Predicting and recommending collaborations : an author-, institution-, and country-level analysis

This is a postprint of an article published in *Journal of Informetrics*. Please cite as follows:

> Yan, E., & Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics*, 8(2), 295-309. http://dx.doi.org/10.1016/j.joi.2014.01.008

Predicting and recommending collaborations: An author-, institution-, and country-level analysis

Erjia Yan[1]

College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA. Email: erjia.yan@drexel.edu

Raf Guns

University of Antwerp, Institute of Education and Information Sciences, IBW, Venusstraat 35, 2000 Antwerpen, Belgium. Email: raf.guns@uantwerpen.be

Abstract

This study examines collaboration dynamics with the goal to predict and recommend collaborations starting from the current topology. Author-, institution-, and country-level collaboration networks are constructed using a ten-year data set on library and information science publications. Different statistical approaches are applied to these collaboration networks. The study shows that, for the employed data set in particular, higher-level collaboration networks (i.e., country-level collaboration networks) tend to yield more accurate prediction outcomes than lower-level ones (i.e., institution- and author-level collaboration networks). Based on the recommended collaborations of the data set, this study finds that neighbor-information-based approaches are more clustered on a 2-D multidimensional scaling map than topology-based ones. Limitations of the applied approaches on sparse collaboration networks are also discussed.

[1] Corresponding author

Introduction

Social networks have the propensity to evolve over time. Every second, new friendships are established and old friendships are updated in Facebook connections, new collaborations are formed and populated in academic databases, and Twitter follower-following relationships are constantly subject to changes and updates. To capture such evolving features, earlier studies mainly employed a macro-perspective to model network growths and simulate network behaviors (e.g., Albert & Barabási, 2000; Jeong, Néda, & Barabási, 2003; Barrat, Barthélemy, & Vespignani, 2004; Sakaki, Okazaki, & Matsuo, 2010). Later on, studies that focused on individuals' growth patterns and behaviors in social networks were also introduced (e.g., Kretschmer, 2004; Liu et al., 2005; Yan & Ding, 2009). These micro-level analyses have complemented the scholarship of social network analysis. They are specialized in examining individuals' power, stratification, ranking, and inequality in various sociological settings (Wasserman & Faust, 1994).

Micro-level analyses have been one of the foci in informetric research. Studies in this field have typically employed authors, research communities, and institutions as the unit of analysis. Informetric studies have applied various indicators to collaboration networks. These studies have revealed the most "central" authors through centrality measures (e.g., Liu et al., 2005; Yin et al., 2006; Fiala, Rousselot, & Ježek, 2008; Yan & Ding, 2011), identified factors that are associated with collaboration and citation (e.g., Yan & Sugimoto, 2011), and examined the relationship between geographic location and collaboration (e.g., Ponds, Van Oort, & Frenken, 2007). However, these studies mainly used static approaches, and consequently did not inform the dynamic characteristics of collaborations. The goal of this study is to fill this gap by probing into collaboration dynamics using a ten-year data set on library and information science publications.

Specifically, we aim to predict and recommend collaborations based on the structure of current collaboration networks. This topology-based prediction is also known as link prediction (Liben-Nowell & Kleinberg, 2007). Link prediction recommends collaborations purely based on the intrinsic collaboration topology. This method does not rely on any data concerning the complex social, cognitive, institutional, or geographical factors (e.g., Ponds, Van Oort, & Frenken, 2007; Yan & Sugimoto, 2011). These factors are indirectly accounted for, because they may influence the network topology through mechanisms like homophily or the Matthew effect.

The performance of link predictors determines the effectiveness of collaboration recommendations. In the past, various link predictors were proposed and applied (e.g., Liben-Nowell & Kleinberg, 2007; Sharan & Neville, 2008; Guns & Rousseau, 2013). These studies, however, focused largely on author collaborations. Consequently, we have limited understanding of collaboration dynamics of other major collaborative entities, such as institutions and countries. These collaborative entities should not be neglected. Rather, they should be systematically examined. They deliver unique perspectives to examine collaborations that author-level analysis may be inadequate to afford. For instance, institutions can be used as proxies to delve into

authors' collective collaboration behaviors (Hoekman, Frenken, & van Oort, 2009). Country-level collaboration analysis can provide "a tool for high-level scrutiny of the quality and quantity of the research enterprise" (Holton, 1978, p. 200). Both institution- and country-level analysis can signify spatial-temporal discoveries of knowledge production and innovation (Havermann, Heinz, & Kretschmer, 2006; Yan & Sugimoto, 2010).

This study is thus motivated to further our understanding of collaboration dynamics. It investigates collaboration prediction and recommendation at author-, institution-, and country-levels. Through the application of several link predictors, the following research questions are addressed:

- To what extent do different levels of aggregation, i.e. author-, institution-, or country-level, affect the performance of link predictors?
  Previous studies have mainly examined the dynamics of author collaborations. A systematic analysis of all three levels of collaboration (i.e., author, institution, and country) has not yet been carried out. To fill this gap, the current study conducts an integrated examination of collaboration dynamics at the levels of authors, institutions and countries using link prediction methods.

- Based on true/false positive and true/false negative statistics, what are the between-object distances of different link predictors?
  A set of evaluation methods (i.e., precision-recall, receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), and top *k* evaluation) is triangulated to ensure the highest level of validity. Specifically, the current study also uses multidimensional scaling to visualize the between-object distances among eight predictors on a two-dimensional map.

- Starting from past collaboration relations in library and information science, what new collaborations are most probable to establish at author-, institution-, and country-levels? And what approach can be used to integrate the recommended collaborations obtained from multiple link predictors?
  Previous efforts on link prediction mainly relied on one predictor to recommend collaborations (e.g., Guns, 2009, 2011). Nonetheless, different link predictors may capture different collaboration characteristics. We propose a straightforward approach to merge the collaboration recommendations obtained from different predictors.

Note that these research questions are addressed using a data set on library and information science publications. Thus, findings of this study do not necessarily generalize to other research fields. Nevertheless, this study should inform dynamic analyses of collaborations in general and assist scholars trying to discern collaboration characteristics at different collaboration levels. This study also contributes to micro-level informetric research by providing ways to assess individuals' collaboration potentials.

Literature review

Collaboration networks

Studies of collaboration networks have a long standing in information science. Collaboration networks furnish an important medium to examine scholarly communication (e.g., Logan & Shaw, 1991; Luukkonen, Persson, & Sivertsen, 1992). Although early studies of coauthorship networks helped information scientists gain an in-depth understanding of the socio-cognitive structure of several author communities, these studies were limited to a small scale and the employed approaches were largely constrained to descriptive statistics.

In the last decade, we have witnessed a new movement in network analysis. The focus has shifted to large-scale statistical properties of graphs (Newman, 2001a, 2001b). In particular, the discoveries of small-world (Watts & Strogatz, 1998) and scale-free (Barabási & Albert, 1999) properties have promoted studies of collaboration networks. Recently, collaboration networks have been used to evaluate various clustering techniques, such as modularity-based techniques (Newman & Girvan, 2004), Clique Percolation Method (Farkas, Ábel, Palla, & Vicsek, 2007), link communities (Ahn, Bagrow, & Lehmann, 2010), and community kernel (Wang, Lou, Tang, & Hopcroft, 2011). These meso-level techniques have reshaped the research landscape of scientific collaboration and have propelled its analysis toward a more granular level. They have provided insights into interdisciplinarity (e.g., Moody, 2004), teams of science (e.g., Wuchty, Jones, & Uzzi, 2007; Börner et al., 2010), and even human mobile communications (e.g., Blondel et al., 2008).

In addition to macro- and meso-level analyses, at the micro-level, the predominant theme of analysis is situated on identifying the "status" of authors in a given research community. The status typically represents authors' ability in forming research synergies. A set of standard centrality measures and variants were used to approximate author status in collaboration networks (e.g., Liu et al., 2005; Sidiropoulos & Manolopoulos, 2006; Fiala, Rousselot, & Ježek, 2008; Yan & Ding, 2009). Yet, approaches of these studies remained static and yielded only retrospective perspectives on the status of authors.

The link prediction problem

Simply put, link prediction addresses the following question: given the topology of a collaboration network at time *t*, what collaborations will be formed in a future time *t'* (Liben-Nowell & Kleinberg, 2007). The link prediction problem attempts to model the evolving mechanism of social networks through their intrinsic features (i.e., their topology). Studies have used a number of link predictors to predict collaborations in fields such as physics (Liben-Nowell & Kleinberg, 2007), computer science (Sharan & Neville, 2008), and malaria and tuberculosis research (Guns & Rousseau, 2013). Recent studies have generalized the link prediction method, allowing it to work with supervised (Lichtenwalter, Lussier, & Chawla, 2010),

directed (e.g., Shibata, Kajikawa, & Sakata, 2012; Guo, Yang, & Zhou, 2013), weighted (e.g., Lü & Zhou, 2010), and multi-relational data sets (e.g., Ströele, Zimbrão, & Souza, 2013).

Previous efforts on link prediction were largely confined to author collaborations. An investigation on multiple collaboration levels (i.e., authors, institutions and countries) is lacking. It is thus unclear from the previous literature how different collaboration levels affect prediction results. To address this, this paper uses an empirical data set and applies eight link predictors to all three collaboration levels: author-, institution-, and country-levels. Prediction results of the three levels are compared and contrasted using a set of triangulated evaluative methods.

Methods

Link predictors

To date, more than 30 algorithms have been used to address the link prediction problem (e.g., Liben-Nowell & Kleinberg, 2007; Sarkar, Chakrabarti, & Moore, 2010; Guns, 2011). Eight link predictors were selected for this study, due to their algorithmic transparency, ease of implementation, and marked performance (Guns, 2011; Liben-Nowell & Kleinberg, 2007; Lichtenwalter & Chawla, 2011). They can be grouped into two categories: predictors that consider only nodes' neighboring information, including Adamic/Adar (Adamic & Adar, 2003), Common Neighbors, Preferential Attachment, and Jaccard; and predictors that consider the whole topology of collaboration networks, including Katz (Katz, 1953), Rooted PageRank, Weighted Rooted PageRank, and SimRank. These eight algorithms are applied to a data set of library and information science publications.

The data set was divided into two slices based on the papers' years of publication. A two-step approach was adopted: first, all eight predictors were applied to the data set of the first time slice; and the data set of the second time slice was used to evaluate the performance of these predictors. In the second step, the two best performing predictors were applied to the whole data set to predict the most likely collaborations (among those who have not collaborated yet). The evaluation is not simply a comparison between two disconnected collaboration networks, but rather, we intend to verify whether the predicted collaborations based on the first network have been established *in addition to* existing ones.

The eight link prediction algorithms are introduced here, using the following notations provided in Liben-Nowell and Kleinberg (2007). $\Gamma(x)$ denotes the set of neighbors of x. $|A|$ is the cardinality of set A. An example of calculating these link predictors is illustrated in Figure 1.

Common Neighbors: $|\Gamma(x) \cap \Gamma(y)|$. The formula of common neighbors denotes the number of neighbors that node *x* and node *y* share.

Jaccard: $\frac{|\Gamma(x)\cap\Gamma(y)|}{|\Gamma(x)\cup\Gamma(y)|}$. Jaccard similarity is commonly used in information retrieval to measure the probability that two sets share certain features. In this case, the feature is all neighbors that node *x* and node *y* have.

Adamic/Adar: $\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\log|\Gamma(z)|}$. Adamic and Adar (2003) proposed a measure based on the rarity of traits shared by two Web users. Their hypothesis is that if two users share a very rare trait (e.g., an unusual hobby), it indicates a social connection with more confidence than a very common trait. In the context of link prediction, the Adamic/Adar measure further normalizes the calculation of common neighbors by giving higher weights to less usual common features between node *x* and node *y*. Suppose that *x* and *y* have two common neighbors *a* and *b*, where *a* has only two neighbors *x* and *y* but *b* has ten neighbors in addition to *x* and *y*; then *a* contributes more to the likelihood that *x* and *y* will have a connection in the future.

Preferential Attachment: $|\Gamma(x)|\cdot|\Gamma(y)|$. According to Barabási and colleagues (2002), the probability of collaboration of *x* and *y* is proportional to the product of the number of collaborators of x and y.

Katz: $\sum_{k=1}^{\infty}\beta^k\mathbf{A}_{ij}^{(k)}$, where $\mathbf{A}_{ij}^{(k)}$ is the element corresponding to nodes i and j in the k-th power of the adjacency matrix **A**, i.e., the number of walks with length k from i to j (an unweighted version is used in this study where the element equals 1 if *i* and *j* collaborates; for a weighted version, please see Guns & Rousseau, 2013). The parameter β ($0<\beta<1$) represents the effectiveness of a single link. Thus, each path with length k has a probability β^k of effectiveness. According to empirical results (Liben-Nowell & Kleinberg, 2007), choosing a small β (at the 0.001 level) will yield more accurate predictions. In this study, β is set at 0.001. For reasons of computational feasibility, we do not consider walks where $k > 10$.

Rooted PageRank is inspired by Google's PageRank. Like standard PageRank, rooted PageRank can be understood using the concept of a random Web surfer, who, at each time step, either moves to a node adjacent to the current one or teleports. However, unlike standard PageRank, rooted PageRank does not teleport the surfer to a random node; instead, teleportation always puts the surfer back at a fixed root node. The rooted PageRank value of a node is the stationary probability that the random surfer will be located at that node. As such, it represents a node's proximity to the root node. According to empirical results (Liben-Nowell & Kleinberg, 2007), choosing an α value between 0.1 and 0.5 yields more accurate predictions. In this study, α is set at 0.3. **Weighted Rooted PageRank** uses edge weights to calculate the transition probability from one node to another.

SimRank follows a recursive definition: two nodes are similar if they are connected by similar neighbors. It is defined as

$$\mathrm{Sim}(x,y) = c\frac{\sum_{\alpha\in\Gamma(x)}\sum_{b\in\Gamma(y)}\mathrm{Sim}(a,b)}{|\Gamma(x)|\cdot|\Gamma(y)|}$$

where $\mathrm{Sim}(x,x) = 1$ and c is a constant between 0 and 1. In this study it is set as 0.5. Like PageRank, SimRank can be understood in terms of random walks on a network. Jeh and Widom (2002) prove that the SimRank score $Sim(u,v)$ can be interpreted as the time before two random walkers meet on the network if they start at nodes u and v and randomly walk the network.

Figure 1 shows the calculation of eight link predictors for two sample collaboration networks. Our goal is to estimate the prospect of author *i* collaborating with author *j* as highlighted in Figure 1.

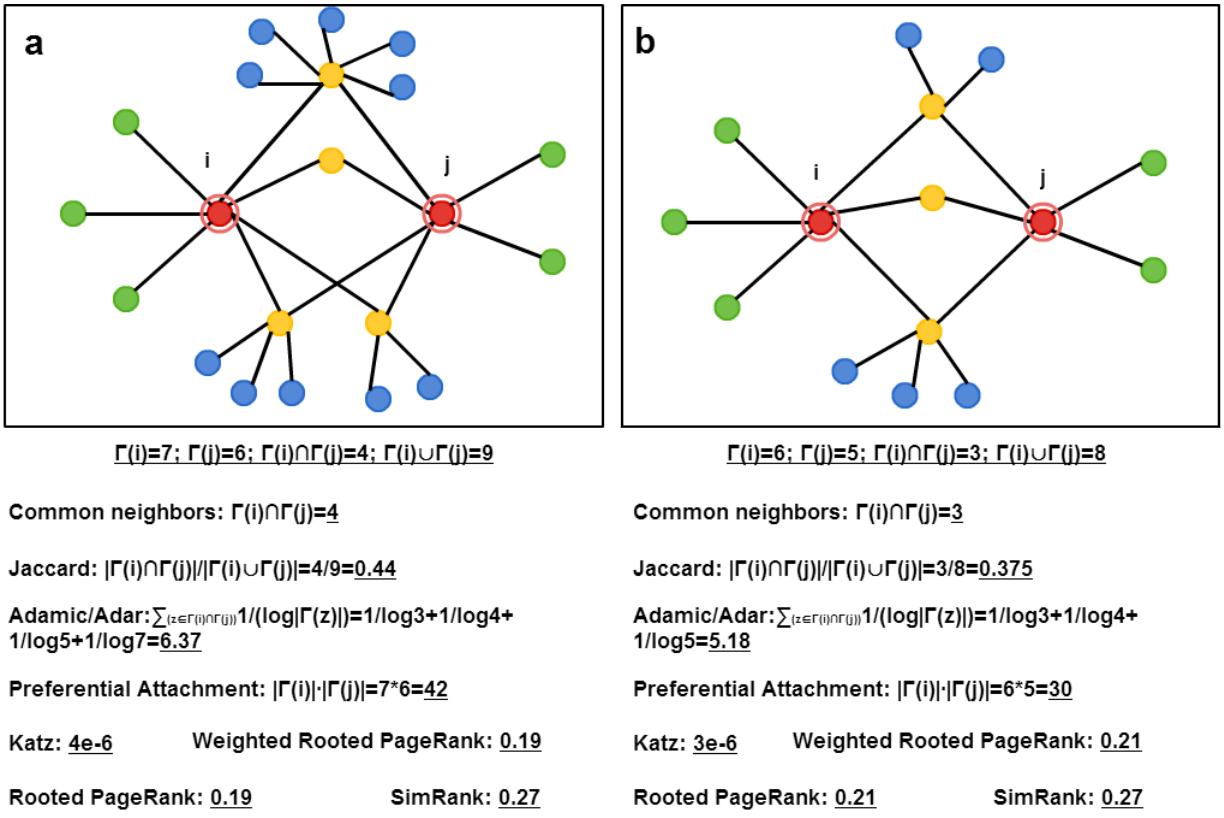

Figure 1. An example of applying link predictors

In Figure 1(a), author *i* has 7 coauthors and author *j* has 6 coauthors; thus, based on its topology, the intersection of *i* and *j* is 4 and the union of *i* and *j* is 9. For the four authors in the intersection, they have three, four, five, and seven coauthors respectively; thus, Γ(z) equals 3, 4, 5, and 7. Using such information, Common Neighbors, Jaccard, and Preferential Attachment can be obtained. The calculation of topology-based predictors is less upfront and thus only the final result is given (because the two networks are binary ones, Rooted PageRank and Weighted Rooted PageRank yield the same results). Results in Figure 1(a) and Figure 1(b) show that: while neighbor-information-based predictors produced more consistent results (i.e., author *i* and *j* in

Figure 1(a) are more likely to become coauthors than author *i* and *j* in Figure 1(b)), topology-based predictors generated less consistent outcomes. The inconsistencies in Figure 1 illustrate the marked differences between different predictors and the need for systematic comparison and evaluation.

To effectively evaluate the performance of link predictors, we first applied eight link predictors to three collaboration levels: author-, institution-, and country-levels. The employed evaluation methods include precision-recall, receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), and top *k* evaluation. The evaluation thus addresses the first research question. We then examined the relationship among link predictors through multidimensional scaling. This addresses the second research question. Last, we use the Borda count method (Aslam & Montague, 2001) to merge recommendation results obtained from different predictors. This addresses the third research question.

Data

The data set contains all publications from the 59 journals indexed in the 2008 version of the Journal Citation Reports in the Information Science & Library Science category. All document types published within these journals from January 2001 to February 2010 were downloaded for analysis. Name disambiguation for author names and institution names was implemented. Situations where authors have multiple affiliations were considered and these affiliations were treated equally. Please refer to Yan and Sugimoto (2011) for detailed information on the used name disambiguation method.

The dataset was then divided into two time slices based on papers' years of publication. Table 1 shows the sizes and densities of author collaboration networks, institution collaboration networks, and country/state collaboration networks (worldwide countries and U.S. states). Both author and institution collaboration networks are quite sparse, which suggests that authors from a given institution tend to collaborate with a limited number of authors in other institutions. Note that the densities of the author and institution collaboration networks decrease between the first and second period, whereas the density of the country collaboration network almost doubles.

Table 1. Size and density of collaboration networks

Time	Author collaboration networks			Institution collaboration networks			Country collaboration networks		
	Nodes	Links	Density	Nodes	Links	Density	Nodes	Links	Density
2001-2005	9,659	10,509	2.2e-4	3,010	530	1.2e-4	149	1,479	0.13
2006-2010	12,766	16,588	2.0e-4	3,783	785	1.1e-4	151	2,836	0.25
2001-2010	16,657	21,527	1.6e-4	4,836	1,223	1.0e-4	164	3,012	0.23

Results

Precision-recall graphs and ROC curves for the eight predictors

Two graphical evaluation methods are used to evaluate the eight predictors: the precision-recall graph and the receiver operating characteristic (ROC) curve which plots true positive rate (equal to recall) against false positive rates (the ratio of false positives versus the total number of non-collaborations). The two measures have previously been applied to predictor evaluation as well as a variety of information retrieval and machine learning algorithms (e.g., Bradley, 1997; Clauset, Moore, & Newman, 2008; Lichtenwalter, Lussier, & Chawla, 2010; Pencina, et al., 2008).

Figures 2 and 3 show the precision-recall graph and ROC curve of the eight predictors for author collaboration networks. For topology-based predictors, Katz has the best performance, followed by Weighted Rooted PageRank, Rooted PageRank, and SimRank. For neighbor-information-based predictors, Common Neighbor has the best performance based on precision and recall, followed by Preferential Attachment, Adamic/Adar, and Jaccard Coefficient; Preferential Attachment has the best performance based on the ROC curve, followed by Adamic/Adar, Common Neighbor, and Jaccard Coefficient.

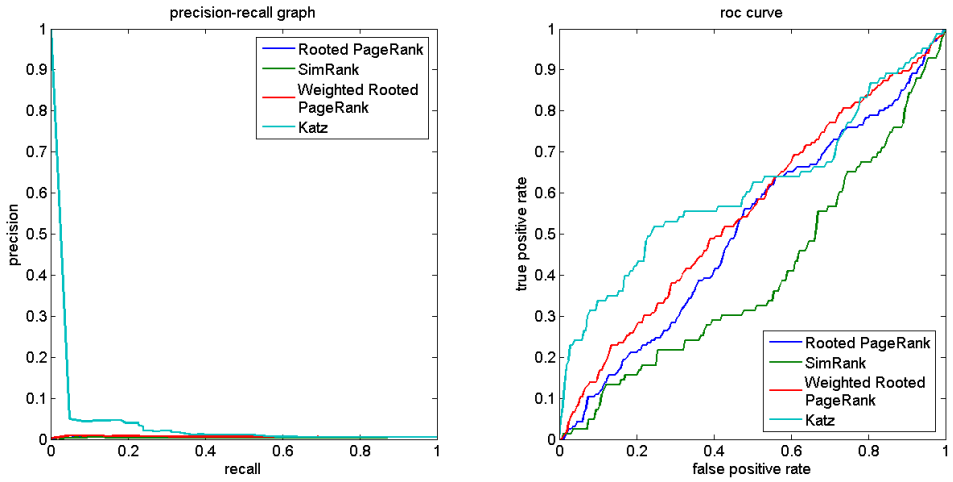

Figure 2. Performance of topology-based link predictors (author collaboration networks)

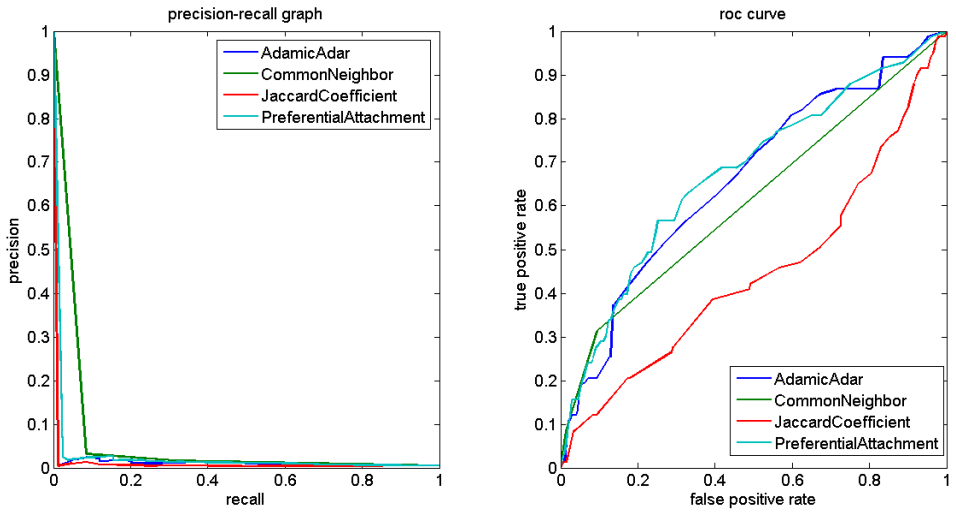

Figure 3. Performance of neighbor-information-based link predictors (author collaboration networks)

Figures 4 and 5 show the precision-recall graph and ROC curve for institution collaboration networks. For topology-based predictors, Katz has the best performance, followed by Weighted Rooted PageRank, Rooted PageRank, and SimRank. For neighbor-information-based predictors, Preferential Attachment has the best performance, followed by Adamic/Adar, Common Neighbor, and Jaccard Coefficient.

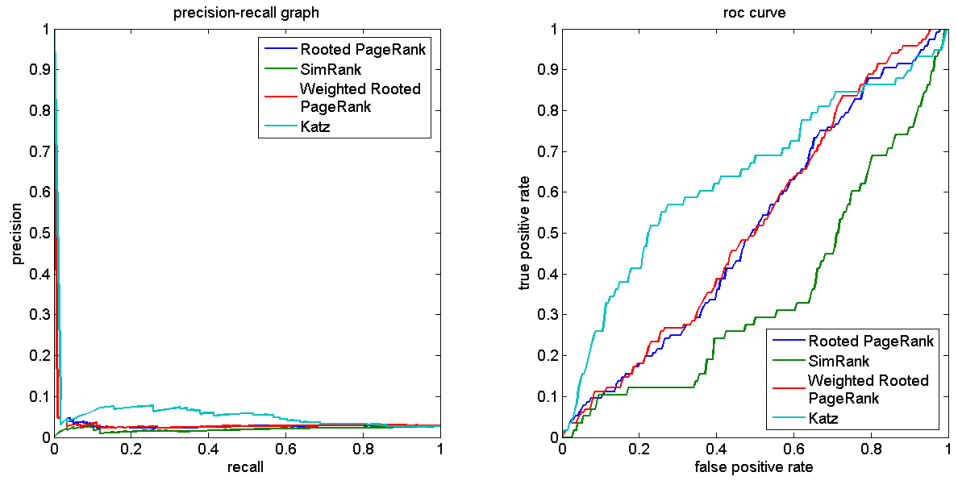

Figure 4. Performance of topology-based link predictors (institution collaboration networks)

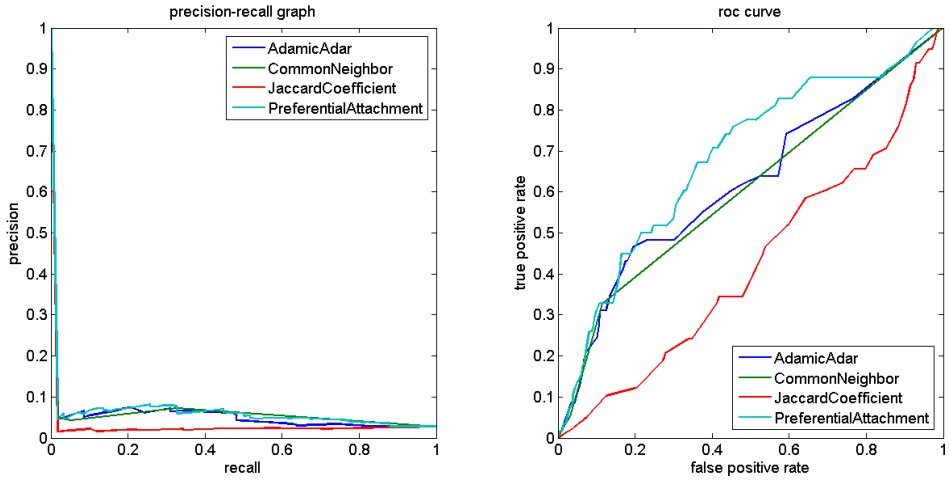

Figure 5. Performance of neighbor-information-based link predictors (institution collaboration networks)

Figures 6 and 7 show the precision-recall graph and ROC curve for state/country collaboration networks. For topology-based predictors, Katz has the best performance, followed by Rooted PageRank, Weighted Rooted PageRank, and SimRank. For neighbor-information-based predictors, Preferential Attachment has the best performance, followed by Adamic/Adar, Common Neighbor, and Jaccard Coefficient.

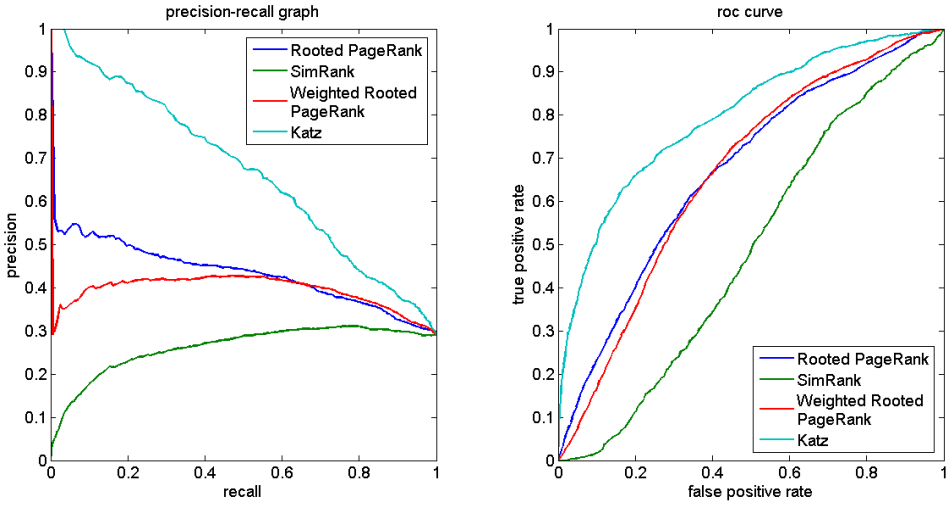

Figure 6. Performance of topology-based link predictors (country collaboration networks)

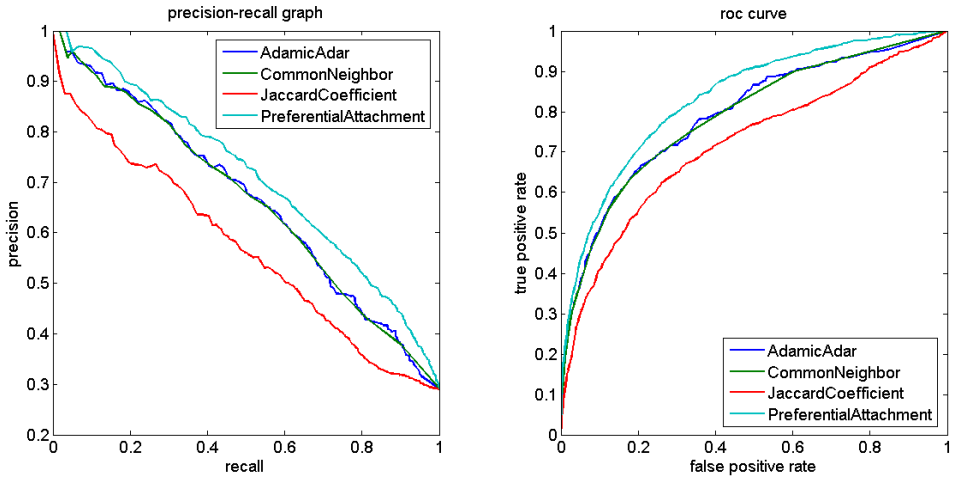

Figure 7. Performance of neighbor-information-based link predictors (country collaboration networks)

Overall, across the three aggregation levels we find that Katz is the best performing topology-based predictor and Preferential Attachment is the best performing neighbor-based predictor.

Area under ROC (AUC) and top k evaluations

If the precision–recall curve or ROC curve of one predictor is completely above the curve of another, we can safely conclude that the former has a better performance than the latter. In some cases, however, two curves may overlap, making it less clear which of the two should be preferred. For this reason we also consider two single number indicators: Area under ROC curve (AUC) and top *k* relevance with normalized discounted cumulative gain (nDCG).

Because the vertical axis of the ROC charts denotes true positive rate, the higher AUC, the better the performance. An AUC of 1 denotes a perfect predictor. Because a random method (e.g., getting either head or tail when flipping a coin) would yield an AUC of 0.5, a score below 0.5 can thus be considered as worthless. The result is shown in Table 2 where the highest AUCs for both topology-based and neighbor-information-based predictors are displayed in bold.

Table 2. Area under ROC (AUC) for eight predictors

	Author	Institution	State/Country
Rooted PageRank	0.5110	0.5108	0.6713
SimRank	0.4022	0.3519	0.4857
Weighted Rooted PageRank	0.5547	0.5212	0.6643
Katz	**0.6123**	**0.6399**	**0.7971**
Adamic/Adar	0.6598	0.6209	0.7899
Common Neighbors	0.6113	0.6037	0.7878

Jaccard Coefficient	0.4488	0.4229	0.7180
Preferential Attachment	**0.6739**	**0.6826**	**0.8337**

Katz has the highest AUC among topology-based predictors, followed by Weighted Rooted PageRank and Rooted PageRank (except in the case of countries, where Rooted PageRank performs slightly better than Weighted Rooted PageRank). SimRank has the lowest AUC. For neighbor-information-based predictors, Preferential Attachment has the highest AUC, followed by Adamic/Adar and Common Neighbor. Jaccard Coefficient has the lowest AUC. This is consistent across all three levels of aggregation. For both topology-based and neighbor-information-based predictors, Preferential Attachment has the best performance, Katz has the second best performance, and Adamic/Adar has the third best performance (except for authors, where Adamic/Adar outperforms Katz).

The results are generally consistent with findings in previous studies. For instance, Huang, Li, and Chen (2005) have found that Katz and Preferential Attachment have the best performance among several link predictors. A study by Liben-Nowell and Kleinberg (2007) also found that Katz and Adamic/Adar are among the best performing predictors. The same study also concluded that Preferential Attachment performs better when applied to denser networks. In this case, the country collaboration networks are much denser than author and institution collaboration networks. Therefore, the finding that Preferential Attachment performs exceptionally well on country collaboration networks is consistent with Liben-Nowell and Kleinberg's (2007) study. Contrary to our findings here, Guns (2009) obtained poor results for Preferential Attachment. This can be explained by the low density and the multidisciplinary nature of the network used in that study.

The Jaccard Coefficient is a normalization of Common Neighbors but clearly performs worse; this is consistent with findings elsewhere that Common Neighbors outperforms normalized forms like Jaccard (Guns, 2011; Liben-Nowell & Kleinberg, 2007). Predictors' performances are largely consistent across author, institution, and country collaboration networks.

Top *k* relevance was implemented to evaluate predictors' performances at different top levels. The results were first ranked based on respective predictor scores (e.g., Rooted PageRank scores, Katz scores, number of common neighbors). The following values for *k* were examined: 20, 50, 100, 200, and 500. Normalized discounted cumulative gain (nDCG) is used as the relevance measurement. Specifically, for the top *k* predicted collaboration pairs, true positive predictions have a raw score of 1 and false positive predictions have a raw score of 0. The raw scores are processed to give discounted scores at lower ranks. The processed scores are normalized so that an nDCG of 1 is the upper bound (all true positive for all top *k* pairs) and 0 for the lower bound (all false positive for top *k* pairs). The results are presented in Table 3 where the highest nDCGs are displayed in bold for each level of aggregation and at each different *k* value.

Table 3. Top *k* relevance for eight predictors

		Author	Institution	State/Country
Rooted PageRank	k=20	0	0.0336	0.3770
	k=50	0	0.0343	0.4625
	k=100	0	0.0432	0.4904
	k=200	0	0.0340	0.4982
	k=500	0.0016	0.0272	0.5085
SimRank	k=20	0	0	0.4422
	k=50	0.0299	0.0262	0.4259
	k=100	0.0188	0.0466	0.3991
	k=200	0.0115	0.0591	0.3522
	k=500	0.0159	0.0565	0.2808
Weighted Rooted PageRank	k=20	0	**0.1280**	0.3382
	k=50	0	**0.0876**	0.3196
	k=100	0	**0.0695**	0.3118
	k=200	0	0.0500	0.3456
	k=500	0.0034	0.0385	0.3623
Katz	k=20	0	0.0648	**1**
	k=50	0.0305	0.0369	**1**
	k=100	**0.0617**	0.0375	**1**
	k=200	0.0377	**0.0687**	0.9525
	k=500	**0.0484**	0.0690	0.9033
Adamic/Adar	k=20	0	0.0682	**1**
	k=50	0.0279	0.0388	**1**
	k=100	0.0176	0.0550	0.9720
	k=200	0.0107	0.0533	0.9523
	k=500	0.0182	0.0683	0.9067
Common Neighbor	k=20	0	0	**1**
	k=50	0.0283	0.0430	**1**
	k=100	0.0248	0.0421	0.9721
	k=200	**0.0424**	0.0488	0.9520
	k=500	0.0296	0.0695	0.9031
Jaccard Coefficient	k=20	0	0.0551	**1**
	k=50	0	0.0473	0.9384
	k=100	0.0070	0.0298	0.9023
	k=200	0.0083	0.0255	0.8706
	k=500	0.0057	0.0244	0.8189
Preferential Attachment	k=20	0	0.0727	**1**
	k=50	**0.0547**	0.0544	**1**
	k=100	0.0344	0.0490	**1**
	k=200	0.0210	0.0649	**0.9760**
	k=500	0.0237	**0.0725**	**0.9355**

In general, Preferential Attachment has the highest nDCG, especially for dense country collaboration networks; Katz and Weighted Rooted PageRank also have high nDCG. For different *k* levels, nDCG varies across different predictors and different aggregation levels. For instance, nDCG scores increase for Rooted PageRank and Weighted Rooted PageRank when

applied to country collaboration networks as *k* increases; however, it decreases when applied to institution collaboration networks. For Katz, Adamic/Adar, and Preferential Attachment, no noticeable trend can be identified between different *k* levels and nDCG scores. For different levels of aggregation, country collaboration networks yield the highest nDCG scores, followed by institution collaboration networks and author collaboration networks, suggesting that network densities affect prediction results – denser networks may yield more precise collaboration predictions. This is reinforced by the fact that the density of the country collaboration network increases – and many new links are formed, whereas the density of the other two decreases (and fewer new links are formed).

Despite the large differences between the three levels of aggregation, we find that strong predictors at one aggregation level (e.g., Katz and Preferential Attachment) tend to yield good results at the other levels as well. The results indicate that scientific collaboration is a complex research activity: in addition to the topological factor, geographical, topical, policy-related, or accidental factors also contribute to and thus inform the formation of future collaborations.

Relationships between the eight predictors

Using prediction results of country collaboration networks as an example, Table 4 shows the matching results of true positive predictions among eight predictors; for each predictor, the best matched predictor is displayed in bold. The last row shows the total number of overlapped predictions for each predictor.

Table 4. Overlapping of true positive predictions

	Rooted PageRank	SimRank	Weighted Rooted PageRank	Katz	Adamic/ Adar	Common Neighbor	Jaccard Coefficient	Preferential Attachment
Rooted PageRank	*2780*							
SimRank	793	*2780*						
Weighted Rooted PageRank	921	**848**	*2780*					
Katz	1089	764	1036	*2780*				
Adamic/Adar	1051	757	1026	1341	*2780*			
Common Neighbor	1044	806	993	1363	1340	*2780*		
Jaccard Coefficient	1004	762	983	1223	1200	1225	*2780*	
Preferential Attachment	**1111**	767	**1048**	**1366**	**1380**	**1382**	**1262**	*2780*
Sum	7013	5497	6855	8182	8095	8153	7659	8316

Rooted PageRank, Weighted Rooted PageRank, Katz, Adamic/Adar, Common Neighbor, and Jaccard Coefficient have the highest numbers of overlapping true positive predictions with Preferential Attachment. SimRank has the highest number of overlapping true positive predictions with Weighted Rooted PageRank. The results suggest that Preferential Attachment is more similar to other predictors and SimRank is the least similar to other predictors. These

findings are confirmed through multidimensional scaling (MDS). The binary true/false positive and true/false negative prediction results were used to form a Euclidean distance matrix for MDS. The Kruskal's stress value is 0.15 and R squared is 0.95, suggesting that the two dimensions (as illustrated in Figure 8) deliver a sound representation of the prediction results.

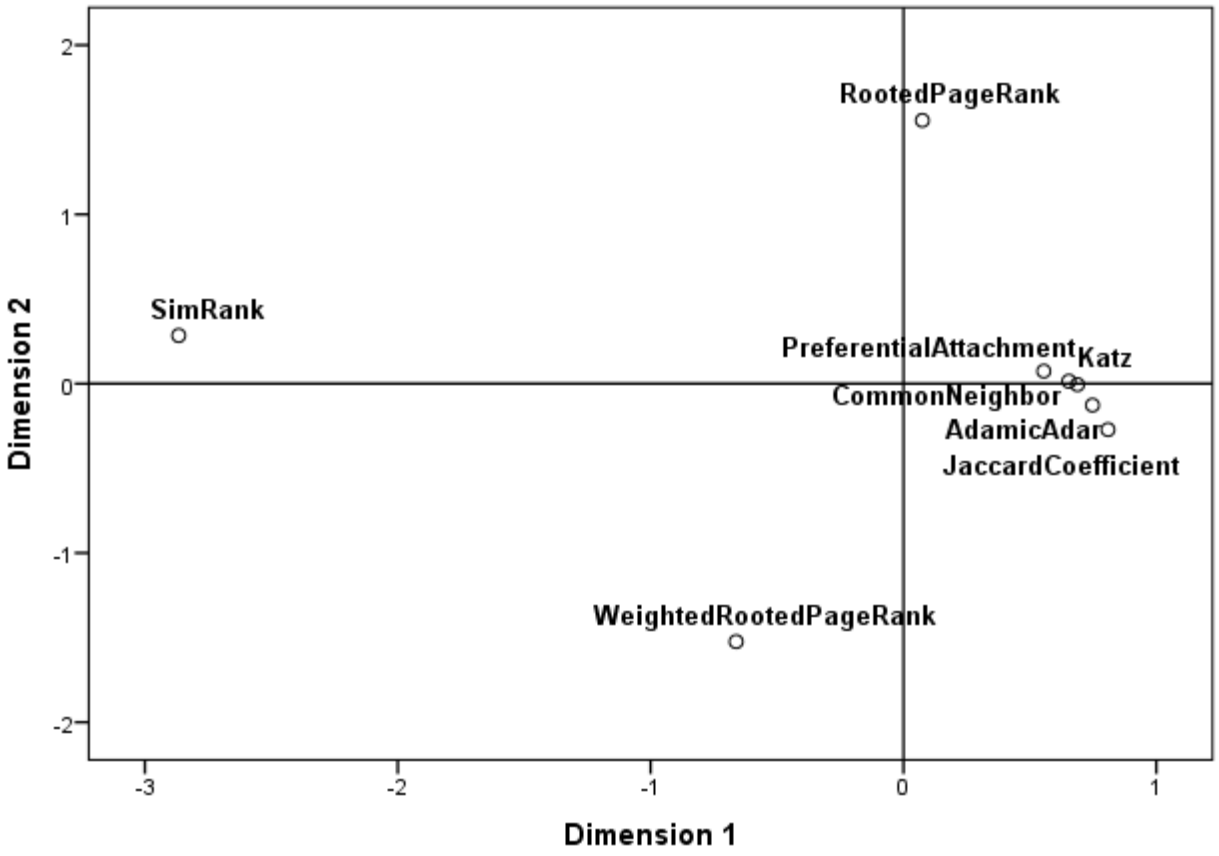

Figure 8. Graphic presentation of multidimensional scaling of the eight predictors

As indicated by MDS, SimRank is the most different from all other link predictors. Weighted Rooted PageRank and Rooted PageRank, surprisingly, are on the two ends of dimension 2, suggesting that collaboration intensity (i.e., edge weight) has a great impact on prediction results. While topology-based predictors are scattered in different corners in Figure 8, neighbor-information-based predictors are more collocated and thus producing more consistent prediction outcomes. Among all topology-based predictors, Katz yielded results the most similar to neighbor-information-based predictors; among all neighbor-information-based predictors, Preferential Attachment yielded results the most similar to topology-based predictors.

Collaboration recommendation

Two predictors, Katz (the best performing topology-based predictor) and Preferential Attachment (the best performing neighbor-information-based predictor), were found to perform

well across all three levels of aggregation. In this section they are applied to the integrated 2001-2010 collaboration networks to generate recommendations for collaboration between authors, institutions, and countries. Top 10 predicted collaboration pairs are presented in Tables 5 to 7.

Table 5. Top 10 predicted author collaboration pairs

Katz		Preferential Attachment	
SHAW BR – LANDUCCI G	1.61e-05	ROUSSEAU R – OPPENHEIM C	2891
ROWLEY J – COOPER J	1.42e-05	TENOPIR C – OPPENHEIM C	2596
RAY K – COOPER J	1.42e-05	OPPENHEIM C – NICHOLAS D	2419
COULSON G – COOPER J	1.42e-05	OPPENHEIM C – LEYDESDORFF L	2183
COOPER J – BANWELL L	1.42e-05	TENOPIR C – ROUSSEAU R	2156
ROWLEY J – LIGHT A	1.42e-05	OPPENHEIM C – BOOTH A	2065
ROWLEY J – BARKER A	1.42e-05	OPPENHEIM C – DAVIS GB	2065
RAY K – BARKER A	1.42e-05	ROUSSEAU R – NICHOLAS D	2009
RAY K – LIGHT A	1.42e-05	OPPENHEIM C – BENBASAT I	2006
BARKER A – BANWELL L	1.42e-05	OPPENHEIM C – NUNAMAKER JF	1947

Predictors Katz and Preferential Attachment yielded quite different top 10 results for author collaboration networks. On the Preferential Attachment column, results suggest that Charles Oppenheim and Ronald Rousseau may form a collaboration relation which may complement their research; other suggested collaborations are Carol Tenopir and Charles Oppenheim, Loet Leydesdorff and Charles Oppenheim, David Nicholas and Charles Oppenheim, and Carol Tenopir and Ronald Rousseau.

For savvy information scientists, some of these recommended collaborations may not be feasible. Our goal here is not merely to predict and verify who will collaborate with whom in the future, nor to designate authors to work with others based on the recommendations. But rather, we intend to suggest a small set of potential collaborators for their scrutiny. The suggested collaborators may be improbable and in other cases trivial, but they may also be latent from scholars' daily activities. Especially in the latter case, these recommended collaborations can be beneficial.

Table 6 shows predicted institutional collaborations.

Table 6. Top 10 predicted institution collaboration pairs

Katz		Preferential Attachment	
INDIANA UNIV,BLOOMINGTON – GEORGIA STATE UNIV,ATLANTA	2.21e-05	INDIANA UNIV,BLOOMINGTON – HARVARD UNIV,CAMBRIDGE	1564
UNIV WISCONSIN,MADISON – UNIV MARYLAND,COLLEGE PK	1.82e-05	HARVARD UNIV,CAMBRIDGE – GEORGIA STATE UNIV,ATLANTA	1426
UNIV GEORGIA,ATHENS – INDIANA UNIV,BLOOMINGTON	1.81e-05	PENN STATE UNIV,UNIVERSITY PK – HARVARD UNIV,CAMBRIDGE	1196

UNIV ARIZONA,TUCSON – INDIANA UNIV,BLOOMINGTON	1.61e-05	UNIV ARIZONA,TUCSON – HARVARD UNIV,CAMBRIDGE	1150
UNIV PITTSBURGH,PITTSBURGH – MICHIGAN STATE UNIV,E LANSING	1.42e-05	KATHOLIEKE UNIV LEUVEN,BELGIUM – HARVARD UNIV,CAMBRIDGE	1058
UNIV PITTSBURGH,PITTSBURGH – UNIV ILLINOIS,CHICAGO	1.42e-05	UNIV MARYLAND,COLLEGE PK – HARVARD UNIV,CAMBRIDGE	1058
INDIANA UNIV,BLOOMINGTON – DUKE UNIV,DURHAM	1.42e-05	INDIANA UNIV,BLOOMINGTON – GEORGIA STATE UNIV,ATLANTA	1054
UNIV PITTSBURGH,PITTSBURGH – BRIGHAM &38; WOMENS HOSP,BOSTON	1.41e-05	UNIV PITTSBURGH,PITTSBURGH – UNIV ARIZONA,TUCSON	975
UNIV WASHINGTON,SEATTLE – INDIANA UNIV,BLOOMINGTON	1.41e-05	MICHIGAN STATE UNIV,E LANSING – HARVARD UNIV,CAMBRIDGE	920
UNIV WASHINGTON,SEATTLE – UNIV MARYLAND,COLLEGE PK	1.41e-05	UNIV PITTSBURGH,PITTSBURGH – KATHOLIEKE UNIV LEUVEN,BELGIUM	897

Predictors Katz and Preferential Attachment yield somewhat more consistent results for institution collaboration networks. It is suggested that Indiana University and Georgia State University may benefit from forming collaborations. Although the top 10 pairs of Katz and Preferential Attachment have no other predictions in common, it can be seen that the nodes (institutions) involved are to a larger extent the same, which is not the case for authors (Table 5).

Table 7 shows predicted collaborations between U.S. states and/or countries.

Table 7. Top 10 predicted state/country collaboration pairs

Katz		Preferential Attachment	
SCOTLAND – GEORGIA,USA	1.39e-03	SCOTLAND – GEORGIA,USA	6230
GERMANY – COLORADO,USA	1.37e-03	GERMANY – COLORADO,USA	6216
WISCONSIN,USA – FRANCE	1.28e-03	WISCONSIN,USA – FRANCE	5904
WASHINGTON,USA – SPAIN	1.28e-03	WASHINGTON,USA – SPAIN	5776
SOUTH KOREA – DENMARK	1.26e-03	DENMARK – ARIZONA,USA	5740
DENMARK – ARIZONA,USA	1.26e-03	SOUTH KOREA – DENMARK	5576
OKLAHOMA,USA – GERMANY	1.26e-03	OKLAHOMA,USA – GERMANY	5544
OKLAHOMA,USA – INDIA	1.22e-03	SPAIN – ARIZONA,USA	5320
NEBRASKA,USA – INDIA	1.22e-03	NEVADA,USA – GERMANY	5208
NEVADA,USA – GERMANY	1.20e-03	HUNGARY – ARIZONA,USA	5180

Predictors Katz and Preferential Attachment yield more consistent results for country collaboration networks: eight out of 10 recommendations belong to the top 10 of both predictors. Cross state/country collaborations may promote knowledge sharing and stimulate innovation (Jones, Wuchty, & Uzzi, 2008).

Discussion

Integrating prediction results

In our opinion, each predictor has its strengths and weaknesses. This can be illustrated by considering the Preferential Attachment results in Table 5. While some of these collaborators share similar research expertise, others are less so; for instance, Carol Tenopir and Ronald Rousseau are both prolific authors in information science, but they possess different research interests: one on information access and the other on informetrics. The strong results of Preferential Attachment in the previous section indicate that it captured certain aspects of network evolution which may be overlooked by other predictors. Hence, a promising approach seems to be *merging* the results of different predictors, each with their own conceptual and algorithmic properties. This approach can be seen as an application of the principle of polyrepresentation (Ingwersen, 1994; Larsen, Ingwersen, & Lund, 2009).

A simple merging procedure is the Borda method (Aslam & Montague, 2001), where – for each predictor – the highest ranked prediction gets a score of n (the number of possible predictions), the second highest a score of $n - 1$, and so on. The merged prediction is then obtained by summing the Borda score over all predictors for each item (in this case, each collaboration pair). Table 8 displays the results of applying this procedure to the recommendations for author collaboration according to Katz and PreferentialAttachment. This introduces several new recommendations into the top 10 (compare with Table 5). At the same time, some earlier recommendations, including the recommended collaboration between Tenopir and Rousseau, remain. Future research may also consider other merging procedures.

Table 8. Top 10 predicted author collaboration pairs with merged predictions

Prediction	Score
THELWALL M – LEYDESDORFF L	4965869
TENOPIR C – SMITH A	4965053
SMITH A – HUNTINGTON P	4964992
TENOPIR C – BAWDEN D	4964846
YEN DC – JIANG JJ	4964775
TENOPIR C – ROUSSEAU R	4964140
ROUSSEAU R – LEYDESDORFF L	4964125
HUNTINGTON P – BAWDEN D	4964060
ROUSSEAU R – GLANZEL W	4964002
YEN DC – KLEIN G	4963846

Toward a multi-level analysis of collaboration

Scientific collaboration is a complex societal phenomenon. One school of thought believes that collaboration is predicated upon institutions and institutions are governed by epistemological cultures (e.g., Mulkay, Gilbert, & Woolgar, 1975). Others hold a different belief that collaboration is "transepistemic" as scientific inquiries are conducted in an environment where

both scientists and non-scientists work together, and thus collaboration possesses both technical and non-technical nature (Knorr-Cetina, 1982): "[T]he situational contingencies observed in the laboratory are traversed and sustained by relationships which constantly transcend the site of research" (p.102).

In order to examine these beliefs, one must situate all essential collaborative entities (i.e., authors, institutions, and countries) in a systematic context (Figure 9).

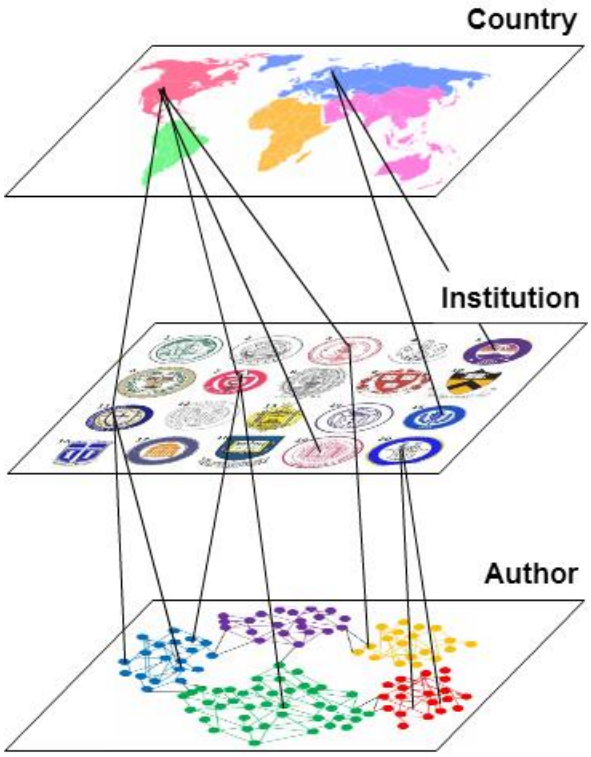

Figure 9. Collaboration levels

Author-level analysis is helpful for understanding individuals' collaboration motivations; it may help to reconcile the debate over the inclination of authors to collaborate with domain exports vs. scholars with similar academic standing, authors with similar research specialties vs. authors with diverse specialties, authors with close physical proximity vs. authors with an extended geographical distance. Institution collaboration studies help the examination of the role of research institutions in shaping disciplines' landscape (e.g., Boschma, 2005). It also provides social and institutional aspects to study disciplinarity (e.g., Abbot, 2001). Country-level collaboration studies are useful for identifying macro-level factors that contribute to collaboration. Factors such as language, science policy, and culture can all affect country-level collaborations. For instance, Thelwall (2012) found that language is a deciding factor for university website interlinking among European countries.

In this paper, we have evaluated eight link predictors in a well-defined data set. Based on the evaluation results, we applied the two best performing ones, Katz and PreferentialAttachment, to recommend collaborations at author-, institution-, and country-levels. The same predictors tend to perform well at all collaboration levels. Most variability among the author-, institution- and country-levels appears to be attributed to differences in network density. Our results indicate that all predictors perform better when applied to the dense country collaboration network instead of the sparse author or institution collaboration networks. This is especially the case for PreferentialAttachment, which yields the highest level of accuracy for country collaboration networks.

The results show that most likely collaborations comprise authors with similar research specialties (e.g., Leydesdorff – Thelwall, Leydesdorff – Rousseau). At the same time, institution-level predictions show that geographic distance is another factor shaping future collaborations, as the recommended collaborations seem to maintain shorter geographic distances or are located in the same country/territory (e.g., UNIV MARYLAND,COLLEGE PK – HARVARD UNIV,CAMBRIDGE or INDIANA UNIV,BLOOMINGTON – GEORGIA STATE UNIV,ATLANTA).

Link prediction and teams of science

Science is monotonically becoming more collaborative (e.g., Babchuk, Keith, & Peters, 1999; Wuchty, Jones, Uzzi, 2007): research teams are becoming predominant "cross nearly all fields" (Wuchty, Jones, Uzzi, 2007, p. 1036). Research teams are formed comprising scientists and scholars of diverse expertise (e.g., Fiore, 2008), from multiple universities (e.g., Jones, Wuchty, & Uzzi, 2008), between post-docs and colleagues (e.g., Horta, 2009), and between doctoral students and advisors (e.g., Moody, 2004). Teams produce high impact research, especially for those of multi-university collaborations (Jones, Wuchty, & Uzzi, 2008). Teams also help researchers engage more actively in information exchange with domestic and international peers and allow them to integrate into international scholarly communities (e.g., Horta, 2009; Lambiotte, R., & Panzarasa, 2009).

Because the current analysis is based on a data set on library and information science publications, collaborations outside this field are not included. Yet, the recommendations do include suggestions for cross-expertise collaboration, such as between Carol Tenopir (a renowned scholar on information access) and Ronald Rousseau (a renowned scholar on informetrics), and between Indiana University (a university that has a department specialized in informetrics) and Georgia State University (a university that has a department specialized in information systems). Such cross-expertise recommendation is attributed to the way that link predictors work: link predictors do not differentiate whether an author is associated with any particular field; as long as there are instances of collaborations, either cross-expertise or within-expertise, link predictors will follow such topology and recommend collaborations that may pertain to it. In this regard, link prediction is capable of recommending teams of science,

provided that there are precedents of teams in the existing topology. Future analysis may benefit from using more interdisciplinary data sets (e.g., data sets on energy and brain science) and evaluating the performance of link predictors on recommending teams of science.

Limitation

The link prediction method relies only on topology. Such an approach brings forward conveniences and benefits (e.g., low requirement of data sources and ease of implementation); however, it may also result in limitations. As scientific collaboration is a complex socio-cognitive process, multiple factors can contribute to collaboration decisions (e.g., Moody, 2004; Yan & Sugimoto, 2011). These factors may not always be effectively captured and assimilated by topology. Consequently, false positive rates are high and some of the recommended collaborations may be undesirable or unrealistic. As stated earlier, the goal of this study is not simply evaluating different link predictors, but illustrating how to use these predictors to reveal latent information and to recommend potential collaborations. As long as the prediction results can inspire some authors (and thereby, potentially, institutions and countries) to establish new collaborations, the link prediction method is worth investigating. Future studies may benefit from incorporating machine learning methods (e.g., Al Hasan et al., 2006; Lichtenwalter, Lussier, & Chawla, 2010) with link prediction to enhance prediction performance.

Conclusion

This study has explored collaboration dynamics through the link prediction method. Author-, institution-, and country-level collaboration networks have been constructed using a ten-year data set on library and information science publications. Eight link predictors have been applied to these collaboration networks. They have been grouped into two categories: neighbor-information-based (Adamic/Adar, Common Neighbors, Preferential Attachment, and Jaccard Coefficient) and topology-based (Katz, Rooted PageRank, Weighted Rooted PageRank, and SimRank).

The study has revealed that, for the employed data set in particular, higher-level collaboration networks (i.e., country-level collaboration networks) tend to yield more accurate prediction outcomes than lower-level ones (i.e., institution- and author-level collaboration networks). Based on the recommended collaborations of the data set, this study also finds that prediction results by SimRank are the least consistent with other link predictors. In the meantime, neighbor-information-based approaches are more collocated than topology-based ones on the multidimensional scaling map. Additionally, this study has proposed a succinct way to integrate collaboration recommendations obtained from multiple predictors.

Note that the findings only refer to the application of the employed link predictors on one particular data set. Results may vary for different data sets. By relating the findings of this study with those obtained from previous ones (e.g., Huang, Li, Chen, 2005; Liben-Nowell & Kleinberg, 2007; Guns, 2011), we have discovered that the following observations are consistent across

these studies: (1) Katz and Preferential Attachment are among the best performing link predictors; (2) Common Neighbors outperforms normalized forms like Jaccard; and (3) link predictors perform better when applied to denser networks at higher levels of aggregation. Although these observations are subject to further verifications, they are not trivial and should inform the selection of link predictors for ongoing studies of collaboration dynamics.

References

Abbot, A. (2001). *Chaos of disciplines*. Chicago: University of Chicago Press.

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

Ahn, Y.-Y., Bagrow, J. P., & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature, 466*(7307), 761-764.

Al Hasan, M., Chaoji, V., Salem, S., & Zaki, M. (2006). Link prediction using supervised learning. In *SDM'06: Workshop on Link Analysis, Counter-terrorism and Security*.

Albert, R., & Barabási, A. L. (2000). Topology of evolving networks: local events and universality. *Physical review letters, 85*(24), 5234-5237.

Aslam, J.A., & Montague, M. (2001). Models for metasearch. In D. H. Kraft et al. (Eds.), *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)* (pp. 276–284). New York: ACM.

Babchuk, N., Keith, B., & Peters, G. (1999). Collaboration in sociology and other scientific disciplines: A comparative trend analysis of scholarship in the social, physical, and mathematical sciences. *The American Sociologist, 30*(3), 5-21.

Barabási, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications, 311*(3-4), 590-614.

Barabási, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509-512.

Barrat, A., Barthélemy, M., & Vespignani, A. (2004). Weighted Evolving Networks: Coupling Topology and Weight Dynamics. *Physical Review Letters, 92*(22), 228701.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.

Börner, K., Contractor, N., Falk-Krzesinski, H. J., Fiore, S. M. , Hall, K. L. , Keyton, J., Spring, B., Stokols, D. , Trochim , W., & Uzzi, B. (2010). A multi-level systems perspective for the science of team science. *Science Translational Medicine, 2*(49), cm24. DOI: 10.1126/scitranslmed.3001399

Boschma, R. A. (2005). Proximity and innovation: A critical assessment. *Regional Studies, 39*(1), 61-74.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition, 30*(7), 1145-1159.

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

Farkas, I., Ábel, D., Palla, G., & Vicsek, T. (2007). Weighted network modules. *New Journal of Physics, 9*(6), 180-209.

Fiala, D., Rousselot, F., & Ježek, K. (2008). PageRank for bibliographic networks. *Scientometrics, 76*(1), 135-158.

Fiore, S. M. (2008). Interdisciplinarity as teamwork how the science of teams can inform team science. *Small Group Research, 39*(3), 251-277.

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

Guns, R. (2011). Bipartite networks for link prediction: Can they improve prediction performance? In E. Noyons, P. Ngulube, & J. Leta (Eds.), *Proceedings of the ISSI 2011 Conference* (pp. 249-260). Leiden: Leiden University.

Guns, R., & Rousseau, R. (2013). Predicting and recommending potential research collaborations. In *Proceedings of ISSI 2013* (pp. 1409–1418). Vienna: AIT.

Guo, F., Yang, Z., & Zhou, T. (2013). Predicting link directions via a recursive subgraph-based ranking. *Physica A: Statistical Mechanics and its Applications, 392*, 3402-3408.

Havermann, F., Heinz, M., & Kretschmer, H. (2006). Collaboration and distances between German immunological institutes: A trend analysis. *Journal of Biomedical Discovery and Collaboration, 1*(6). DOI: 10.1186/1747-5333-1-6

Hoekman, J., Frenken, K., & van Oort, F. (2009). The geography of collaborative knowledge production in Europe. *The Annals of Regional Science, 43*(3), 721-738.

Holton, G. (1978). Can Science Be measured? In *Scientific Imaginations: Case Studies* (pp. 199-228). Cambridge, UK: Cambridge University Press.

Horta, H. (2009). Holding a post-doctoral position before becoming a faculty member: does it bring benefits for the scholarly enterprise? *Higher Education, 58*(5), 689-721.

Huang, Z., Li, X., & Chen, H. (2005). Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries* (pp. 141-142). New York: ACM Press.

Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 101–110). New York: Springer-Verlag.

Jeh, G., & Widom, J. (2002). SimRank: a measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 538–543). New York: ACM.

Jeong, H., Néda, Z., & Barabási, A. L. (2003). Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters), 61*(4), 567-572.

Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: shifting impact, geography, and stratification in science. *Science, 322*(5905), 1259-1262.

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

Knorr-Cetina, K. D. (1982). Scientific communities or transepistemic arenas of research? A critique of quasi-economic models of science. *Social studies of Science, 12*(1), 101-130.

Kretschmer, H. (2004). Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web. *Scientometrics, 60*(3), 409-420.

Lambiotte, R., & Panzarasa, P. (2009). Communities, knowledge creation, and information diffusion. *Journal of Informetrics, 3*(3), 180-190.

Larsen, B., Ingwersen, P., & Lund, B. (2009). Data fusion according to the principle of polyrepresentation. *Journal of the American Society for Information Science and Technology*, 60(4), 646–654.

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.

Lichtenwalter, R. N., & Chawla, N. V. (2011). Lpmade: Link prediction made easy. *The Journal of Machine Learning Research, 12*, 2489-2492.

Lichtenwalter, R. N., Lussier, J. T., & Chawla, N. V. (2010). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 243-252). New York: ACM Press.

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

Logan, E. L., & Shaw, W. M. (1991). A bibliometric analysis of collaboration in a medical specialty. *Scientometrics, 20*(3), 417-426.

Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications, 390*(6), 1150-1170.

Luukkonen, T., Persson, O., & Sivertsen, G. (1992). Understanding patterns of international scientific collaboration. *Science Technology & Human Values, 17*(1), 101-126.

Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review, 69*(2), 213-238.

Mulkay, M., Gilbert, G. N., & Woolgar, S. (1975). Problem areas and research networks in science. *Sociology, 9*(2), 187-203.

Newman, M. E. J. (2001a). Scientific collaboration networks: I. Network construction and fundamental results. *Physical Review E, 64*(1), 016131.

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

Newman, M., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*(2), 026113.

Pencina, M. J., D'Agostino Sr, R. B., D'Agostino Jr, R. B., & Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine, 27*(2), 157-172.

Ponds, R., Van Oort, F., & Frenken, K. (2007). The geographical and institutional proximity of research collaboration. *Papers in Regional Science, 86*(3), 423-443.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851-860). New York: ACM Press.

Sarkar, P., Chakrabarti, D., & Moore, A. W. (2010). Theoretical justification of popular link prediction heuristics. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, July 16-22, 2011, Barcelona, Spain. Retrieved June 8, 2012 from http://www.aaai.org/ocs/index.php/IJCAI/IJCAI11/paper/viewFile/3327/3664

Sharan, U., & Neville, J. (2008). Temporal-Relational classifiers for prediction in evolving domains. In *IEEE International Conference on Data Mining* (pp. 540–549). Los Alamitos, CA: IEEE Computer Society.

Shibata, N., Kajikawa, Y., & Sakata, I. (2012). Link prediction in citation networks. *Journal of the American society for information science and technology*, *63*(1), 78-85.

Sidiropoulos, A., & Manolopoulos, Y. (2006), A Generalized comparison of graph-based ranking algorithms for publications and authors. *Journal of Systems and Software, 79*(12), 1679-1700.

Ströele, V., & Souza, J. M. (2013). Group and link analysis of multi-relational scientific social networks. *Journal of Systems and Software, 86,* 1819-1830.

Thelwall, M. (2012). Webometrics: The evolution of a digital social science research field. *Social Science and Digital Research: Interdisciplinary Insights*. University of Oxford, UK. March 12, 2012.

Wang, L., Lou, T., Tang, J., & Hopcroft, J. E. (2011). Detecting community kernels in large social networks. In *Proceedings of 2011 IEEE International Conference on Data Mining (ICDM 2011).* December 11-14, 2011, Vancouver, Canada.

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

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

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036-1039.

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.

Yan, E., & Ding, Y. (2011c). Discovering author impact: A PageRank perspective. *Information Processing and Management, 47*(1), 125-134.

Yan, E., & Sugimoto, C. R. (2011). Institutional interactions: Exploring social, cognitive, and geographic relationships between institutions as demonstrated through citation networks. *Journal of the American Society for Information Science and Technology*, *62*(8), 1498-1514.

Yin, L., Kretschmer, H., Hanneman, R. A., & Liu, Z. (2006). Connection and stratification in research collaboration: An analysis of the COLLNET network. *Information Processing and Management*, *42*, 1599-1613.