Comparing Apples and Oranges: Normalized PageRank for Evolving Graphs

Klaus Berberich  Srikanta Bedathur  Gerhard Weikum  Michalis Vazirgiannis*
Max-Planck Institute for Informatics  Saarbrücken, Germany  INRIA-Futurs  Paris, France
{kberberi, bedathur, weikum}@mpi-inf.mpg.de  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 applications 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, PageRank values differ between different graphs.

Categories and Subject Descriptors: H.4.m [Information Systems]: Miscellaneous
General Terms: Algorithms, Measurement
Keywords: PageRank, web dynamics, web graph

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

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

with \( \text{out}(u) \) denoting the out-degree of node \( u \) and \( \epsilon \) 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:

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

\[
\hat{r}(v) = \frac{r(v)}{r_{\text{low}}}.
\]

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 \( \mathcal{O}(m) \) operations (i.e., edges emanating from or pointing to the node must be removed first), additions to the graph (i.e., edges emanating from or pointing to the node must be removed first), and stored on snapshots of a large evolving graph (e.g., the Web).

We use this refined lower bound for normalizing the normalized PageRank scores.

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

\[
\hat{r}(v) = \sum_{u \in V} \sum_{t : u \rightarrow v} \left( \prod_{i=1}^{l(t)} \frac{1}{\text{out}(\omega_i)} \right) (1 - e)^{l(t)}.
\]

Here, \( t : u \rightarrow v \) denotes the set of tours (i.e., paths possibly containing cycles) that start from \( u \) and end in \( v \). For a single tour \( t = (\omega_1, \ldots, \omega_n) \) (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., \( \bigcup_{u \in V : u \rightarrow v} t \)) changes, or, the out-degree of any predecessor of \( v \) changes. Let us now distinguish the two cases addition/removal of 1) a node \( u \) and 2) an edge \( (u, v) \).

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 \( u \rightarrow v \).
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 \( v \)'s out-degree does not affect the normalized PageRank score, 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 \( \{S_1, S_2, \ldots, S_n\} \) 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 \( v \).

**REFERENCES**


