The Link Prediction Problem for Social Networks

David Liben-Nowell[*] Jon Kleinberg[†]

ABSTRACT

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 develop approaches to link prediction based on measures of the "proximity" of nodes in a network. Experiments on large co-authorship 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.

General Terms

Algorithms

Keywords

Social networks, link analysis, link prediction

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining; J.4 [**Social and Behavioral Sciences**]: Sociology; G.2.2 [**Graph Theory**]: Network Problems

1. 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 co-authored papers; 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 availability of large, detailed datasets encoding such networks has stimulated extensive study of their properties and the identification of recurring structural features. (For a thorough recent survey, see [11].)

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. Understanding 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 co-authorship network among scientists, for example. There are many reasons, exogenous to the network, why two scientists who have never written a paper 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 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 paper is in the area of network evolution models. While there has been a proliferation of such models in recent years (again see [11]), they 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 natural basis for 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 [10],

[*] Laboratory for Computer Science, Massachusetts Institute of Technology. Email: dln@theory.lcs.mit.edu. Supported in part by an NSF Graduate Research Fellowship.

[†] Department of Computer Science, Cornell University. Email: kleinber@cs.cornell.edu. Supported in part by a David and Lucile Packard Foundation Fellowship and NSF ITR Grant IIS-0081334.

CIKM'03, November 3–8, 2003, New Orleans, Louisiana, USA.
Copyright 2003 ACM 1-58113-723-0/03/0011 ...$5.00.

who considers the correlation between certain network growth models and data on the appearance of edges of co-authorship networks. Concurrently with the present work, Popescul and Ungar [13] have also investigated a related formulation of the link prediction problem.

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 AI and data mining have argued that a large organization, such as a company, can benefit from the interactions within the informal social network among its members; these serve to supplement the official hierarchy imposed by the organization itself [8, 14]. Effective methods for link prediction could be used to analyze such a social network, and suggest promising interactions that have not yet been utilized 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 interacting even though their interaction has not been directly observed.

2. DATA AND EXPERIMENTAL SETUP

We model a social network as a graph $G = \langle V, E \rangle$ in which each edge $e \in E$ represents an interaction between its endpoints at a particular time $t(e)$. We record multiple interactions by parallel edges with different time-stamps. For times $t < t'$, let $G[t, t']$ denote the subgraph of G restricted to edges with time-stamps between t and t'. To formulate the link prediction problem, we choose a *training interval* $[t_0, t_0']$ and a *test interval* $[t_1, t_1']$ where $t_0' < t_1$, 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 $G[t_1, t_1']$.

For our experiments, we use co-authorship networks G obtained from papers found in five sections of the physics e-Print arXiv, `www.arxiv.org`. (See Figure 1.) Occasional syntactic anomalies were handled heuristically, and authors were identified by first initial and last name; this appears to introduce only a small amount of error due to ambiguous identifiers. Our training interval is the period $[1994, 1996]$, and the test interval is $[1997, 1999]$. Denote the training interval subgraph $G[1994, 1996]$ by $G_{collab} := \langle A, E_{old} \rangle$, and let E_{new} denote the set of edges $\langle u, v \rangle$ where u and v co-author a paper during the test interval but not the training interval—these are the new interactions we are seeking to predict.

In evaluating link prediction methods, we focus on links between authors who have each written at least a minimum number of papers: we define the set Core to be all nodes incident to at least κ_{train} edges in the training interval and at least κ_{test} edges in the test interval, where κ_{train} and κ_{test} are both set to 3. Each link predictor p outputs a ranked list L_p of pairs in $A \times A - E_{old}$; these are predicted new collaborations, in decreasing order of confidence. Define $E^*_{new} := E_{new} \cap (\mathsf{Core} \times \mathsf{Core})$ 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 in $\mathsf{Core} \times \mathsf{Core}$, and determine the size of the intersection of this set of pairs with the set E^*_{new}.

3. METHODS FOR LINK PREDICTION

In this section, we survey an array of methods for link prediction. Each assigns a connection weight $\mathsf{score}(x, y)$ to pairs of nodes, producing a ranked list in decreasing order of $\mathsf{score}(x, y)$. A predictor can thus be viewed as computing a measure of proximity or "similarity" between nodes x and y, relative to the network topology. These predictors are adapted from techniques used in graph theory and social network analysis, and many must be modified from their original purposes to measure node-to-node similarity.

	training period			Core		
	auths.	papers	edges	auths.	$\lvert E_{old}\rvert$	$\lvert E_{new}\rvert$
`astro-ph`	5343	5816	41852	1561	6178	5751
`cond-mat`	5469	6700	19881	1253	1899	1150
`gr-qc`	2122	3287	5724	486	519	400
`hep-ph`	5414	10254	17806	1790	6654	3294
`hep-th`	5241	9498	15842	1438	2311	1576

Figure 1: ArXiv sections from which networks were constructed: astrophysics, condensed matter, general relativity/quantum cosmology, and high energy physics (phenomenology and theory).

Perhaps the most basic approach is to rank pairs by the length of the shortest path between them in G_{collab}. Such a measure follows the notion that collaboration networks are "small worlds," in which individuals are related through short chains [11]. (We predict a random subset of pairs at distance two in G_{collab}; distance-one pairs are edges in the training set E_{old}.)

Methods based on node neighborhoods. For a node x, let $\Gamma(x)$ be the set of neighbors of x in G_{collab}. Several approaches are based on the idea that two nodes x and y are more likely to form a link if $\Gamma(x)$ and $\Gamma(y)$ have large overlap; this follows the natural intuition that such node pairs represent authors with many colleagues in common, and hence are more likely to come into contact themselves [6].

• *Common neighbors.* One can directly use this idea by setting $\mathsf{score}(x, y) := |\Gamma(x) \cap \Gamma(y)|$, the number of common neighbors of x and y. In collaboration networks, Newman [10] has verified 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, commonly used in information retrieval [15], measures the number of features that *both* x and y have compared to the number of features that *either* x or y has. Taking "features" as neighbors in G_{collab}, this leads to $\mathsf{score}(x, y) := |\Gamma(x) \cap \Gamma(y)| / |\Gamma(x) \cup \Gamma(y)|$. Adamic and Adar [1] consider a related measure, in the context of deciding when two personal home pages are strongly "related." They compute features of the pages, and define the similarity between two pages to be $\sum_{z:\ \text{feature shared by } x,\, y} \frac{1}{\log(\mathsf{frequency}(z))}$. This refines the simple counting of common features by weighting rarer features more heavily. This suggests the measure $\mathsf{score}(x, y) := \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$.

• *Preferential attachment* has received considerable attention as a model of network growth [11]. The basic premise is that the probability a new edge involves node x is proportional to $|\Gamma(x)|$. Barabasi et al. [2] and Newman [10] have further proposed, on the basis of empirical evidence, that the probability of co-authorship of x and y is correlated with the product of the number of collaborators of x and y, corresponding 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 [7]* defines a measure that directly sums over this collection of paths, exponentially damped by length to count short paths more heavily. This leads to the measure $\mathsf{score}(x, y) := \sum_{\ell=1}^{\infty} \beta^{\ell} \cdot |\mathsf{paths}^{\langle \ell \rangle}_{x,y}|$, where $\mathsf{paths}^{\langle \ell \rangle}_{x,y}$ is the set of all length-ℓ paths from x to y. 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 *weighted* Katz, where $\mathsf{paths}^{\langle 1 \rangle}_{x,y} = \ell$ if there are ℓ parallel edges $\langle x, y \rangle$, and *unweighted* Katz, where parallel edges are ignored.

• *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. 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. We also consider the symmetric *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 is that $H_{x,y}$ is quite small whenever y is a node with a large *stationary probability* π_y, regardless of the identity of x. Thus we also consider *normalized* measures $-H_{x,y} \cdot \pi_y$ or $-(H_{x,y} \cdot \pi_y + H_{y,x} \cdot \pi_x)$. 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 is to allow the random walk from x to y to periodically "reset," returning to x with a fixed probability α at each step; in this way, distant parts of the graph will almost never be explored. Random resets form the basis of the *PageRank* measure for Web pages [3], and we can adapt it for link prediction as follows: Define the *rooted PageRank* measure to be the stationary probability of y in a random walk that returns to x with probability α each step, moving to a random neighbor with probability $1 - \alpha$.

• *SimRank* [5] is a fixed point of the following recursive definition: two nodes are similar insofar as they are joined to similar neighbors. Numerically, we define $\mathsf{score}(x,x) := 1$ and, for some $\gamma \in [0,1]$,

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

SimRank can be interpreted in terms of a random walk on G_{collab}: it is the expected value of γ^{ℓ}, where ℓ 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 above methods.

• *Low-rank approximation.* All our link prediction methods can be formulated in terms of the adjacency matrix M. For example, common neighbors of two nodes can be computed as the inner product between the two corresponding rows of M. A common technique when analyzing a large matrix M is to choose a relatively small number k and compute the rank-k matrix M_k that best approximates M under any of a number of standard matrix norms. This can be done efficiently using the singular value decomposition, and it forms the core of methods like *latent semantic analysis* [4]. Intuitively, this can be viewed as a type of "noise-reduction" technique that preserves most of the structure in the matrix. We consider three applications of low-rank approximation: (i) the Katz measure, using M_k rather than M in the underlying formula; (ii) common neighbors, using inner products of rows in M_k rather than M; and—most simply of all—(iii) defining $\mathsf{score}(x,y)$ to be the (x,y) entry in the matrix M_k.

• *Unseen bigrams.* Link prediction is akin to the problem of estimating frequencies of *unseen bigrams* in language modeling—pairs of words that co-occur in a test corpus, but not in the corresponding training corpus (see, e.g., [9]). Following ideas in that literature, we can improve $\mathsf{score}(x,y)$ using values of $\mathsf{score}(z,y)$ for nodes z that are "similar" to x. Suppose we have values $\mathsf{score}(x,y)$ computed under one of the measures above. Let $S_x^{\langle\delta\rangle}$ denote the δ nodes most related to x under $\mathsf{score}(x,\cdot)$, for a parameter $\delta > 0$. We then define enhanced scores in terms of these nodes: $\mathsf{score}^*(x,y) := |\{z : z \in \Gamma(y) \cap S_x^{\langle\delta\rangle}\}|$ or $\mathsf{score}^*_{wtd}(x,y) := \sum_{z \in \Gamma(y) \cap S_x^{\langle\delta\rangle}} \mathsf{score}(x,z)$.

• *Clustering.* We can also try to improve the quality of a predictor by deleting the more "tenuous" edges in G_{collab} by a clustering procedure, and then running the predictor on the resulting "cleaned-up" subgraph. Specifically, 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. We now re-compute $\mathsf{score}(x,y)$ for all pairs $\langle x,y \rangle$ on this subgraph.

4. RESULTS AND DISCUSSION

In Figure 2, we show each predictor's performance on each arXiv section, in terms of the factor improvement over random predictions. (Many collaborations form for reasons outside the scope of the network, so improvement over random is arguably more meaningful here than raw performance.) A number of methods significantly outperform random, suggesting that the network topology alone does contain useful information; the Katz measure and its variants perform consistently well, and some of the very simple measures (e.g., common neighbors and the Adamic/Adar measure) also perform well. At the same time, 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 [12].

• The fact that collaboration networks form a small world—i.e., there are short paths connecting almost all pairs of scientists [11]—is normally viewed as vital to the scientific community. In our context, though, this implies that there are often very short (and very tenuous) paths between two scientists in unrelated disciplines; this suggests why 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.

• Performance of the low-rank approximation methods tends to be best at an intermediate rank, but on `gr-qc` they perform best at rank 1. This suggests a sense in which the collaborations in `gr-qc` have a much "simpler" structure. One also observes the apparent importance of node degree in the `hep-ph` collaborations: the preferential attachment predictor does uncharacteristically well on this dataset, outperforming the basic graph distance predictor.

• Certain of the methods show high overlap in the predictions they make; one such cluster of methods is Katz, low-rank inner product, and Adamic/Adar. It would be interesting to understand the generality of these overlap phenomena, especially since some of the large overlaps (such as the one just mentioned) do not seem to follow obviously from the definitions of the measures.

• Given the low performance of the predictors on `astro-ph` (and the fact that none beats simple ranking by common neighbors), it is an interesting challenge is to formalize a sense in which it is a "difficult" dataset. By running our predictors on some other datasets, we have discovered that performance swells dramatically as the topical focus of the dataset widens. In a narrow field, almost anyone can collaborate with anyone else, and new collaborations are largely random. It would be interesting to make precise a sense in which such new collaborations are simply not predictable from the training data.

Acknowledgements. We thank Tommi Jaakkola, Lillian Lee, Frank McSherry, and Grant Wang for helpful discussions, and Paul Ginsparg for generously providing arXiv bibliographic data.

5. REFERENCES

[1] L. Adamic, E. Adar. Friends and neighbors on the web. *Soc. Networks*, 25(3), 2003.

[2] A. Barabasi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek. Evolution of the social network of scientific collaboration. *Physica A*, 311(3–4), 2002.

[3] S. Brin, L. Page. The anatomy of a large-scale hypertextual Web search engine. *Comput. Networks ISDN*, 1998.

[4] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, R. Harshman. Indexing by latent semantic analysis. *J. Am. Soc. Inform. Sci.*, 41(6), 1990.

[5] G. Jeh, J. Widom. SimRank: A measure of structural-context similarity. In *KDD*, 2002.

		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-two pairs)		*9.6*	*25.3*	*21.4*	*12.2*	*29.2*
common neighbors		**18.0**	**41.1**	**27.2**	**27.0**	**47.2**
preferential attachment		4.7	6.1	7.6	*15.2*	7.5
Adamic/Adar		*16.8*	**54.8**	**30.1**	**33.3**	**50.5**
Jaccard		*16.4*	**42.3**	19.9	**27.7**	*41.7*
SimRank	$\gamma = 0.8$	*14.6*	*39.3*	*22.8*	*26.1*	*41.7*
hitting time		6.5	23.8	*25.0*	3.8	13.4
hitting time—normed by stationary distribution		5.3	23.8	11.0	11.3	21.3
commute time		5.2	15.5	**33.1**	*17.1*	23.4
commute time—normed by stationary distribution		5.3	16.1	11.0	11.3	16.3
rooted PageRank	$\alpha = 0.01$	*10.8*	*28.0*	**33.1**	*18.7*	*29.2*
	$\alpha = 0.05$	*13.8*	*39.9*	**35.3**	*24.6*	*41.3*
	$\alpha = 0.15$	*16.6*	**41.1**	**27.2**	**27.6**	*42.6*
	$\alpha = 0.30$	*17.1*	**42.3**	*25.0*	**29.9**	*46.8*
	$\alpha = 0.50$	*16.8*	**41.1**	*24.3*	**30.7**	*46.8*
Katz (weighted)	$\beta = 0.05$	3.0	21.4	19.9	2.4	12.9
	$\beta = 0.005$	*13.4*	**54.8**	**30.1**	*24.0*	**52.2**
	$\beta = 0.0005$	*14.5*	**54.2**	**30.1**	**32.6**	**51.8**
Katz (unweighted)	$\beta = 0.05$	*10.9*	**41.7**	**37.5**	*18.7*	**48.0**
	$\beta = 0.005$	*16.8*	**41.7**	**37.5**	*24.2*	**49.7**
	$\beta = 0.0005$	*16.8*	**41.7**	**37.5**	*24.9*	**49.7**
Low-rank approximation: Inner product	$\mathsf{rank} = 1024$	*15.2*	**54.2**	**29.4**	**34.9**	**50.1**
	$\mathsf{rank} = 256$	*14.6*	**47.1**	**29.4**	**32.4**	**47.2**
	$\mathsf{rank} = 64$	*13.0*	**44.7**	**27.2**	**30.8**	**47.6**
	$\mathsf{rank} = 16$	*10.1*	21.4	**31.6**	**27.9**	*35.5*
	$\mathsf{rank} = 4$	8.8	15.5	**42.6**	*19.6*	23.0
	$\mathsf{rank} = 1$	6.9	6.0	**44.9**	*17.7*	14.6
Low-rank approximation: Matrix entry	$\mathsf{rank} = 1024$	8.2	16.7	6.6	*18.6*	21.7
	$\mathsf{rank} = 256$	*15.4*	*36.3*	8.1	*26.2*	*37.6*
	$\mathsf{rank} = 64$	*13.8*	**46.5**	16.9	**28.1**	*40.9*
	$\mathsf{rank} = 16$	9.1	21.4	*26.5*	*23.1*	*34.2*
	$\mathsf{rank} = 4$	8.8	15.5	**39.7**	*20.0*	22.5
	$\mathsf{rank} = 1$	6.9	6.0	**44.9**	*17.7*	14.6
Low-rank approximation: Katz ($\beta = 0.005$)	$\mathsf{rank} = 1024$	*11.4*	*27.4*	**30.1**	**27.1**	*32.1*
	$\mathsf{rank} = 256$	*15.4*	**42.3**	11.0	**34.3**	*38.8*
	$\mathsf{rank} = 64$	*13.1*	**45.3**	19.1	**32.3**	*41.3*
	$\mathsf{rank} = 16$	9.2	21.4	**27.2**	*24.9*	*35.1*
	$\mathsf{rank} = 4$	7.0	15.5	**41.2**	*19.7*	23.0
	$\mathsf{rank} = 1$	0.4	6.0	**44.9**	*17.7*	14.6
unseen bigrams (weighted)	common neighbors, $\delta = 8$	*13.5*	*36.9*	**30.1**	*15.6*	**47.2**
	common neighbors, $\delta = 16$	*13.4*	*39.9*	**39.0**	*18.6*	**48.8**
	Katz ($\beta = 0.005$), $\delta = 8$	*16.9*	*38.1*	*25.0*	*24.2*	**51.3**
	Katz ($\beta = 0.005$), $\delta = 16$	*16.5*	*39.9*	**35.3**	*24.8*	**50.9**
unseen bigrams (unweighted)	common neighbors, $\delta = 8$	*14.2*	*40.5*	**27.9**	*22.3*	*39.7*
	common neighbors, $\delta = 16$	*15.3*	*39.3*	**42.6**	*22.1*	*42.6*
	Katz ($\beta = 0.005$), $\delta = 8$	*13.1*	*36.9*	**32.4**	*21.7*	*38.0*
	Katz ($\beta = 0.005$), $\delta = 16$	*10.3*	*29.8*	**41.9**	*12.2*	*38.0*
clustering: Katz ($\beta_1 = 0.001, \beta_2 = 0.1$)	$\rho = 0.10$	7.4	*37.5*	**47.1**	**33.0**	*38.0*
	$\rho = 0.15$	*12.0*	**46.5**	**47.1**	*21.1*	*44.2*
	$\rho = 0.20$	4.6	*34.5*	19.9	*21.2*	*35.9*
	$\rho = 0.25$	3.3	*27.4*	20.6	*19.5*	17.5

Figure 2: Performance of link predictors on the task defined in Section 2. For each predictor and each arXiv section, the given number specifies the factor improvement over random prediction. 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.

[6] E. Jin, M. Girvan, M. Newman. The structure of growing social networks. *Phys. Rev. E*, 64(046132), 2001.

[7] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1), March 1953.

[8] H. Kautz, B. Selman, M. Shah. ReferralWeb: Combining social networks and collaborative filtering. *CACM*, 1997.

[9] L. Lee. Measures of distributional similarity. In *ACL*, 1999.

[10] M. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64(025102), 2001.

[11] M. Newman. The structure and function of complex networks. *SIAM Review* 45:167-256, 2003.

[12] C. Palmer, P. Gibbons, C. Faloutsos. ANF: A Fast and Scalable Tool for Data Mining in Massive Graphs. In *KDD*, 2002.

[13] A. Popescul, L. Ungar. Statistical Relational Learning for Link Prediction. *Workshop on Learning Statistical Models from Relational Data,* IJCAI 2003.

[14] P. Raghavan. Social networks: From the web to the enterprise. *IEEE Internet Comp.*, Jan/Feb 2002.

[15] G. Salton, M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.