# The Link-Prediction Problem for Social Networks

**David Liben-Nowell**
*Department of Computer Science, Carleton College, Northfield, MN 55057. E-mail: dlibenno@carleton.edu*

**Jon Kleinberg**
*Department of Computer Science, Cornell University, Ithaca, NY 14853. E-mail: kleinber@cs.cornell.edu*

**Given a snapshot of a social network, can we infer which new interactions among its members are likely to occur in the near future? We formalize this question as the *link-prediction problem*, and we develop approaches to link prediction based on measures for analyzing the "proximity" of nodes in a network. Experiments on large coauthorship networks suggest that information about future interactions can be extracted from network topology alone, and that fairly subtle measures for detecting node proximity can outperform more direct measures.**

## Introduction

As part of the recent surge of research on large, complex networks and their properties, a considerable amount of attention has been devoted to the computational analysis of *social networks*—structures whose nodes represent people or other entities embedded in a social context, and whose edges represent interaction, collaboration, or influence between entities. Natural examples of social networks include the set of all scientists in a particular discipline, with edges joining pairs who have coauthored articles; the set of all employees in a large company, with edges joining pairs working on a common project; or a collection of business leaders, with edges joining pairs who have served together on a corporate board of directors. The increased availability of large, detailed datasets encoding such networks has stimulated extensive study of their basic properties, and the identification of recurring structural features (e.g., see the work of Adamic & Adar, 2003; Grossman, 2002; Newman, 2002; Watts, 1999; Watts & Strogatz, 1998; for a thorough recent survey, see Newman, 2003).

Social networks are highly dynamic objects; they grow and change quickly over time through the addition of new edges, signifying the appearance of new interactions in the underlying social structure. Identifying the mechanisms by which they evolve is a fundamental question that is still not well understood, and it forms the motivation for our work here. We define and study a basic computational problem underlying social-network evolution, the *link-prediction problem*: Given a snapshot of a social network at time $t$, we seek to accurately predict the edges that will be added to the network during the interval from time $t$ to a given future time $t'$.

In effect, the link-prediction problem asks: To what extent can the evolution of a social network be modeled using features *intrinsic to the network itself*? Consider a coauthorship network among scientists, for example. There are many reasons exogenous to the network why two scientists who have never written an article together will do so in the next few years: For example, they may happen to become geographically close when one of them changes institutions. Such collaborations can be hard to predict. But one also senses that a large number of new collaborations are hinted at by the topology of the network: Two scientists who are "close" in the network will have colleagues in common and will travel in similar circles; this social proximity suggests that they themselves are more likely to collaborate in the near future. Our goal is to make this intuitive notion precise and to understand which measures of "proximity" in a network lead to the most accurate link predictions. We find that a number of proximity measures lead to predictions that outperform chance by factors of 40 to 50, indicating that the network topology does indeed contain latent information from which to infer future interactions. Moreover, certain fairly subtle measures—involving infinite sums over paths in the network—often outperform more direct measures such as shortest-path distances and numbers of shared neighbors.

We believe that a primary contribution of the present article is in the area of network-evolution models. While there has been a proliferation of such models in recent years—e.g., see the work of Barabasi et al., 2002; Davidsen, Ebel, and Bornholdt, 2002; Jin, Girvan, and Newman, 2001; for recent work on collaboration networks, or the survey of Newman, 2003—these models have generally been evaluated only by asking whether they reproduce certain global structural features observed in real networks. As a result, it has been difficult to evaluate and compare different approaches on a principled footing. Link prediction, on the other hand, offers a very natural basis for

Received June 16, 2005; revised July 5, 2006; accepted August 25, 2006

 ● Published online 26 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20591

such evaluations: *A network model is useful to the extent that it can support meaningful inferences from observed network data*. One sees a related approach in recent work of Newman (2001a), who considers the correlation between certain network-growth models and data on the appearance of edges of coauthorship networks.

In addition to its role as a basic question in social-network evolution, the link-prediction problem could be relevant to a number of interesting current applications of social networks. Increasingly, for example, researchers in artificial intelligence and data mining have argued that a large organization (e.g., a company) can benefit from the interactions within the informal social network among its members; these ties serve to supplement the official hierarchy imposed by the organization itself (Kautz, Selman, & Shah, 1997; Raghavan, 2002). Effective methods for link prediction could be used to analyze such a social network to suggest promising interactions or collaborations that have not yet been identified within the organization. In a different vein, research in security has recently begun to emphasize the role of social-network analysis, largely motivated by the problem of monitoring terrorist networks; link prediction in this context allows one to conjecture that particular individuals are working together even though their interaction has not been directly observed (Krebs, 2002).

The link-prediction problem also is related to the problem of inferring missing links from an observed network: In a number of domains, one constructs a network of interactions based on observable data and then tries to infer additional links that, while not directly visible, are likely to exist (Goldberg & Roth, 2003; Popescul & Ungar, 2003; Taskar, Wong, Abbeel, & Koller, 2003). This line of work differs from our problem formulation in that it works with a static snapshot of a network rather than considering network evolution; it also tends to take into account specific attributes of the nodes in the network rather than evaluating the power of prediction methods that are based purely on the graph structure.

We turn to a description of our experimental setup in the next section. Our primary focus is on understanding the relative effectiveness of network-proximity measures adapted from techniques in graph theory, computer science, and the social sciences, and we review a large number of such techniques. Finally, we discuss the results of our experiments.

## Data and Experimental Setup

Suppose that we have a social network $G = \langle V, E \rangle$ in which each edge $e = \langle u, v \rangle \in E$ represents an interaction between $u$ and $v$ that took place at a particular time $t(e)$. We record multiple interactions between $u$ and $v$ as parallel edges, with potentially different timestamps. For two times $t < t'$, let $G[t, t']$ denote the subgraph of $G$ consisting of all edges with a timestamp between $t$ and $t'$. Here, then, is a concrete formulation of the link-prediction problem. We choose four times $t_0 < t'_0 < t_1 < t'_1$ and give an algorithm access to the network $G[t_0, t'_0]$; it must then output a list of edges not present in $G[t_0, t'_0]$ that are predicted to appear in the network $G[t_1, t'_1]$. We refer to $[t_0, t'_0]$ as the *training interval* and $[t_1, t'_1]$ as the *test interval*.

Of course, social networks grow through the addition of nodes as well as edges, and it is not sensible to seek predictions for edges whose endpoints are not present in the training interval. Thus, in evaluating link-prediction methods, we will generally use two parameters, $\kappa_{training}$ and $\kappa_{test}$, and define the set Core to consist of all nodes that are incident to at least $\kappa_{training}$ edges in $G[t_0, t'_0]$ and at least $\kappa_{test}$ edges in $G[t_1, t'_1]$. We will then evaluate how accurately the new edges between elements of Core can be predicted.

We now describe our experimental setup more specifically. We work with five coauthorship networks $G$, obtained from the author lists of articles contained in five sections of the physics e-Print arXiv, `www.arxiv.org` (see Figure 1 for statistics on the sizes of each of these five networks). Some heuristics were used to deal with occasional syntactic anomalies, and authors were identified by first initial and last name, a process that introduces a small amount of noise due to multiple authors with the same identifier (Newman, 2001b). The errors introduced by this process appear to be minor.

Now consider any one of these five graphs. We define the training interval to be the 3 years from 1994 through 1996, and the test interval to be the 3 years from 1997 through 1999. We denote the subgraph $G[1994, 1996]$ on the training interval by $G_{collab} := \langle A, E_{old} \rangle$ and use $E_{new}$ to denote the set of edges $\langle u, v \rangle$ such that $u, v \in A$, and $u, v$ coauthor an article during the test interval, but not the training interval—these are the new interactions we are seeking to predict. In our experiments on the arXiv, we can identify which authors are active throughout the entire period on the basis of the number of articles published and not on the number of coauthors. Thus, here we define the set Core to consist of all authors who have written at least $\kappa_{training} := 3$ articles during the training period and at least $\kappa_{test} := 3$ articles during the test period.

| | Training Period | | | Core | | |
|---|---|---|---|---|---|---|
| | Authors | Articles | Collaborations[a] | Authors | $E_{old}$ | $E_{new}$ |
| `astro-ph` | 5,343 | 5,816 | 41,852 | 1,561 | 6,178 | 5,751 |
| `cond-mat` | 5,469 | 6,700 | 19,881 | 1,253 | 1,899 | 1,150 |
| `gr-qc` | 2,122 | 3,287 | 5,724 | 486 | 519 | 400 |
| `hep-ph` | 5,414 | 10,254 | 47,806 | 1,790 | 6,654 | 3,294 |
| `hep-th` | 5,241 | 9,498 | 15,842 | 1,438 | 2,311 | 1,576 |

[a]A collaboration is an ordered pair of authors who have written at least one article together during the training period. This number is odd in the `cond-mat` dataset because in that arXiv section there were three (an odd number) instances of "self-collaboration"—where 2 authors of the same article have the same first initial and last name; the 2 researchers are therefore conflated into a single node $x$, and a collaboration between $x$ and $x$ is recorded. These self-collaborations are examples of the rare errors that are introduced because multiple authors are mapped to the same identifier.

FIG. 1. The five sections of the arXiv from which coauthorship networks were constructed: `astro-ph` (astrophysics), `cond-mat` (condensed matter), `gr-qc` (general relativity and quantum cosmology), `hep-ph` (high energy physics–phenomenology), and `hep-th` (high energy physics–theory). The set Core is the subset of the authors who have written at least $\kappa_{training} = 3$ articles during the training period 1994–1996 and $\kappa_{test} = 3$ articles during the test period 1997−1999. The sets $E_{old}$ and $E_{new}$ denote undirected edges between Core authors that first appear during the training and test periods, respectively.


DOI: 10.1002/asi

*Evaluating a Link Predictor*

Each link predictor $p$ that we consider outputs a ranked list $L_p$ of pairs in $A \times A - E_{old}$; these are predicted new collaborations, in decreasing order of confidence. For our evaluation, we focus on the set Core, so we define $E^*_{new} := E_{new} \cap (\text{Core} \times \text{Core})$ and $n := |E^*_{new}|$. Our performance measure for Predictor $p$ is then determined as follows: From the ranked list $L_p$, we take the first $n$ pairs that are in Core × Core, and determine the size of the intersection of this set of pairs with the set $E^*_{new}$.

## Methods for Link Prediction

In this section, we survey an array of methods for link prediction. All the methods assign a connection weight score($x$, $y$) to pairs of nodes $\langle x, y \rangle$, based on the input graph $G_{collab}$, and then produce a ranked list in decreasing order of score($x$, $y$). Thus, they can be viewed as computing a measure of proximity or "similarity" between nodes $x$ and $y$, relative to the network topology. In general, the methods are adapted from techniques used in graph theory and in social-network analysis; in a number of cases, these techniques were not designed to measure node-to-node similarity and hence need to be modified for this purpose. Figure 2 summarizes most of these measures; we discuss them in more detail later. Note that some of these measures are designed only for connected graphs; because each graph $G_{collab}$ that we consider has a *giant component*—a single component containing most of the nodes—it is natural to restrict the predictions for these measures to this component.

Perhaps the most basic approach is to rank pairs $\langle x, y \rangle$ by the length of their shortest path in $G_{collab}$. Such a measure follows the notion that collaboration networks are "small worlds," in which individuals are related through short chains (Newman, 2001b) (In keeping with the notion that we rank pairs in *decreasing* order of score($x$, $y$), we define score($x$, $y$) here to be the negative of the shortest path length.) Pairs with shortest-path distance equal to one are joined by an edge in $G_{collab}$, and hence they belong to the training edge set $E_{old}$. For all of our graphs $G_{collab}$, there are well more than $n$ pairs at shortest-path distance two, so our shortest-path predictor simply selects a random subset of these distance-two pairs.

*Methods Based on Node Neighborhoods*

For a node $x$, let $\Gamma(x)$ denote the set of neighbors of $x$ in $G_{collab}$. A number of approaches are based on the idea that two nodes $x$ and $y$ are more likely to form a link in the future if their sets of neighbors $\Gamma(x)$ and $\Gamma(y)$ have large overlap; this

| | |
|---|---|
| graph distance | (negated) length of shortest path between $x$ and $y$ |
| common neighbors | $\lvert\Gamma(x) \cap \Gamma(y)\rvert$ |
| Jaccard's coefficient | $\dfrac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x) \cup \Gamma(y)\rvert}$ |
| Adamic/Adar | $\Sigma_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log\lvert\Gamma(z)\rvert}$ |
| preferential attachment | $\lvert\Gamma(x)\rvert \cdot \lvert\Gamma(y)\rvert$ |
| Katz$_\beta$ | $\Sigma_{\ell=1}^{\infty} \beta^\ell \cdot \lvert\text{paths}_{x,y}^{\langle\ell\rangle}\rvert$<br>where $\text{paths}_{x,y}^{\langle\ell\rangle}$ := {paths of length exactly $\ell$ from $x$ to $y$}<br>weighted: $\text{paths}_{x,y}^{\langle 1\rangle}$ := number of collaborations between $x$, $y$.<br>unweighted: $\text{paths}_{x,y}^{\langle 1\rangle}$ := 1 iff $x$ and $y$ collaborate. |
| hitting time<br>stationary-normed<br>commute time<br>stationary-normed | $-H_{x,y}$<br>$-H_{x,y} \cdot \pi_y$<br>$-(H_{x,y} + H_{y,x})$<br>$-H_{x,y} \cdot \pi_y + H_{y,x} \cdot \pi_x)$<br>where $H_{x,y}$ := expected time for random walk from $x$ to reach $y$<br>$\pi_y$ := stationary-distribution weight of $y$ (proportion of time the random walk is at node $y$) |
| rooted PageRank$_\alpha$ | stationary distribution weight of $y$ under the following random walk:<br>with probability $\alpha$, jump to $x$.<br>with probability $1 - \alpha$, go to a random neighbor of current node. |
| SimRank$_\gamma$ | $\begin{cases} 1 & \text{if } x = y \\ \gamma \cdot \dfrac{\Sigma_{a \in \Gamma(x)} \Sigma_{b \in \Gamma(y)} \text{score}(a, b)}{\lvert\Gamma(x)\rvert \cdot \lvert\Gamma(y)\rvert} & \text{otherwise} \end{cases}$ |

FIG. 2. Values for score($x$, $y$) under various predictors; each predicts pairs $\langle x, y \rangle$ in descending order of score($x$, $y$). The set $\Gamma(x)$ consists of the neighbors of the node $x$ in $G_{collab}$.


DOI: 10.1002/asi

approach follows the natural intuition that such nodes $x$ and $y$ represent authors who have many colleagues in common and hence who are more likely to come into contact themselves. Jin et al. (2001) and Davidsen et al. (2002) defined abstract models for network growth using this principle, in which an edge $\langle x, y\rangle$ is more likely to form if edges $\langle x, z\rangle$ and $\langle z, y\rangle$ are already present for some $z$.

*Common neighbors.* The most direct implementation of this idea for link prediction is to define $\mathsf{score}(x, y) := |\Gamma(x) \cap \Gamma(y)|$, the number of neighbors that $x$ and $y$ have in common. Newman (2001a) computed this quantity in the context of collaboration networks, verifying a correlation between the number of common neighbors of $x$ and $y$ at time $t$ and the probability that they will collaborate in the future.

*Jaccard's coefficient and Adamic/Adar.* The Jaccard coefficient—a commonly used similarity metric in information retrieval (Salton & McGill, 1983)—measures the probability that both $x$ and $y$ have a feature $f$, for a randomly selected feature $f$ that *either* $x$ or $y$ has. If we take "features" here to be neighbors in $G_{collab}$, this approach leads to the measure $\mathsf{score}(x,y) := |\Gamma(x) \cap \Gamma(y)| / |\Gamma(x) \cup \Gamma(y)|$.

Adamic and Adar (2003) considered a similar measure, in the context of deciding when two personal home pages are strongly "related." To do this, they computed features of the pages and defined the similarity between two pages to be

$$\sum_{z\,:\,\text{feature shared by } x,\,y} \frac{1}{\log(\mathsf{frequency}(z))}.$$

This quantity refines the simply counting of common features by weighting rarer features more heavily. This idea suggests the measure $\mathsf{score}(x, y) := \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log|\Gamma(z)|}$.

*Preferential attachment.* This has received considerable attention as a model of the growth of networks (Barabási & Albert, 1999; Mitzenmacher, 2004). The basic premise is that the probability that a new edge has node $x$ as an endpoint is proportional to $|\Gamma(x)|$, the current number of neighbors of $x$. Newman (2001a) and Barabási et al. (2002) further proposed, on the basis of empirical evidence, that the probability of coauthorship of $x$ and $y$ is correlated with the product of the number of collaborators of $x$ and $y$. This proposal corresponds to the measure $\mathsf{score}(x, y) := |\Gamma(x)| \cdot |\Gamma(y)|$.

### *Methods Based on the Ensemble of All Paths*

A number of methods refine the notion of shortest-path distance by implicitly considering the ensemble of *all* paths between two nodes.

*Katz (1953).* Katz defined a measure that directly sums over this collection of paths, exponentially damped by length to count short paths more heavily. This notion leads to the measure

$$\mathsf{score}(x, y) := \sum_{\ell=1}^{\infty} \beta^{\ell} \cdot |\mathsf{paths}_{x,y}^{\langle\ell\rangle}|,$$

where $\mathsf{paths}_{x,y}^{\langle\ell\rangle}$ is the set of all length-$\ell$ paths from $x$ to $y$, and $\beta > 0$ is a parameter of the predictor (A very small $\beta$ yields predictions much like common neighbors because paths of length 3 or more contribute very little to the summation.) One can verify that the matrix of scores is given by $(I - \beta M)^{-1} - I$, where $M$ is the adjacency matrix of the graph. We consider two variants of this Katz measure: (a) *unweighted*, in which $\mathsf{paths}_{x,y}^{\langle 1\rangle} = 1$ if $x$ and $y$ have collaborated and 0 otherwise, and (b) *weighted*, in which $\mathsf{paths}_{x,y}^{\langle 1\rangle}$ is the number of times that $x$ and $y$ have collaborated.

*Hitting time, PageRank, and variants.* A *random walk* on $G_{collab}$ starts at a node $x$ and iteratively moves to a neighbor of $x$ chosen uniformly at random from the set $\Gamma(x)$. The *hitting time* $H_{x,y}$ from $x$ to $y$ is the expected number of steps required for a random walk starting at $x$ to reach $y$. Because the hitting time is not in general symmetric, it also is natural to consider the *commute time* $C_{x,y} := H_{x,y} + H_{y,x}$. Both of these measures serve as natural proximity measures and hence (negated) can be used as $\mathsf{score}(x, y)$.

One difficulty with hitting time as a measure of proximity is that $H_{x,y}$ is quite small whenever $y$ is a node with a large *stationary probability* $\pi_y$, regardless of the identity of $x$. To counterbalance this phenomenon, we also consider *normalized* versions of the hitting and commute times, by defining $\mathsf{score}(x,y) := -H_{x,y} \cdot \pi_y$ or $\mathsf{score}(x,y) := -(H_{x,y} \cdot \pi_y + H_{y,x} \cdot \pi_x)$.

Another difficulty with these measures is their sensitive dependence to parts of the graph far away from $x$ and $y$, even when $x$ and $y$ are connected by very short paths. A way of counteracting this dependence is to allow the random walk from $x$ to $y$ to periodically "reset," returning to $x$ with a fixed probability $\alpha$ at each step; in this way, distant parts of the graph almost never will be explored. Random resets form the basis of the *PageRank* measure for Web pages (Brin & Page, 1998), and we can adapt it for link prediction as follows. Define $\mathsf{score}(x, y)$ under the *rooted PageRank* measure with parameter $\alpha \in [0, 1]$ to be the stationary probability of $y$ in a random walk that returns to $x$ with probability $\alpha$ each step, moving to a random neighbor with probability $1 - \alpha$. Similar approaches have been considered for *personalized PageRank*, in which one wishes to rank Web pages based both on their overall importance, the core of PageRank, and their relevance to a particular topic or individual, by biasing the random resets toward topically relevant or bookmarked pages (Haveliwala, 2003; Haveliwala, Kamvar, & Jeh, 2003; Jeh & Widom, 2003; Kamvar, Haveliwala, Manning, & Golub, 2003).

*SimRank (Jeh & Widom, 2002).* SimRank is a fixed point of the following recursive definition: Two nodes are similar to the extent that they are joined to similar neighbors. Numerically, this quantity is specified by defining $\mathsf{similarity}(x, x) := 1$ and


DOI: 10.1002/asi

$$\mathsf{similarity}(x, y) := \gamma \cdot \frac{\Sigma_{a \in \Gamma(x)} \Sigma_{b \in \Gamma(y)} \mathsf{similarity}(a, b)}{|\Gamma(x)| \cdot |\Gamma(y)|}$$

for a parameter $\gamma \in [0, 1]$. We then define $\mathsf{score}(x, y) := \mathsf{similarity}(x, y)$. SimRank also can be interpreted in terms of a random walk on the collaboration graph: It is the expected value of $\gamma^{\ell}$, where $\ell$ is a random variable giving the time at which random walks started from $x$ and $y$ first meet.

*Higher Level Approaches*

We now discuss three "meta-approaches" that can be used in conjunction with any of the methods discussed earlier.

*Low-rank approximation.* Because the adjacency matrix $M$ can be used to represent the graph $G_{collab}$, all of our link-prediction methods have an equivalent formulation in terms of this matrix $M$. In some cases, this correspondence was noted explicitly earlier (e.g., in the case of the Katz similarity score), but in many other cases the matrix formulation also is quite natural. For example, the common-neighbors method consists simply of mapping each node $x$ to its row $r(x)$ in $M$, and then defining $\mathsf{score}(x, y)$ to be the inner product of the rows $r(x)$ and $r(y)$.

A common general technique when analyzing the structure of a large matrix $M$ is to choose a relatively small number $k$ and compute the rank-$k$ matrix $M_k$ that best approximates $M$ with respect to any of a number of standard matrix norms. This computation can be done efficiently using the singular-value decomposition, and it forms the core of methods such as *latent semantic analysis* in information retrieval (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). Intuitively, working with $M_k$ rather than $M$ can be viewed as a type of "noise-reduction" technique that generates most of the structure in the matrix, but with a greatly simplified representation.

In our experiments, we investigate three applications of low-rank approximation: (a) ranking by the Katz measure, in which we use $M_k$ rather than $M$ in the underlying formula; (b) ranking by common neighbors, in which we score by inner products of rows in $M_k$ rather than $M$; and—most simply of all—(c) defining $\mathsf{score}(x, y)$ to be the $\langle x, y \rangle$ entry in the matrix $M_k$.

*Unseen bigrams.* Link prediction is akin to the problem of estimating frequencies for *unseen bigrams* in language modeling—pairs of words that co-occur in a test corpus, but not in the corresponding training corpus (e.g., see the work of Essen & Steinbiss, 1992). Following ideas proposed in that literature (e.g., Lee, 1999), we can augment our estimates for $\mathsf{score}(x, y)$ using values of $\mathsf{score}(z, y)$ for nodes $z$ that are "similar" to $x$. Specifically, we adapt this approach to the link-prediction problem as follows. Suppose we have values $\mathsf{score}(x, y)$ computed under one of the measures above. Let $S_x^{(\delta)}$ denote the $\delta$ nodes most related to $x$ under $\mathsf{score}(x, \cdot)$, for a parameter $\delta \in \mathbb{Z}^+$. We then define enhanced scores in terms of the nodes in this set:

$$\mathsf{score}^*_{unweighted}(x, y) := |\{z : z \in \Gamma(y) \cap S_x^{(\delta)}\}|$$

$$\mathsf{score}^*_{weighted}(x, y) := \Sigma_{z \in \Gamma(y) \cap S_x^{(\delta)}} \mathsf{score}(x, z).$$

*Clustering.* One might seek to improve on the quality of a predictor by deleting the more "tenuous" edges in $G_{collab}$ through a clustering procedure, and then running the predictor on the resulting "cleaned-up" subgraph. Consider a measure computing values for $\mathsf{score}(x, y)$. We compute $\mathsf{score}(u, v)$ for all edges in $E_{old}$, and delete the $(1-\rho)$ fraction of these edges for which the score is lowest, for a parameter $\rho \in [0, 1]$. We now recompute $\mathsf{score}(x, y)$ for all pairs $\langle x, y \rangle$ on this subgraph; in this way, we determine node proximities using only edges for which the proximity measure itself has the most confidence.

## Results and Discussion

As discussed in the first section, many collaborations form (or fail to form) for reasons outside the scope of the network; thus, the raw performance of our predictors is relatively low. To more meaningfully represent predictor quality, we use as our baseline a *random predictor*, which simply predicts randomly selected pairs of authors who did not collaborate in the training interval. The probability that a random prediction is correct is just the ratio between $|E_{new}|$, the number of possible correct predictions, and $\binom{|\mathsf{Core}|}{2} - |E_{old}|$, the number of possible predictions that can be made (Any pair chosen from the set Core of core authors is a legal prediction unless they had already collaborated, which occurs for $|E_{old}|$ pairs.) A random prediction is correct with probability between 0.15% (`cond-mat`) and 0.48% (`astro-ph`).

Figures 3 and 4 show each predictor's performance on each arXiv section, in terms of the factor improvement over random. One can use standard tail inequalities (e.g., see the text of Motwani & Raghavan, 1995) to show that the probability of a random predictor's performance exceeding its expectation by a factor of 5 is very small: this probability ranges from about 0.004 for `gr-qc` to about $10^{-48}$ for `astro-ph`. Thus, almost every predictor performs significantly better than random predictions on every dataset.

Figures 5, 6, and 7 show the average relative performance of several different predictors versus three baseline predictors—the random predictor, the graph-distance predictor, and the common-neighbors predictor. There is no single clear winner among the techniques, but we see that a number of methods significantly outperform the random predictor, suggesting that there is indeed useful information contained in the network topology alone. The Katz measure and its variants based on clustering and low-rank approximation perform consistently well; on three of the five arXiv sections, a variant of Katz achieves the best performance. Some of the very simple measures also perform surprisingly well, including common neighbors and the Adamic/Adar measure.


DOI: 10.1002/asi

| **Predictor** | | astro-ph | cond-mat | gr-qc | hep-ph | hep-th |
|---|---|---|---|---|---|---|
| probability that a random prediction is correct | | 0.475% | 0.147% | 0.341% | 0.207% | 0.153% |
| graph distance (all distance-2 pairs) | | *9.4* | *25.1* | *21.3* | *12.0* | *29.0* |
| common neighbors | | **18.0** | **40.8** | **27.1** | **26.9** | **46.9** |
| preferential attachment | | 4.7 | 6.0 | 7.5 | *15.2* | 7.4 |
| Adamic/Adar | | *16.8* | **54.4** | **30.1** | **33.2** | **50.2** |
| Jaccard | | *16.4* | **42.0** | 19.8 | **27.6** | *41.5* |
| SimRank | $\gamma = 0.8$ | *14.5* | *39.0* | *22.7* | *26.0* | *41.5* |
| hitting time | | 6.4 | 23.7 | *24.9* | 3.8 | 13.3 |
| hitting time—normed by stationary distribution | | 5.3 | 23.7 | 11.0 | 11.3 | 21.2 |
| commute time | | 5.2 | 15.4 | **33.0** | *17.0* | 23.2 |
| commute time—normed by stationary distribution | | 5.3 | 16.0 | 11.0 | 11.3 | 16.2 |
| rooted PageRank | $\alpha = 0.01$ | *10.8* | *27.8* | **33.0** | *18.7* | *29.1* |
| | $\alpha = 0.05$ | *13.8* | *39.6* | **35.2** | *24.5* | *41.1* |
| | $\alpha = 0.15$ | *16.6* | **40.8** | **27.1** | **27.5** | *42.3* |
| | $\alpha = 0.30$ | *17.1* | **42.0** | *24.9* | **29.8** | *46.5* |
| | $\alpha = 0.50$ | *16.8* | **40.8** | *24.2* | **30.6** | *46.5* |
| Katz (weighted) | $\beta = 0.05$ | 3.0 | 21.3 | 19.8 | 2.4 | 12.9 |
| | $\beta = 0.005$ | *13.4* | **54.4** | **30.1** | *24.0* | **51.9** |
| | $\beta = 0.0005$ | *14.5* | **53.8** | **30.1** | **32.5** | **51.5** |
| Katz (unweighted) | $\beta = 0.05$ | *10.9* | **41.4** | **37.4** | *18.7* | **47.7** |
| | $\beta = 0.005$ | *16.8* | **41.4** | **37.4** | *24.1* | **49.4** |
| | $\beta = 0.0005$ | *16.7* | **41.4** | **37.4** | *24.8* | **49.4** |

FIG. 3. Performance of various predictors on the link-prediction task defined in the section on Data and Experimental Setup. For each predictor and each arXiv section, the given number specifies the factor improvement over random prediction. Two predictors in particular are used as baselines for comparison: graph distance and common neighbors (see Methods for definitions). *Italicized* entries have performance at least as good as the graph-distance predictor; **bold** entries are at least as good as the common-neighbors predictor (see also Figure 4).

| **Predictor** | | astro-ph | cond-mat | gr-qc | hep-ph | hep-th |
|---|---|---|---|---|---|---|
| probability that a random prediction is correct | | 0.475% | 0.147% | 0.341% | 0.207% | 0.153% |
| graph distance (all distance-2 pairs) | | *9.4* | *25.1* | *21.3* | *12.0* | *29.0* |
| common neighbors | | **18.0** | **40.8** | **27.1** | **26.9** | **46.9** |
| Low-rank approximation: | rank = 1024 | *15.2* | **53.8** | **29.3** | **34.8** | **49.8** |
| Inner product | rank = 256 | *14.6* | **46.7** | **29.3** | **32.3** | **46.9** |
| | rank = 64 | *13.0* | **44.4** | **27.1** | **30.7** | **47.3** |
| | rank = 16 | *10.0* | 21.3 | **31.5** | **27.8** | *35.3* |
| | rank = 4 | 8.8 | 15.4 | **42.5** | *19.5* | 22.8 |
| | rank = 1 | 6.9 | 5.9 | **44.7** | *17.6* | 14.5 |
| Low-rank approximation: | rank = 1024 | 8.2 | 16.6 | 6.6 | *18.5* | 21.6 |
| Matrix entry | rank = 256 | *15.4* | *36.1* | 8.1 | *26.2* | *37.4* |
| | rank = 64 | *13.7* | **46.1** | 16.9 | **28.1** | *40.7* |
| | rank = 16 | 9.1 | 21.3 | *26.4* | *23.1* | *34.0* |
| | rank = 4 | 8.8 | 15.4 | **39.6** | *20.0* | 22.4 |
| | rank = 1 | 6.9 | 5.9 | **44.7** | *17.6* | 14.5 |
| Low-rank approximation: | rank = 1024 | *11.4* | *27.2* | **30.1** | **27.0** | *32.0* |
| Katz ($\beta = 0.005$) | rank = 256 | *15.4* | **42.0** | 11.0 | **34.2** | *38.6* |
| | rank = 64 | *13.1* | **45.0** | 19.1 | **32.2** | *41.1* |
| | rank = 16 | 9.2 | 21.3 | **27.1** | *24.8* | *34.9* |
| | rank = 4 | 7.0 | 15.4 | **41.1** | *19.7* | 22.8 |
| | rank = 1 | 0.4 | 5.9 | **44.7** | *17.6* | 14.5 |
| unseen bigrams | common neighbors, $\delta = 8$ | *13.5* | *36.7* | **30.1** | *15.6* | **46.9** |
| (weighted) | common neighbors, $\delta = 16$ | *13.4* | *39.6* | **38.9** | *18.5* | **48.6** |
| | Katz ($\beta = 0.005$), $\delta = 8$ | *16.8* | *37.9* | *24.9* | *24.1* | **51.1** |
| | Katz ($\beta = 0.005$), $\delta = 16$ | *16.5* | *39.6* | **35.2** | *24.7* | **50.6** |
| unseen bigrams | common neighbors, $\delta = 8$ | *14.1* | *40.2* | **27.9** | *22.2* | *39.4* |
| (unweighted) | common neighbors, $\delta = 16$ | *15.3* | *39.0* | **42.5** | *22.0* | *42.3* |
| | Katz ($\beta = 0.005$), $\delta = 8$ | *13.1* | *36.7* | **32.3** | *21.6* | *37.8* |
| | Katz ($\beta = 0.005$), $\delta = 16$ | *10.3* | *29.6* | **41.8** | *12.2* | *37.8* |
| clustering: | $\rho = 0.10$ | 7.4 | *37.3* | **46.9** | **32.9** | *37.8* |
| Katz ($\beta_1 = 0.001$, $\beta_2 = 0.1$) | $\rho = 0.15$ | *12.0* | **46.1** | **46.9** | *21.0* | *44.0* |
| | $\rho = 0.20$ | 4.6 | *34.3* | 19.8 | *21.2* | *35.7* |
| | $\rho = 0.25$ | 3.3 | *27.2* | 20.5 | *19.4* | 17.4 |

FIG. 4. Performance of various meta-approaches on the link-prediction task defined in the Data and Experimental Setup section. As before, for each predictor and each arXiv section, the given number specifies the factor improvement over random predictions (see Figure 3).


DOI: 10.1002/asi

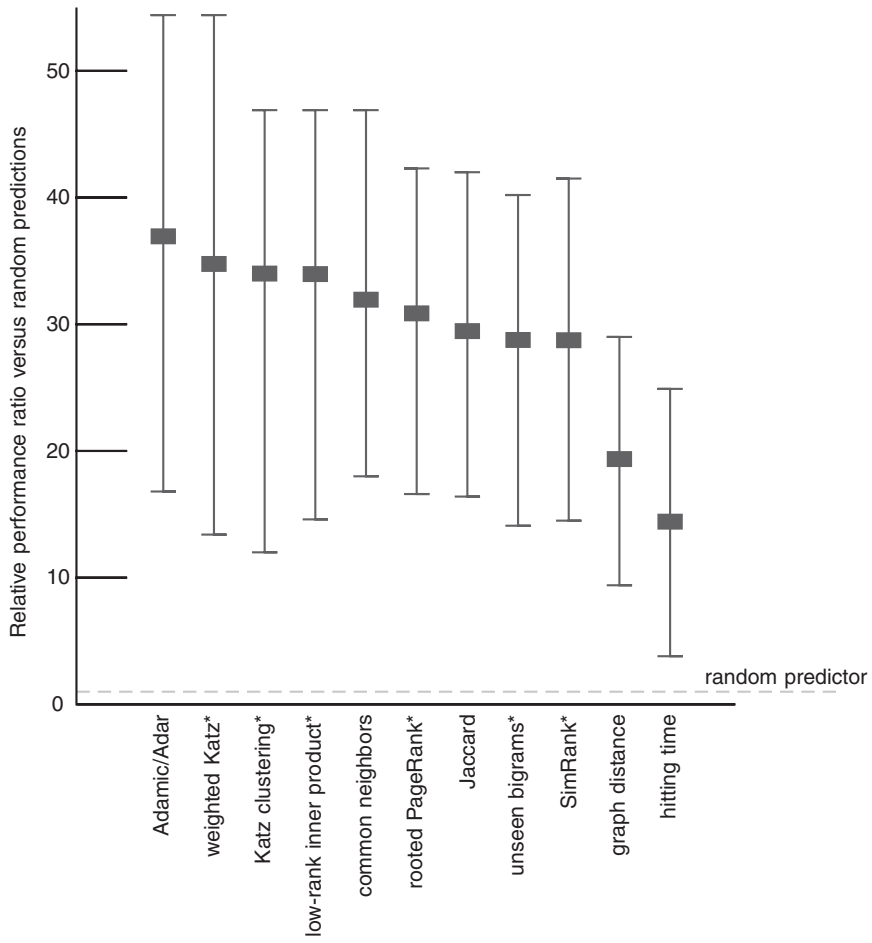



FIG. 5. Relative average performance of various predictors versus random predictions. The value shown is the average ratio over the five datasets of the given predictor's performance versus the random predictor's performance. The error bars indicate the minimum and maximum of this ratio over the five datasets. The parameters for the starred predictors are as follows: (a) for weighted Katz, $\beta = 0.005$; (b) for Katz clustering, $\beta_1 = 0.001$, $\rho = 0.15$, $\beta_2 = 0.1$; (c) for low-rank inner product, rank $= 256$; (d) for rooted Pagerank, $\alpha = 0.15$; (e) for unseen bigrams, unweighted common neighbors with $\delta = 8$; and (f) for SimRank, $\gamma = 0.8$.

*Similarities Among the Predictors and the Datasets*

Not surprisingly, there is significant overlap in the predictions made by the various methods. In Figure 8, we show the number of common predictions made by 10 of the most successful measures on the `cond-mat` graph. We see that Katz, low-rank inner product, and Adamic/Adar are quite similar in their predictions, as are (to a somewhat lesser extent) rooted PageRank, SimRank, and Jaccard. Hitting time is remarkably unlike any of the other nine in its predictions, despite its reasonable performance. The number of common *correct* predictions shows qualitatively similar behavior (see Figure 9). It would be interesting to understand the generality of these overlap phenomena, especially because certain of the large overlaps do not seem to follow obviously from the definitions of the measures.

It is harder to quantify the differences among the datasets, but their relationship is a very interesting issue as well. One perspective is provided by the methods based on low-rank approximation: On four of the datasets, their performance tends to be best at an intermediate rank while on `gr-qc` they perform best at rank 1 (see Figure 10, e.g., for a plot of the change in the performance of the low-rank matrix-entry predictor as the rank of the approximation varies). This fact suggests a sense in which the collaborations in `gr-qc` have a much "simpler" structure than those in the other four. One also observes the apparent importance of node degree in the `hep-ph` collaborations: The preferential-attachment predictor—which considers only the number (and not the identity) of a scientist's coauthors—does uncharacteristically well on this dataset, outperforming the basic graph-distance predictor. Finally, it would be interesting to make precise a sense in which `astro-ph` is a "difficult" dataset, given the low performance of all methods relative to random and the fact that none beats simple ranking by common neighbors. We will explore this issue further when we consider collaboration data drawn from other fields.

Because almost all of our experiments were carried out on social networks formed via the collaborations of physicists, it is difficult to draw broad conclusions about link prediction in social networks in general. The culture of physicists and of physics collaboration (e.g., see the work of Katz & Martin, 1997) plays a role in the quality of our results. The considerations discussed earlier suggest that there are some important differences even within physics


DOI: 10.1002/asi

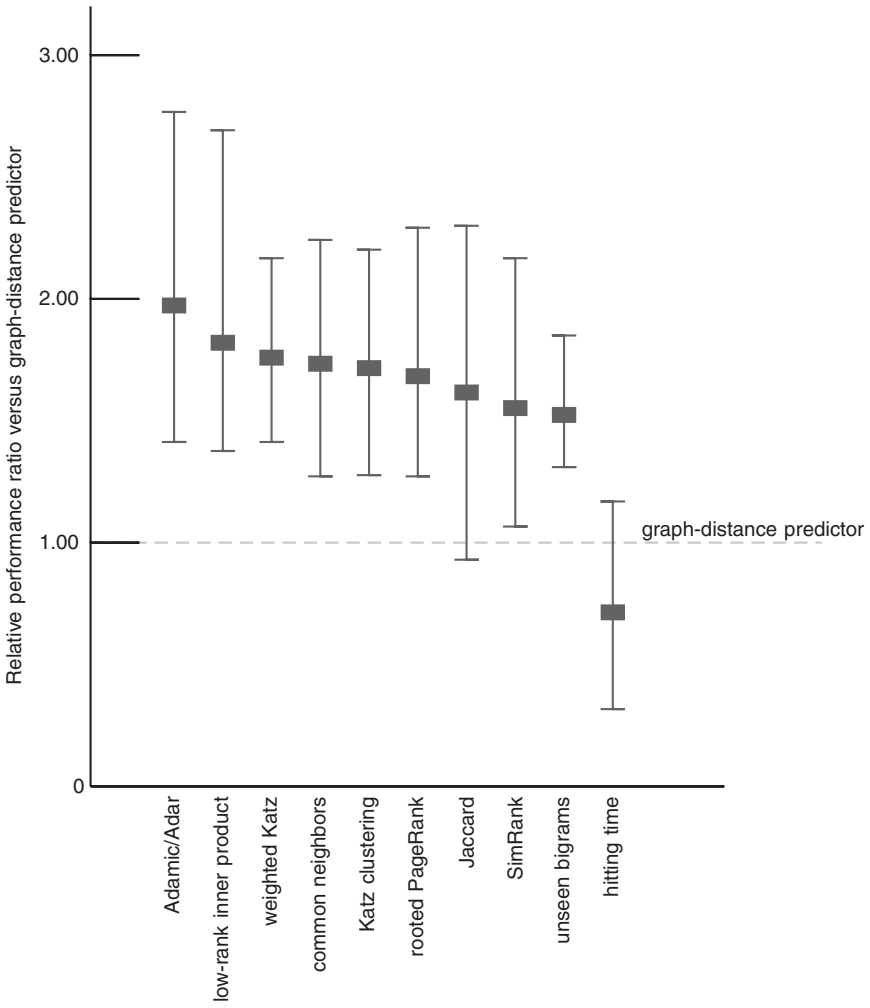



FIG. 6. Relative average performance of various predictors versus the graph-distance predictor. The plotted value shows the average taken over the five datasets of the ratio of the performance of the given predictor versus the graph-distance predictor; the error bars indicate the range of this ratio over the five datasets. All parameter settings are as in Figure 5.

(depending on the subfield), and an important area for future study is to understand how other social networks differ from the ones that we studied here.

### *Small Worlds*

It is reassuring that even the basic graph-distance predictor handily outperforms random predictions, but this measure has severe limitations. Extensive research has been devoted to understanding the so-called *small-world problem* in collaboration network; that is, accounting for the existence of short paths connecting virtually every pair of scientists (Newman, 2001b). This property is normally viewed as a vital fact about the scientific community (New ideas spread quickly, and every discipline interacts with—and gains from—other fields.), but in the context of our prediction task, we come to a different conclusion: The small-world problem is really a problem. The shortest path between two scientists in wholly unrelated disciplines is often very short (and very tenuous). To take one particular, but not atypical, example, the developmental psychologist Jean Piaget has as small an Erdös number—3 (Castro & Grossman, 1999)—as most mathematicians and computer scientists. Overall, the basic graph-distance predictor is not competitive with most of the other approaches studied; our most successful link predictors can be viewed as using measures of proximity that are robust to the few edges that result from rare collaborations between fields.

### *Restricting to Distance 3*

The small-world problem suggests that there are many pairs of authors separated by a graph distance of 2 who will not collaborate, but we also observe the dual problem: Many pairs who collaborate are at distance greater than 2. Between 71 `(hep-ph)` and 83% `(cond-mat)` of new edges form between pairs at distance 3 or greater (see Figure 11).

Because most new collaborations are not at distance 2, we also are interested in how well our predictors perform when we disregard all distance-2 pairs. Clearly, nodes at distances greater than 2 have no neighbors in common, and hence this


DOI: 10.1002/asi

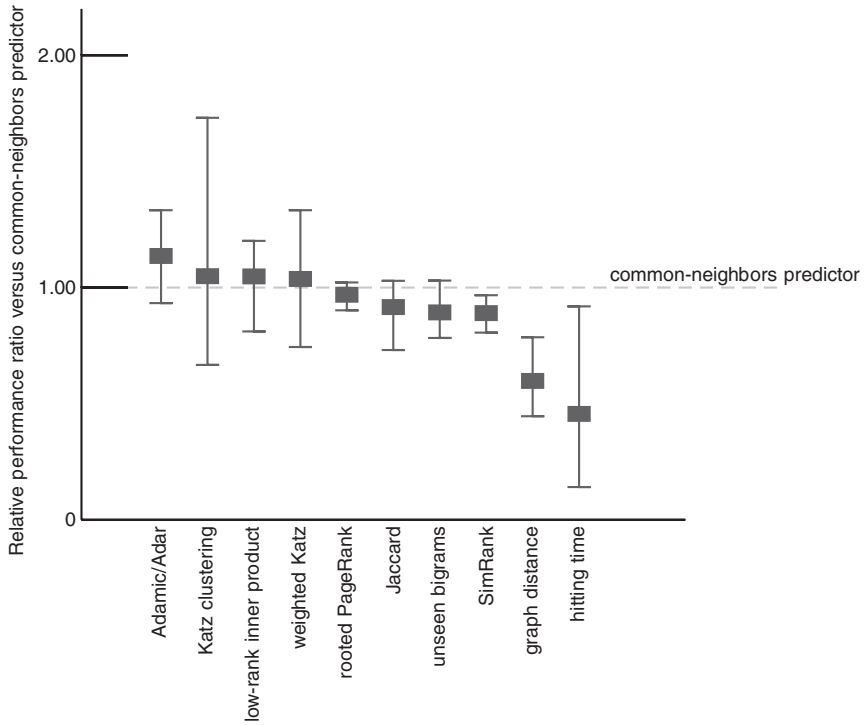



FIG. 7. Relative average performance of various predictors versus the common-neighbors predictor, as in Figure 6. Error bars display the range of the performance ratio of the given predictor versus common neighbors over the five datasets; the displayed value gives the average ratio. Parameter settings are as in Figure 5.

| | Adamic/ Adar | Katz clustering | common neighbors | hitting time | Jaccard's coefficient | weighted Katz | low-rank inner product | rooted Pagerank | SimRank | unseen bigrams |
|---|---|---|---|---|---|---|---|---|---|---|
| Adamic/Adar | 1,150 | 638 | 520 | 193 | 442 | 1,011 | 905 | 528 | 372 | 486 |
| Katz clustering | | 1,150 | 411 | 182 | 285 | 630 | 623 | 347 | 245 | 389 |
| common neighbors | | | 1,150 | 135 | 506 | 494 | 467 | 305 | 332 | 489 |
| hitting time | | | | 1,150 | 87 | 191 | 192 | 247 | 130 | 156 |
| Jaccard's coefficient | | | | | 1,150 | 414 | 382 | 504 | 845 | 458 |
| weighted Katz | | | | | | 1,150 | 1,013 | 488 | 344 | 474 |
| low-rank inner product | | | | | | | 1,150 | 453 | 320 | 448 |
| rooted Pagerank | | | | | | | | 1,150 | 678 | 461 |
| SimRank | | | | | | | | | 1,150 | 423 |
| unseen bigrams | | | | | | | | | | 1,150 |

FIG. 8. The number of common predictions made by various predictors on the `cond-mat` dataset, of 1,150 predictions. Parameter settings are as in Figure 5.

| | Adamic/ Adar | Katz clustering | common neighbors | hitting time | Jaccard's coefficient | weighted Katz | low-rank inner product | rooted Pagerank | SimRank | unseen bigrams |
|---|---|---|---|---|---|---|---|---|---|---|
| Adamic/Adar | 92 | 65 | 53 | 22 | 43 | 87 | 72 | 44 | 36 | 49 |
| Katz clustering | | 78 | 41 | 20 | 29 | 66 | 60 | 31 | 22 | 37 |
| common neighbors | | | 69 | 13 | 43 | 52 | 43 | 27 | 26 | 40 |
| hitting time | | | | 40 | 8 | 22 | 19 | 17 | 9 | 15 |
| Jaccard's coefficient | | | | | 71 | 41 | 32 | 39 | 51 | 43 |
| weighted Katz | | | | | | 92 | 75 | 44 | 32 | 51 |
| low-rank inner product | | | | | | | 79 | 39 | 26 | 46 |
| rooted Pagerank | | | | | | | | 69 | 48 | 39 |
| SimRank | | | | | | | | | 66 | 34 |
| unseen bigrams | | | | | | | | | | 68 |

FIG. 9. The number of *correct* common predictions made by various predictors on the `cond-mat` dataset, of 1,150 predictions. The diagonal entries indicate the number of correct predictions for each predictor. Parameter settings are as in Figure 5.


DOI: 10.1002/asi

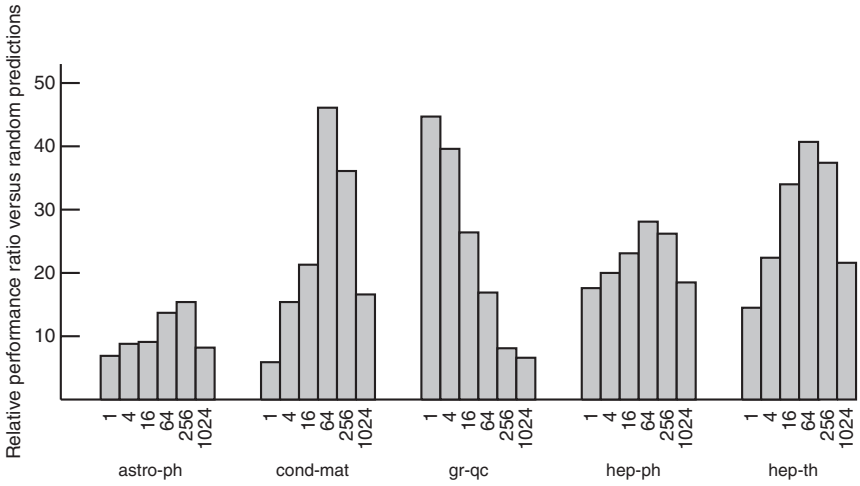



FIG. 10. Relative performance of the low-rank matrix-entry predictor for various ranks on each arXiv section. For each arXiv section, the performance of this predictor, measured by the factor of improvement over random predictions, is shown for ranks 1, 4, 16, 64, 256, and 1024. Notice that for all arXiv sections save `gr-qc`, predictor performance is maximized by an intermediate rank; for that dataset, performance continues to improve as the rank decreases all the way to rank 1.

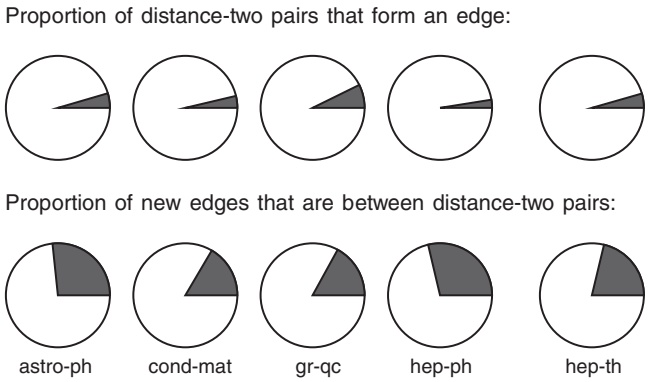
task essentially rules out the use of methods based on common neighbors. The performance of the other measures is shown in Figure 12. The graph-distance predictor (i.e., predicting all distance-3 pairs) performs between about three and nine times random and is consistently beaten by virtually all of the predictors: SimRank, rooted PageRank, Katz, and the low-rank and unseen-bigram techniques. The unweighted Katz and unseen-bigram predictors have the best performance (as high as about 30 times random, on `gr-qc`), followed closely by weighted Katz, SimRank, and rooted PageRank.

### *The Breadth of the Data*

We also have considered three other datasets: (a) the proceedings of two conferences in theoretical computer science, *Symposium on the Theory of Computing* (*STOC*) and *Foundations of Computer Science* (*FOCS*), (b) the articles found in the Citeseer (`www.citeseer.com`) online database, which finds articles by crawling the Web for any files in postscript format, and (c) all five of the arXiv sections merged into one. Consider the performance of the common-neighbors predictor compared to random on these datasets:

| STOC/FOCS | arXiv sections | combined arXiv sections | Citeseer |
|---|---|---|---|
| 6.1 | 18.0–46.9 | 71.2 | 147.0 |

Performance versus random swells dramatically as the topical focus of our dataset widens. That is, when we consider a more diverse collection of scientists, it is fundamentally easier




| | astro-ph | cond-mat | gr-qc | hep-ph | hep-th |
|---|---|---|---|---|---|
| No. pairs at distance two | 33,862 | 5,145 | 935 | 37,687 | 7,545 |
| No. new collaborations at distance two | 1,533 | 190 | 68 | 945 | 335 |
| No. new collaborations | 5,751 | 1,150 | 400 | 3,294 | 1,576 |

FIG. 11. Relationship between new collaborations and graph distance.


DOI: 10.1002/asi

| **Predictor** | | astro-ph | cond-mat | gr-qc | hep-ph | hep-th |
|---|---|---|---|---|---|---|
| graph distance (all distance-three pairs) | | 2.8 | 5.4 | 7.7 | 4.0 | 8.6 |
| preferential attachment | | 3.2 | 2.6 | 8.6 | 4.7 | 1.4 |
| SimRank | $\gamma = 0.8$ | 5.9 | 14.3 | 10.6 | 7.6 | 21.9 |
| hitting time | | 4.4 | 10.1 | 13.7 | 4.5 | 4.7 |
| hitting time—normed by stationary distribution | | 2.0 | 2.5 | 0.0 | 2.5 | 6.6 |
| commute time | | 3.8 | 5.9 | 21.1 | 5.9 | 6.6 |
| commute time—normed by stationary distribution | | 2.6 | 0.8 | 1.1 | 4.8 | 4.7 |
| rooted PageRank | $\alpha = 0.01$ | 4.6 | 12.7 | 21.1 | 6.5 | 12.6 |
| | $\alpha = 0.05$ | 5.3 | 13.5 | 21.1 | 8.7 | 16.6 |
| | $\alpha = 0.15$ | 5.4 | 11.8 | 18.0 | 10.7 | 19.9 |
| | $\alpha = 0.30$ | 5.8 | 13.5 | 8.4 | 11.6 | 19.9 |
| | $\alpha = 0.50$ | 6.3 | 15.2 | 7.4 | 12.7 | 19.9 |
| Katz (weighted) | $\beta = 0.05$ | 1.5 | 5.9 | 11.6 | 2.3 | 2.7 |
| | $\beta = 0.005$ | 5.5 | 14.3 | 28.5 | 4.2 | 12.6 |
| | $\beta = 0.0005$ | 6.2 | 13.5 | 27.5 | 4.2 | 12.6 |
| Katz (unweighted) | $\beta = 0.05$ | 2.3 | 12.7 | 30.6 | 9.0 | 12.6 |
| | $\beta = 0.005$ | 9.1 | 11.8 | 30.6 | 5.1 | 17.9 |
| | $\beta = 0.0005$ | 9.2 | 11.8 | 30.6 | 5.1 | 17.9 |
| Low-rank approximation: | rank = 1024 | 2.3 | 2.5 | 9.5 | 4.0 | 6.0 |
| Inner product | rank = 256 | 4.8 | 5.9 | 5.3 | 9.9 | 10.6 |
| | rank = 64 | 3.8 | 12.7 | 5.3 | 7.1 | 11.3 |
| | rank = 16 | 5.3 | 6.7 | 6.3 | 6.8 | 15.3 |
| | rank = 4 | 5.1 | 6.7 | 32.7 | 2.0 | 4.7 |
| | rank = 1 | 6.1 | 2.5 | 32.7 | 4.2 | 8.0 |
| Low-rank approximation: | rank = 1024 | 4.1 | 6.7 | 6.3 | 5.9 | 13.3 |
| Matrix entry | rank = 256 | 3.8 | 8.4 | 3.2 | 8.5 | 19.9 |
| | rank = 64 | 2.9 | 11.8 | 2.1 | 4.0 | 10.0 |
| | rank = 16 | 4.4 | 8.4 | 4.2 | 5.9 | 16.6 |
| | rank = 4 | 4.9 | 6.7 | 27.5 | 2.0 | 4.7 |
| | rank = 1 | 6.1 | 2.5 | 32.7 | 4.2 | 8.0 |
| Low-rank approximation: | rank = 1024 | 4.3 | 6.7 | 28.5 | 5.9 | 13.3 |
| Katz ($\beta = 0.005$) | rank = 256 | 3.6 | 8.4 | 3.2 | 8.5 | 20.6 |
| | rank = 64 | 2.8 | 11.8 | 2.1 | 4.2 | 10.6 |
| | rank = 16 | 5.0 | 8.4 | 5.3 | 5.9 | 15.9 |
| | rank = 4 | 5.2 | 6.7 | 28.5 | 2.0 | 4.7 |
| | rank = 1 | 0.3 | 2.5 | 32.7 | 4.2 | 8.0 |
| unseen bigrams (weighted) | common neighbors, $\delta = 8$ | 5.8 | 6.7 | 14.8 | 4.2 | 23.9 |
| | common neighbors, $\delta = 16$ | 7.9 | 9.3 | 28.5 | 5.1 | 19.3 |
| | Katz ($\beta = 0.005$), $\delta = 18$ | 5.2 | 10.1 | 22.2 | 2.8 | 17.9 |
| | Katz ($\beta = 0.005$), $\delta = 16$ | 6.6 | 10.1 | 29.6 | 3.7 | 15.3 |
| unseen bigrams (unweighted) | common neighbors, $\delta = 8$ | 5.4 | 5.1 | 13.7 | 4.5 | 21.3 |
| | common neighbors, $\delta = 16$ | 6.3 | 8.4 | 25.3 | 4.8 | 21.9 |
| | Katz ($\beta = 0.005$), $\delta = 8$ | 4.1 | 7.6 | 22.2 | 2.0 | 17.3 |
| | Katz ($\beta = 0.005$), $\delta = 16$ | 4.3 | 4.2 | 28.5 | 3.1 | 16.6 |
| clustering: | $\rho = 0.10$ | 3.2 | 4.2 | 31.7 | 7.1 | 8.6 |
| Katz ($\beta_1 = 0.001, \beta_2 = 0.1$) | $\rho = 0.15$ | 4.6 | 4.2 | 32.7 | 7.6 | 6.6 |
| | $\rho = 0.20$ | 2.3 | 5.9 | 7.4 | 4.5 | 8.0 |
| | $\rho = 0.25$ | 2.0 | 11.8 | 6.3 | 6.8 | 5.3 |

FIG. 12. The distance-three task: performance of predictors only on edges in $E_{new}$ for which the endpoints were at distance three or more in $G_{collab}$. Methods based on common neighbors are not appropriate for this task (see Results and Discussion).


DOI: 10.1002/asi

to group scientists into fields of study (and therefore outperform the random predictor, which usually will make guesses between fields). When we consider a sufficiently narrow set of researchers (e.g., STOC/FOCS), almost any author can collaborate with almost any other author, and there seems to be a strong random component to new collaborations (In extensive experiments on the STOC/FOCS data, we could not beat random guessing by a factor of more than about 7.) It is an interesting challenge to formalize the sense in which the STOC/FOCS collaborations are truly intractable to predict; that is, to what extent information about new collaborations is simply not present in the old collaboration data.

### *Future Directions*

While the predictors that we have discussed perform reasonably well, even the best (Katz clustering on `gr-qc`) is correct on only about 16% of its predictions. There is clearly much room for improvement in performance on this task, and finding ways to take better advantage of the information in the training data is an interesting open question. Another issue is to improve the efficiency of the proximity-based methods on very large networks; fast algorithms for approximating the distribution of node-to-node distances may be one approach (Palmer, Gibbons, & Faloutsos, 2002).

The graph $G_{collab}$ is a lossy representation of the data; we also can consider a bipartite collaboration graph $B_{collab}$, with a vertex for every author and article, and an edge connecting each article to each of its authors. The bipartite graph contains more information than $G_{collab}$, so we may hope that predictors can use it to improve performance. The size of $B_{collab}$ is much larger than that of $G_{collab}$, making experiments prohibitive, but we have tried using the SimRank and Katz predictors on smaller datasets (`gr-qc`, or shorter training periods). Their performance does not seem to improve, but perhaps other predictors can fruitfully exploit the additional information in $B_{collab}$.

Similarly, our experiments treat all training-period collaborations equally. Perhaps one can improve performance by treating more recent collaborations as more important than older ones. One also could tune the parameters of the Katz predictor, for example, by dividing the training set into temporal segments, training $\beta$ on the beginning, and then using the end of the training set to make final predictions.

One also might try to use additional information such as the titles of articles or the institutional affiliations of the authors to identify the specific research area or geographic location of each scientist, and then use areas/locations to predict collaborations. In the field of bibliometrics, for example, Katz (1994), Melin and Persson (1996), and Ding, Foo, and Chowdhury (1999), among others, have observed institutional and geographic correlations in collaboration; a natural further direction would be to attempt to use geographic location, for instance, as a component of a predictor. To some extent, such geographic information, or indeed any other relevant properties of the nodes, is latently present in the graph $G_{collab}$—precisely because such factors already have played a role in the formation of old edges in the training set; however, direct access to such information may well confer additional predictive power, and it is an interesting open question to better understand the strength of such information in link prediction.

Finally, there has been relevant work in the machine-learning community on *estimating distribution support* (Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 1999): Given samples from an unknown probability distribution $P$, we must find a "simple" set $S$ so that $\Pr_{x \sim P}[x \notin S] < \varepsilon$. We can view training-period collaborations as samples drawn from a probability distribution on pairs of scientists; our goal is to approximate the set of pairs that have positive probability of collaborating. There also has been some potentially relevant work in machine learning on classification when the training set consists only of a relatively small set of positively labeled examples and a large set of unlabeled examples, with no labeled negative examples (Yu, Zhai, & Han, 2003). It is an open question whether these techniques can be fruitfully applied to the link-prediction problem.

## Acknowledgments

We thank Jon Herzog, Tommi Jaakkola, David Karger, Lillian Lee, Frank McSherry, Mike Schneider, Grant Wang, and Robert Wehr for helpful discussions and comments on earlier drafts of this article. We thank Paul Ginsparg for generously providing the bibliographic data from the arXiv.

An abbreviated preliminary version of this article appears in the *Proceedings of the 12th Annual ACM International Conference on Information and Knowledge Management* (*CIKM'03*), November 2003, pp. 556–559.

David Liben-Nowell was supported in part by an NSF Graduate Research Fellowship. Jon Kleinberg was supported in part by a David and Lucile Packard Foundation Fellowship and NSF ITR Grant IIS-0081334.

## References

Adamic, L.A., & 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, 509–512.

Barabási, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaboration. Physica A, 311(3–4), 590–614.

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.

Castro, R.D., & Grossman, J.W. (1999). Famous trails to Paul Erdös. Mathematical Intelligencer, 21(3), 51–63.

Davidsen, J., Ebel, H., & Bornholdt, S. (2002). Emergence of a small world from local interactions: Modeling acquaintance networks. Physical Review Letters, 88(128701).

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407.

Ding, Y., Foo, S., & Chowdhury, G. (1999). A bibliometric analysis of collaboration in the field of information retrieval. International Information and Library Review, 30, 367–376.


DOI: 10.1002/asi

Essen, U., & Steinbiss, V. (1992). Cooccurrence smoothing for stochastic language modeling. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 1, pp. 161–164). IEEE Computer Society.

Goldberg, D.S., & Roth, F.P. (2003). Assessing experimentally derived interactions in a small world. Proceedings of the National Academy of Sciences, 100(8), 4372–4376.

Grossman, J.W. (2002). The evolution of the mathematical research collaboration graph. Congressus Numerantium, 158, 201–212.

Haveliwala, T., Kamvar, S., & Jeh, G. (2003). An analytical comparison of approaches to personalizing PageRank (Technical Report). Stanford, CA: Stanford University.

Haveliwala, T.H. (2003). Topic-sensitive PageRank: A context-sensitive ranking algorithm for Web search. IEEE Transactions on Knowledge and Data Engineering, 15(4), 784–796.

Jeh, G., & Widom, J. (2002). SimRank: A measure of structural-context similarity. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 271–279). New York: ACM Press.

Jeh, G., & Widom, J. (2003). Scaling personalized Web search. In Proceedings of the 12th International World Wide Web Conference (WWW12) (pp. 271–279). New York: ACM Press.

Jin, E.M., Girvan, M., & Newman, M.E.J. (2001). The structure of growing social networks. Physical Review Letters E, 64(046132).

Kamvar, S.D., Haveliwala, T.H., Manning, C.D., & Golub, G.H. (2003). Exploiting the block structure of the Web for computing PageRank (Technical Report). Stanford, CA: Stanford University.

Katz, J.S. (1994). Geographic proximity and scientific collaboration. Scientometrics, 31(1), 31–43.

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

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

Kautz, H., Selman, B., & Shah, M. (1997). ReferralWeb: Combining social networks and collaborative filtering. Communications of the ACM, 40(3), 63–65.

Krebs, V. (2002). Mapping networks of terrorist cells. Connections, 24(3), 43–52.

Lee, L. (1999). Measures of distributional similarity. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (pp. 25–32). Morristown, NJ: Association for Computational Linguistics.

Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. Scientometrics, 36(3), 363–377.

Mitzenmacher, M. (2004). A brief history of lognormal and power law distributions. Internet Mathematics, 1(2), 226–251.

Motwani, R., & Raghavan, P. (1995). Randomized algorithms. New York: Cambridge University Press.

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

Newman, M.E.J. (2001b). The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences, 98, 404–409.

Newman, M.E.J. (2002). The structure and function of networks. Computer Physics Communications, 147, 40–45.

Newman, M.E.J. (2003). The structure and function of complex networks. SIAM Review, 45, 167–256.

Palmer, C., Gibbons, P., & Faloutsos, C. (2002). ANF: A fast and scalable tool for data mining in massive graphs. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 81–90). New York: ACM Press.

Popescul, A., & Ungar, L. (2003). Statistical relational learning for link prediction. In Workshop on Learning Statistical Models From Relational Data at the International Joint Conference on Artificial Intelligence (pp. 81–90). New York: ACM Press.

Raghavan, P. (2002). Social networks: From the Web to the enterprise. IEEE Internet Computing, 6(1), 91–94.

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

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., & Williamson, R.C. (1999). Estimating the support of a high-dimensional distribution (Technical Report. No. MSR-TR-99–87). Microsoft Research.

Taskar, B., Wong, M.-F., Abbeel, P., & Koller, D. (2003). Link prediction in relational data. In Proceedings of Neural Information Processing Systems (pp. 659–666). Cambridge, MA: MIT Press.

Watts, D.J. (1999). Small worlds. Princeton, NJ: Princeton University Press.

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

Yu, H., Zhai, C., & Han, J. (2003). Text classification from positive and unlabeled documents. In Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM'03) (pp. 232–239). New York: ACM Press.


DOI: 10.1002/asi