' links that were already present in the training network). Common Neighbors again outperforms Cosine, – at least for the highest ranked predictions, which are the most important. This is reflected in a higher Precision at *k* for Common Neighbors. The Informetrics case study on the other hand paints a rather different picture. Here, bipartite Common Neighbors does worse than regular Common Neighbors. Moreover, in this case bipartite Cosine clearly outperforms all other predictors considered, both in the recall–precision chart (Figure 3) and for Precision at *k* (Table 2). The relatively poor performance of bipartite Common Neighbors is due to the fact that in this data set the number of joint papers is not very discriminating, since in virtually all cases it is quite low. Indeed, 91% of all author collaborations in the training network consists of three or less joint publications (UA: 76.3%), and over all collaborations we find only 9 different numbers of joint papers in the training network (UA: 35). We observe that in such cases, normalization by (for instance) the cosine measure – thereby correcting for the total number of publications of the authors and increasing the predictor's discriminatory value – can considerably improve accuracy. This observation has, to the best of our knowledge, not been made previously in the literature.

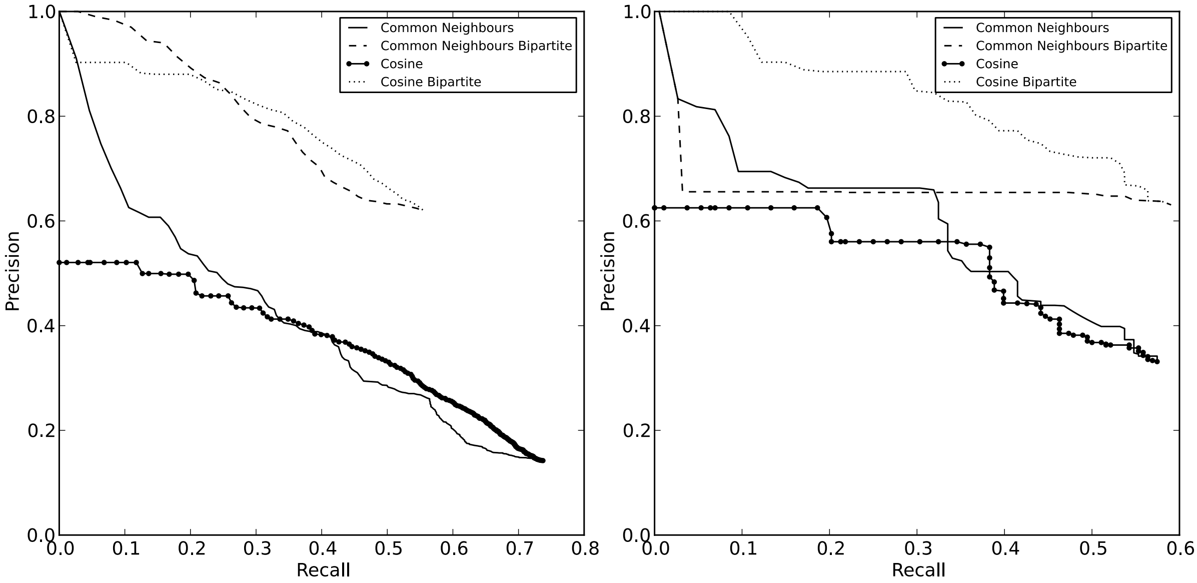

Figure 3. Recall–precision of unipartite and bipartite Common Neighbors and Cosine predictors (left: UA, right: Informetrics)

Table 2. Precision at *k* for all predictors (UA: *k* = 60; Informetrics: *k* = 20)

Predictor	*UA*	*Informetrics*
Common Neighbors	0.952381	0.761905
Common Neighbors BP	1	0.657895
Cosine	0.562963	0.625
Cosine BP	0.918367	0.956522
Adamic/Adar	0.983607	0.863636
Adamic/Adar BP	1	0.65
Resource Allocation	0.9625	0.926829
Resource Allocation BP	0.99115	0.65
Katz ($\lambda = 0.001$)	1	0.680672
Katz BP ($\lambda = 0.05$)	1	0.648148
Rooted PageRank ($\alpha = 0.85$)	0.754098	0.772727
Rooted PageRank BP ($\alpha = 0.85$)	0.915094	0.757576
Rooted PageRank BP ($\alpha = 0.2$)	0.933333	0.818182

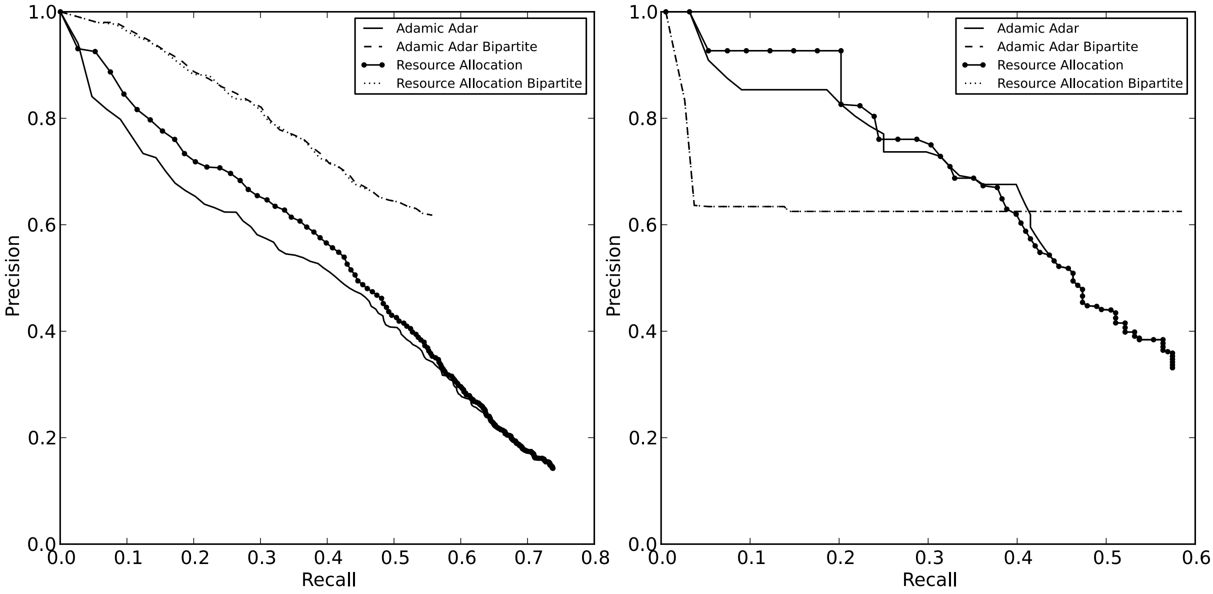

Figure 4. Recall–precision of unipartite and bipartite Adamic/Adar and Resource Allocation predictors (left: UA, right: Informetrics)

We now turn to the Adamic/Adar and Resource Allocation predictors, in essence a subgroup of neighbor-based predictors that gives preference to ‘rare’ common traits (co-authors or publications). Their results are shown in Figure 4. For the unipartite predictors we find that Resource Allocation has overall slightly better performance than Adamic/Adar, because Resource Allocation ‘punishes’ harder for common neighbors with high degree centrality (Zhou, Lü & Zhang, 2009). Note, however, that the UA case study reports a somewhat lower Precision at *k* for Resource Allocation than for Adamic/Adar.

The bipartite variants of Adamic/Adar and Resource Allocation yield virtually identical rankings and hence almost identical plots. Again we can see a remarkable difference between the two case studies: whereas the bipartite variants perform better on UA, they perform much worse on Informetrics. It is no coincidence that these plots strongly resemble the ones for bipartite Common Neighbors in Figure 3: just like bipartite Common Neighbors, they suffer from the fact that the number of joint papers is not discriminating enough for this case study. Accounting for the ‘rarity’ of these papers even results in a slightly worse Precision at *k*.

The Katz predictor involves an extra parameter λ. Through experimentation, we found that the optimal (i.e., resulting in the most accurate predictions) values for λ are 0.001 in the unipartite case and 0.05 in the bipartite case. These optimal values are remarkably similar for both case studies. The value for the unipartite case is much lower, essentially giving very low weight to paths of length 2 or more (cf. eq. (7)). Since the shortest possible path between two authors in the bipartite case has length 2, the optimal value for λ is naturally higher there.

The results in Figure 5 and Table 2 are based on the optimal values for λ. For both case studies, the unipartite and bipartite variants result in fairly similar plots. For UA, both Katz variants achieve a perfect Precision at *k*, whereas the results for Informetrics are relatively poor (slightly worse for the bipartite variant). Because the UA case study contains a sizable portion of publications with many authors (the largest number of authors for a single publication is 69), we expected to find relatively worse performance for bipartite Katz there. The results did, however, not confirm this expectation.

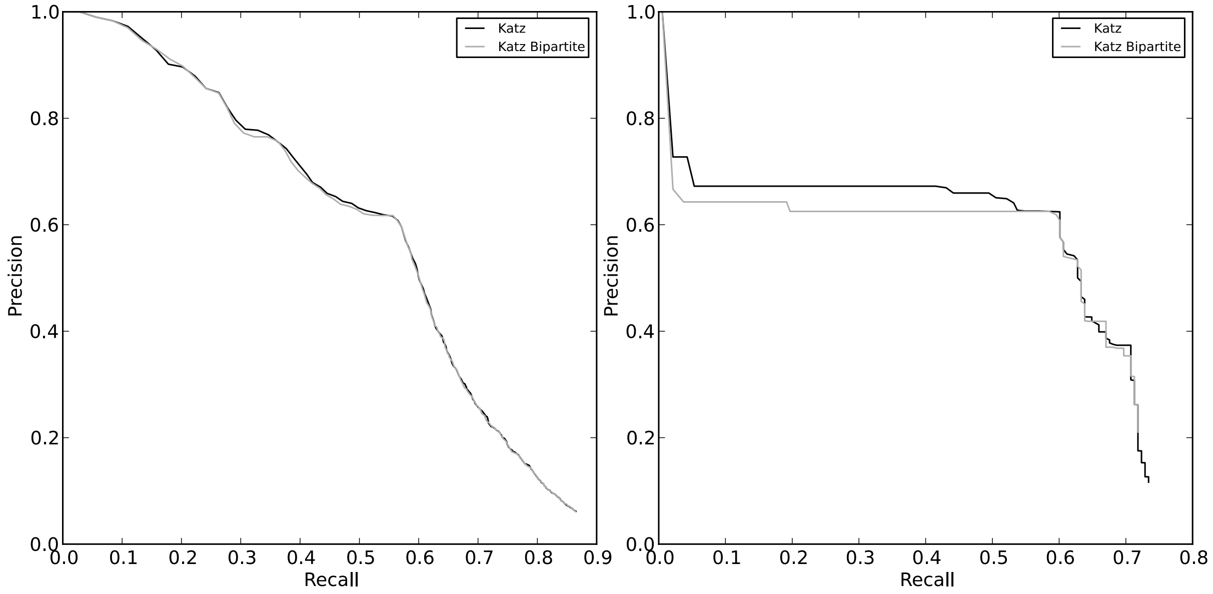

Figure 5. Recall–precision of unipartite ($\lambda = 0.001$) and bipartite ($\lambda = 0.05$) Katz predictors (left: UA, right: Informetrics)

Finally, we consider the Rooted PageRank predictor. Figure 6 shows the results for Rooted PageRank with α = 0.85. In both case studies, there is an obvious advantage to the bipartite predictor, although the difference is greater for UA than for Informetrics. Indeed, in the latter case we see that the first part of the bipartite Rooted PageRank plot (up to about 0.4 on the horizontal axis) is a bit similar to the bipartite Common Neighbors plot in Figure 3: an initial sharp decrease and then a steady part around 0.7 precision (although the precision of Rooted PageRank is a bit better than Common Neighbors'). This caused us to wonder whether other α values would yield better predictions. It turns out that the optimal value for α in the unipartite case lies around 0.85, whereas its optimal value in the bipartite case lies around 0.2 (cf. the last row of Table 2).

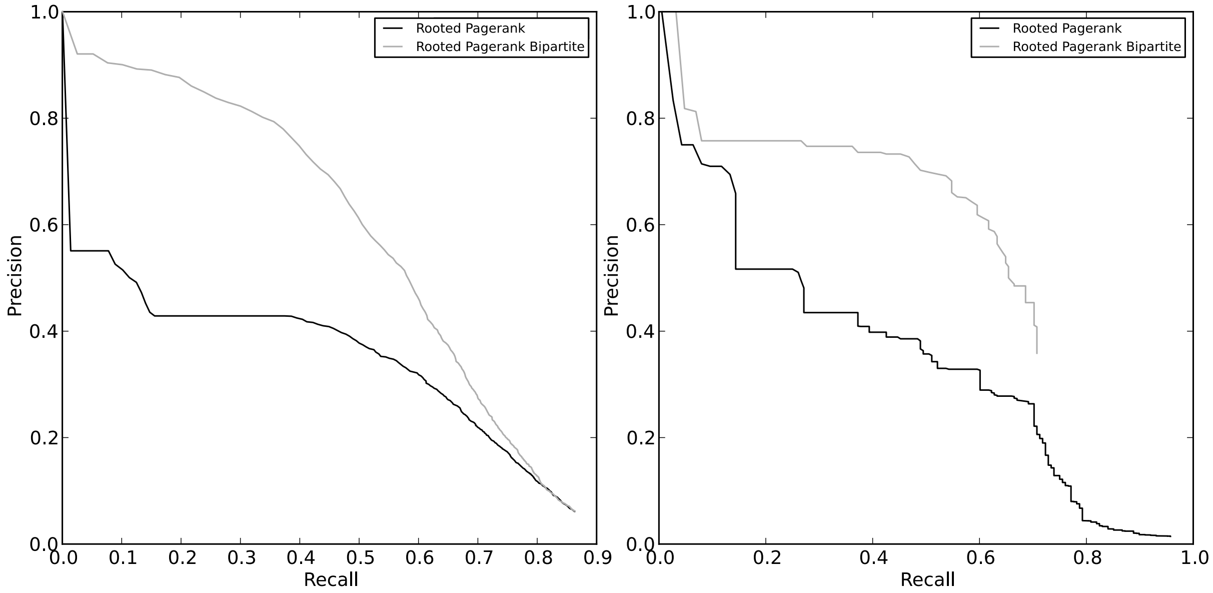

Figure 6. Recall–precision of unipartite and bipartite Rooted PageRank predictors (left: UA, right: Informetrics)

Conclusions

Networks are central to many aspects of informetric research. It is therefore important to understand how and why such networks evolve. Link prediction may help to understand the factors that influence link formation or disappearance.

Many informetric networks are derived from a bipartite network: this is the case for collaboration (co-authorship) networks, but also co-citation, bibliographic coupling and co-word networks. In this paper, we have examined the question whether link prediction on the basis of the original bipartite network can improve prediction performance and which indicators are best suited for prediction on the basis of a bipartite training network.

The results indicate that bipartite link prediction can indeed improve prediction performance, although much depends on the predictor used. Some predictors, like Rooted PageRank, benefit significantly from the bipartite structure. For others, like Katz, there is hardly no difference between the unipartite and the bipartite variant.

Some of the bipartite neighbor-based predictors achieve very good performance, but one should keep in mind that they are limited to ‘predicting’ which links from the training network are likely to sustain. Contrary to the accepted view, we have found that the Cosine predictor can in some cases significantly outperform Common Neighbors. Note that the results of bipartite Common Neighbors can be obtained from the unipartite training network, whereas the remarkable improvement of bipartite Cosine is only possible on the basis of the bipartite network.

Some predictors achieve very different results, depending on the case study. This is often glossed over, but in fact an important point. Indeed, it is perhaps logical that networks with different characteristics also call for different predictors, bipartite or otherwise. This is an important direction for future research: which predictor is best suited for which kind of network?

References

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

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

Bradley, A. P.(1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30, 1145–1159.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7), 107–117.

Clauset, A., Moore, C. & Newman, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191), 98–101.

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

Erdős, P. & Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae*, 6(26), 290–297.

Guns, R. (2009). Generalizing link prediction: Collaboration at the University of Antwerp as a case study. In Grove, A (ed.), *ASIST 2009. Proceedings of the 72nd ASIS&T Annual Meeting*, vol. 46, Silver Spring, Md. ASIS&T.

Huang, Z., Li, X., & Chen, H. (2005). Link prediction approach to collaborative filtering. In *Proc. of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL ’05)*, 141–142. NY: ACM Press.

Jacsó, P. (2009). Errors of omission and their implications for computing scientometric measures in evaluating the publishing productivity and impact of countries. *Online Information Review*, 33(2), 376–385.

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

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.

Langville, A. N. & Meyer, C. D. (2005). A survey of eigenvector methods for web information retrieval. *SIAM Review*, 47(1), 135–161.

Leydesdorff, L. & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. *Journal of the American Society for Information Science and Technology*, 57(12), 1616–1628.

Liben-Nowell, D. & Kleinberg, J. (2003). The link-prediction problem for social networks. In *Proc. of the 12th International Conference on Information and Knowledge Management (CIKM)*, 556–559.

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.

Liu, X., Bollen, J., Nelson, M.L. & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing & Management*, 41(6), 1462–1480.

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: University Press.

Morris, S. A. (2005). Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology*, 56(12), 1250–1273.

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

Salton, G. & McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. NY: McGraw-Hill.

Schneider, J. W. & Borlund, P. (2007). Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal of the American Society for Information Science and Technology*, 58(11), 1586–1595.

Yan, E. & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, 60(10), 2107–2118.

Zhou, T., Lü, L. & Zhang, Y.-C. (2009). Predicting missing links via local information. *The European Physical Journal B - Condensed Matter and Complex Systems*, 71(4), 623–630.