# Anchor-based Proximity Measures

Amruta Joshi Ravi Kumar Benjamin Reed Andrew Tomkins
Yahoo! Research, 701 First Ave, Sunnyvale, CA 94089.
{amrutaj,ravikumar,breed,atomkins}@yahoo-inc.com

## ABSTRACT

We present a family of measures of proximity of an arbitrary node in a directed graph to a pre-specified subset of nodes, called the *anchor*. Our measures are based on three different propagation schemes and two different uses of the connectivity structure of the graph. We consider a web-specific application of the above measures with two disjoint anchors — *good* and *bad* web pages — and study the accuracy of these measures in this context.

**Categories and Subject Descriptors:** H.3.m [**Information Search and Retrieval**]: Miscellaneous

**General Terms:** Algorithms, Experimentation.

**Keywords:** link propagation, proximity, harmonic rank

## Introduction

Methods for producing a static (query-independent) ranking of web pages, hosts, or domains using graph analysis have been used employed successfully to combine information from multiple perhaps-distant neighbors. These methods represent one of our most successful tools for cross-page analysis of the web, in particular because they are efficiently computable while allowing any node potentially to influence any other, depending on the nature of the graph. Perhaps due to the constraint on processing time, such schemes are typically straightforward in nature, and make little use of domain knowledge regarding the structure of the graph. The dominant paradigm is the following: a random walk is initiated from a set of seed pages, and with some probability at each step either continues forward, or restarts. The score of a node is taken to be the steady state probability of the node in this process.

In this note we present a family of measures of proximity of an arbitrary node in a directed graph to a pre-specified subset of nodes, called the *anchor*. Our measures are based on three different propagation schemes and two different uses of the connectivity structure of the graph. The interpretation and presentation of the propagation measures in the context of proximity to an anchor is novel.

We then consider a web-specific application of the above measures with two disjoint anchors: *good* and *bad*. The key assumption is that good web pages are highly unlikely to link to bad web pages. The goal is to assign a goodness quality score to all web pages. While the key assumption (like all assumptions) is violated on the web in many ways, it remains largely true, and it gives us a starting point from which to evaluate the quality of unknown web pages. Our measures are especially applicable to combating web spam.


*WWW 2007,* May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

## Preliminaries

Let $G = (V, E)$ be a directed graph with $|V| = n$. Let $M$ be the matrix associated with the graph, i.e., $m_{u,v} = 1$ if there is an edge from $u$ to $v$, and 0 otherwise. Let $\mathrm{od}_u$ be the out-degree of node $u$, i.e., $\mathrm{od}_u = \sum_{v=1}^{n} m_{u,v}$. Likewise, let $\mathrm{id}_u$ be the in-degree of $u$, i.e., $\mathrm{id}_u = \sum_{v=1}^{n} m_{v,u}$. Let $\mathrm{out}(u)$ denote the out-neighbors of $u$.

Let $R = [r_{u,v}]$ be the row-normalized version of $M$: $r_{u,v} = m_{u,v}/\mathrm{od}_u$. Similarly, let $C = [c_{u,v}]$ be the column-normalized version of $M$: $c_{u,v} = m_{u,v}/\mathrm{id}_v$. $R$ defines a Markov process on the graph whose one-step update rule for any probability distribution $\pi$ over the $n$ nodes is $\pi \leftarrow R^T \cdot \pi$. The matrix $C$ may be seen as the transpose of the row-normalized form of $M^T$. Thus, there is another natural Markov process on $G$ defined by walking backwards on the edges rather than forwards. The update rule for this process is $\pi \leftarrow C \cdot \pi$.

## Proximity to an anchor

Let $S \subseteq V$ be a subset of nodes in the graph, called the *anchor*. We propose various notions of *proximity* of a given node to this anchor. All of these notions compute a real-valued score $\pi(S; u) \in [0, 1]$ for every node $u \in V \setminus S$; $\pi(S; u) = 1$ for $u \in S$.

The most natural way to define the proximity of $u$ to $S$ would be to look at the connectivity of $u$ to $S$. In this, we have two options: either use the forward connectivity of $u$ to $S$ or use the backward connectivity of $S$ to $u$. We denote the former by $\pi(S, f; u)$ and the latter by $\pi(S, b; u)$. For simplicity, we present only the forward connectivity approaches. The backward connectivity approaches can be easily realized either by reversing all edges in $G$ or working with the transpose of $M$.

A first cut approach would be to take the shortest path from $u$ to any node in $S$. Unfortunately, this is not robust since it does not take into account multiple connections from $u$ to $V$. An alternate approach would be to compute the maximum flow from $u$ to the anchor $S$, realized by hooking all nodes in $S$ to a node $v$ and then computing a $u$–$v$ flow. This has the advantage that parallel paths from $v$ to $S$ are taken into account; however, the lengths of these paths are ignored. Compromising these extremes, we consider various natural propagation methods that take into account both the length and the number of paths from $u$ to $S$.

PERSONALIZED PAGERANK. We assume that the reader is familiar with the topic-sensitive PageRank notion [1]. We take $S$ to specify the personalization vector. Consider the Markov chain of the following random walk on the nodes of $G$. At each step, with probability $1 - \alpha = 0.85$, the walk proceeds to a neighbor of $u$ (if any), chosen uniformly at random and if there are no neighbors, then the walk jumps to a uniformly chosen node in $V$. With probability $\alpha = 0.15$, the walk jumps to a uniformly chosen node in $S$. The proximity of $u$ to $S$ is then given by the stationary probability of $u$ in the Markov chain; thus we have the column vector

$$\pi(S, f, \mathrm{pr}) = (1 - \alpha) \cdot R^T \cdot \pi(S, f, \mathrm{pr}) + (\alpha/|S|) \cdot \chi_S,$$

where $\chi_S$ is the characteristic vector of $S$.

HARMONIC RANK. Consider the simple random walk given by the matrix $R$. We will modify the walk to begin at a specific start node $u$. We will then create two absorbing states: $s$, corresponding to the anchor $S$, and $r$, corresponding to the notion of "re-start" of the random process. We will modify all pages to link with probability $\alpha$ to $r$; for pages with no outlinks, we will modify them to link to $r$ with probability 1. We will then modify the result so that all nodes in $S$ link with probability 1 instead to the state $s$. This new walk must be absorbed into either the state $s$ or the restart state $r$. If the start node $u$ has many short paths to $S$, then it is much more likely to be absorbed into $s$. Let $\pi(S, f, \mathrm{hr}; u)$ be the probability that the walk is absorbed into $s$, so with probability $1 - \pi(S, f, \mathrm{hr}; u)$, the walk is absorbed into $r$ instead.

We show this quantity can be computed efficiently. Let $F = [f_{u,v}]$ be the matrix for the walk described above. $F$ is a row stochastic matrix with $f_{u,v}$ representing forward walk from node $u$. $F$ incorporates a certain probability $\alpha$ to jump to restart state $s$. The remaining probability is evenly distributed over all outgoing links starting from $u$. Thus, $f_{u,v} = \alpha$ if $v = r$ and $(1 - \alpha)/\mathrm{od}_u$ otherwise.

The proximity score of a node is related to those of its out-neighbors by the following harmonic equation:

$$\pi(S, f, \mathrm{hr}; u) = \begin{cases} 0 & u = r \\ 1 & u \in S \cup \{s\} \\ \sum_{v \in \mathrm{out}(u)} f_{u,v} \cdot \pi(S, f, \mathrm{hr}; v) & \text{otherwise} \end{cases}$$

Consider a distribution $\pi$ over the nodes in which $\pi(r) = 0$ and $\pi(s) = 1$. Then the harmonic equation given above may be rewritten as $\pi \leftarrow F \cdot \pi$. Observe that this steady-state equation is quite different from the steady-state equation for a single step in the random walk: $\pi \leftarrow F^T \cdot \pi$. The solutions to this latter equation are non-zero in only the states $r$ and $s$, and the values depend on the start location of the walk. The former equation, which is of interest to us, does not represent a walk and is expressed as a column-stochastic rather than a row-stochastic matrix. Thus, the absorption probabilities of $n$ markov processes, each tailored to a single start state, may be efficiently computed simultaneously, due to the special structure of this matrix.

NON-CONSERVING RANK. Consider a propagation rule in which each node $u$ begins with some initial score $p_u$, and the score is updated by the rule $\pi \leftarrow \pi + \gamma M^T \pi$, where $\gamma$ is an attenuation parameter that controls how much a particular score decays as it propagates. Generally, we may perform this propagation infinitely many steps, resulting in a final equation for $\pi$ based on some initial vector $p$:

$$\pi(S, f, \mathrm{nr}) = \sum_{j=0}^{\infty} \gamma^i (M^T)^i v = (I - \gamma M^T)^{-1} p.$$

If $M$ is stochastic, observe that this equation is similar to the equation for personalized pagerank with reset distribution given by $p$, and reset probability given by $(1 - \gamma)$. If $M$ is not row-stochastic, we must check that the sum converges; but as long as this is the case, the new measure is a natural generalization of personalized pagerank.

Non-conserving rank has a desirable property in the context of spam resilience: if a spammer's destination page is marked as spam, then all pages created by the spammer to direct traffic towards this destination page will also be marked as spam. Even if the spammer is able to manipulate the graph by adding other links, the score of the inlinking pages will never be demoted by this manipulation.

## Experiments and results

We illustrate a primary application of this technique in the detection of spam (bad) pages. We are given two anchors, namely, *good* pages and *bad* pages. The crucial assumption we deploy is: a good page will not typically choose to link to a bad page. Therefore, pages with links from good pages are more likely to be good, and pages that link to bad pages are more likely to be bad.

We consider a graph of 48 million web domains (nodes) and obtained two non-overlapping anchors of three million *bad* nodes and two million *good* nodes. For evaluation, we used leave-one -out validation with 1000 random nodes from each anchor removed before propagation. The success of a technique is measured as the fraction of the 2000 nodes classified correctly. We consider the proximity measures that are consistent with the above assumption and computed the accuracies. The results are tabulated below.

| Measure | Acc. | Measure | Acc. |
|---|---|---|---|
| $\pi(G, f, \mathrm{pr})$ | 82.06 | $\pi(B, b, \mathrm{pr})$ | 82.06 |
| $\pi(G, b, \mathrm{hr})$ | 83.89 | $\pi(B, f, \mathrm{hr})$ | **85.71** |
| $\pi(G, f, \mathrm{nr})$ | **84.49** | $\pi(B, b, \mathrm{nr})$ | 83.89 |

Clearly, harmonic rank works best for the bad anchor and non-conserving rank works best for the good anchor. We then used logistic regression and multilayer perceptron to combine multiple proximity measures. The following table presents the performance of the combinations. Harmonic rank achieves the best performance.

| Measure | Acc. |
|---|---|
| $\pi(\cdot, \cdot, \mathrm{pr})$ | 82.97 |
| $\pi(\cdot, \cdot, \mathrm{hr})$ | **86.93** |
| $\pi(\cdot, \cdot, \mathrm{nr})$ | 85.71 |
| $\pi(\cdot, \cdot, \cdot)$ | 86.93 |

Above, we illustrated an application of the proximity measures to web spam. However, the techniques are quite general, and will apply for other definitions of the good and bad anchors. For example, we may employ a set of known pornographic pages as the bad anchor. Or we may select a set of high-caliber blogs as the good anchor, and a set of lower-caliber blogs as the bad anchor, in order to determine the likely caliber of a set of unknown blogs.

## Reference

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