Countering Web Spam with Credibility-Based Link Analysis*

James Caverlee
Department of Computer Science
Texas A&M University
College Station, TX
caverlee@gmail.com

Ling Liu
College of Computing
Georgia Institute of Technology
Atlanta, GA
lingliu@cc.gatech.edu

ABSTRACT

We introduce the concept of link credibility, identify the conflation of page quality and link credibility in popular Web link analysis algorithms, and discuss how to decouple link credibility from page quality. Our credibility-based link analysis exhibits three distinct features. First, we develop several techniques for semi-automatically assessing link credibility for all Web pages. Second, our link credibility assignment algorithms allow users to assess credibility in a personalized manner. Third, we develop a novel credibility-based Web ranking algorithm – CredibleRank – which incorporates credibility information directly into the quality assessment of each page on the Web. Our experimental study shows that our approach is significantly and consistently more spam-resilient than both PageRank and TrustRank.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *Information Networks*

General Terms: Algorithms, Experimentation

Keywords: Web algorithms, link analysis, credibility, spam, PageRank

1. INTRODUCTION

With millions of Web servers supporting the autonomous sharing of billions of Web pages, the Web is arguably the most pervasive and successful distributed computing application today. Web spam refers to the type of attacks that manipulate how users view and interact with the Web, degrade the quality of information on the Web and place the users at risk for exploitation by Web spammers. Recent studies suggest that Web spam affects a significant portion of all Web content, including 8% of pages [5] and 18% of sites [10].

*This work is partially supported by grants from NSF ITR, CSR, CyberTrust and SGER, an AFOSR grant, an IBM SUR grant and a recent IBM Faculty Partnership award.

PODC'07, August 12–15, 2007, Portland, Oregon, USA.

Most of the popular link-based Web ranking algorithms, like PageRank [14], HITS [13], and TrustRank [10], all rely on a fundamental assumption that the quality of a page and the quality of a page's links are strongly correlated: a page ranked higher will be unlikely to contain lower quality links. This assumption, however, also opens doors for spammers to create link-based Web spam that manipulate links to the advantage of the Web spammers. Consider the following two common link-spam scenarios:

• **Hijacking:** Spammers hijack legitimate reputable pages and insert links that point to a spammer-controlled page, so that it appears to link analysis algorithms that the reputable page endorses the spam page. For example, in January 2006, a reputable computer science department's web page for new PhD students was *hijacked* by a Web spammer, and over 50 links to pornography-related Web sites were added to the page.

• **Honeypots:** Instead of directly hijacking a link from a reputable page and risking exposure, spammers often create legitimate-appearing websites (*honeypots*) to induce reputable pages to voluntarily link to these spammer-controlled pages. A honeypot can then pass along its accumulated authority by linking to a spam page.

Both scenarios show how spammers can take advantage of the tight quality-credibility coupling to subvert popular link-based Web ranking algorithms and why the assumption that the quality of a page and the quality of a page's links are highly correlated is vulnerable to link-based Web spam.

In this paper we advocate a clean separation of page quality and link (or reference) quality and argue that the intrinsic quality of a page should be distinguished from its intrinsic link credibility. Our goal is to assign each page a link credibility score defined in terms of link quality, not in terms of page quality. To guide our understanding of this problem, we address a number of important research questions.

• Can we formally define the concept of credibility to provide a degree of separation between page quality and link quality?

• What are the factors impacting the computation of credibility, and to what degree do these factors impact the application semantics of credibility-based link analysis?

• Can and how will credibility be impacted by the scope of linkage information considered? E.g., a page's local links, it's neighbor's links, its neighbor's neighbor's links, and so on.

This paper addresses each of these questions in detail to provide an in-depth understanding of link credibility. We develop a CredibleRank algorithm that incorporates credibility into an enhanced spam-resilient Web ranking algorithm. Concretely, we make three unique contributions: First, we introduce the concept of link credibility, identify the conflation of page quality and link credibility in popular link-based algorithms, and discuss how to decouple link credibility from page quality. Second, we develop several tech-

niques for semi-automatically assessing link credibility for all Web pages, since manually determining the credibility of every page on the Web is infeasible. Another unique property of our link credibility assignment algorithms is that they allows users with different risk tolerance levels to assess credibility in a personalized manner. Third, we present a novel credibility-based Web ranking algorithm - CredibleRank - which incorporates credibility information directly into the quality assessment of each page on the Web.

In addition, we develop a set of metrics for measuring the spam resilience properties of ranking algorithms, and show how the credibility information derived from a small set of known spam pages can be used to support high accuracy identification of new (previously unknown) spam pages. We have conducted an extensive experimental study on the spam resilience of credibility-based link analysis over a Web dataset of over 100 million pages, and we find that our proposed approach is significantly and consistently more spam-resilient than both PageRank and TrustRank.

2. REFERENCE MODEL

In this section, we present the Web graph model and discuss several popular approaches for link-based Web ranking.

2.1 Web Graph Model

Let $\mathcal{G} = \langle \mathcal{P}, \mathcal{L} \rangle$ denote a graph model of the Web, where the vertexes in $\mathcal{P}$ correspond to Web pages and the directed edges in $\mathcal{L}$ correspond to hyperlinks between pages. For convenience, we assume that there are a total of n pages ($|\mathcal{P}| = n$) and that pages are indexed from 1 to n. A page $p \in \mathcal{P}$ sometimes is referred to by its index number i. A hyperlink from page p to page q is denoted as the directed edge $(p, q) \in \mathcal{L}$, where $p, q \in \mathcal{P}$. We denote the set of pages that p points to as $Out(p)$, and the set of pages that point to p as $In(p)$. Typically, each edge $(p, q) \in \mathcal{L}$ is assigned a numerical weight $w(p, q) > 0$ to indicate the strength of the association from one page to the other, where $\sum_{q \in Out(p)} w(p, q) = 1$. A common approach assigns each edge an equal weight (i.e. $w(p, q) = \frac{1}{|Out(p)|}$). Other approaches that favor certain edges are possible.

A Web graph $\mathcal{G}$ can be represented by an $n \times n$ transition matrix $\mathbf{M}$ where the ij^{th} entry indicates the edge strength for an edge from page i to page j. The absence of an edge from one page to another is indicated by an entry of 0 in the transition matrix:

$$M_{ij} = \begin{cases} w(i,j) & \text{if } (i,j) \in \mathcal{L} \\ 0 & \text{otherwise} \end{cases}$$

2.2 Link-Based Ranking: Overview

A number of link-based ranking algorithms have been proposed over the Web graph, including the popular HITS [13], PageRank [14], and TrustRank [10]. Most of these algorithms assume that a link from one page to another is counted as a "recommendation vote" by the originating page for the target page. To illustrate the core of link-based rank analysis, we below outline PageRank and TrustRank.

PageRank: PageRank provides a single global authority score to each page on the Web based on the linkage structure of the entire Web. PageRank assesses the importance of a page by recursively considering the authority of the pages that point to it via hyperlinks. This formulation counts both the number of pages linking to a target page *and* the relative quality of each pointing page for determining the overall importance of the target page.

For n Web pages we can denote the PageRank authority scores as the vector $\mathbf{r_p} = (r_{p1}, r_{p2}, ..., r_{pn})$. The PageRank calculation considers the transition matrix $\mathbf{M}$ as well as an n-length static score vector $\mathbf{e}$, which is typically taken to be the uniform vector $\mathbf{e} = \left(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}\right)$. We can write the PageRank equation as a combination of these two factors according to a mixing parameter α:

$$\mathbf{r_p} = \alpha \mathbf{M}^T \mathbf{r_p} + (1 - \alpha)\mathbf{e} \quad (1)$$

which can be solved using a stationary iterative method like Jacobi iterations [6]. To ensure convergence, pages with no outlinks are modified to include virtual links to all other pages in the Web graph (forcing $\mathbf{M}$ to be row stochastic).

TrustRank: The recently proposed TrustRank algorithm [10] suggests biasing the PageRank calculation toward pre-trusted pages in an effort to suppress Web spam (similar to PageRank personalization suggested in [14] and more fully explored in [11]). Instead of considering the uniform static score vector $\mathbf{e}$, the TrustRank algorithm considers an n-length vector $\mathbf{v}$ that reflects a priori trust in each page. For example, in a web graph of 5 pages, if pages 1 and 3 are pre-trusted then $\mathbf{v}$ could be set to $\mathbf{v} = \left(\frac{1}{2}, 0, \frac{1}{2}, 0, 0\right)$. For n Web pages the TrustRank scores can be denoted by the vector $\mathbf{r_t} = (r_{t1}, r_{t2}, ..., r_{tn})$, and we can write the TrustRank equation as:

$$\mathbf{r_t} = \alpha \mathbf{M}^T \mathbf{r_t} + (1 - \alpha)\mathbf{v} \quad (2)$$

Determining the a priori trust vector $\mathbf{v}$ is of critical importance, and a number of techniques have been suggested, including the use of expert-selected whitelists, high PageRank pages, and topically-segmented trusted pages [10, 19].

In summary, link-based ranking algorithms like HITS, PageRank, and TrustRank attempt to estimate a page's intrinsic quality by analyzing the hyperlink structure of the Web. Fundamentally, the quality of a page and the quality of its links are tightly coupled in each of these ranking algorithms. Returning to Equation 1, a page with a PageRank score of r contributes $\alpha \cdot r$ to the pages that it links to, where α is typically in the range $[0.75, 0.95]$. We have discussed that such tight coupling is not only inadequate in practice but also creates Web spam vulnerabilities.

3. LINK CREDIBILITY

In this section, we formally introduce the concept of credibility in terms of *k-Scoped Credibility*. We shall ground our discussion of link credibility in the context of Web spam and explore how to algorithmically determine a page's credibility. Concretely, let C be a *credibility function* that instantaneously evaluates the link quality of a Web page p at time t. A score of $C(p, t) = 0$ indicates that the page p is not credible in terms of its links at time t. In contrast, a score of $C(p, t) = 1$ indicates that the page p is perfectly credible in terms of its links at time t. We observe that a desirable credibility function should have the following qualities:

• First, we observe that a page's link quality should depend on its own outlinks and perhaps is related to the outlink quality of its neighbors up to some small number (k) of hops away. Hence, link credibility of pages should be a function of the local characteristics of a page and its place in the Web graph, and not the global properties of the entire Web (as in a PageRank-style approach).

• Second, we observe that relying heavily on a set of known good pages (a whitelist) may be problematic. Spammers may attempt to mask their low quality outlinks to spam pages by linking to known whitelist pages. Also, relying too heavily on a whitelist for link credibility assignment makes these pages extremely valuable for spammers to corrupt.

• Third, we observe that a page's credibility should be related to

its distance to known spam pages (a blacklist) to penalize pages for poor quality outlinks.

Generally speaking, the Web is too large and too quickly growing to manually label each page as either spam or not spam. We shall assume that the set $\mathcal{P}$ of all pages can be divided into the set of known good pages, denoted by $\mathcal{P}_w$ (the whitelist), the set of known spam pages, denoted by $\mathcal{P}_b$ (the blacklist), and the set of pages for which the user has no experience or judgment, denoted by $\mathcal{P}_u$ (the unknown pages), such that $\mathcal{P} = \mathcal{P}_w \cup \mathcal{P}_b \cup \mathcal{P}_u$. In practice, only a fraction of all pages on the Web will belong to either the whitelist or the blacklist($|\mathcal{P}_w|, |\mathcal{P}_b| << |\mathcal{P}|$).

3.1 Naive Credibility

We begin our analysis of credibility functions by considering a simple approach that illustrates some of our observations above and serves as a comparison to the proposed k-Scoped Credibility function. The naive credibility function assigns a whitelist page a perfect credibility score of value one, a blacklist page no credibility (value zero), and an unknown page a default credibility value θ, $(0 < \theta < 1)$:

$$C_{naive}(p,t) = \begin{cases} 0 & \text{if } p \in P_b \\ \theta & \text{if } p \in P_u \\ 1 & \text{if } p \in P_w \end{cases}$$

The advantage of this naive credibility assignment is its ease of evaluation. However it has several apparent drawbacks – (i) it makes no effort to evaluate credibility in terms of the links of a page; (ii) the majority of all pages (P_u) receive a default credibility value; and (iii) whitelist pages, though generally high-quality, may not necessarily be perfectly credible in reality at all times.

3.2 k-Scoped Credibility

We next introduce k-Scoped Credibility, which evaluates the credibility of a page in terms of the quality of a random walk originating from the page and lasting for up to k steps. Critical to this *k-Scoped Credibility* is the notion of a *path*.

Definition 1 (path) Consider a directed web graph $\mathcal{G} = \langle \mathcal{P}, \mathcal{L} \rangle$, an originating page p and a destination page q. A *path* in the directed graph $\mathcal{G}$ from page p to page q is a sequence of nodes: $path(p,q) = \langle n_0, n_1, ..., n_j \rangle$ (where $p = n_0$ and $q = n_j$) such that there exists a directed edge between successive nodes in the path, $(n_i, n_{i+1}) \in \mathcal{L}$ for $0 \leq i \leq j-1$. The length $|path(p,q)|$ of a path is j, the number of edges in the path. There may exist multiple paths from p to q.

We refer to the set of all paths of a specified length (say, k) that originate from a page p as $Path_k(p)$. We will sometimes refer to a specific path of length k originating from p using the notation $path_k(p)$, where $path_k(p) \in Path_k(p)$.

Our notion of k-Scoped Credibility relies on a special type of path that we call a *bad path*.

Definition 2 (bad path) Consider a directed web graph $\mathcal{G} = \langle \mathcal{P}, \mathcal{L} \rangle$, an originating page p and a destination page q. We say that a path in the directed graph $\mathcal{G}$ from page p to page q is a *bad path* if the destination page is a spam page, $q \in \mathcal{P}_b$, and no other page in the path is a spam page. $path(p,q) = \langle n_0, n_1, ..., n_j \rangle$ (where $p = n_0$ and $q = n_j$) and $q \in \mathcal{P}_b$ and $n_i \notin \mathcal{P}_b$, for $0 \leq i \leq j-1$.

We refer to the set of all bad paths of a specified length (say, k) that originate from a page p as $BPath_k(p)$. The probability of a random walker traveling along a k-length path from page p is denoted by $Pr(path_k(p))$, and is determined by the probabilistic edge weights for each hop of the path:

$$Pr(path_k(p)) = \prod_{i=0}^{k-1} w(n_i, n_{i+1})$$

Formally, we define the *k-Scoped Credibility* of a page in terms of the probability that a random walker *avoids* spam pages after walking up to k hops away from the originating page. For $k = 1$, the k-Scoped Credibility is simply the fraction of a page's links that point to non-spam pages. Increasing k extends the scope of this credibility function by considering random walks of increasing length. For an originating page $p \in \mathcal{P}$, if p is a spam page, we set its link credibility to be 0, regardless of the characteristics of the pages it links to.

Definition 3 (k-Scoped Credibility) Let $\mathcal{G} = \langle \mathcal{P}, \mathcal{L} \rangle$ be a directed web graph, k be a maximum walk radius where $k > 0$, and $p \in \mathcal{P}$ be a page in the Web graph. The *k-Scoped Credibility* of page p at time t, denoted by $C_k(p,t)$, is defined as follows*:*

$$C_k(p,t) = 1 - \sum_{l=1}^{k} \left(\sum_{path_l(p) \in BPath_l(p)} Pr(path_l(p)) \right)$$

For the special case when $p \in \mathcal{P}_b$, let $C_k(p,t) = 0$.

In the case that there are no spam pages within k hops of page p, then p is perfectly credible: $C_k(p,t) = 1$. In the case that p itself is a spam page or in the case that all paths originating at page p hit a spam page within k hops, then p is not credible at all: $C_k(p,t) = 0$. Intuitively, the k-Scoped Credibility function models a random walker who when arriving at a spam page, becomes stuck and ceases his random walk, and for all other pages the walker continues to walk, for up to k hops.

4. COMPUTING CREDIBILITY

In practice, of course, the k-Scoped Credibility function can only have access to some portion of the entire Web graph, due to the size of the Web, its evolution, and the cost of crawling all pages. Additionally, only some spam pages will be known to the credibility function through the blacklist. In order to correct the inaccuracy in computing k-Scoped Credibility due to the presence of an incomplete Web graph and a partial blacklist, in this section we introduce the concept of tunable k-Scoped Credibility, which augments the basic k-Scoped Credibility computation by including a credibility penalty factor as a control knob. Our goals are to better approximate the k-Scoped Credibility under realistic constraints and understand how different parameters may influence the quality of a credibility function.

4.1 Tunable k-Scoped Credibility

The tunable k-Scoped Credibility is a function of two components: a random-walk component with respect to the known bad paths (based on the blacklist) and a penalty component. The penalty component is intended to compensate for the bad paths that are unknown to the credibility function. We first define the tunable k-Scoped Credibility and then focus our discussion on alternative approaches for assigning the credibility discount factor to offset the problem of an incomplete Web graph and a partial blacklist.

Definition 4 (Tunable k-Scoped Credibility) Let $\mathcal{G} = \langle \mathcal{P}, \mathcal{L} \rangle$ be a directed web graph, k be a maximum walk radius where $k > 0$,

and $\gamma(p)$ be the credibility penalty factor of a page $p \in \mathcal{P}$ where $0 \leq \gamma(p) \leq 1$. We define the *tunable k-Scoped Credibility* of page p, denoted by $C_k(p)$, in two steps: when $p \notin \mathcal{P}_b$:

$$C_k(p) = \left(1 - \sum_{l=1}^{k} \left(\sum_{path_l(p) \in BPath_l(p)} Pr(path_l(p)) \right) \right) \cdot \gamma(p)$$

In the case of $p \in \mathcal{P}_b$, let $C_k(p) = 0$.

The penalty factor $\gamma(p)$ is an important tunable parameter of the credibility assignment and can be used as the credibility discount knob. Since the blacklist $\mathcal{P}_b$ provides only a partial list of all spam pages in the Web graph at a given point of time, the penalty factor can be used to update the random walk portion of the credibility calculation to best reflect the possible spam pages that are not yet on the blacklist. We propose to use a hop-based approach for determining the proper credibility discount factor in computing k-Scoped Credibility for each page in the Web graph. To better understand the advantage of our hop-based approach, we also discuss the optimistic and pessimistic approaches as two extremes for selecting the penalty factor for each page.

4.1.1 The Optimistic Approach

This approach defines the credibility penalty factor for a page by assigning no penalty at all. In other words, for all pages, we assign a credibility discount factor of 1:

$$\gamma_{opt}(p) = 1, \forall p \in \mathcal{P}$$

meaning the random walk component of the tunable k-Scoped Credibility is not penalized at all. We call this approach an *optimistic* one since it is equivalent to assuming that all spam pages are on the blacklist. The optimistic approach will tend to over-estimate the credibility of pages that link to the spam pages not on the blacklist.

4.1.2 The Pessimistic Approach

In contrast, a *pessimistic* approach treats a page with *any* j-length path ($1 \leq j \leq k$) leading to a blacklist page as *not credible* within the k-hop scope in the sense that *all* paths originating from such a page are considered bad paths.

$$\gamma_{pess}(p) = \begin{cases} 0 & \text{if } |BPath_j(p)| \geq 1 \text{ for any j, } 0 < j \leq k \\ 1 & \text{otherwise} \end{cases}$$

A pessimistic approach may be appropriate in circumstances when links to spam pages are highly correlated (e.g., if the presence of a link to a blacklist page is always accompanied by another link to a spam page that is not on the blacklist). In many circumstances, the presence of a single bad path originating at a page may be the result of a temporary linking mistake (as in the hijacking example discussed in the introduction) or truly indicative that the page has only a fraction of links leading to spam pages. Hence, the pessimistic approach may be too draconian.

4.1.3 The Hop-Based Approach

The third approach for determining the credibility discount factor balances the extremes of the optimistic and pessimistic approaches by considering the number and the length of the bad paths for a page. A bad path is treated as evidence that there are other bad paths for an originating page that have been overlooked due to the partial nature of the blacklist and the incompleteness of the Web graph.

For a bad path of length j originating at page p, we associate a hop-based discount factor $\gamma_{hop,j}(p)$, where $0 \leq \gamma_{hop,j}(p) \leq 1$. By default, we let $\gamma_{hop,j}(p) = 1$ if there are no bad paths of length j originating from p (i.e. $BPath_j(p) = \emptyset$). The hop-based discount factor can then be calculated as the product of the constituent discount factors: $\gamma_{hop}(p) = \prod_{j=1}^{k} \gamma_{hop,j}(p)$.

Determining the choice of hop-based discount factors is critical to the quality of tunable k-scoped credibility calculation. We below discuss three alternative ways to compute $\gamma_{hop,j}(p)$. We begin with a user-defined discount factor ψ ($0 < \psi < 1$) to set the initial hop-based discount for bad paths of length 1, i.e., $\gamma_{hop,1}(p) = \psi$. Then we introduce three approaches for damping the user-defined discount factor that determine how quickly the discount factor approaches 1 as path length increases. Setting ψ close to 0 will result in a more pessimistic credibility penalty, whereas ψ close to 1 is intuitively more optimistic. By tuning ψ and the damping function we can balance these extremes.

Constant: One way for damping the initial setting of the user-defined discount factor ψ for bad paths of increasing length is to penalize all paths of varying lengths emanating from a page equally if there exists one bad path, i.e., $BPath_j(p) \neq \emptyset$. We refer to this approach as a *constant* discount factor since the hop-based discount does not vary with bad path length:

$$\gamma_{hop,i}(p) = \psi$$

Using a constant damping factor, a page that directly links to a spam page results in a credibility penalty that is the same as the penalty for a more distant path to a spam page.

Linear: The second approach to set the discount factor is *linear* in the length of a bad path up to some pre-specified path length L. Paths of distance L or greater are considered too distant to provide additional evidence of other bad paths, and so the discount factor is 1 for those paths.

$$\gamma_{hop,i}(p) = \begin{cases} \frac{(i-1)}{L-1}(1-\psi) + \psi & \text{if } i < L \\ 1 & \text{otherwise} \end{cases}$$

Using a linear damping factor, a path to a spam page that is farther away from the originating page results in a less severe hop-based discount than the credibility penalty for a direct link or short path from the originating page p to a spam page.

Exponential: The third approach for setting the discount factor is *exponential* in the length of the path, meaning that the initial discount factor ψ for bad paths of length 1 is quickly damped close to 1 as the bad path length increases.

$$\gamma_{hop,i}(p) = 1 - (1-\psi)\psi^{i-1}$$

Compared with the linear damping factor, the exponential damping factor allows the credibility discount for a spam page to be quickly damped close to 1. Put differently, when a spam page is closer to the originating page p, the link credibility of p is discounted less than the linear case with respect to the hop count.

4.2 Implementation Strategy

The k-Scoped Credibility is a local computation, requiring only an originating page and a forward crawl of all pages within k hops of the originating page. Hence, the credibility of a page can be updated in a straightforward fashion and as often as the k-hop neighbors are refreshed via Web crawls.

In practice, we anticipate computing the k-Scoped Credibility in batch for all Web pages in the current Web graph state after

each Web crawl. The main cost of computing the tunable k-Scoped Credibility is the cost of identifying the set of bad paths for each page and the cost of explicitly computing the path probabilities (recall Section 3.2). We propose to calculate the tunable k-Scoped Credibility for all pages using an equivalent iterative approach that is cheaper and faster. Let $\mathcal{G} = \langle \mathcal{P}, \mathcal{L} \rangle$ denote a graph model of the Web and $|\mathcal{P}| = n$ (recall Section 2.1). We first construct an n-length indicator vector $\mathbf{d} = (d_1, d_2, \ldots, d_n)$ to reflect whether a page is in the blacklist or not: $d_i = 1$ if $p_i \in P_b$, and $d_i = 0$ otherwise.

We next construct an $n \times n$ transition matrix $\mathbf{B}$ that replicates the original Web graph transition matrix $\mathbf{M}$, but with transition probabilities exiting a blacklist page of 0, to indicate the random walker stops when he arrives at a blacklist page.

$$B_{ij} = \begin{cases} M_{ij} & \text{if } d_i \notin \mathcal{P}_b \\ 0 & \text{otherwise} \end{cases}$$

In practice, the matrix $\mathbf{B}$ need not be explicitly created. Rather, the original matrix $\mathbf{M}$ can be augmented with rules to disregard blacklist entries.

The penalty factors can be encoded in an $n \times n$ diagonal matrix $\mathbf{\Gamma}$, where the diagonal elements correspond to the per-page penalty factors:

$$\Gamma_{ij} = \begin{cases} \gamma(i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Finally, the n-length tunable k-Scoped Credibility vector $\mathbf{C}_{[k]}$ can be computed:

$$\mathbf{C}_{[k]} = \left(\mathbf{1} - \sum_{j=0}^{k} \mathbf{B}^{(j)} \mathbf{d}^T \right) \cdot \mathbf{\Gamma}$$

where $\mathbf{1}$ is an n-length vector of numerical value 1s. Note that the matrix multiplication of $\mathbf{\Gamma}$ can be implemented as an element-wise vector product, so the expense of a matrix-by-matrix multiplication can be largely avoided.

5. CREDIBILITY-BASED WEB RANKING

We have presented the design of several credibility functions for evaluating Web page link quality. In this section, we use this decoupled credibility information to augment the page quality assessment of each page on the Web with a goal of suppressing Web spam. Concretely, we demonstrate how link credibility information can improve PageRank and TrustRank-style approaches through a credibility-based Web ranking algorithm called CredibleRank.

Returning to PageRank (see Equation 1), there are several avenues for incorporating link credibility information. We outline four alternatives below:

• First, the initial score distribution for the iterative PageRank calculation (which is typically taken to be a uniform distribution) can be seeded to favor high credibility pages. While this modification may impact the convergence rate of PageRank, it has no impact on ranking quality since the iterative calculation will converge to a single final PageRank vector regardless of the initial score distribution.

• Second, the graph structure underlying the transition matrix $\mathbf{M}$ can be modified to remove low credibility pages and edges to low credibility pages. While this modification may eliminate some Web spam pages, it could also have the negative consequence of eliminating legitimate pages that are merely of low credibility.

• Third, the edge weights in the transition matrix $\mathbf{M}$ can be adjusted to favor certain edges, say edges to high-credibility pages. While this change may have some benefit, a low credibility page p's overall influence will be unaffected (since $\sum_{q \in Out(p)} w(p,q) = 1$).

• Finally, the static score vector $\mathbf{e}$ can be changed to reflect the link credibility information, much like in TrustRank and Personalized PageRank [11, 10]. By skewing $\mathbf{e}$ toward high credibility pages (or away from low credibility pages) we can give a ranking boost to these pages, which could have the undesired consequence of ranking a low-quality high-credibility page over a high-quality low-credibility page.

Alternatively, we propose a credibility-augmented Web ranking algorithm that uses credibility information to impact the size of the vote of each page. CredibleRank asserts that a page's quality be determined by two criteria: (1) the quality of the pages pointing to it; and (2) the credibility of each pointing page. A link from a high-quality/high-credibility page counts more than a link from a high-quality/low-credibility page. Similarly, a link from a low-quality/high-credibility page counts more than a link from a low-quality/low-credibility page. By decoupling link credibility from the page's quality, we can determine the credibility-augmented quality of each page through a recursive formulation.

Recall that $In(p)$ denotes the set of pages linking to p. We compute the CredibleRank score $r_c(p)$ for a page p as:

$$r_c(p) = \sum_{q \in In(p)} C(q) \cdot r_c(q) \cdot w(q,p)$$

This formula states that the CredibleRank score (quality) of page p is determined by the quality ($r_c(q)$) and the link credibility ($C(q)$) of the pages that point to it, as well as the strength of the link $w(q,p)$. In this sense, the link weights are used to determine how a page's "vote" is split among the pages that it points to, but the credibility of a page impacts how large or small is the page's vote.

We can extend this formulation to consider n Web pages, where we denote the CredibleRank authority scores by the vector $\mathbf{r_c} = (r_{c1}, r_{c2}, ..., r_{cn})$. Recall that $\mathbf{M}$ denotes the $n \times n$ Web transition matrix, and $\mathbf{v}$ is an n-length static score vector. We can construct an $n \times n$ diagonal credibility matrix $\mathbf{CR}$ from the link credibility vector $\boldsymbol{\gamma}$, where the elements of the credibility matrix are defined as:

$$CR_{ij} = \begin{cases} C(i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

We can then write the CredibleRank vector $\mathbf{r_c}$ as:

$$\mathbf{r_c} = \alpha(\mathbf{CR} \cdot \mathbf{M})^T \mathbf{r_c} + (1 - \alpha)\mathbf{v} \quad (3)$$

which, like PageRank and TrustRank, can be solved using a stationary iterative method like Jacobi iterations. The matrix multiplication of $\mathbf{CR}$ and $\mathbf{M}$ can be implemented as an element-wise vector product to avoid the expense of a matrix-by-matrix multiplication.

6. EXPERIMENTAL EVALUATION

In this section, we report the results of an experimental study of credibility-based link analysis over a Web dataset of over 100 million pages. We report two sets of experiments – (1) an evaluation of tunable k-Scoped Credibility and the factors impacting it (like scope k, discount factor, blacklist size, and damping function); and (2) a study of the spam-resilience characteristics of CredibleRank, where we show that our proposed approach is significantly and consistently more spam-resilient than both PageRank and TrustRank.

Figure 1: Credibility Coverage

Scope (k)	Blacklist Size		
	Small	Medium	Large
1	7%	27%	39%
2	5%	33%	46%
3	26%	73%	79%
4	75%	94%	95%
5	95%	98%	98%
10	99%	99%	99%

6.1 Setup

The experiments reported in this paper use a Stanford WebBase dataset consisting of 118 million pages and 993 million links. The dataset was originally collected in 2001 and includes pages from a wide variety of top-level-domains.

Defining what exactly constitutes *spam* is an open question, and so as a baseline for our experiments we considered pornography related pages in the dataset as spam. Naturally, this is one of many possible spam definitions and we anticipate revisiting this topic in our continuing research. Since manually inspecting all 118 million pages is an onerous task, we applied a simple procedure to identify spam pages. We first identified all sites with a URL containing a pornography related keyword (where we define a site by the host-level information embedded in each page's URL). This resulted in 11,534 sites and over 1.5 million pages. For these 11,534 sites, we then sampled a handful of pages from each site and kept only those sites that we judged to be spam. Applying this filter yielded 9,034 sites consisting of 1,202,004 pages. We refer to these pages as the *Spam Corpus*.

We generated three blacklists by randomly selecting sites from the Spam Corpus. The first blacklist (referred to as **Large**) contains 20% of the sites in the Spam Corpus (1807 sites, or 0.24% of all sites); the second blacklist (**Medium**) contains 10% of the Spam Corpus (903 sites); the third blacklist (**Small**) contains just 1% (90 sites).

For the whitelist, we manually selected 181 sites from the top-5000 sites (as ranked by PageRank). These whitelist sites are each maintained by a clearly legitimate real-world entity, either a major corporation, university, or organization. We additionally ensured that each of these 181 sites was not within two-hops of any site in the Spam Corpus.

We grouped pages in the entire dataset into sites (again, by the host information of each page's URL), resulting in 738,626 sites. We constructed a site graph where each site is a node in the graph. If a page in one site points to a page in another site we included an edge in the site graph, excluding self-edges. The result is 11,816,108 edges in the site-level Web graph. We adopted a fairly standard approach for defining the edge strength for a site-level edge as the fraction of page-level hyperlinks pointing from the originating site to the target site (e.g., [12, 16]), and constructed the transition matrix $\mathbf{M}$ based on these edge weights. For all ranking calculations, we relied on the standard mixing parameter $\alpha = 0.85$ used in the literature (e.g., [10, 14]), and we terminated the Jacobi method after 50 iterations.

6.2 Credibility Assignment Evaluation

In the first set of experiments, we evaluate tunable k-Scoped Credibility and the many factors impacting it.

6.2.1 Credibility Coverage

In our first experiment (shown in Figure 1), we examine how widely the tunable k-Scoped Credibility can assign link credibility scores to sites beyond the pre-labelled blacklist. By increasing the tunable scope parameter k, the credibility function will consider paths of increasing length, meaning that there will be a greater opportunity for identifying bad paths. We measure the *coverage* of k-Scoped Credibility in terms of the scope parameter k, a complete blacklist b (the Spam Corpus), and a partial blacklist b', where $b' \subset b$:

$$cov(k, b, b') = \frac{|\{p \in \mathcal{P} | \exists j, 1 \leq j \leq k \text{ s.t. } BPath_j(p, b') \neq \emptyset\}|}{|\{p \in \mathcal{P} | \exists j, 1 \leq j \leq k \text{ s.t. } BPath_j(p, b) \neq \emptyset\}|}$$

where $BPath_k(p, b)$ denotes the set of all bad paths to sites in blacklist b of length k that originate from a page p. The numerator corresponds to the count of all sites with at least one path to a site on the partial blacklist. The denominator corresponds to the count of all sites with at least one path to a site on the complete blacklist (the Spam Corpus). So, for $k = 1$, there are 18,305 sites that are either spam sites or directly link to spam sites; of these 27% are on the Medium blacklist or directly link to a site on the Medium blacklist.

There are three interesting observations. First, the size of the blacklist is important. A larger blacklist leads to more evidence of bad paths, and hence will give our tunable k-Scoped Credibility function the opportunity to make higher-quality credibility assessments, even for small k. Second, even for the Small blacklist – with just 90 sites – we find fairly reasonable coverage (26%) for $k = 3$. Third, for large k, nearly all pages have at least one path to a spam page. Thus, pages that are quite distant from an originating page likely have little impact over the credibility of the originating page. The choice of k should be made with care.

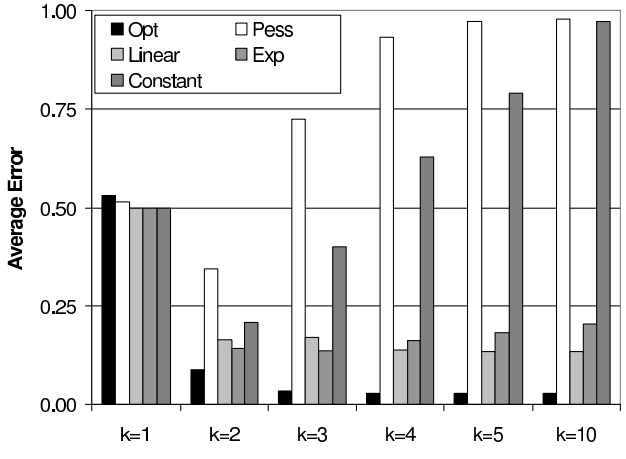
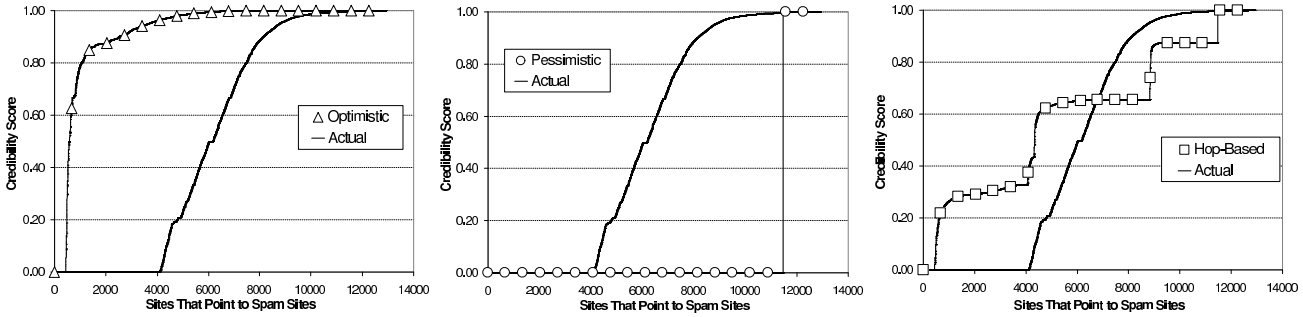


Figure 2: Average Credibility Error - Varying k

6.2.2 Credibility Quality

We next study the quality of the tunable k-Scoped Credibility function over different settings of the credibility penalty factor (Figures 2, 3, and 4). Recall that the penalty factor is used to update the random walk portion of the credibility calculation to reflect the possible spam pages (or sites, in this case) not yet on the blacklist. Our goal is to understand how well the tunable k-Scoped Credibility functions perform as compared to the k-Scoped Credibility with access to the full Spam Corpus.

We consider five different settings for the penalty factor of the k-Scoped Credibility – the Optimistic, Pessimistic, and three Hop-Based approaches. For each of the Hop-Based approaches – con-



(a) Optimistic (b) Pessimistic (c) Hop-Based (exponential $\psi = 0.5$)

Figure 3: Distribution of Credibility Scores for Sites That Point to Spam Sites (vs. Actual) [k=3]

stant, linear, and exponential – we report the results for an initial credibility discount factor $\psi = 0.5$. For each of these 5 settings, we calculated the credibility for each site using only 10% of all spam sites (the Medium blacklist b').

We evaluate the error for each of these credibility functions over the Medium blacklist b' versus the actual credibility computed over the entire Spam Corpus b. We measure the overall error for a tunable credibility function C over b' as the average of the pair-wise credibility differences with the actual credibility function C^* over b:

$$error(C, b, b') = \frac{1}{|X|} \sum_{p \in X} |C^*(p) - C(p)|$$

where X is the set of sites with at least one bad path to a site in the Spam Corpus: $X = \{p \in \mathcal{P} | \exists j, 1 \leq j \leq k \text{ s.t. } BPath_j(p, b) \neq \emptyset\}$.

In Figure 2, we report the average credibility error for each of the five tunable k-Scoped Credibility functions evaluated over increasing values of the scope parameter k.

There are three interesting observations. First, the Optimistic penalty factor performs very well, resulting in the lowest average credibility error for $k >= 2$. This indicates that the credibility scores assigned by the Optimistic approach are the closest to the scores assigned by the credibility function with access to the entire Spam Corpus.

Second, the Pessimistic and Constant penalty factors perform well for $k = 2$, and then increasingly worse as the scope parameter k increases. These two approaches are very pessimistic, assigning 0 or low credibility to sites with even a single bad path. For $k = 2$, only sites within a close radius of sites on the blacklist are penalized. Thus we see a fairly low error rate. As k increases, most sites have at least one path to a blacklist site (recall Figure 1), and are assigned a 0 or low credibility score, resulting in a high error rate.

Third, the Exponential and Linear approaches result in better performance than Pessimistic and Constant, but worse than Optimistic. As k increases, the error increase observed in the Constant approach is avoided since the Exponential and Linear penalty factors treat long paths less severely. On further inspection, we discovered that only these two approaches balance the credibility over-estimation of the Optimistic approach and the underestimation of the Pessimistic and Constant approaches.

To further illustrate this over- and underestimation balance, we next report the distribution of credibility scores based on the Medium blacklist for the Optimistic, Pessimistic, and Hop-Based (exponential) approaches for all sites that point to sites in the Spam Corpus. Figure 3 reports the distribution of credibility scores versus the actual credibility scores based on the entire Spam Corpus. The Optimistic approach assigns very high credibility for nearly all sites that point to the Spam Corpus, whereas the Pessimistic approach assigns 0 credibility to most sites. Only the Hop-Based approach balances these over and under estimation errors. As we will see in our spam resilience experiments in the following sections, this balance will lead to better spam-resilience than the Optimistic approach in all cases, and to better spam-resilience than the Pessimistic approach for $k > 2$.

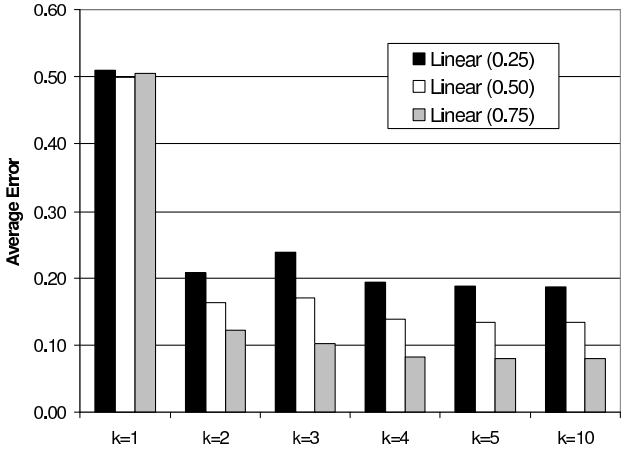


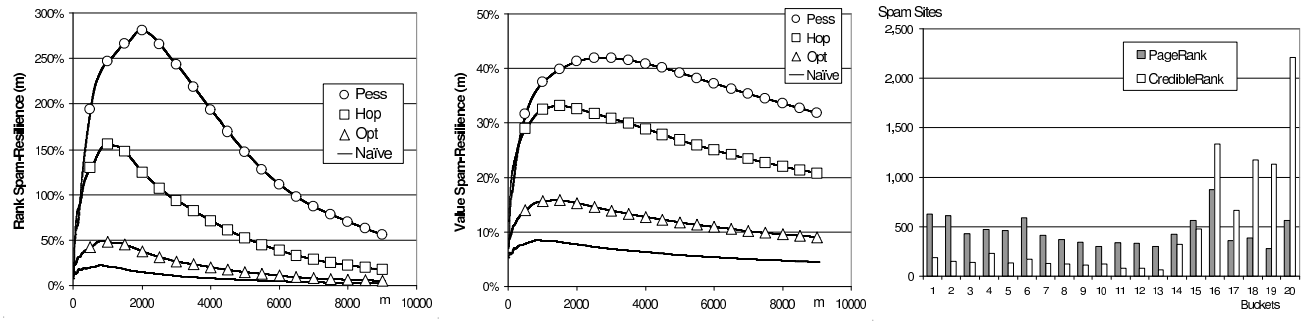
Figure 4: Average Credibility Error - Varying ψ

The linear and exponential Hop-Based approaches are also impacted by the choice of the initial discount factor ψ. In Figure 4, we report the average credibility error for the linear case (for $L = 4$) for three setting of ψ (0.25, 0.50, and 0.75). It is encouraging to see that the error drops significantly for $k = 2$ and is fairly stable for increasing values of k (in contrast to the Constant and Pessimistic approaches reported in Figure 2).

We have also studied the impact of the partial blacklist size on credibility quality. We find that for the Optimistic, Linear, and Exponential approaches the error rate is fairly stable, whereas the Pessimistic and Constant approaches degrade severely as the blacklist size increases. For these two cases, a larger blacklist leads to more sites within a few hops of the blacklist, resulting in more 0 scores, even when the random walk probability of such a bad path is low.

6.3 Spam Resilience Evaluation

In the following sections we evaluate the quality of each credibility assignment approach through Web ranking experiments, and



(a) $SR_{Rank}(m)$ (b) $SR_{Value}(m)$ (c) Spam Distribution

Figure 5: CredibleRank vs. PageRank: Rank Spam Resilience, Value Spam Resilience, and Spam Distribution

compare these results with PageRank and TrustRank. We measure the effectiveness of a Web ranking approach by its spam resilience. To quantify spam resilience, we introduce two metrics. Each evaluates the quality of a candidate ranking algorithm versus a baseline ranking algorithm over a set of spam sites $\mathcal{X}$. We refer to $\mathcal{X}$ as a *portfolio* of spam sites. In this paper, we use the Spam Corpus as the portfolio $\mathcal{X}$. We consider the baseline ranking for a portfolio of $|\mathcal{X}|$ sites: $B = (B_1, ..., B_{|\mathcal{X}|})$, and a ranking induced by the candidate ranking algorithm $E = (E_1, ..., E_{|\mathcal{X}|})$.

Rank-Based Spam Resilience

The first spam resilience metric SR_{Rank} measures the relative change of the *ranks* of the portfolio sites:

$$SR_{Rank}(m) = \frac{\sum_{i=1}^{m} R(E_i)}{\sum_{i=1}^{m} R(B_i)} - 1$$

where $R(E_i)$ returns the rank of site E_i according to the candidate ranking algorithm and $R(B_i)$ returns the rank of site B_i according to the baseline ranking algorithm. By evaluating $SR_{Rank}(m)$ for different values of m, we may assess the spam resilience of a ranking algorithm at different levels (e.g., for the top-100 pages, the top-1000, and so on). A candidate ranking algorithm that induces a ranking that exactly matches the baseline ranking, will result in $SR(m)$ values of 0 for all choices of k. A ranking algorithm that induces a more spam-resilient ranking will result in positive $SR_{Rank}(m)$ values, meaning that the rank of the portfolio will have been reduced. Negative values indicate that the candidate algorithm is less spam-resilient than the baseline.

Value-Based Spam Resilience

The second spam resilience metric is based on the change in *value* of the spam portfolio. Let us assume that each rank position has an associated value (say, in dollars), and that these values are monotonically decreasing as the rank position increases. That is, for a value function $V(\cdot)$, we have $R(i) < R(j) \Rightarrow V(R(i)) > V(R(j))$. Hence, we can measure the spam resilience by considering the relative change in the value of the spam portfolio under the candidate ranking algorithm versus the baseline ranking algorithm:

$$SR_{Value}(m) = 1 - \frac{\sum_{i=1}^{m} V(R(E_i))}{\sum_{i=1}^{m} V(R(B_i))}$$

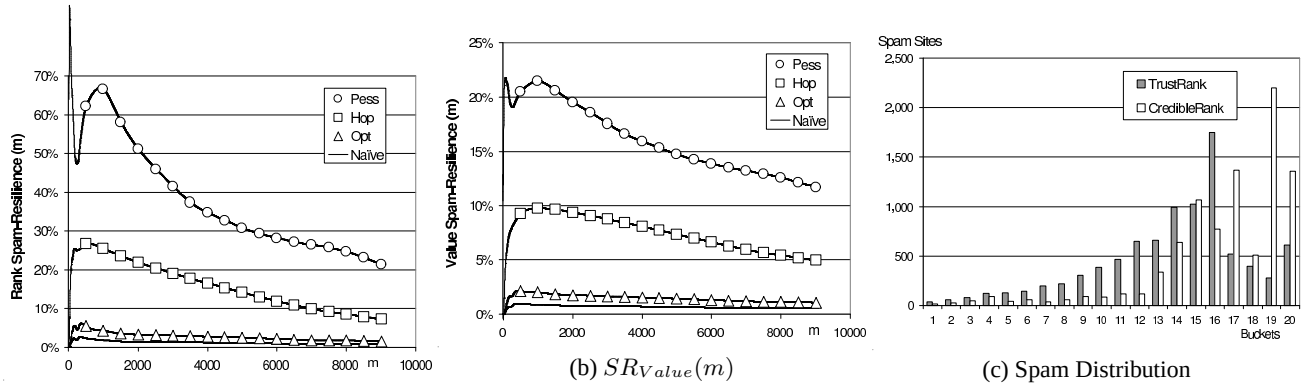
where $V(R(E_i))$ and $V(R(B_i))$ are the values of applying a value function to the rank $R(E_i)$ and $R(B_i)$ respectively. A positive SR_{Value} value means that the candidate algorithm is more spam-resilient than the baseline algorithm since the overall value of the spam portfolio has been reduced. We consider a power-law rank value function since there is nice intuitive support for it, that is, the top-rank positions are quite valuable, while most positions have relatively low value. Concretely, the arbitrary value of rank x is $V(x) = 1,000,000x^{-0.5}$, meaning that the top-ranked site has value of $\$1m$, the 100th-ranked site is worth $\$100k$, the 10,000th-ranked site is worth $\$10k$, and so on. The key here is not to estimate the actual value precisely, but to provide a relative value of different rank positions.

6.3.1 PageRank versus CredibleRank

Given the above two metrics, we now compare the effectiveness of CredibleRank to that of PageRank with respect to spam resilience. Here PageRank is used as the baseline ranking. For fairness of comparison, we do not incorporate any whitelist information into the CredibleRank calculation, so the static score vector in Equation 3 is set to the uniform vector, as it is in PageRank (Equation 1).

For CredibleRank, we consider the Naive approach and three tunable k-Scoped Credibility assignment approaches – Optimistic, Pessimistic, and Hop-Based (exponential $\psi = 0.5$) – using the medium blacklist for scope of $k = 2$. In Figures 5(a) and 5(b), we report the $SR_{Rank}(m)$ and $SR_{Value}(m)$ spam resilience scores for $m = 1$ to $m = 9,034$ (the size of the Spam Corpus) for the four candidate CredibleRank rankings (i.e., Opt, Pess, Hop, and Naive) versus the baseline PageRank ranking. We are encouraged to see that for both rank-based and value-based spam resilience that all CredibleRank approaches result in more spam-resilient rankings versus PageRank, with the Pessimistic performing the best, closely followed by the Hop-Based approach. The spam resilience rapidly increases and then peaks over the top-2000 spam sites, indicating that CredibleRank performs well over these top-ranked spam sites. As k increases to consider more sites in the spam resilience measurement, more lower-ranked sites are considered, which have less downward space to move, meaning that the overall spam resilience decreases relative to the top-ranked sites.

To further demonstrate how CredibleRank demotes spam sites relative to PageRank, we sorted the sites by rank order for the Hop-Based CredibleRank and PageRank ranking vectors, and divided the sites into 20 buckets of an equal number of sites. Along the x-axis of Figure 5(c) we consider these 20 buckets, from the bucket of top-ranked sites (bucket 1) to the bucket of the bottom-ranked sites (bucket 20). Along the y-axis, we plot the number of Spam Corpus sites (of the 9,034 total spam sites) in each bucket. What is immediately obvious is that CredibleRank penalizes spam sites considerably more than PageRank by demoting spam sites to lower-ranked buckets, even when only 10% of the spam sites have been explicitly assigned to the blacklist.



(a) $SR_{Rank}(m)$ (b) $SR_{Value}(m)$ (c) Spam Distribution

Figure 6: CredibleRank vs. TrustRank: Rank Spam Resilience, Value Spam Resilience, and Spam Distribution

Our results so far have measured the effectiveness of CredibleRank with respect to its spam resilience. We also would like to show that CredibleRank does not negatively impact known good sites. Hence, we compared the ranking of each whitelist site under PageRank versus its ranking under CredibleRank. We find that the average rank movement is only 26 spots, meaning that we can feel fairly confident that CredibleRank is not unduly punishing good sites.

6.3.2 *TrustRank versus CredibleRank*

Recall that TrustRank incorporates whitelist information into the ranking calculation to favor whitelist sites and the sites that they point to over other sites. In this experiment, we compare CredibleRank to TrustRank, where TrustRank is used as the baseline ranking and for fairness, the CredibleRank approach relies on the same whitelist-based static score vector used in TrustRank (Equation 2).

For CredibleRank, we again consider the four link credibility assignment approaches – Naive, Optimistic, Pessimistic, and Hop-Based – using the medium blacklist. In Figures 6(a) and 6(b), we report the $SR_{Rank}(m)$ and $SR_{Value}(m)$ spam resilience scores for $m = 1$ to $m = 9,034$ for the three candidate CredibleRank rankings versus the baseline TrustRank ranking. As in the PageRank comparison, we see that all CredibleRank approaches result in more spam-resilient rankings comparing to TrustRank, with the Pessimistic and Hop-Based performing the best.

For the Hop-Based CredibleRank and the TrustRank ranking vectors, we report the bucket-based spam site distribution in Figure 6(c). We find that CredibleRank penalizes spam sites considerably more than TrustRank, pushing most sites into the bottom-ranked buckets.

We wish to note that the choice of whitelist is extremely important for TrustRank. Since links from whitelist sites are favored over links from other sites, a spammer has a great incentive to induce links from a whitelist site. In our experiments, we find choosing a whitelist with sites that either link directly to spam sites or are within several hops of spam sites results in very poor spam resilience for TrustRank. We find for one poor quality whitelist that CredibleRank has a rank-based spam resilience achieving a maximum improvement of 107% over TrustRank, with a 32% improvement over the entire spam corpus. We have also evaluated CredibleRank and TrustRank using only blacklist information and no whitelist information (by skewing the static score vector to non-blacklist sites). Since CredibleRank distinguishes page (or site) quality from link credibility, we find that it achieves rank-based spam resilience of up to 134% over TrustRank, with a 16% improvement over the entire spam corpus.

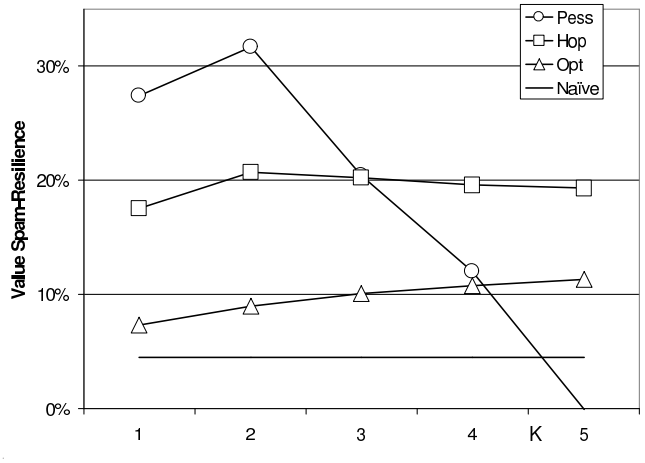


Figure 7: Impact of Scope [K] (CR vs. PR)

6.3.3 *Impact of Scope (K)*

Our results so far have measured the effectiveness of CredibleRank with respect to the tunable k-Scoped Credibility function for $k = 2$. But what are the implications of changing the scope parameter on the spam-resilience of CredibleRank? In Figure 7 we report the value-based spam resilience for $k = 1$ to $k = 5$ for the Naive approach and the three tunable k-Scoped Credibility assignment approaches – Optimistic, Pessimistic, and Hop-Based (exponential $\psi = 0.5$). The Naive approach does not consider scope and so its spam-resilience is unaffected by changes in k. The Hop-Based and Optimistic approaches are fairly stable with increasing k. For $k = 1$ and $k = 2$, the Pessimistic approach performs well, since sites that either directly link or are within 2 hops of blacklist sites have no ranking influence over the sites that they point to. The Pessimistic approach severely degrades in spam-resilience for increasing values of k until it performs even worse than PageRank for $k = 5$. When $k = 5$, nearly all sites have at least one path to a blacklist site, resulting in a Pessimistic credibility score of 0. In the CredibleRank interpretation, this means that nearly all links in the Web graph are disregarded and so the resulting rankings are essentially random.

6.3.4 *Impact of Blacklist Size*

We have also explored the impact of the blacklist size on the

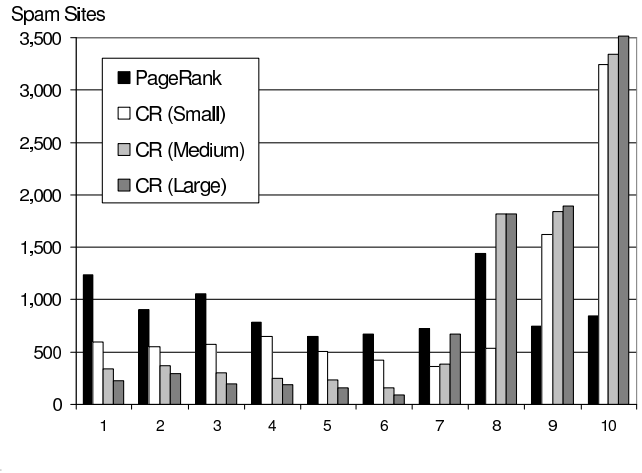


Figure 8: Impact of Blacklist Size (CR vs. PR)

spam resilience of CredibleRank. For the three blacklists – small (1% of the Spam Corpus), medium (10%), and large (20%) – we report in Figure 8 the ranking distribution of the Spam Corpus for the Hop-Based CredibleRank versus PageRank. For presentation clarity, we divide all sites into 10 buckets in this experiment. CredibleRank based on the small blacklist (containing just 90 sites) results in a remarkable improvement over PageRank. The advantage increases as more spam sites are added to the blacklist.

We also evaluated CredibleRank's spam-resilience versus TrustRank for varying choices of scope parameter (k) and blacklist size, and we find results of a similar spirit to the ones reported for PageRank in Sections 6.3.3 and 6.3.4, but are omitted here due to the space constraint.

7. RELATED WORK

For an introduction to Web spam, we refer the interested reader to [9]. Some previous techniques suggested for dealing with Web spam include the statistical analysis of Web properties [5], the identification of nepotistic links [4], and several attempts to propagate a "bad" rank to pages based on linking patterns [2, 17]. Several researchers have studied collusive linking arrangements with respect to PageRank, including [1] and [21].

Several researchers have suggested identifying and penalizing pages that derive a large amount of ranking benefit from spam links, e.g., [2], [8], and [17]. With respect to PageRank, previous researchers have suggested varying the random walk mixing parameter to favor pages with few links versus pages with many links [15]. It is important to note that most of this previous research is complementary to credibility-based link analysis; since CredibleRank integrates spam-resilience into the ranking model, rather than attempting to identify Web spam outright, it may be augmented with these alternative algorithms to further enhance its effectiveness.

Our notion of link credibility has some analogues in trust network research, in which computational models are developed for measuring trust. The authors of [3] argued for distinguishing between direct trust and recommendation trust. In the context of peer-to-peer networks, the PeerTrust system models the believability (or credibility) of peer feedback to guide the trust calculation of nodes in the network [20]. Link credibility is also somewhat related to the notion of distrust, which has recently received increasing attention (e.g., [7], [18]). For example, in [7], the authors argue for a trust propagation technique in which the recommendations of distrusted nodes are discounted completely. Note that our link credibility model allows for a continuum of credibility scores.

8. CONCLUSIONS

We have explored the concept of link credibility, presented several techniques for semi-automatically assessing link credibility for all Web pages, and presented an efficient and yet spam-resilient credibility-based Web ranking algorithm. We also introduced a set of metrics to measure the spam resilience properties of credibility-based link analysis, and have shown that our credibility-based ranking algorithm outperforms both PageRank and TrustRank. We have made a first step towards credibility-based link analysis for countering Web spam, and we believe that this work will trigger more research and discussions on this important topic.

9. REFERENCES

[1] R. Baeza-Yates, C. Castillo, and V. Lopez. PageRank increase under different collusion topologies. In *AIRWeb*, 2005.

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

[3] T. Beth, M. Borcherding, and B. Klein. Valuation of trust in open networks. In *ESORICS*, 1994.

[4] B. Davison. Recognizing nepotistic links on the Web. In *Workshop on AI for Web Search*, 2000.

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

[6] D. Gleich, L. Zhukov, and P. Berkhin. Fast parallel PageRank: A linear system approach. Technical report, Yahoo!, 2004.

[7] R. Guha et al. Propagation of trust and distrust. In *WWW*, 2004.

[8] Z. Gyöngyi et al. Link spam detection based on mass estimation. In *VLDB*, 2006.

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

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

[11] T. H. Haveliwala. Topic-sensitive PageRank. In *WWW*, 2002.

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

[13] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

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

[15] X. Wang, A. Shakery, and T. Tao. Dirichlet PageRank. In *Poster Proceedings of SIGIR*, 2005.

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

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

[18] B. Wu, V. Goel, and B. Davison. Propagating trust and distrust to demote web spam. In *Models of Trust for the Web (MTW)*, 2006.

[19] B. Wu, V. Goel, and B. Davison. Topical TrustRank: Using topicality to combat Web spam. In *WWW*, 2006.

[20] L. Xiong and L. Liu. Supporting reputation-based trust for peer-to-peer electronic communities. *TKDE*, 16(7), 2004.

[21] H. Zhang et al. Improving eigenvector-based reputation systems against collusions. In *Algorithms and Models for the Web Graph*, 2004.