Algorithms for Estimating Relative Importance in Networks

Scott White, Padhraic Smyth
Information and Computer Science
University of California, Irvine
CA 92697-3425, USA
{scott, smyth}@ics.uci.edu

ABSTRACT

Large and complex graphs representing relationships among sets of entities are an increasingly common focus of interest in data analysis—examples include social networks, Web graphs, telecommunication networks, and biological networks. In interactive analysis of such data a natural query is "which entities are most important in the network relative to a particular individual or set of individuals?" We investigate the problem of answering such queries in this paper, focusing in particular on defining and computing the importance of nodes in a graph relative to one or more root nodes. We define a general framework and a number of different algorithms, building on ideas from social networks, graph theory, Markov models, and Web graph analysis. We experimentally evaluate the different properties of these algorithms on toy graphs and demonstrate how our approach can be used to study relative importance in real-world networks including a network of interactions among September 11th terrorists, a network of collaborative research in biotechnology among companies and universities, and a network of co-authorship relationships among computer science researchers.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

graphs, Markov chains, PageRank, social networks, relative importance

1. INTRODUCTION

Many data sets can be described in the form of graphs or networks where nodes in the graph represent entities and edges in the graph represent relationships between pairs of entities. The Web can be viewed as a very large graph of this form, where nodes represent Web pages and (directed) edges represent hyperlinks between pages. In social networks the nodes typically represent individuals (or "actors") and the edges represent relationships among individuals such as social or professional relationships. Citation graphs can be constructed with papers as nodes and references as directed edges. Graph models for biological data are also of increasing interest, modeling for example the interactions between proteins.

In this context there is increasing interest in developing algorithms and software tools for exploratory and interactive analysis of graph data. While visualization techniques such as graph-drawing can be very useful for gaining qualitative intuition about the structure of small graphs, there is also a need for quantitative tools for characterizing graph properties beyond simple lists of "who is connected to who," particularly as graphs become too large and complex for manual analysis. A number of different approaches have been developed in the search for such tools. In the research field of social networks there is a long tradition of developing quantitative frameworks to characterize the importance of a node in a graph relative to all other nodes. For example, a variety of measures have been proposed by sociologists to determine the "centrality" of a node in a social network (Katz, 1953; Freeman, 1979; Stephenson and Zelen, 1989; Wasserman and Faust, 1994). Statisticians have also developed general methods for quantitative graph modeling, such as the embedding of social network data in latent Euclidean spaces (Hoff, Raftery, and Handcock, 2002). In the area of Web graphs, computer scientists have proposed a number of algorithms (such as HITS (Kleinberg, 1999) and PageRank (Brin and Page, 1998; Page et al, 1998)) for automatically determining the "importance" of Web pages.

Virtually all of these techniques focus on global measures of node importance in that each node is ranked relative to all other nodes in the graph. In this paper we propose and investigate a number of algorithms that focus on a related but somewhat different problem, that of determining the relative importance of nodes in a graph with respect to a set of root nodes R. This is a very natural form of query to be able to answer in the context of interactive exploration of graph data.

One approach to answering such queries might be to use a standard "global" algorithm (such as PageRank) to rank all nodes in a subgraph surrounding the root nodes R of interest (e.g., all nodes within some specific path-length of the root nodes). The problem with such an approach is that the root nodes are not given any preferential treatment in the resulting ranking—in effect, one is ranking nodes in the local

SIGKDD '03 August 24-27, 2003 Washington D.C., USA

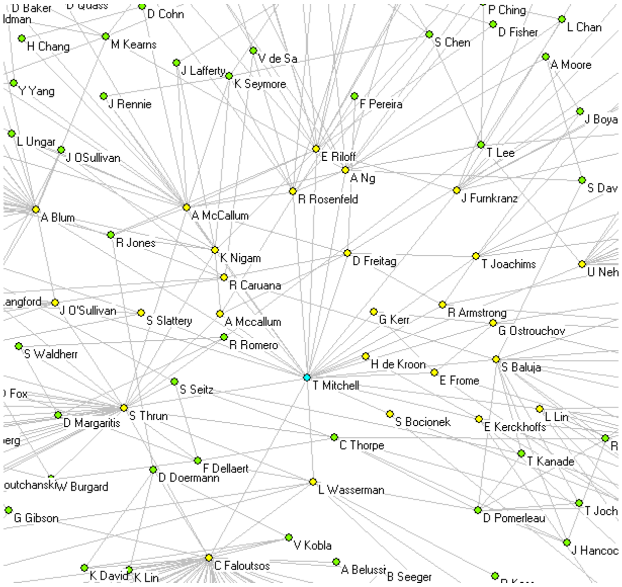

Figure 1: Coauthorship network centered around Tom Mitchell of CMU.

subgraph around R rather than ranking them relative to R. There are some other limitations of the "global" approach in this context which we do not elaborate on here and we do not pursue the "global ranking on local subgraphs" approach further in this paper.

Figure 1 shows an example of a subgraph based on coauthorship patterns among computer science researchers. Here the root set consists of the single author Tom Mitchell, a well-known researcher in machine learning. The problem is to rank all other nodes in the graph given that Tom Mitchell is defined as the root set. The root set R can be thought of as representing the data analyst's prior knowledge or bias in terms of which nodes are considered important in the graph. In effect we are interested in answering a form of conditional query: given that the nodes in R are assumed to be important, rank the other nodes in the graph. In the limit, as the set of root nodes encompasses the entire graph, the relative importance approaches the global importance.

The primary novel contribution of this paper is a new framework and class of techniques for measuring relative importance in graphs. We introduce newly defined measures of node importance (such as Markov centrality) and extensions and generalizations of techniques previously proposed. We also illustrate how our proposed techniques can be applied to a broader set of graphs than is typical in the research literature on network analysis (where typically only one type of graph, such as social networks or Web graphs, is the focus of a given study.)

The paper proceeds as follows. In Section 2 we briefly define our general notation followed by a discussion of related work in Section 3. A precise definition of the problem of determining relative importance in a graph then follows in Section 4. Section 5 introduces a general class of algorithms for relative importance based on weighted paths and motivated by graph-theoretic ideas. Section 6 proposes a different class of algorithms for relative importance based on Markov chain models, including a technique based on mean-first passage times, extensions of the PageRank and HITS algorithms, as well as a technique based on k-step Markov Chain arrival probabilities. Section 7 discusses experimental results on small toy graphs to provide some intuition of how these algorithms operate. Section 8 discusses experimental results on three real-world networks, including a network of September 11th 2001 terrorists, a network describing collaborative partnerships among biotechnology companies, and a large network consisting of computer science researchers and their co-author relations. Correlation of ranks from different algorithms are analyzed in Section 9 and conclusions are presented in Section 10.

2. NOTATION

A directed graph, or digraph, $G = (V, E)$ is comprised of two sets, the set of nodes V and the set of edges E. We define each edge e to be an ordered pair of nodes (u, v) representing a directed connection from u to v. The graphs in this paper are assumed to be directed, unweighted and simple, i.e., without self-loops or parallel edges. We treat undirected graphs as a special case of directed graphs where for each edge from u to v there exists a corresponding edge $e' = (v, u)$. A walk from vertex u to vertex v is a sequence of edges $(u, u_1), (u_1, u_2), \ldots, (u_k, v)$. For brevity we also write this as $u - u_1 - u_2 - \ldots - u_k - v$ where u is the start node and v is the terminal node.

A walk is a path if no nodes are repeated in the sequence. Two (or more) paths are node-disjoint if they have no common intermediate nodes. We define the k-short paths as the set of all paths less than some length k. We will use $\mathcal{P}(u, v)$ to denote a particular set of paths between vertices u and v. For any collection of elements Z, we define $|Z|$ as the number of elements in that collection, e.g., $|V|$ is the number of vertices in V. We define $S_{\text{out}}(u)$ as the set of distinct outgoing edges emanating from node u. $d_{\text{out}}(u) = |S_{\text{out}}(u)|$ is the out-degree. Similarly, $S_{\text{in}}(u)$ is the set of incoming edges connecting to node u, and $d_{\text{in}}(u) = |S_{\text{in}}(u)|$ is the in-degree.

3. RELATED WORK

Although no general framework or methodology has previously been proposed for ranking nodes in a graph relative to a set of root nodes (to our knowledge), the problem of identifying nodes in a network according to some operational notion of "importance" has been an active area of research in many different fields, primarily in the study of social networks, Web graph analysis, and bibliometrics. Here we briefly review work that most closely relates to ideas discussed in this paper.

Freeman (1979) described a set of measures for computing the global importance of each actor in a network where importance is defined in terms of how "central" the actor is in a network. Freeman's work consolidated much of the prior literature along similar lines in social networks, dating as far back as the 1930s. He proposed three measures of centrality, one based on degree, and the other two based on measuring shortest paths to quantify the effect of one node on another. In Section 5 we will discuss the advantages and limitations of shortest paths in computing relative importance.

The idea of using weighted paths to approximate global measures of importance has been used for a long time in the social networks literature although not in the generalized form we describe in this paper. Katz (1953) described a measure of the standing of an actor, in terms of weighted paths. This is close to a special case of a class of algorithms we propose where the set $\mathcal{P}$, defined later in the paper, is the set of all paths between two nodes. Stephenson and Zelen (1989) defined a similar approach called Information Centrality, which also used the set of all paths between two nodes weighted by an information-based weighting for each path that is derived from the inverse of its length. In addition, much work has been carried out in social networks and bibliometrics to define importance in terms of the principal eigenvector of a matrix derived from the underlying graph. Much of this work can be seen as precursors to eigenvector methods proposed recently in the Web ranking literature.

The two seminal contributions to ranking nodes in a Web graph are the PageRank algorithm of Brin and Page (1998), and the HITS algorithm of Kleinberg (1999). Research in this area has since been very active in developing a variety of extensions and new algorithms. Lempel and Moran (2000) described a variation of HITS called SALSA that can be understood as a random walk on a bipartite graph of hubs. Borodin et al. (2001) described a number of algorithms for ranking nodes in a Web graph, including extensions of both SALSA and HITS. Their approximation algorithms based on weighted paths can be viewed as closely related to the class of weighted path algorithms introduced later in this paper.

Almost all of this prior work, from social networks, bibliometrics, and Web graph analysis, is either implicitly or explicitly focused on global rankings of nodes. There are two notable exceptions. The first is the work of Haveliwala (2002) and Jeh and Widom (2002), who developed personalized variations of PageRank, extending earlier ideas of Page et al. (1998). Their goals are similar to ours (but for the specific context of PageRank and Web pages), namely, to bias the standard PageRank rankings towards a set of prior topics (or root set). In Section 6 we discuss this approach in more detail and we include it as one of the algorithms we evaluate in our experimental results later in the paper. The second exception to global rankings is the work of Chang, Cohn and McCallum (2000) that describes a personalized variant of HITS (different than the personalized version of HITS we desribe in this paper) based on iteratively alternating between running the original version of HITS and then performing gradient ascent on each element of the adjacency matrix to update the weights with respect to the preferred nodes. Apart from this work, there has been relatively little attention paid to the problem of computing importance in graphs relative to a subset of nodes in the graph. This paper focuses directly on the definition and computation of relative importance for graph data. We propose a general family of techniques for computing relative importance and evaluate and compare these techniques with existing approaches such as personalized PageRank.

4. PROBLEM FORMULATION

We now describe four successive problem formulations, each building upon the next, that defines our approach to ranking nodes in an unweighted digraph $G(V, E)$:

1. Given G and r and t, where $\{r, t\} \subset G$, compute the "importance"[1] of t with respect to the root node r. We denote this as $I(t|r)$, in general a non-negative quantity.

2. Given G and node $r \in G$, rank all vertices in $T(G), T \subseteq V$, with respect to r. To do this we compute $I(t|r)$ for all $t \in T$ and then sort the values so that the largest values can be said to have highest importance and conversely for the smallest values.

3. Given G, a set of nodes $T(G)$ to rank, and a set of root nodes $R(G)$ where $R \subseteq V$, rank all vertices in T with respect to R. This is the same as case 2 except instead of computing $I(t|r)$ for a single node r, we compute $I(t|R)$ as a function of all nodes $r \in R$, where $I(t|R)$ is generally defined as a function of the set of individual rankings $\{I(t|r) : r \in R\}$. For example, we can use the average importance relative to the set R:

$$I(t|R) = \frac{1}{|R|} \sum_{r \in R} I(t|r). \qquad (1)$$

   Instead of averaging the individual $I(t|r)$ terms we could (for example) define $I(t|R) = \min\{I(t|r) : r \in R\}$, which requires that a node t have high importance relative to *all* nodes in R in order to be ranked highly overall. In this paper we will use the "average" (as in Equation 1 above) for $I(t|R)$ in most of our examples, but in general other functional forms can be used when appropriate.

4. Given G, rank all nodes. This is a special case of the last step where $R = T = V$.

We restrict our attention in this paper to ranking algorithms that lie within this general framework.

5. COMPUTING RELATIVE IMPORTANCE USING WEIGHTED PATHS

We begin by describing a class of algorithms that use explicit definitions of relative importance for $I(t|r)$ based on various graph-theoretic notions of distance. The two main properties we model using this approach are: 1) two nodes are related according to the paths that connect them and 2) the longer a path is, the less importance is conferred along that path. To achieve these two ends we define $I(t|r)$ as follows:

$$I(t|r) = \sum_{i=1}^{|\mathcal{P}(r,t)|} \lambda^{-|p_i|} \qquad (2)$$

where $\mathcal{P}(r, t)$ is a set of paths from r to t, p_i is the ith path in $\mathcal{P}$, and λ is a scalar coefficient, $1 \leq \lambda \leq \infty$, that determines how much importance is conferred from r to t. In this model, the amount of importance that is conferred along a path decays exponentially with path length, a realistic assumption for most real-world networks. Borodin et al. (2001) proposed what can be seen as a special case of this formulation with $\lambda = 2$ and their choice for $\mathcal{P}$ involving choosing paths that can be constructed by alternating

[1] Depending on the specific entities and relationships that the nodes and edges represent, more nuanced words may be appropriate to describe what is being calculated, e.g., authority, prestige, energy, mass, etc. However, for the purposes of discussing a broad framework that can encompass multiple definitions of $I(t|s)$ we purposely use the less specific term "importance" throughout the paper.

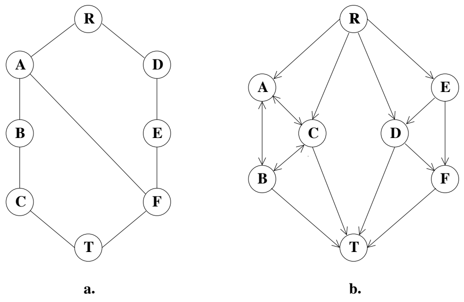

Figure 2: Two simple toy graphs.

between moving forward along the out-links and backward along the in-links starting at a given node. In the results in this paper we will also use $\lambda = 2$.

Figure 2 shows two different graphs with terminal vertices R and T, and several paths connecting them along vertices A, B, C, D, E and F in each graph. Two examples of possible choices of $\mathcal{P}$ are $\mathcal{P}_1(R,T) = \{R-A-B-C-T, R-A-F-T, R-D-E-F-T\}$ and $\mathcal{P}_2(R,T) = \{R-A-F-T\}$. How we wish to define importance will drive our choice of $\mathcal{P}$. To see that this is the case, we examine several plausible candidates for $\mathcal{P}$ and analyze their strengths and weaknesses using the examples in Figure 2.

5.1 Shortest Paths

Shortest paths, often called geodesics, can provide useful measures of importance if it is assumed that all the vertices that do not lie on the geodesics but yet are reachable from both R and T play a negligible role. For example, if one imagines a transportation network with R trying to transport "cargo" to T using as few intermediate vertices as possible, using shortest paths would be well suited to model the set of paths that best characterize the importance that R confers to T. But in many types of graphs, such as Web graphs or citation graphs, using shortest paths can yield poor approximations. For example, Figure 2a shows a situation where an assessment based on the two shortest paths between R and T, $\{R-C-T, R-D-T\}$ fails to capture much of the connectivity between R and T, by ignoring the importance of vertices A, B, E, and F. These vertices could in principle add more importance to T, relative to R, than if they did not exist. Despite these limitations, shortest paths are an important measure that is widely used to measure pair-wise relations in a graph. For example, many of the centrality measures in social networks such as "closeness" and "betweenness" (Freeman, 1979; Wasserman and Faust, 1994) use shortest paths to measure how two nodes are related.

5.2 K-Short Paths

Using K-short paths as a choice for $\mathcal{P}$, for a prespecified value K and assuming K is large enough, addresses the problem we just described for shortest paths where there are often longer paths than the very shortest path that are important to take into account. For example in Figure 2a, the set of 3-short paths would be $\{R-C-T, R-D-T, R-A-B-T, R-C-B-T, R-A-C-T, R-E-F-T, R-E-D-T, R-D-F-T\}$. Although K-short paths, even for small K, can succeed in detecting the important paths that connect R to T, they have one major drawback. They do not take into account "capacity constraints" that may exist on nodes or edges. Because nodes and edges may occur multiple times on different paths, they can be seen as having infinite capacity and as such can be used to double-count the importance that is conferred on a node. For networks where the notion of a capacity constraint or bottleneck has no meaning, this might be acceptable. However, for many real world networks this is an unrealistic assumption.

5.3 K-Short Node-Disjoint Paths

K-short node-disjoint paths are sets of K-short paths that have neither edges nor nodes in common, i.e., no node or edge can be used more than once in the set of multiple paths from R to T. In Figure 2b a set of 3-short node-disjoint paths is $\{R-C-T, R-D-T, R-A-B-T, R-E-F-T\}$. For our experimental evaluation of different graphs, we use sets of K-short node-disjoint paths as our choice for the set of paths $\mathcal{P}$ in the definition of $I(u|v)$ using weighted paths. This specific choice implicitly enforces capacity constraints on nodes and edges (no node or edge can be double-counted in the definition of relative importance). We chose this definition of $\mathcal{P}$ as we found it gives quite good approximations of the relative importance of nodes based, in general, on a relatively small set of paths in the neighborhood of radius K from the root nodes. In addition, we chose it in order to contrast it with the Markov chain methods described in the next section where, in its unconstrained form, the connectivity of the entire graph, by virtue of traversing the set of all possible walks, is used to compute the importance of a node. The use of weighted K-short node-disjoint paths above is in contrast motivated more by a "mass flow" analogy where importance can "flow" along disjoint paths in the graph from R to T.

To compute importance using K-short node-disjoint paths, we use a heuristic breadth-first search algorithm to find a good set of paths $\mathcal{P}$. Using paths of length less than or equal to K is motivated by the fact that for $\lambda = 2$ (for example), the weights die off rather quickly as a function of K, and one can generally get a good approximation of $I(u|v)$ with relatively small integer values of K, avoiding the computational expense of exploring longer paths that contribute very little to $I(u|v)$. In the general case, the choice of the path set $\mathcal{P}$ will depend on what aspect of "importance" the data analyst is interested in.

6. COMPUTING RELATIVE IMPORTANCE USING MARKOV CHAINS

In the previous section we used graph-theoretic notions of distance defined explicitly on the graph as a general framework for estimating relative importance. A conceptually different approach is to view the graph as representing a stochastic process, more specifically, a first-order Markov chain. Intuitively one can imagine that there is a single "token" traversing the graph in a stochastic manner for an infinitely long time, where the next node that the token moves to is a stochastic function of properties of the current node. The fraction of time that the token spends at any single node (the stationary distribution of the corresponding Markov chain, under appropriate assumptions) can then be interpreted as being proportional to an estimate of the global importance or importance of this node relative to all other nodes in the graph.

Perhaps the most well-known and successful example of this general idea is the PageRank algorithm (Brin and Page, 1998; Page et al., 1998). The Markov analogy in the PageRank algorithm is quite clear and has been well-described in the literature, namely a "random surfer" surfing the Web based on a suitably-defined transition matrix. For other types of (non-Web) graphs, the Markov analogy is not quite as obvious. For example, in a graph where nodes are authors and edges correspond to co-authorship, a loose analogy would be that there is a single publication ("the book of knowledge") that is being circulated ad infinitum (in Markov fashion) through the entire author graph. The stationary distribution (or importance) could be thought of as the fraction of time that each author gets to contribute to the book of knowledge as it passes among authors. In more general social networks, where nodes are human actors and edges represent specific types of interactions such as friendship, institutional membership, etc., the Markov analogy is further strained, but one can imagine that there is a finite amount of some currency that is being circulated through the entire graph. Based on the success of PageRank (and similar) algorithms, we will take it at face-value that this type of Markov-chain analogy can lead to useful definitions for importance across a much broader set of graphs than just Web graphs. Thus, our focus in this paper in the context of Markov-based algorithms is to define and evaluate different techniques for calculating the relative importance of t relative to r, in contrast to the more usual global importance that is calculated by algorithms such as PageRank.

In what follows below we assume that the graph can be converted into an equivalent Markov chain—specifically, the probability of transitioning to node i from node j is defined (unless stated otherwise) as $p(i|j) = 1/d_{out}(j)$ for nodes i that have an edge from j to i, and 0 otherwise. Other choices for transition probabilities could also be used in the algorithms discussed below if information is available on the relative weights of different edges in the network.

6.1 Markov Centrality

The first approach we examine is the inverse of the mean first-passage time in the Markov chain. The mean first passage time m_{rt} from r to t is defined as the expected number of steps taken until the first arrival at node t starting at node r (Kemeny and Snell, 1976):

$$m_{rt} = \sum_{n=1}^{\infty} n f_{rt}^{(n)} \qquad (3)$$

where n denotes the number of steps taken, and $f_{rt}^{(n)}$ denotes the probability that the chain first returns to state t in exactly n steps. A useful property of using mean first passage times, besides having a natural Markov interpretation, is that one can directly compute a mean first passage matrix giving the mean first passage times for all pairs of nodes. The mean first passage matrix is given by

$$\mathbf{M} = (\mathbf{I} - \mathbf{Z} + \mathbf{E}\mathbf{Z}_{dg})\,\mathbf{D} \qquad (4)$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{E}$ is a matrix containing all ones, and $\mathbf{D}$ is the diagonal matrix with elements $d_{vv} = \frac{1}{\pi(v)}$ where $\pi(v)$ is the stationary distribution (in the Markov chain) of node v. $\mathbf{Z}$ is known as the fundamental matrix and $\mathbf{Z}_{dg}$ agrees with $\mathbf{Z}$ on the diagonal but is 0 everywhere else. The fundamental matrix is defined as

$$\mathbf{Z} = \left(\mathbf{I} - \mathbf{A} - \mathbf{e}\boldsymbol{\pi}^T\right)^{-1} \qquad (5)$$

where $\mathbf{A}$ is the Markov transition probability matrix, $\mathbf{e}$ is a column vector of all ones, and $\boldsymbol{\pi}$ is a column vector of the stationary probabilities for the Markov chain.

In the results reported in this paper we use the inverse of the average mean first passage times for defining the importance of node t given a root set R, i.e.,

$$I(t|R) = \frac{1}{\frac{1}{|R|}\sum_{r \in R} m_{rt}}$$

If we set $R = T = V$ we get a global "objective" ranking function, yielding a ranking algorithm where nodes that are more "central" in a network, i.e., closer to the center of mass, have higher ranking than those that are less central. We can see that this is the case by observing that nodes that are more central will take less time to reach on average from all other nodes. As $|R|$ gets smaller with respect to $|V|$, the rankings will be more biased towards the nodes in our root set yet still preferring nodes that are "central" in the network. We refer to this as "Markov centrality" a concept that reflects the notion of how central a node t is in a network relative to a particular node r, analogous to centrality measures developed by sociologists for analyzing social networks (Katz, 1953; Freeman, 1979; Stephenson and Zelen, 1989; Wasserman and Faust, 1994).

Finally, observe that a disadvantage of this methodology is that the computational complexity of a direct method scales as $O(|V|^3)$ since we must invert a matrix of size $|V| \times |V|$ to solve Equation 5, and the space complexity is $O(|V|^2)$. Consequently we do not apply the method to the two larger data sets in our experimental results section. It may be possible to reduce the complexity by taking advantage of both the sparsity of $\mathbf{A}$ and the size of R (since we only need first-passage times relative to the set R, not relative to all nodes).

6.2 PageRank with Priors

Haveliwala (2002) and Jeh and Widom (2002) show that the PageRank algorithm can be extended to generate "personalized" ranks. Borrowing from their work, we demonstrate how PageRank can be retrofitted into our framework. We define a vector $\mathbf{p}_R$ of prior probabilities $\mathbf{p}_R = \{p_1, \ldots, p_{|V|}\}$ such that the probabilities sum to 1 and where p_v denotes the relative importance (or "prior bias") we attach to node v. In this paper we use $p_v = \frac{1}{|R|}$ for $v \in R$, $p_v = 0$ otherwise, i.e., all nodes in the root set have equal prior probability. In addition to $\mathbf{p}_R$, we also define a "back probability" $\beta, 0 \leq \beta \leq 1$ which determines how often we jump back to the set of root nodes in R. Integrating these two extensions into the original formula for PageRank, we get iterative stationary probability (or rank) equations of the form:

$$\pi(v)^{(i+1)} = (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} p(v|u)\pi^{(i)}(u)\right) + \beta p_v. \qquad (6)$$

We use the resulting "ranks", biased towards the set R, as our definition of relative importance, i.e., $I(v|R) = \pi(v)$ after convergence.

Intuitively this equation represents a Markov chain for a random surfer who transitions "back" to the root set R with probability β at each time-step. This is similar in spirit to the use of weighted paths as follows: we are evaluating the probability of landing on a node in the modified Markov chain where a random graph surfer starts in the set R (with appropriate prior probabilities) and executes a random walk that ends stochastically with probability β (at which point the process restarts). This process defines an (infinite) set of walks of variable length starting at the root set (in fact they will follow a geometric distribution with mean $\frac{1}{\beta}$). The "rank" equation above estimates the relative probability of landing on any particular node during this set of walks. The computational complexity is the same as that of the standard PageRank algorithm, namely sparse matrix-vector multiplications to implement each iteration in the recursion defined by Equation 6.

6.3 HITS with Priors

We can also extend the HITS algorithm (Kleinberg, 1999) to fit into our proposed framework using logic similar in spirit to the preceding discussion. To do so we borrow the same extensions of defining a vector $\mathbf{p}_R = \{p_1, \ldots, p_{|V|}\}$ of prior probabilities and a "back probability" β yielding:

$$a^{(i+1)}(v) = (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} \frac{h^{(t)}(u)}{H^{(i)}}\right) + \beta p_v \qquad (7)$$

$$h^{(i+1)}(v) = (1-\beta)\left(\sum_{u=1}^{d_{out}(v)} \frac{a^{(t)}(u)}{A^{(i)}}\right) + \beta p_v \qquad (8)$$

where $d_{\text{in}}(v)$ and $d_{\text{out}}(v)$ are the indegree and outdegree of v, respectively, and where $H(t)$ and $A(t)$ are defined as:

$$H^{(i)} = \sum_{v=1}^{|V|} \sum_{u=1}^{d_{in}(v)} h^{(i)}(u) \qquad (9)$$

$$A^{(i)} = \sum_{v=1}^{|V|} \sum_{u=1}^{d_{out}(v)} a^{(i)}(u) \qquad (10)$$

We call this extension "HITS with Priors" and its intuition is as follows: let there be a random surfer who starts from a randomly chosen page in G and visits a new page at every time step. At each time step, the surfer tosses a coin with bias β. If the coin lands heads and it is an even time step, then the surfer follows a random in-link. If the coin lands heads and it is an odd time step, then the surfer follows a random out-link. If the coin lands tails, then the surfer jumps to a page in R chosen according to the distribution $\mathbf{p}_R$. As before, this process defines a random walk and the resulting stationary distribution of each page is then used to define $I(t|R)$. For the purposes of this discussion we assume R is the same set for both hubs and authorities although one could make the model more complicated by having two sets R_A and R_H, one for authorities and one for hubs. At even time steps when the coin lands tails, the random surfer would jump to a page chosen randomly in R_H, at odd time steps the surfer would jump to a page chosen randomly in R_H. However, we will not consider this model further in this paper. As with PageRank with Priors, as β approaches 1, the rankings are more biased towards R.

6.4 The K-Step Markov Approach

The "back probability" interpretation for personalized PageRank and HITS also suggests a slightly different algorithm, one that also generates random walks starting from R, but where now the walks are of fixed-length K (whereas PageRank and HITS with Priors in effect generate walks where the length is stochastic). Now we are computing the relative probability that the system will spend time at any particular node, given that it starts in R and ends after K steps. This is an estimate of the transient distribution of states in the Markov chain, starting from R: as K gets larger we will (under appropriate assumptions on the Markov transition matrix $\mathbf{A}$) converge to the steady-state distribution used by PageRank. Thus, the value of K controls the relative tradeoff between a distribution "biased" towards R and the steady-state distribution which is independent of where the Markov process started. We call this algorithm K-Step Markov or KSMarkov for short. In this case, $I(t|R)$ can be computed using the equation:

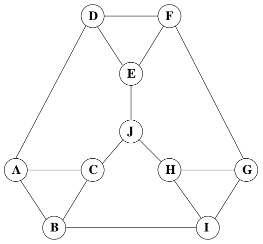

Figure 3: An undirected toy graph.

$$I(t|R) = \left[\mathbf{A}\mathbf{p}_R + \mathbf{A}^2\mathbf{p}_R \ldots \mathbf{A}^K\mathbf{p}_R\right]_t \qquad (11)$$

where $\mathbf{A}$ is the transition probability matrix of size $n \times n$, $\mathbf{p}_R$ is an $n \times 1$ vector of initial probabilities for the root set R, and $I(t|R)$ is the t-th entry in this sum vector. The computational complexity for sparse graphs is approximately that of sparse matrix-matrix multiplication performed K times.

7. EVALUATION ON SIMULATED DATA

Given that these algorithms are intended as exploratory data analysis tools, it is not clear that there are well-defined quantitative metrics for evaluating the quality of one algorithm's rankings relative to another. With this in mind, the interpretation of the ranking results for a particular graph will depend on the goals of the data analyst who requests the ranking. In a similar vein, the choices for parameters such as β and K are inherently subjective and different values for these parameters reflect different biases in the ranking process. Despite these general reservations, in this section and the next we examine the top-10 rankings of the algorithms for a variety of graphs and root sets with the goal of obtaining a better understanding of what kinds of rankings are produced by each algorithm as well as illustrating how the methods work on both real and simulated data.

We compare five different ranking algorithms described in the previous sections: weighted K-short node-disjoint paths (WKPaths), Markov Centrality (MarkovC), PageRank with Priors (PRankP), HITS with Priors (HITSP), and K-step Markov (KSMarkov), with abbreviations for each in parentheses. For both PRankP and HITSP, we use $\beta = 0.3$. For both WKPaths and KSMarkov we use $K = 6$. For undirected graphs the importance and hubs scores are the same so there is no need to display both scores in this case.

We first consider a simple undirected graph with ten nodes, each with degree three, and fifteen edges as shown in Figure 3. In the resulting ranking PRankP, HITSP, and KSMarkov

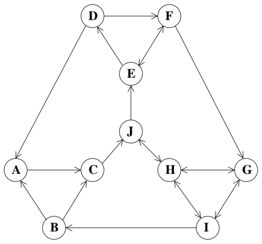

Figure 4: A directed toy graph.

are perfectly correlated with a ranking based on degree, and each node gets the same score. For the graph in Figure 3, these algorithms do not discern important structural differences between nodes. All three algorithms give the same rank to J (the central hub connecting the three sub graphs A-B-C, D-E-F, and G-H-I) as they do to each of the rest of the nodes. The other two algorithms can distinguish between J and the others. With MarkovC, the central hub J is ranked highest with a normalized ranking score of 0.112 while all other nodes get the same score of 0.098. This is consistent with the earlier observation that this algorithm prefers nodes that are central in the graph. WKPaths makes a further distinction between C,E, and H, the middle ring, and the other nodes on the outside ring A,B,D,F,G,I. In this case, C,E,H are ranked equally high and above J because of the fact that there are more higher weighted node-disjoint paths, in general, to C,E, and H than there are to J.

Figure 4 shows a directed graph version of the previous graph (Figure 3) where relationships between nodes are now more complicated. For example, the shortest path from A to B which used to be A-B is now A-C-J-H-I-B. and the shortest path from D to J which used to be D-E-J is now D-A-C-J. In addition, the in-degrees and out-degrees are no longer the same for all nodes. As a consequence, hubs and importance scores for HITSP will no longer be the same and so both scores are displayed.

Consider an example with more than one root node. Table 1 shows the results when both A and F are selected as root nodes for the directed graph in Figure 4. In this example, which typifies the kinds of complexities we find in real world network data, the algorithms produce more diverse results in terms of ranking which nodes are more important. Given that each algorithm has a different understanding of what "importance" means this is to be expected. HITSP, for example, gives results that are intuitive when it is understood that nodes that are closer to the root nodes should generally be ranked higher, that nodes which send out edges to other important nodes should be ranked higher as hubs, while nodes which receive edges from important nodes should be ranked higher as authorities. We can see this is the case by observing, for example, that F is ranked higher as a hub than A (and all other nodes) since it sends out two edges (while A only sends out one), one of which is G which is part of the highest interconnected subgraph of nodes G,H,I,J. Of the nodes that aren't root nodes, D is ranked highest as a hub since it is the only node sending out edges to both root nodes. We can see also how MarkovC gives roughly intuitive results by noticing that it ranks highest the nodes that are most central in the graph while still being close to the root nodes. Finally, we observe that both KSMarkov and MarkovC rank the root nodes near the bottom of the rankings. This might seem rather bizarre if one expects the root nodes by default to be near the top by virtue of the fact that rankings are relative to them. However, this need not be the case if one thinks of relative importance as a sphere of influence where for PRankP, HITSP, and WKPaths the sphere's density diminishes exponentially away from the root nodes whereas for MarkovC and KSMarkov the sphere is centered at the root nodes but the sphere's density is much more uniform. For MarkovC and KSMarkov the root nodes will be ranked highest only if they are the most important nodes in their vicinity. In the case of KSMarkov, H is ranked highest because it is in the center of the densest subgraph G,H,I,J and so is easiest to reach on average for a walk of 6 steps away from the root nodes.

8. EVALUATION ON REAL-WORLD DATA

We use three real world graph data sets to illustrate the applicability of our algorithms to large, complex graphs: a graph of the September 11th terrorist network, a network of biotechnology collaborations, and a coauthorship network constructed from the Citeseer database of scientific literature (Lawrence, Giles, and Bollacker, 1999).

8.1 September 11th Terrorist Network

The terrorist network graph consists of 63 nodes (terrorists) and 308 edges representing known interactions between terrorists (Krebs, 2001). This graph includes the 19 September 11th hijackers and their associates.

Table 4 show the top ten authorities (shortened last names only) relative to two associates of the September 11th hijackers that were known to be part of European Al Qaeda terrorist cells: Essid Sami Ben Khemais who is known to have been part of an Italian cell and Djamal Beghal who is known to have been a leader in the Al Qaeda European network. The results are interesting insofar as they show how each of the different algorithms are able to measure different aspects of the terrorists' roles in Al Qaeda and the September 11th hijacking. In all of the algorithms except MarkovC, we can discern many of the important terrorists, for example, Tarek Maaroufi, Kamel Daoudi, and Zacarias Moussaoui, that were known to have had strong ties with Khemais and Beghal as well as playing a major part in European operations of Al Qaeda. The fact that KSMarkov agrees strongly with PRankP, HITSP and WKPaths in ranking the root nodes highest indicates that Khemais and Beghal were indeed important players in the Al Qaeda network in general. Furthermore, these four algorithms are able to identify terrorists that fall on critical paths between the two root nodes and who themselves were embedded in locally cohesive networks, including Tarek Maaroufi and Abu Qatada. Finally, MarkovC identifies the terrorists who were very central overall in the network, for example Mohammed Atta who it ranks highest and is known to have been the leader of the hijackers, while still being within a few points of contact from the root nodes.

8.2 Biotech Collaboration Network

The biotech network data set consists of 2700 nodes (biotech firms and their collaborators) and 8690 edges (collaborations). The biotech firms encompass 482 international pub-

Table 1: Importance rankings for the nodes in Figure 3 with respect to nodes A and F.

Rank	PRankP		HITSPa		HITSPh		WKPaths		MarkovC		KSMarkov	
1	F	0.200	A	0.252	F	0.225	F	0.206	J	0.180	H	.146
2	A	0.167	F	0.241	A	0.186	A	0.206	C	0.133	G	.142
3	C	0.122	G	0.128	D	0.162	E	0.116	G	0.130	E	.142
4	E	0.107	C	0.110	B	0.119	C	0.108	H	0.129	J	.140
5	J	0.105	E	0.099	E	0.090	G	0.095	E	0.111	C	.120
6	G	0.103	H	0.052	I	0.067	J	0.068	I	0.101	I	.098
7	H	0.086	D	0.032	H	0.061	H	0.066	F	0.069	F	.087
8	I	0.056	I	0.032	J	0.050	I	0.052	D	0.051	D	.061
9	D	0.037	J	0.025	G	0.028	D	0.052	A	0.047	A	.034
10	B	0.013	B	0.024	C	0.008	B	0.026	B	0.044	B	.024

Table 2: Importance rankings for the terrorist network with respect to nodes Khemais and Beghal.

Rank	PRankP		HITSP		WKPaths		MarkovC		KSMarkov	
1:	Khemais	0.221	Khemais	0.173	Beghal	0.045	Atta	0.063	Khemais	0.115
2:	Beghal	0.218	Beghal	0.166	Khemais	0.045	Al-Shehhi	0.041	Beghal	0.108
3:	Moussaoui	0.044	Atta	0.038	Moussaoui	0.045	al-Shibh	0.037	Moussaoui	0.065
4:	Maaroufi	0.039	Moussaoui	0.029	Maaroufi	0.044	Moussaoui	0.036	Maaroufi	0.059
5:	Qatada	0.036	Maaroufi	0.026	Bensakhria	0.037	Jarrah	0.030	Qatada	0.052
6:	Daoudi	0.035	Qatada	0.025	Daoudi	0.037	Hanjour	0.028	Daoudi	0.049
7:	Courtaillier	0.032	Bensakhria	0.023	Qatada	0.036	Al-Omari	0.026	Bensakhria	0.045
8:	Bensakhria	0.031	Daoudi	0.023	Walid	0.031	Khemais	0.025	Courtaillier	0.045
9:	Walid	0.030	Courtaillier	0.022	Courtaillier	0.031	Qatada	0.025	Walid	0.040
10:	Khammoun	0.025	Khammoun	0.021	Khammoun	0.029	Bahaji	0.024	Khammoun	0.034

licly and privately held companies involved in human therapeutic and diagnostic applications of biotechnology (Powell et al., 2002). The data set covers collaborations over the 12-year period, 1988-99. Collaborations include such relationships as finance, R&D, and commercial ventures. A portion of the network is shown in Figure 5.

We examine the top relative authorities in the biotech collaboration network relative to two British universities that are known to have expertise in biotechnology: Cambridge University and Oxford University. Table 5 shows the results (names are abbreviated) of ranking the top ten authorities relative to them. Perhaps not surprisingly in this case, the authorities that are selected tend to be British companies, some of them located in the same cities as the universities themselves. This seems reasonable given the assumption that universities tend more often than not to engage in research with companies and other institutions located near them geographically. Nevertheless, given that this data set includes many different types of collaborations, not just commercial partnerships, it is interesting that geographic location so strongly influences the results. For example, British Biotech, Cantab Pharmaceuticals now acquired by Xenova, Oxford GlycoSciences, and Glaxo Pharmaceuticals are all British biotech companies. (It is an interesting side note that one of the non-British companies, Cortecs International, has since this data was collected been bought out by British pharmaceutical company Provalis). Finally, it should also not be surprising that major players in the biotech arena, of which NIH is by far the largest (in terms of funding), show up in these results as well.

8.3 The CITESEER Co-Authorship Network

The CiteSeer data set we used consists of 387,703 papers that were published between 1991 and 2002. From this data set we extracted a co-authorship graph where the nodes are authors in the data set and the edges are all of the pair-wise co-authorships between authors. An edge exists between two authors if they have been co-authors on 1 or more papers together.

Table 7 shows the most important nodes relative to Tom Mitchell (Figure 1). The rankings from different algorithms correlate strongly, due in large part to the strong interconnected network involving Tom Mitchell, Sebastian Thrun, Andrew McCallum, Thorsten Joachims, etc. All of these authors were in the Computer Science department at CMU for several years during the time this data set was collected and most of them have written papers with one another numerous times. So the authorities we see here are highly influenced by the fact that there was a strong interconnected community of machine learning researchers in the same location.

Table 5 shows the top authorities relative to a different root node set consisting of Sergey Brin and Larry Page, who developed the PageRank algorithm and Jon Kleinberg who developed HITS. As with previous experiments, the results for each algorithm correlate quite strongly. For example, all four algorithms seem to suggest that relative to the three root nodes, Rajeev Motwani, Prabhakar Raghavan, Jeff Ullman and Craig Silverstein are the most important. These results are intuitive for several reasons: 1) Ullman is a well known researcher in the field of databases and has coauthored papers with Brin, Motwani and Silverstein, 2) Silverstein and Motwani have also coauthored papers with Brin and Page, 3) Motwani and Raghavan are close collaborators having, among other things, written the well known book *Randomized Algorithms* together, and 4) Motwani and Raghavan have also coauthored several papers with Kleinberg on Web graph analysis and thus can be seen as a key

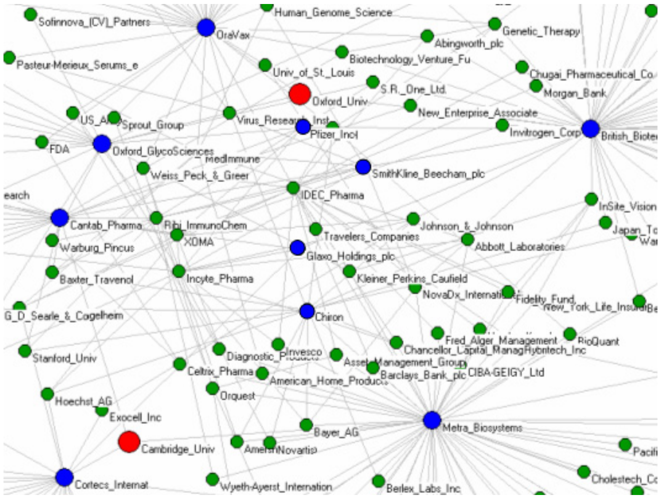

Figure 5: A portion of the biotechnology network.

Table 3: Importance rankings for the biotechnology network with respect to nodes Cambridge University and Oxford University.

Rank	PRankP		HITSP		WKPaths		KSMarkov	
1:	CambridgeU	0.1537	OxfordU	0.1510	OxfordU	0.0020	Cortecs	0.0616
2:	OxfordU	0.1531	CambridgeU	0.1510	CambridgeU	0.0020	Cantab	0.0559
3:	Cortecs	0.0480	Metra	0.0088	OxfordGlyco	0.0016	BritishBio	0.0550
4:	Cantab	0.0453	BritishBio	0.0084	Cantab	0.0016	Metra	0.0532
5:	BritishBio	0.0451	OraVax	0.0080	OraVax	0.0016	OraVax	0.0510
6:	Metra	0.0443	Cantab	0.0075	BritishBio	0.0016	OxfordGlyco	0.0428
7:	OraVax	0.0432	OxfordGlyco	0.0072	Glaxo	0.0015	Pfizer	0.0069
8:	OxfordGlyco	0.0395	Cortecs	0.0072	Metra	0.0015	Glaxo	0.0066
9:	Pfizer	0.0046	NIH	0.0068	SmithKline	0.0014	Incyte	0.0066
10:	Glaxo	0.0044	Chiron	0.0055	Pfizer	0.0014	CambridgeU	0.0056

Table 6: Correlations of top-10 rankings in Table 2.

	PRankP	HITSP	WKPaths	MarkovC	KSMarkov
PRankP	1	0.80	0.87	0.47	0.98
HITSP	0.80	1	0.76	0.52	0.82
WKPaths	0.87	0.76	1	0.44	0.89
MarkovC	0.47	0.52	0.44	1	0.43
KSMarkov	0.98	0.82	0.89	0.43	1

bridge connecting Brin and Page to Kleinberg.

9. CORRELATIONS OF RANKED LISTS

We used the K-Min (minimizing Kendall distance) metric of Fagin et al. (2003) to measure distances between the top-10 lists produced by the different relative importance algorithms. We normalize the K-min metric so that K-min$(s1, s2) = 1$ when both lists are identical and K-min$(s1, s2) = 0$ when one list is the reverse ordering of the other (most dissimilar). Table 6 shows one such example of pairwise correlations (corresponding to the terrorist network in Table 2).

Experimental results across all of the data sets used in this paper indicate that all five algorithms tend to be strongly correlated (according to K-min distance) in their top 10 rankings. However, there is some variation depending on the topology of the graph and whether the graph is directed or not. For example, for both directed and undirected graphs, PRankP and WKPaths tend to very highly correlated (often within 0.9 to 0.8 K-Min distance from another) while MarkovC tends to be the most dissimilar from the others (usually no more than 0.5 to 0.7). For undirected graphs, HITSP tends to be highly correlated with PRankP and WKPaths (usually within 0.9 to 0.8) but for directed graphs the correlation is often much weaker (usually within 0.8 to 0.6). KSMarkov appears to be slightly less correlated with PRankP, WKPaths, and HITSP and marginally more correlated with MarkovC than the others (although KSMarkov is different in Table 6 in this respect).

10. CONCLUSIONS

In this paper we provided a first step in addressing the problem of answering "relative importance" queries on graph data sets. We described a general framework that can be used to estimate the importance of nodes in a graph relative to a root set R and proposed several new algorithms within this framework based on both graph-theoretic notions of weighted paths and Markov chain models. In ongoing work, we are addressing a number of limitations we have identified—for example, how weighted edges can be incorpo-

Table 4: Importance rankings for the coauthorship network with respect to the Tom Mitchell node.

Rank	PRankP		HITSP		WKPaths		KSMarkov	
1	Mitchell	0.342	Mitchell	0.322	Mitchell	0.005	McCallum	0.070
2	Freitag	0.054	Thrun	0.038	Thrun	0.004	Freitag	0.067
3	McCallum	0.054	McCallum	0.038	Freitag	0.003	Mitchell	0.067
4	Thrun	0.051	Freitag	0.035	McCallum	0.003	Thrun	0.064
5	Joachims	0.050	Nigam	0.034	Nigam	0.002	Joachims	0.061
6	Armstrong	0.046	Blum	0.032	Joachims	0.002	Armstrong	0.054
7	Nigam	0.040	Joachims	0.031	Armstrong	0.002	Nigam	0.046
8	Blum	0.036	Armstrong	0.031	Blum	0.002	Blum	0.041
9	O'Sullivan	0.035	O'Sullivan	0.030	O'Sullivan	0.002	O'Sullivan	0.038
10	Seymore	0.011	Seymore	0.006	Caruana	0.001	Seymore	0.019

Table 5: Importance rankings for the coauthorship network with respect to nodes Brin, Page, and Kleinberg.

Rank	PRankP		HITSP		WKPaths		KSMarkov	
1:	Brin	0.2014	Brin	0.1119	Kleinberg	0.0023	Brin	0.1045
2:	Page	0.1352	Kleinberg	0.1107	Brin	0.0019	Motwani	0.0627
3:	Kleinberg	0.1137	Page	0.1087	Motwani	0.0017	Ullman	0.0536
4:	Motwani	0.0474	Motwani	0.0184	Raghavan	0.0016	Silverstein	0.0467
5:	Ullman	0.0429	Raghavan	0.0147	Page	0.0014	Page	0.0394
6:	Silverstein	0.0392	Ullman	0.0136	Silverstein	0.0014	Kleinberg	0.0194
7:	Raghavan	0.0111	Silverstein	0.0119	Ullman	0.0014	Raghavan	0.0138
8:	Lynch	0.0086	Williamson	0.0113	Williamson	0.0012	Zhang	0.0109
9:	Kedem	0.0086	Papadimitriou	0.0110	Vempala	0.0012	Guibas	0.0106
10:	Williamson	0.0085	Lynch	0.0108	Indyk	0.0010	Robertson	0.0101

rated (for example, for authors who author multiple papers with another author) and how methods such as Markov Centrality be scaled up computationally to very large graphs.

Acknowledgements

The data used were generously made availably by Steve Lawrence (Citeseer) and Walter Powell (biotech). The research in this paper was supported by the National Science Foundation under Grants NSF-IRI-9703120 and NSF-IIS-0083489 (PS) and by a National Defense Science and Engineering Graduate Fellowship (SW). We also thank Carter Butts and Douglas White for useful discussions and Joshua O' Madadhain for extensive assistance in producing the results in this paper.

References

Borodin, A., Roberts, G., Rosenthal, J., and Tsparas, P. (2001). Finding authorities and hubs from the link structures of the World Wide Web. *Proceedings of the 10th International World Wide Web Conference*, Hong Kong, pp. 415–429.

Brin, S., and Page, L. (1998) The anatomy of a large-scale hypertextual Web search engine. *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, pp. 107–117.

Chang, H., Cohn, D., and McCallum, A. (2000) Creating customized authority lists. *Proceedings of the 17th International Conference on Machine Learning*, San Francisco, CA: Morgan Kaufmann, pp. 167–174.

Fagin R., Kumar R., Sivakumar D. (2003) Comparing top k lists. *Proceedings of the 14th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 28–36.

Freeman, L. (1979) Centrality in social networks: I. conceptual clarification. *Social Networks*, 1, pp. 215-239.

Haveliwala, T. (2002) Topic-sensitive PageRank. *Proceedings of the 11th International World Wide Web Conference*, Honolulu, Hawaii, pp. 517–526.

Hoff, P. D., Raftery, A. E., Handcock, M. S. (2002) Latent space approaches to social network analysis. *Journal of The American Statistical Association*, 97, pp. 1090–1098.

Jeh, G. and Widom J. (2002) Scaling personalized Web search. Stanford University, Computer Science Department Technical Report.

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

Kemeny, J. and Snell, J. (1976) *Finite Markov Chains.* Springer Verlag.

Kleinberg, J. (1999) Authoritative sources in a hyperlinked environment. *Journal of ACM*, 46(5), pp. 604–632.

Krebs, V. E. (2001) Mapping networks of terrorist cells. *Connections*, 24(3), pp. 43–52.

Lawrence, S., Giles, C. L., and Bollacker, K. (1999) Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6), pp. 67–71.

Lempel, R., and Moran, S. (2000) The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Proceedings of the 9th International World Wide Web Conference*, Amsterdam, NL, pp. 387–401.

Page, L., Brin, S., Motwani, R., Winograd, T. (1998) The PageRank citation ranking: bringing order to the web. Stanford University, Computer Science Department Technical Report.

Powell, W. W., White, D. R., Koput, K. W., and Owen-Smith, J. (2002) Network dynamics and field evolution: the growth of interorganizational collaboration in the life sciences. Submitted.

Stephenson, K., and Zelen, M. (1989) Rethinking centrality: methods and examples. *Social Networks*, 11, pp. 1–37.

Wasserman, S., and Faust, K. (1994) *Social Network Analysis: Methods and Applications.* Cambridge: Cambridge University Press.