Link-Based Ranking of the Web with Source-Centric Collaboration

(Invited Paper)

James Caverlee
Georgia Institute of Technology
Atlanta, Georgia 30332 USA
caverlee@cc.gatech.edu

Ling Liu
Georgia Institute of Technology
Atlanta, GA 30332 USA
lingliu@cc.gatech.edu

William B. Rouse
Georgia Institute of Technology
Atlanta, GA 30332 USA
brouse@isye.gatech.edu

***Abstract*— Web ranking is one of the most successful and widely used collaborative computing applications, in which Web pages collaborate in the form of varying degree of relationships to assess their relative quality. Though many observe that links display strong source-centric locality, for example, in terms of administrative domains and hosts, most Web ranking analysis to date has focused on the flat page-level Web linkage structure. In this paper we develop a framework for link-based collaborative ranking of the Web by utilizing the strong Web link structure. We argue that this source-centric link analysis is promising since it captures the natural link-locality structure of the Web, can provide more appealing and efficient Web applications, and reflects many natural types of structured human collaborations. Concretely, we propose a generic framework for source-centric collaborative ranking of the Web. This paper makes two unique contributions. First, we provide a rigorous study of the set of critical parameters that can impact source-centric link analysis, such as source size, the presence of self-links, and different source-citation link weighting schemes (e.g., uniform, link count, source consensus). Second, we conduct a large-scale experimental study to understand how different parameter settings may impact the time complexity, stability, and spam-resilience of Web ranking. We find that careful tuning of these parameters is vital to ensure success over each objective and to balance the performance across all objectives.**

I. Introduction

From its earliest days, the Web has been the subject of intense focus for organizing, sorting, and understanding its massive amount of data. One of the most popular and effective Web analysis approaches is collaborative Web ranking, in which link relationships on the Web are used to assess the importance of Web pages. By considering the number and nature of link relationships among Web pages, each page can be ranked according to the overall view of all Web pages. The essence of this collaborative approach to ranking has been adapted to power community-based Web discovery tools like Digg and to organize massive collaborative review and feedback systems like those used by eBay and Amazon.

Interestingly, most link-based ranking algorithms to date have been based on the most basic Web element – Web pages. Page-based link analysis relies on a fundamentally flat view of the Web, in which all pages are treated as equal nodes in a Web graph. In contrast, a number of recent studies have noted a strong Web link structure, in which links display strong source-centric locality in terms of domains and hosts (e.g., [3], [18]). This link-locality naturally suggests the importance of source-centric link analysis. Complementary to the page-based view, the source-centric view relies on a hierarchical abstraction of the flat page-level view and reflects many natural types of structured human collaborations. For example, we could imagine ranking all database students at a university according to the views of the database community alone. We could then move up the hierarchy to rank the database department relative to the other departments at the university. Finally, we could rank the entire university relative to all other universities. Hence, this structured collaborative approach allows us to treat the nature of relationships at each level differently.

Research on source-centric link analysis has shown some initial success, however, most studies over the past years have focused exclusively on a single goal – improving the efficiency of page-based ranking algorithms (e.g., [7], [18], [22]). All of the approaches have explored only a fraction of the parameter space, leaving many important questions unanswered. We argue that fully exploring source-centric link analysis can have a profound impact on link-based algorithms and our general understanding of the Web.

In this paper, we propose a parameterized framework to support the systematic study and evaluation of source-centric link analysis of the Web in a variety of application settings. We address the following three important open questions:

- What are the most important parameters for guiding source-centric link analysis?
- How should these parameters be set to achieve the specific objectives of the source-centric link analysis?
- What impact do the parameter settings have on the effectiveness of the analysis? Do certain parameter settings conflict or correlate with the objectives?

Concretely, we identify a set of critical parameters that can impact the effectiveness of source-centric link analysis, including source size, the presence of self-links, and different source-citation link weighting schemes (e.g., uniform, link count, source consensus). We provide a rigorous study on the set of critical parameters, especially with respect to the above three open questions. All previously proposed approaches are instances of our parameterized framework and have certain drawbacks in terms of their applicability to different source-

centric link analysis objectives. We conduct a large-scale comparative study of different parameter settings of source-centric link analysis over three large Web datasets against multiple and possibly competing objectives. Through experimental evaluation of our parameterized framework over three objectives – time complexity, stability, and spam-resilience – we show how the parameters should be tuned to ensure efficient, stable, and robust Web ranking.

II. Key Parameters for SLA

In this section, we identify important parameters for guiding source-centric link analysis (**SLA**), and discuss how these parameters impact the effectiveness of link analysis. Source-centric link analysis relies on a source view of the Web. Just as the page graph $\mathcal{G}_\mathcal{P} = \langle \mathcal{P}, \mathcal{L}_\mathcal{P} \rangle$ models the Web as a directed graph where the nodes of the graph correspond to Web pages $\mathcal{P}$ and the set of directed edges $\mathcal{L}_\mathcal{P}$ correspond to hyperlinks between pages, the *source graph* has nodes that correspond to sources and edges that denote the linkage between sources. We use the term *source edge* to refer to the notion of source-centric citation. A source s_1 has a source edge to another source s_2 if one page in s_1 has a hyperlink to a page in s_2. We call s_1 the originating source and s_2 the target source.

In general, a source graph can consist of multiple levels of source-hierarchy; that is, a page may belong to a source that belongs to a larger source, and so on. In the rest of this paper we shall require that each page in the page graph belong to one and only one source in the source graph, meaning that the hierarchal view of the Web consists of two-levels: a page level and a source level. Hence, a Web source graph $\mathcal{G}_\mathcal{S} = \langle \mathcal{S}, \mathcal{L}_\mathcal{S} \rangle$ is a directed graph where the nodes of the graph correspond to Web sources in $\mathcal{S}$ and the set of directed edges $\mathcal{L}_\mathcal{S}$ corresponds to source edges as described above.

A. Overview

Given the source view of the Web, we next discuss the choice of parameters for guiding source-centric link analysis. The choice of parameters and their specific settings are greatly impacted by the particular application of the link analysis (e.g., ranking, categorization, clustering). In this paper, we focus our parameter discussion on three objectives that are fundamental across link analysis applications and of particular importance to Web ranking:

- Time complexity: Since the Web is incredibly large, our first objective is to leverage the higher source-abstraction level to improve the time complexity relative to page-based approaches.
- Stability: The second objective is to study the stability of source-centric link analysis in the face of the Web's constant evolution.
- Spam-resilience: Finally, since Web spammers deliberately manipulate link analysis algorithms, our third objective is to understand the spam-resilience properties of source-centric link analysis.

We identify four key parameters that impact source-centric link analysis with respect to these three objectives:

- Source Definition (Γ): The first and most important parameter is the source definition. The determination of how sources are organized is at the heart of source-centric link analysis and all other parameter settings are entirely dependent on the source definition.
- Source-Centric Citation (Θ): The second parameter we consider is the nature of the citation-based association between sources. We study the presence and strength of the linkage arrangements from one source to another.
- Source Size (Ξ): Since sources may vary greatly in the number of constituent pages, the third parameter we study is source size and how this non-linkage information may be directly incorporated into the analysis.
- Application-Specific Parameters (Υ): Finally, there may be some additional application-specific parameters that are necessary, e.g., the number of iterations to run a ranking algorithm until sufficient convergence.

We describe source-centric link analysis in terms of an application, a specific objective, and as a combination of these four parameters: $SLA_{<app,obj>}(\Gamma; \Theta; \Xi; \Upsilon)$.

In the following sections, we discuss the first three of these important parameters, present some of their possible settings, and provide insight into how best these parameters may be assigned based on the ultimate objectives of the link analysis. We examine the fourth parameter in the context of Web ranking in Section III. As we will see in our evaluation in Section IV, a careful approach to these parameters is necessary to ensure high-quality results across objectives.

B. Parameter 1: Source Definition

How does the source definition impact the quality of source-centric link analysis with respect to the three objectives? Clearly, the determination of how sources are organized should have a profound impact on the quality and value of source-centric link analysis. To understand the importance of source definition, we consider five different approaches – at one extreme we treat each page as a unique source, meaning that the source view of the Web corresponds directly to the page view; at another extreme we disregard all page relationships and randomly assign pages to sources. The other approaches rely on the link-locality of the Web and assign pages based on their administrative organization – by domain, host, or directory.

To illustrate the locality-based linking phenomenon on the Web, we consider three large real-world Web datasets. The first dataset – **UK2002** – is derived from a 2002 crawl of the `.uk` top-level-domain by UbiCrawler [4]. The second dataset – **IT2004** – is derived from a 2004 crawl of the `.it` top-level-domain, again by UbiCrawler. The third dataset – **WB2001** – was originally collected by the Stanford WebBase project[1] and

[1] `http://dbpubs.stanford.edu:8090/~testbed/doc2/WebBase/`

includes pages from a wide variety of top-level-domains. All three datasets are available at `http://webgraph-data.dsi.unimi.it/`. For each dataset we report the number of pages and links in Table I, where the data has been cleaned by the typical pre-processing step of removing all self-links.

TABLE I
SUMMARY OF DATASETS (IN MILLIONS)

Dataset	Pages	Links
UK2002	18.5	292.2
IT2004	41.3	1,135.7
WB2001	118.1	992.8

In Table II, we report four classes of links over these three datasets. We report the fraction of all links that point from pages in one domain to pages in the *same* domain (intra-domain links), the fraction that point from pages in one host to pages in the *same* host (intra-host links), and the fraction that point from pages in one directory to pages in the *same* directory or lower in the directory hierarchy (intra-directory links). Note that we consider intra-directory links from the first directory level only. Since the WB2001 dataset includes pages from many domains, we also report the fraction of pages in WB2001 that point from pages in one top-level domain (TLD) to pages in the *same* TLD (intra-TLD links).

TABLE II
FRACTION OF PAGE LINKS

Dataset	Intra-TLD	Intra-Domain	Intra-Host	Intra-Directory
UK2002	–	94.6%	92.3%	66.9%
IT2004	–	91.0%	90.8%	67.9%
WB2001	97.9%	95.5%	94.1%	62.7%

These statistics consistently show that the Web exhibits a strong locality-based link structure. Given this phenomenon, it is natural to assign pages to sources based on one of these administrative organizations. Hence, we study five different settings for the source definition parameter Γ – by domain, by host, by directory, as well as the extremes of by page, and by random assignment.[2]

As we shall see in Section IV, the analysis quality depends heavily on the presence of link locality and the source definition. We find that a lack of locality results in poor time complexity, but that even moderate locality ($\sim 65\%$) leads to good time complexity and stability results that are comparable with source definitions with extremely high locality.

C. Parameter 2: Source-Centric Citation

Unlike the straightforward notion of linkage in the page graph, source edges are derived from the page edges in the underlying page graph. Different page edges often carry different significance with respect to the sources involved. Careful design that takes these factors into account is critical, and so the second parameter we study is the nature and strength of source-centric citation from one source to another.

Given the directed source graph $\mathcal{G}_\mathcal{S} = \langle \mathcal{S}, \mathcal{L}_\mathcal{S} \rangle$, our goal is to understand the source-centric citation in terms of the appropriate edge weights for the set of directed edges $\mathcal{L}_\mathcal{S}$. Let $w(s_i, s_j)$ denote the weight assigned to the source edge $(s_i, s_j) \in \mathcal{L}_\mathcal{S}$. We consider source-centric citation as a scalar value in the range $[0, 1]$, where the outgoing edge weights for any source sum to 1. In cases where the normalization is not explicit, we will require the normalization of the raw edge weights. We consider six edge weighting schemes.

1. Uniform: This is the simplest case where all source edges pointing out from an originating source are treated equally. This *uniform (u)* weighting is defined as:

$$w_u(s_i, s_j) = \frac{1}{\sum_{s_k \in \mathcal{S}} \mathcal{I}\left[(s_i, s_k) \in \mathcal{L}_\mathcal{S}\right]}$$

where the indicator function $\mathcal{I}(\cdot)$ resolves to 1 if the argument to the function is true, and 0 otherwise.

Since each node in the source graph is an aggregation of one or more pages, treating each source edge equally may not properly capture the citation strength between two sources. With this in mind, we next introduce three source edge weighting schemes that are based on the hyperlink information encoded in the page graph $\mathcal{G}_\mathcal{P} = \langle \mathcal{P}, \mathcal{L}_\mathcal{P} \rangle$.

2. Link Count: The link count scheme assigns edge weights based on the count of *page links* between pages that belong to sources. Such an edge weighting is effective when we would like to reward sources that have strong linkage at the page level. We define the *link count (lc)* weighting as:

$$w_{lc}(s_i, s_j) = \sum_{p_i | s(p_i) = s_i; p_j | s(p_j) = s_j} \mathcal{I}\left[(p_i, p_j) \in \mathcal{L}_\mathcal{P}\right]$$

where the source to which page p_i belongs is denoted $s(p_i)$.

3. Source Consensus: This edge weighting scheme counts the number of *unique pages* within an originating source that point to a target source. The main motivation behind the design of this scheme is to address the weakness of the link count weighting scheme. For example, we may wish to differentiate between the case where a single page within the originating source is contributing all n links to the target, and the case where there are n pages in the originating source and each has a single link to the target. We capture this notion of *source consensus (sc)* in the following edge weighting definition:

$$w_{sc}(s_i, s_j) = \sum_{p_i | s(p_i) = s_i} \left(\bigvee_{p_j | s(p_j) = s_j} \mathcal{I}\left[(p_i, p_j) \in \mathcal{L}_\mathcal{P}\right] \right)$$

4. Target Diffusion: In contrast to how many pages in the originating source are responsible for the page links between sources, another factor that is of interest when evaluating source-citation strength is the number of different target pages that are pointed to by the originating source. The *target diffusion (td)* weighting is defined as:

$$w_{td}(s_i, s_j) = \sum_{p_j | s(p_j) = s_j} \left(\bigvee_{p_i | s(p_i) = s_i} \mathcal{I}\left[(p_i, p_j) \in \mathcal{L}_\mathcal{P}\right] \right)$$

Each of the previous three alternatives to the uniform edge weighting scheme – *link count*, *source consensus*, and *target diffusion* – relies exclusively on the page linkage between the

[2]Of course, not all pages grouped by domain, host, or directory will always form a coherent Web source. It may also make sense to assign pages to sources based on their topical locality as identified in [10]. We are pursuing these issues in our continuing research.

component pages in each source. In addition to these purely link-based approaches, we also consider two approaches that rely on both the page links and the *quality* of the pages that provide the linking, where we denote page p_i's quality score by $q(p_i)$. There are a number of ways for assigning a quality value to each page, including the PageRank score for the page or by using a simple heuristic like the page's relative depth in the directory tree. Additionally, content-based factors may be incorporated in the quality component to reward certain source associations, like those between topically-similar sources.

5. Quality-Weighted Link Count: This edge weighting scheme directly integrates the page quality score into the *link count* weighting scheme. Let (q) denote the use of a page quality metric. We define the quality-weighted link count scheme as follows:

$$w_{lc(q)}(s_i, s_j) = \sum_{p_i|s(p_i)=s_i; p_j|s(p_j)=s_j} q(p_i) \cdot \mathcal{I}\left[(p_i, p_j) \in \mathcal{L}_{\mathcal{P}}\right]$$

6. Quality-Weighted Source Consensus: Similarly, we can integrate the page quality score into the *source consensus* edge weighting scheme to produce the quality-weighted source consensus edge weighting scheme:

$$w_{sc(q)}(s_i, s_j) = \sum_{p_i|s(p_i)=s_i} q(p_i) \cdot \left(\bigvee_{p_j|s(p_j)=s_j} \mathcal{I}\left[(p_i, p_j) \in \mathcal{L}_{\mathcal{P}}\right]\right)$$

Interesting to note is that there is not a natural quality-weighted extension to the *target diffusion* edge weighting scheme since this edge weighting scheme is not focused on which page in the source is providing the forward linkage.

Another factor that can influence source-centric citation is whether we take into account self-edges. Given a particular edge weighting scheme, there may be some applications that require self-edges, while others do not. For example, in a ranking context, a self-edge may be interpreted as a self-vote by the source, meaning that the source could manipulate its own rank. In the case where self-edges are eliminated, we will require the edge weight $w(s_i, s_i) = 0$ for all $s_i \in \mathcal{S}$. On the other hand, it may be reasonable to include self-edges since the locality-based structure of Web links indicates a strong degree of association between a source and itself.

Hence, we shall consider twelve different settings for the source citation parameter Θ – the looped and loop-less versions of the six association strength edge weighting schemes. We find that some edge weighting schemes are extremely vulnerable to spam manipulation, while others are much less vulnerable. In terms of stability, we find that self-edges have a very strong impact.

D. Parameter 3: Source Size

Since sources may vary greatly in size, from a source of a single page to a source encompassing millions of pages, what is the impact of source size on the underlying objectives of source-centric link analysis? For many applications it may be reasonable to distinguish between sources based on the per-source size discrepancy. Source size is one example of non-linkage information that can be incorporated into the link analysis. Of course, there could be other non-link information of interest (like source topic or source trustworthiness), but in this paper we shall restrict our examination to source size. The parameter Ξ considers two options – the size in pages of each source s_i (denoted by $|s_i|$) and no size information. As we shall see in Section IV, source size is a very important parameter for the stability of the algorithm, but results in the least satisfactory in terms of spam-resilience. In our experiments we further explore this fundamental tension.

III. Applying SLA to Web Ranking

The parameters introduced in the previous section can be combined in a number of way to achieve a particular objective with respect to a link-based application (e.g., ranking, clustering). To more fully examine source-centric link analysis, we select one application area – Web ranking – and examine the parameter settings with respect to the three objectives – time complexity, stability, and spam-resilience. Source-centric ranking has intuitive appeal since many users may be interested in identifying highly-ranked sources of information (e.g., CNN or ESPN) rather than specific pages.

Here, we adopt a ranking approach that is similar in spirit to the "random surfer" model often used to describe PageRank, but adapted to source-centric link analysis. Just as PageRank provides a single global authority score to each page on the Web based on the linkage structure of the entire Web, the source-centric ranking approach (SLA_{Rank}) can be used to rank all sources. In general, a source will be ranked highly if many other high-ranking sources point to it. We denote source s_i's authority score as σ_i, where $\sigma_i > \sigma_j$ indicates that the i^{th} source is more important than the j^{th} source. We write the authority score for all sources using the vector notation $\boldsymbol{\sigma}$, where all $|\mathcal{S}|$ sources are assigned a score.

The random walk over the source graph proceeds as follows. For each source $s \in \mathcal{S}$:

- With probability α, the random source walker follows one of the source edges of source s;
- With probability $1-\alpha$, the random source walker teleports to a randomly selected source.

We refer to the first option as the *edge following factor* and the second option as the *teleportation factor*. Associated with the *edge following factor* is an $|\mathcal{S}| \times |\mathcal{S}|$ transition matrix $\mathbf{T}$, where the ij^{th} entry indicates the probability that the random source walker will navigate from source s_i to source s_j. Associated with the *teleportation factor* is an $|\mathcal{S}|$-length teleportation probability distribution $\mathbf{c}$, where c_i indicates the probability that the random walker will teleport to source s_i. Such a random walk may be modelled by a time-homogenous Markov Chain and written in terms of the stochastic transition matrix $\hat{\mathbf{T}}$, where $\hat{\mathbf{T}}$ is a combination of both the *edge following factor* and the *teleportation factor* according to the mixing

parameter α:[3]

$$\hat{\mathbf{T}} = \alpha \cdot \mathbf{T} + (1 - \alpha) \cdot \mathbf{1} \cdot \mathbf{c}^T$$

Since the source graph may have disconnected components and to gracefully deal with nodes that have no out-links, the teleportation factor is included as a "fix" to guarantee that the transition matrix associated with the Markov chain be both aperiodic and irreducible, which ensures convergence to a stationary distribution. The stationary distribution of $\hat{\mathbf{T}}$ (which is its principal eigenvector) encodes the long-term probability of a random walker being at each particular source. We can interpret this distribution as $\boldsymbol{\sigma}$, encoding the authority scores for all sources.

Given the source-centric ranking model (SLA_{Rank}), we next address two questions: (1) How do the source-centric link analysis parameters map to the Web ranking context? and (2) How do we evaluate the objectives of link analysis in the context of Web ranking?

1. Mapping Parameters: All four parameters – Source Definition (Γ), source-centric Citation (Θ), Source Size (Ξ), and the Application-Specific Parameters (Υ) – impact Web ranking. Clearly, the source definition is critically important since it determines the fundamental unit of ranking. The source-centric citation is necessary to construct the transition matrix $\mathbf{T}$ according to the edge weights determined by Θ, that is $T_{ij} = w(s_i, s_j)$). The source size parameter can be used to guide the teleportation factor – that is $c_i = |s_i| / \sum_{j=1}^{|\mathcal{S}|} |s_j|$ – which intuitively captures the behavior of a random surfer being more likely to jump to large sources. Alternatively, source size can be disregarded so the teleportation factors defaults to a uniform distribution: $c_i = 1/|\mathcal{S}|$. For Web ranking, there are two application-specific parameters – the mixing parameter α and the convergence criterion for terminating the algorithm.

2. Evaluating Objectives: We briefly discuss each of the objectives and their necessary parameters. In addition to the time complexity, stability, and spam-resilience objectives, we also consider a fourth objective that is specific to Web ranking – approximating PageRank.

- **Time Complexity:** To measure time complexity, we examine the calculation efficiency of the source-centric ranking approach in terms of the time it takes to calculate each ranking vector: $SLA_{Rank;Time}(\Gamma; \Theta; \Xi; \Upsilon)$.
- **Stability:** We consider two flavors of stability. First, we evaluate the stability of the ranking algorithm as the Web graph evolves, and new pages and sources are discovered. Second, we investigate the stability in terms of the similarity of rankings induced by the various parameter settings: $SLA_{Rank;Stab}(\Gamma; \Theta; \Xi; \Upsilon)$.
- **Spam-Resilience:** Web spammers spend a considerable effort on manipulating Web-based ranking algorithms, and recent studies suggest that it affects a significant portion of all Web content, including 8% of pages [15] and 18% of sites [17]. To evaluate the spam-resilience properties, we measure the impact of several spam scenarios in terms of the ranking impact on a target source: $SLA_{Rank;Spam}(\Gamma; \Theta; \Xi; \Upsilon)$.
- **Approximating PageRank:** Finally, we consider the ranking-specific objective of approximating the traditional global PageRank vector by combining the source-level ranking information with per-source ranking information. Such approximation promises to speed the PageRank calculation considerably: $SLA_{Rank;Approx}(\Gamma; \Theta; \Xi; \Upsilon)$.

Several previous research efforts have considered a source-centric ranking calculation over groups of pages, including [1] and [12]. These approaches have had different ultimate objectives, and each approach has focused exclusively on a handful of parameter settings with respect to a single objective. The first approach sought to bootstrap the calculation of PageRank with an initial starting "guess" derived from a decomposition of the Web into a higher-level block layer and a local level [18]. The second approach has focused on replacing the traditional PageRank vector with an alternative ranking approach by determining a page's authority as a combination of multiple disjoint levels of rank authority (e.g., [7], [21], [22], [25], [30]); the traditional PageRank vector is never computed. The third approach decentralizes the computation of PageRank for use in peer-to-peer networks (e.g., [27], [29]).

Each of these previous approaches relies on only a few parameter settings in the context of a single objective and can be seen as a fairly limited exploration of the parameter space of SLA_{Rank}. For example, the BlockRank [18] and ServerRank [27] algorithms both consider host-level sources, a quality-weighted link count citation weight with self-edges, and disregard source size. By considering the five source definition parameter settings, the 12 source-citation settings, and the two teleportation vectors, we examine 120 different parameter settings for source-centric ranking (SLA_{Rank}), which we evaluate over four distinct objectives. To the best of our knowledge, ours is the first study to consider such a large parameter space and in the context of multiple, possibly competing objectives. In the following section, we shall conduct the first large-scale comparative study to evaluate source ranking in terms of this parameter space and the four objectives. In our continuing research, we are studying the formal properties of these competing objectives. In this paper, we shall evaluate them experimentally.

IV. Experimental Evaluation

In this section, we evaluate source-centric link analysis in the context of Web ranking with respect to four objectives – time complexity, ranking stability, spam-resilience, and approximating PageRank. We are interested to understand what are the most important parameters, how these parameters

[3]We adopt a fairly standard solution for handling sources with no outlinks (so-called dangling sources), whereby we make the transition matrix row stochastic by adding new edges from each dangling source to every other source. Due to the space restriction, we omit the detailed discussion here.

TABLE III
SOURCE GRAPH SUMMARY – BY SOURCE DEFINITION

Dataset	Domain		Host		Dir		Rand		Page	
	Nodes	Links	Nodes	Links	Nodes	Links	Nodes	Links	Nodes	Links
UK2002	81k	1.2m	98k	1.6m	360k	3.5m	98k	286.0m	18.5m	292.2m
IT2004	136k	2.7m	141k	2.8m	505k	8.6m	141k	1,069.3m	41.3m	1,135.7m
WB2001	620k	10.5m	739k	12.4m	3,315k	24.7m	739k	955.4m	118.1m	992.8m

should be set, and what impact these parameter settings have on competing objectives. We find that careful tuning of these parameters is vital to ensure success over each objective, and that some objectives cannot be maximized without negatively impacting other objectives.

A. Experimental Setup

All of our experimental evaluation is over the three Web datasets described in Section II-B. For each dataset we extracted the domain, host, and directory information for each page URL and assigned pages to sources based on these characteristics. We also consider the extreme case when each page belongs to its own source (equivalent to the page graph described in Table I). For the random source definition, we set the number of nodes in the graph to be the same as the number of hosts. In Table III, we present summary information for each of the source graphs (including self-edges). Note that the random source graph displays nearly no link-locality, with less than 1% of all page links being intra-source.

All of the ranking code was written in Java. The data management component was based on the WebGraph compression framework described in [6] for managing large Web graphs in memory. All experiments were run on a dual processor Intel XEON at 2.8GHz with 8GB of memory. We measured the convergence rate for all ranking calculations using the L2-distance of successive iterations of the Power Method. We terminated the ranking calculations once the L2-distance dropped below a threshold of 10e-9.

As a baseline, we computed the global PageRank vector ($\boldsymbol{\pi}$) over each page graph using the standard Power Method and the parameters typically used in the literature (e.g., [24]), including a mixing parameter of 0.85, a uniform teleportation vector, and a uniform link following probability.

For the quality-weighted edge weighting, we measure the quality of each page $q(p_i)$ using the page's PageRank score π_i. Although in practice it may not be reasonable to use the PageRank score as a measure of quality since it is so expensive to calculate, we include these PageRank-weighted options here to more fully understand their impact relative to the edge weighting schemes that do not require PageRank.

For compactness, we shall write a particular SLA_{Rank} parameter combination like $\mathcal{SR}(\mathbf{T}^*_U, \mathbf{c}_u)$, where the transition matrix $\mathbf{T}$ is appended with a subscript to indicate which source edge weighting scheme we use: $\mathbf{T}_U$, $\mathbf{T}_{LC}$, and so on. We shall append an asterisk to the transition matrix to indicate the inclusion of self-edges: $\mathbf{T}^*$. For the choice of teleportation vector $\mathbf{c}$, we consider the standard uniform vector ($\mathbf{c}_u$) and the source-size-based vector ($\mathbf{c}_s$).

B. Measures of Ranking Distance

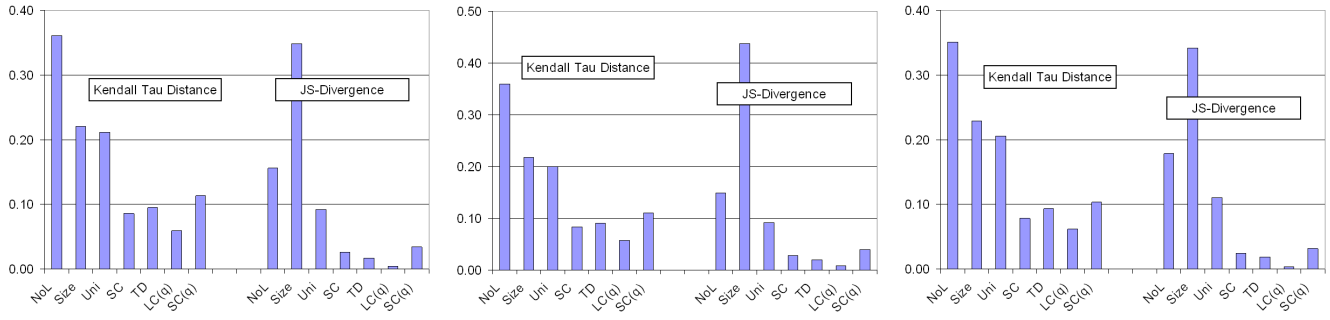
We rely on two distance metrics for comparing ranking vectors. The Kendall Tau Distance Metric [14] is based solely on the relative ordering of the sources in two ranking vectors. In contrast, the Jensen-Shannon Divergence [20] measures the distributional similarity of two vectors, meaning that it considers the magnitude of each source's authority score and not just the relative ordering of the sources.

Kendall Tau Distance Metric. This metric measures the relative ordering of two lists of ranked objects [14]. It is based on the original Kendall Tau correlation described in [19] and provides a notion of how closely two lists rank the same set of objects (or Web sources in our case). The Kendall Tau distance metric takes values in the range [0,1], where two rankings that are exactly the same have a distance of 0, and two rankings in the reverse order have a distance of 1. We rely on a variation of an $O(n \log n)$ version described in [5].

JS-Divergence. The Jensen-Shannon divergence is a measure of the distributional similarity between two probability distributions [20]. It is based on the relative entropy measure (or KL-divergence), which measures the difference between two probability distributions p and q over an event space X: $KL(p,q) = \sum_{x \in X} p(x) \cdot \log(p(x)/q(x))$. If we let p be one of the ranking vectors $\boldsymbol{\sigma}$, and q be the other ranking vector $\boldsymbol{\sigma}'$, then we have $KL(\boldsymbol{\sigma}, \boldsymbol{\sigma}') = \sum_{i \in \mathcal{S}} \sigma_i \cdot \log(\sigma_i/\sigma'_i)$. Intuitively, the KL-divergence indicates the inefficiency (in terms of wasted bits) of using the q distribution to encode the p distribution. Since the KL-divergence is not a true distance metric, the JS-divergence has been developed to overcome this shortcoming, where:
$JS(\boldsymbol{\sigma}, \boldsymbol{\sigma}') = \phi_1 KL(\boldsymbol{\sigma}, \phi_1\boldsymbol{\sigma}+\phi_2\boldsymbol{\sigma}')+\phi_2 KL(\boldsymbol{\sigma}', \phi_1\boldsymbol{\sigma}+\phi_2\boldsymbol{\sigma}')$ and $\phi_1, \phi_2 > 0$ and $\phi_1 + \phi_2 = 1$. In the experiments reported in this paper, we consider $\phi_1 = \phi_2 = 0.5$. The JS-divergence takes values in the range [0,1] with lower values indicating less distance between the two ranking vectors.

C. Objective-Driven Evaluation

We report the most significant results from a total of 360 different ranking vectors we computed by combining the five source definitions, the 12 source-citation edge weights, the two teleportation vectors, and the three datasets. For the 360 ranking vectors we analyze, we fix the mixing parameter α at the commonly adopted value of 0.85. Note that we also varied α in our preliminary experiments, but find that there is no significant change in the results we report here.



(a) UK2002 (b) IT2004 (c) WB2001

Fig. 1. Parameter Tuning: Ranking Distance Versus Baseline Configuration

1) Time Complexity: We begin by examining the ranking efficiency of the source-centric ranking approach in terms of the time it takes to calculate each ranking vector. Recall that the PageRank-style calculation scans the link file for each source graph multiple times until convergence.

Table IV shows the average time per iteration to calculate the ranking vector over the five different source graphs for each of the datasets. We report the results for a ranking based on the uniform edge weight and a uniform teleportation vector: $\mathcal{SR}(\mathbf{T}_U, \mathbf{c}_u)$. These general per-iteration results also hold for the total time to reach the L2 stopping criterion. In our examination of the 360 different ranking vectors, we find that the source definition has the most significant impact on the calculation time, since the source definition directly impacts the size of the source graph. The choice of edge weights and teleportation vector has little discernable impact on the calculation time.

	Source Definition				
Dataset	**Domain**	**Host**	**Dir**	**Rand**	**Page**
UK2002	0.02	0.03	0.07	2.45	2.76
IT2004	0.05	0.05	0.13	9.33	9.44
WB2001	0.21	0.25	0.46	11.32	12.28

TABLE IV
WALLCLOCK TIME (MINUTES) PER ITERATION

As we can see, the directory, host, and domain source definitions result in ranking computation that is one to two orders of magnitude faster than the page-based graph. Since PageRank over a Web graph of billions of nodes takes days or weeks, this improvement is important for source-centric ranking to compensate for PageRank's slow time-to-update. The random source definition performs poorly, even though there are the same number of nodes in the random graph as in the host graph. The key difference is that the random graph has no link locality structure, and hence consists of nearly as many links as in the page graph. We conclude that link locality strongly impacts the degree of source graph size reduction, and hence, the ranking calculation time. Due to its poor performance, we shall drop the random source definition from the rest of our reported experimental results.

2) Stability – Ranking Similarity: We next explore the parameter space to investigate the stability in terms of the similarity of rankings induced by the various parameter settings. Due to its popularity in other works (e.g. [18], [29]), we adopt a baseline ranking based on the link count edge weight with self-edges and a uniform teleportation vector, $\mathcal{SR}(\mathbf{T}^*_{LC}, \mathbf{c}_u)$, and report seven alternative ranking vectors computed by tweaking these baseline parameter settings. We consider a version without self-edges ($\mathcal{SR}(\mathbf{T}_{LC}, \mathbf{c}_u)$), a version including self-edges and the size-based teleportation component ($\mathcal{SR}(\mathbf{T}^*_{LC}, \mathbf{c}_s)$), and five additional versions using the other edge weighting schemes (e.g., $\mathcal{SR}(\mathbf{T}^*_U, \mathbf{c}_u)$) as shown in Table V. We report the results for the host-based graph in this section; we see similar results across the directory and domain source definition settings.

Shorthand	**Version**	**Edge Weight**	**Self-Edges?**	**Telep. Factor**
Baseline	$\mathcal{SR}(\mathbf{T}^*_{LC}, \mathbf{c}_u)$	LC	Yes	Uniform
NoL	$\mathcal{SR}(\mathbf{T}_{LC}, \mathbf{c}_u)$	LC	No	Uniform
Size	$\mathcal{SR}(\mathbf{T}^*_{LC}, \mathbf{c}_s)$	LC	Yes	Size
Uni	$\mathcal{SR}(\mathbf{T}^*_U, \mathbf{c}_u)$	U	Yes	Uniform
SC	$\mathcal{SR}(\mathbf{T}^*_{SC}, \mathbf{c}_u)$	SC	Yes	Uniform
TD	$\mathcal{SR}(\mathbf{T}^*_{TD}, \mathbf{c}_u)$	TD	Yes	Uniform
LC(q)	$\mathcal{SR}(\mathbf{T}^*_{LC(q)}, \mathbf{c}_u)$	LC(q)	Yes	Uniform
SC(q)	$\mathcal{SR}(\mathbf{T}^*_{SC(q)}, \mathbf{c}_u)$	SC(q)	Yes	Uniform

TABLE V
RANKING SIMILARITY: PARAMETER SETTINGS

In Figure 1, we compare the ranking vector resulting from the baseline parameter settings with the ranking vector resulting from each of these seven alternative parameter settings. The y-axis measures the distance between these alternative ranking vectors and the baseline configuration via the Kendall Tau Distance Metric and the JS-Divergence.

As we can see, the exclusion of self-edges (`NoL`) and the choice of teleportation vector (`Size`) are the two factors with the most significant impact on the resulting ranking vector in terms of ranking distance from the baseline setting. Hence, we must be careful when setting these two critical parameters, since the resulting ranking vectors depend so heavily on them. The choice of edge weights has less impact, though we observe

that the uniform edge weighting results in the most dissimilar ranking vector of all the edge weighting schemes. The uniform edge weighting scheme is a less intuitively satisfactory edge weighting scheme, and these results confirm this view. What is interesting here is that the Source Consensus, Target Diffusion, Quality-Weighted Link Count, and Quality-Weighted Source Consensus edge weights have a relatively minor impact on the resulting ranking vector versus the baseline Link Count version. We note that the Quality-Weighted Link Count deviates very little from the Link Count version, in spite of the incorporation of the expensive PageRank scores. This is encouraging since it means we need not rely on PageRank for assessing source quality.

3) Stability – Link Evolution: We next evaluate the stability of SLA_{Rank} as the Web graph evolves for each of the three Web datasets. Since the source view of the Web provides an aggregate view over Web pages, we anticipate that domain, host, and directory-based rankings should be less subject to changes in the underlying page graph than page-based rankings. Our goal is to emulate the gradual discovery of Web pages, similar to how a Web crawler may incrementally discover new pages for ranking.

For each dataset, we randomly selected a fraction of the pages (10%, 30%, ...) and computed the standard PageRank vector over just this fraction of pages, yielding $\boldsymbol{\pi}_{10\%}, \boldsymbol{\pi}_{30\%}$, and so on. Additionally, we computed the ranking vector for the domain, host, and directory-based source graphs derived from the same fraction of all pages, yielding $\boldsymbol{\sigma}_{10\%}, \boldsymbol{\sigma}_{30\%}$, and so on. We then compared the relative page rankings for the pages in $\boldsymbol{\pi}_{10\%}$, $\boldsymbol{\pi}_{30\%}$, ..., to the relative rankings of the *exact same pages* in the PageRank vector for the full Web page graph. Similarly, we compared the relative source rankings for the sources in $\boldsymbol{\sigma}_{10\%}, \boldsymbol{\sigma}_{30\%}, ...$, to the relative rankings of the *exact same sources* in the ranking vector for the full Web source graph. To evaluate the stability, we rely on the Kendall Tau Distance metric as a measure of ranking error.

In Figure 2 we show the ranking error for the WB2001 dataset for PageRank and for three representative parameter settings over the host-based source graph – the baseline version $\mathcal{SR}(\mathbf{T}^*_{LC}, \mathbf{c}_u)$, the loop-less version $\mathcal{SR}(\mathbf{T}_{LC}, \mathbf{c}_u)$, and the size-based teleportation version $\mathcal{SR}(\mathbf{T}^*_{LC}, \mathbf{c}_s)$. Note that these are the three settings that resulted in the most different ranking vectors in our previous experiment. In all cases, the source-centric rankings display significantly less error relative to the rankings over the full Web graph than the PageRank rankings do, meaning that we can rely on source-centric rankings computed over an incomplete Web crawl with substantial confidence. Also note that the size-based version is the most stable, and we find that this stability generally improves as the source definition becomes more inclusive (from page to directory to host to domain).

Since the page and source ranking vectors are of different lengths, we additionally considered a similar stability analysis over just the top-100 and top-1000 page and source rankings. We relied on a variation of the Kendall Tau Distance metric known as the Kendall Min Metric [14] for evaluating top-k ranked lists. These results further validate the source stability, but are omitted here due to the space constraint.

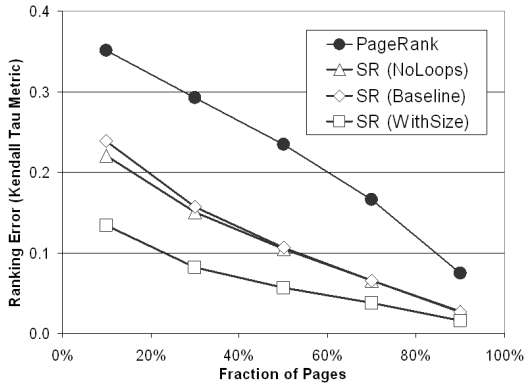


Fig. 2. Ranking Stability – WB2001

4) Approximating PageRank: As we have mentioned, one of the important goals of source-centric ranking is to approximate the traditional global PageRank vector by combining the source-level ranking information with per-source ranking information (the local PageRank scores). Such approximation promises to speed the PageRank calculation considerably. In this experiment we aim to understand under what conditions source-centric ranking may be used to reasonably approximate PageRank.

To approximate PageRank, we decompose the global PageRank of a page into source and local components:

$$\pi(p_i) = \sigma(s_j) \cdot \pi(p_i|s_j) \tag{1}$$

where we denote the local PageRank score for page i in source j as $\pi(p_i|s_j)$. The local PageRank score is calculated based only on local knowledge (e.g., based on the linkage information of pages within the source), takes comparably little time relative to the full PageRank calculation, and forms a probability distribution (i.e., $\sum_{p_k \in s_j} \pi(p_k|s_j) = 1$).

For the PageRank decomposition to hold over all pages, ideally we would have that the local PageRank scores $\pi(p_i|s_j)$ would exactly match the relative global distribution:

$$\pi(p_i|s_j) = \frac{\pi(p_i)}{\sum_{p_k \in s_j} \pi(p_k)} \tag{2}$$

By replacing $\pi(p_i|s_j)$ in Equation 1 with the righthand-side of Equation 2, we find that the source-centric component $\sigma(s_j)$ should equal the sum of the global PageRanks of the constituent pages:

$$\sigma(s_j) = \sum_{p_k \in s_j} \pi(p_k)$$

To test how well the source-centric rankings may be used to approximate PageRank, we compare the rankings induced from various parameter settings with the rankings induced from ranking the sources by the sum of the global PageRanks of their constituent pages. In Figure 3, we report the Kendall

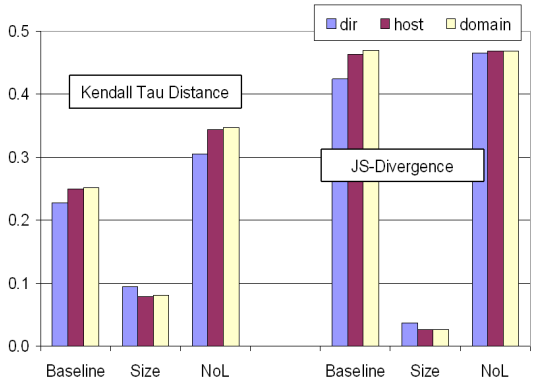


Fig. 3. Approximating PageRank – IT2004

Tau Distance Metric and the JS-Divergence for three representative parameter settings – the baseline version $\mathcal{SR}(\mathbf{T}^*_{LC}, \mathbf{c}_u)$, the loop-less version $\mathcal{SR}(\mathbf{T}_{LC}, \mathbf{c}_u)$, and the size-based teleportation version $\mathcal{SR}(\mathbf{T}^*_{LC}, \mathbf{c}_s)$ – over the domain, host, and directory-based source definitions for the IT2004 dataset. Similar results hold for the other two datasets.

The baseline parameter setting (which has been used elsewhere, e.g., [18], [29]) performs poorly, and is not appropriate for approximating PageRank. Similarly, the loopless version, which disregards the strong evidence of link-locality for setting edge weights, also performs poorly. Only the size-based version is highly correlated with the sum of the actual PageRank values for each source, meaning that source size and the presence of self-edges are critical for approximating PageRank. We also find that high-quality results hold when we replace the link count edge weighting parameter with the source consensus and target diffusion schemes.

5) Spam-Resilience: Finally, we study the spam-resilience properties of source-centric link analysis through a popular spam scenario. We consider a link farm in which a Web spammer generates a large number of colluding pages for the sole purpose of pointing to (and boosting the rank of) a target page. A link farm relies not on the quality of the pointing page to increase the rank of the target page, but on the sheer volume of colluding pages.

We study the impact of a spammer who constructs a link farm in a colluding source for increasing the rank of a target page in a different target source. We randomly selected a target source and a target page within that source. We paired the randomly selected target sources with a randomly selected colluding source, again from the bottom 50% of all sources on the host graph. For each pair, we added a single spam page to the colluding source with a single link to the randomly selected target page within the target source (case A), then 10 pages (B), 100 pages (C), up to 1,000 pages (case D). In Figure 4, we show the influence of the Web spammer in manipulating the rank of the target page and the rank of the target source through the average ranking percentile increase. For example in the WB2001 case, the PageRank of the target page jumped 80 percentile points under case C (from an average rank in the 19th percentile to the 99th), whereas the score of the target source jumped only 4 percentile points for the baseline version (from an average rank in the 27th percentile to the 31st).

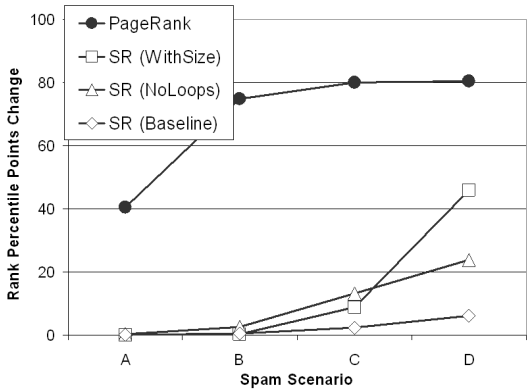


Fig. 4. Inter-Source Link Farm – WB2001

We first note the dramatic increase in PageRank for the target page across all three Web datasets, which indicates that PageRank is extremely susceptible to link manipulation. We are encouraged to observe that all three source-centric versions perform better than PageRank. The baseline version does increase some, but not nearly as much as PageRank. Since the source is an aggregation of many pages, the weighting of the source edges is less susceptible to changes in the underlying page graph. Interestingly, the loop-less version is the least resistant to manipulation for cases A, B, and C. In the loop-less version, external links are the sole determiner of a source's rank, meaning that inter-source manipulation wields more influence than for the looped versions. The size-based teleportation version is the most vulnerable for case D. In fact, under this scenario, a spammer need only add new pages (not links) to increase the score of a source.

As a final note, we want to state that source-centric ranking is still susceptible to source-centric attacks like the creation of dummy sources that all link to a target source. Hence, an interesting next step is the spam-sensitive study of source definition, which we believe can be used to further improve the spam-resilience properties of source-centric link analysis.

D. Summary of Experiments

The evaluation of the parameterized source-centric link analysis framework has yielded several interesting observations:

- Source-centric link analysis heavily depends on the source definition and the degree of link locality. We find that a lack of locality results in poor time complexity, but that even moderate locality (e.g., $\sim 65\%$) leads to good time complexity and stability results that are comparable with source definitions that display extremely high locality.
- In terms of ranking vector stability across parameters, the most important parameters are self-edges and the source-size teleportation component. We also found that incorporating expensive quality information into the edge

weighting schemes resulted in only a slight change to the resulting ranking vector.

- To best approximate PageRank using a layered calculation and for the most stable rankings in the face of Web link evolution, we saw the critical need for using the size-based teleportation component.
- However, using the size-based teleportation component resulted in the most severe vulnerability to spam, although it has these two desirable properties. This fundamental tension in objectives motivates our continuing research.

V. Related Work

In addition to the related work cited elsewhere in this paper, there have been some other efforts to understand higher-level Web abstractions. In [11], the *hostgraph* was explored in terms of various graph properties like indegree and outdegree distribution, and size of connected components. Crawling mechanisms based on the site paradigm, rather than the traditional page-based one, were enumerated in [13]. In [12], the potential spam properties of a HostRank algorithm were observed, and in [26] the ranking quality of several site-level-style PageRank variations was studied. In contrast to page aggregations, other researchers [8] have considered disaggregating Web pages into smaller units for providing ranking over individual components of Web pages.

Source-centric ranking can also take advantage of algorithmic enhancements for speeding PageRank (e.g., [23]).

For an introduction to Web spam, we refer the interested reader to [16]. Previous techniques suggested for dealing with Web spam include a modified version of PageRank that is seeded with expert opinions of trusted sites [17], statistical analysis of Web properties [15], the identification of nepotistic links [9], and several attempts to propagate a "bad" rank to pages based on linking patterns [2], [28].

VI. Conclusion

We have proposed a parameterized framework for source-centric link analysis, explored several critical parameters, and conducted the first large-scale comparative study of source-centric link analysis over multiple large real-world Web datasets and multiple competing objectives. We find that careful tuning of these parameters is vital to ensure success over each objective and to balance the performance across all objectives.

We believe these results are interesting and should lead to further study of source-centric link analysis of the Web. We are currently continuing our study of the Web spam properties of source-centric link analysis by the careful tuning of intra-source and inter-source edge weights. We are also extending the source view to incorporate database-backed Web sites typically underemphasized by page-based approaches.

References

[1] A. Arasu et al. Pagerank computation and the structure of the web: Experiments and algorithms. In *WWW*, 2002.

[2] A. A. Benczur et al. Spamrank - fully automatic link spam detection. In *AIRWeb*, 2005.

[3] K. Bharat et al. Who links to whom: Mining linkage between Web sites. In *ICDM*, 2001.

[4] P. Boldi et al. Ubicrawler. In *WWW*, 2002.

[5] P. Boldi, M. Santini, and S. Vigna. Paradoxical effects in PageRank incremental computations. In *WAW*, 2004.

[6] P. Boldi and S. Vigna. The WebGraph Framework I: Compression techniques. In *WWW*, 2004.

[7] A. Broder et al. Efficient PageRank approximation via graph aggregation. In *WWW*, 2004.

[8] D. Cai et al. Block-level link analysis. In *SIGIR*, 2004.

[9] B. D. Davison. Recognizing nepotistic links on the Web. In *Workshop on Artificial Intelligence for Web Search*, 2000.

[10] B. D. Davison. Topical locality in the web. In *SIGIR*, 2000.

[11] S. Dill et al. Self-similarity in the Web. *ACM Transactions on Internet Technology*, 2(3), 2002.

[12] N. Eiron, K. S. McCurley, and J. A. Tomlin. Ranking the Web frontier. In *WWW*, 2004.

[13] M. Ester, H.-P. Kriegel, and M. Schubert. Accurate and efficient crawling for relevant websites. In *VLDB*, 2004.

[14] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SIAM J. Discrete Mathematics*, 17(1):134–160, 2003.

[15] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics. In *WebDB*, 2004.

[16] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *AIRWeb*, 2005.

[17] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating Web spam with TrustRank. In *VLDB*, 2004.

[18] S. D. Kamvar et al. Exploiting the block structure of the Web for computing PageRank. Technical report, Stanford, 2003.

[19] M. Kendall and J. D. Gibbons. *Rank Correlation Methods*. Edward Arnold, 1990.

[20] J. Lin. Divergence measures based on the shannon entropy. *IEEE Trans. on Inf. Theory*, 37(1):145–151, 1991.

[21] T.-Y. Liu and W.-Y. Ma. Webpage importance analysis using conditional markov random walk. In *Web Intelligence*, 2005.

[22] Y. Lu et al. The PowerRank Web link analysis algorithm. In *WWW*, 2004.

[23] F. McSherry. A uniform approach to accelerated PageRank computation. In *WWW*, 2005.

[24] L. Page et al. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford, 1998.

[25] R. Song et al. Microsoft research asia at web track and terabyte track. In *TREC*, 2004.

[26] M. Thelwall. New versions of PageRank employing alternative Web document models. In *ASLIB*, 2004.

[27] Y. Wang and D. J. DeWitt. Computing PageRank in a distributed Internet search engine system. In *VLDB*, 2004.

[28] B. Wu and B. D. Davison. Identifying link farm spam pages. In *WWW*, 2005.

[29] J. Wu and K. Aberer. Using SiteRank for P2P Web retrieval. Technical report, Swiss Fed. Inst. of Tech., 2004.

[30] G.-R. Xue et al. Exploiting the hierarchical structure for link analysis. In *SIGIR*, 2005.