Comparing Apples and Oranges: Normalized PageRank for Evolving Graphs

Klaus Berberich Srikanta Bedathur Gerhard Weikum
Max-Planck Institute for Informatics
Saarbrücken, Germany
{kberberi, bedathur, weikum}@mpi-inf.mpg.de

Michalis Vazirgiannis*
INRIA/FUTURS
Paris, France
michalis.vazirgiannis@inria.fr

ABSTRACT

PageRank is the best known technique for link-based importance ranking. The computed importance scores, however, are not directly comparable across different snapshots of an evolving graph. We present an efficiently computable normalization for PageRank scores that makes them comparable across graphs. Furthermore, we show that the normalized PageRank scores are robust to non-local changes in the graph, unlike the standard PageRank measure.

Categories and Subject Descriptors: H.4.m [**Information Systems**]: Miscellaneous

General Terms: Algorithms, Measurement

Keywords: PageRank, Web dynamics, Web graph

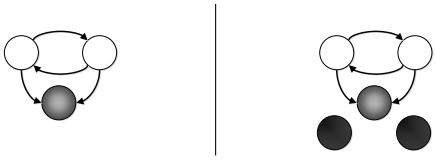
1. MOTIVATION

PageRank [9] is a well known link-based ranking technique, widely adopted both in practice and research. At the core of the method lies a random walk on the (web) graph that can be equivalently represented as a finite Markov chain. The visiting probabilities of the random walk, in other words, the stationary state probabilities of the equivalent Markov chain are represented by the corresponding PageRank scores of nodes in the graph.

As a consequence of its probabilistic foundation and the fact that each node is guaranteed to be visited, PageRank scores are generally *not comparable across different graphs* as the following example demonstrates. Consider the grey and white nodes in the two graphs shown in Figure 1. Intuitively, importance of neither the grey node nor the white nodes should decrease through the addition of the two black nodes, since none of these nodes are "affected" by the graph change. The non-normalized PageRank scores, however, as given in the corresponding table in Figure 1 convey decreases in the importance of the grey node and the white nodes, thus contradicting intuition. These decreases are due to the random jump inherent to PageRank that guarantees the additional black nodes to have non-zero visiting probability.

The need for normalized PageRank scores, which can be compared across different graphs, arises in different contexts. In a bibliographic application, importance of publications from different areas need to be compared. On the Web, growth patterns of PageRank can potentially be used to identify spamming attempts. Finally, in a distributed Web search engine or a metasearch engine like [2, 8], PageRank scores computed on different document collections of varying size must be compared or aggregated.

Previous work [5] has only dealt with the normalization of PageRank scores in an ad-hoc manner. In this work we propose a new principled normalization, describe the rationale behind it, and show its robustness. The proposed normalization has been successfully used in practical applications [3].



Graph A *Graph B*

	PageRank (non-normalized)		PageRank (normalized)	
Node	***A***	***B***	***A***	***B***
White	0.2920	0.2186	1.7391	1.7391
Grey	0.4160	0.3115	2.4781	2.4781
Black	–	0.1257	–	1.0000

Figure 1: Sensitivity of PageRank Values ($\epsilon = 0.15$)

2. PAGERANK NORMALIZATION

We briefly recall the definition of PageRank and introduce some notation. Let $G(V, E)$ be a directed graph, the PageRank score $r(v)$ of a node v is defined as

$$r(v) = (1-\epsilon) \sum_{(u,v) \in E} \frac{r(u)}{out(u)} + \frac{\epsilon}{|V|} ,$$

with $out(u)$ denoting the out-degree of node u and ϵ being the probability of making a random jump (aka. damping factor).

Accordingly, the PageRank score of any node in the graph is lower bounded by

$$r_{low} = \frac{\epsilon}{|V|} ,$$

which is the score assigned to a node without incoming edges. However, this definition does not account for dangling nodes (i.e., nodes without any outgoing edges) – which are shown to form a significant portion of the Web graph crawled by search engines [4]. These nodes are treated by making a random jump whenever the random walk enters a dangling node. Under this model, with $D \subseteq V$ denoting the set of dangling nodes, the modified lower bound for PageRank scores is given by:

$$r_{low} = \frac{1}{|V|}(\epsilon + (1-\epsilon) \sum_{d \in D} r(d))$$

*Michalis Vazirgiannis is on sabbatical leave from Athens University of Economics & Business (Greece). His work is funded by the NGWeMiS Marie Curie Intra European Fellowship (MEIF-CT-2005-011549)

Copyright is held by the author/owner(s).
WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

which is again the score assigned to a node without incoming edges. We use this refined lower bound for normalizing the PageRank scores – for a node v its normalized PageRank score is defined as

$$\hat{r}(v) = \frac{r(v)}{r_{low}} .$$

In contrast to non-normalized PageRank scores that correspond to *visiting probabilities* on the graph and thus depend on its size, the normalized PageRank scores convey *how much more likely a node is to be visited than a node having least possible importance*. The normalization thus eliminates the dependence on the size of the graph. For the earlier example, the normalized PageRank scores of the grey and the white nodes do not change as can be seen from the table in Figure 1.

The computational cost associated with the proposed normalization scheme is low. Identifying the set of dangling nodes is possible in a single scan over the set of edges. Summing up the PageRank scores of the dangling nodes requires another scan over the vector of non-normalized PageRank scores. If PageRank scores are computed as described in [9] (i.e., using a variant of the power iteration method), the value r_{low} can be computed in one additional iteration, noting that for the transition matrix, $\mathbf{M}$, defined by,

$$M_{ij} = \begin{cases} 1/out(i) & : \quad (i,j) \in E \\ 0 & : \quad \text{otherwise} \end{cases}$$

and the vector of PageRank scores $\mathbf{r}$, the following holds

$$\sum_{d \in D} r(d) = \|\mathbf{r}\|_1 - \|\mathbf{r}\,\mathbf{M}\|_1 .$$

Our normalization can be applied separately for each given graph. Thus, for instance, if PageRank scores obtained on two graphs are to be compared, the scores can be normalized separately without knowing the other graph. This property is not common to all normalization schemes and centrality indices as pointed out in [7], but is crucial for applications where PageRank scores are computed, normalized, and stored on snapshots of a large evolving graph (e.g., the Web).

As demonstrated in the example above, non-normalized PageRank scores are not robust given small changes to the graph. In the example, a small change affected the PageRank score of all nodes, regardless of their proximity to the change. We next give a theorem that describes conditions under which the normalized PageRank score of a node is robust to atomic changes in the graph. Before stating the theorem we need two definitions as preliminaries.

DEFINITION 1. *Let $G(V,E)$ be a directed graph, an **atomic change** is an addition/removal of a single node v or a single edge (u,v).*

Clearly, only nodes having neither incoming nor outgoing edges can be removed from the graph (i.e., edges emanating from or pointing to the node must be removed first), and, similarly, added edges must emanate from and point to existing nodes. The graph resulting from an atomic graph change is referred to as $G'(V',E')$ and the corresponding PageRank scores as r' and $\hat{r}'$. The scope of an atomic change includes nodes that are affected by the change and is defined as follows.

DEFINITION 2. *The **scope** $S \subseteq V \cup V'$ of an atomic change is defined as*

- *$S = \{v\}$ for an addition/removal of node v*
- *$S = \{w \in V \cup V' \mid w$ is reachable from u in G or $G'\}$ for an addition/removal of edge (u,v)*

Given these preliminaries we now state the following theorem regarding the robustness of normalized PageRank scores.

THEOREM 1. *Let S be the scope of an atomic change, then $v \notin S \Rightarrow \hat{r}(v) = \hat{r}'(v)$ for a node v.*

Thus, for an atomic change, the normalized PageRank scores of nodes that are not in its scope do not change. For the proof of the theorem we use the following lemma that follows from the results of Avrachenko et al. [1] and Jeh and Widom [6].

LEMMA 1. *For a node $v \in V$ the normalized PageRank score $\hat{r}(v)$ can be written as*

$$\hat{r}(v) = \sum_{u \in V} \sum_{t : u \rightsquigarrow v} \left(\prod_{i=1}^{l(t)} \frac{1}{out(\omega_i)} \right) (1-\epsilon)^{l(t)} .$$

Here, $t : u \rightsquigarrow v$ denotes the set of tours (i.e., paths possibly containing cycles) that start from u and end in v. For a single tour $t = \langle \omega_1, \ldots, \omega_n \rangle$ (with $\omega_1 = u$ and $\omega_n = v$) $l(t) = n-1$ denotes the length of the tour.

PROOF. It can be seen from the preceding lemma that the normalized PageRank score of a node v only changes if either the set of tours leading to v (i.e., $\bigcup_{u \in V} u \rightsquigarrow v$) changes, or, the out-degree of any predecessor of v changes. Let us now distinguish the two cases *addition/removal of 1) a node w and 2) an edge (u,w)*.

1. Since $v \notin S$, clearly $v \neq u$ holds. Apart from that, we know that w has no outgoing edges and thus there is no tour $u \rightsquigarrow v$.
2. Since $v \notin S$, v is not reachable from u in both $G(V,E)$ and $G'(V',E')$. Therefore, the set of tours leading to v is unchanged. Moreover, the change of u's out-degree does not affect the normalized PageRank of v, since u is no predecessor of v.

□

The theorem above considers only atomic changes, but extends naturally to arbitrarily different graphs as the following corollary explains.

COROLLARY 1. *Let $G(V,E)$ and $G'(V',E')$ be two arbitrary graphs and $\langle S_1, \ldots, S_n \rangle$ be the scopes of a series of graph changes that produces the latter from the former graph, then $v \notin \bigcup_{i=1}^{n} S_i \Rightarrow \hat{r}(v) = \hat{r}'(v)$ for a node v.*

3. REFERENCES

[1] K. Avrachenko, N. Litvak, et al. Monte Carlo methods in PageRank computation: When one iteration is sufficient. Tech. rep., INRIA Sophia Antipolis, 2005.

[2] M. Bender, S. Michel, et al. Minerva: Collaborative P2P Search. In *VLDB* 2005.

[3] K. Berberich, S. Bedathur, et al. BuzzRank...and the Trend is Your Friend In *WWW* 2006.

[4] N. Eiron, K.S. McCurley, et al. Ranking the Web Frontier. In *WWW* 2004.

[5] Z. Gyöngyi and H. Garcia-Molina. Link Spam Alliances. In *VLDB* 2005.

[6] G. Jeh and J. Widom. Scaling Personalized Web Search. In *WWW* 2003.

[7] D. Koschützki, K. Lehmann, et al. Advanced Centrality Concepts. In Vol. 3418 of *LNCS*. Springer, 2004.

[8] W. Meng, C. Yu, et al. Building efficient and effective metasearch engines. *ACM Comput. Surv.*, 34(1), 2002.

[9] L. Page, S. Brin, et al. The PageRank Citation Ranking: Bringing Order to the Web. Stanford Tech. rep., 1998.