TruRank: Taking PageRank to the Limit

Sebastiano Vigna
Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Italy
vigna@acm.org

ABSTRACT

PageRank is defined as the stationary state of a Markov chain depending on a *damping factor* α that spreads uniformly part of the rank. The choice of α is eminently empirical, and in most cases the original suggestion $\alpha = 0.85$ by Brin and Page is still used. It is common belief that values of α closer to 1 give a "truer to the web" PageRank, but a small α accelerates convergence. Recently, however, it has been shown that when $\alpha = 1$ all pages in the core component are very likely to have rank 0 [1]. This behaviour makes it difficult to understand PageRank when $\alpha \approx 1$, as it converges to a meaningless value for most pages. We propose a simple and natural modification to the standard preprocessing performed on the adjacency matrix of the graph, resulting in a ranking scheme we call *TruRank*. TruRank ranks the web with principles almost identical to PageRank, but it gives meaningful values also when $\alpha \approx 1$.

Categories and Subject Descriptors: G.2 [**Discrete Mathematics**]: Graph Theory G.3 [**Probability and Statistics**]: Markov processes

General Terms: Algorithms, Experimentation, Measurement.

Keywords: Web graph, PageRank.

1. INTRODUCTION

PageRank [2] is probably the most famous static ranking algorithm for web pages. It is the stationary state of a certain Markov chain obtained by scaling a patched version of the adjacency matrix of a web graph. The patch consists mainly in adding new outlinks towards all nodes to nodes without outlinks. Finally, the matrix is row-normalised. At this point, a *damping factor* is introduced: the matrix is scaled by α, and the final chain is obtained by summing a matrix filled with $1 - \alpha$. PageRank is the stationary state of this final matrix.

The damping factor α is essential in the practical computation of PageRank, as it greatly increases the speed of power methods: however, it is also common belief that when α goes to 1 we get a PageRank that is "truer to the web".

This common belief, however, is false, as shown in [1]: even if all nodes without outlinks are connected to all other nodes, in real-world graphs *most of the nodes will get a PageRank going to 0 as* $\alpha \to 1$ (in particular, this will happen to all nodes in the main component). The problem lies in the very definition of PageRank: the patch applied to the adjacency matrix does *not* create a single component, but, rather, leaves the component tree unchanged.

2. TRURANK

We still believe that understanding PageRank when $\alpha \approx 1$ can be fruitful, but there is no easy way out using the standard PageRank definition. In this poster we make a theoretical proposal to solve this dilemma. Our proposal is presently of difficult, if not impossible, implementation because of lack of sufficient computing power, but we believe it is a step in the right direction.

To describe clearly PageRank's problem, let us set up some definitions from [1]. A node is *terminal* if it does not have outlinks, except possibly for loops. If we want to be specific about the presence of a loop, we shall use the terms *looped* and *loopless*. For instance, the patch to the adjacency matrix tries to fix loopless (and possibly looped) terminal nodes.

However, if we consider the graph of strong components, all nodes belonging to a terminal component will absorb ranking, without giving it back. We call such nodes *rank sinks*.

The problem with PageRank in its current form is that it provides patches for loopless terminal nodes, which are just a kind of rank sink, but not for all rank sinks. Unfortunately, for all aspects related to convergence they are equivalent, as they both represent terminal nodes in the component graph (the only difference being that terminal nodes belong to non-looped terminal components). The effect is not very visible because the values of α currently in use make PageRank dependent on a rather limited in-neighbourood (in the random surfer interpretation, the "average surf" is $1/(1-\alpha)$ links long).

Thus, to make the study of α more meaningful in the region close to 1, we propose a slightly amended ranking algorithm named *TruRank*. The idea of TruRank is that it should work on the same principles of PageRank, but it should give meaningful values when $\alpha \to 1$, resulting in a "true to the web" ranking.

The modification we propose is that the transition matrix P is built from G in the following way (d is the outdegree of a node):

1. arcs going out from nodes that are not rank sinks are weighted $1/d$ (note that in that case necessarily $d > 0$);
2. arcs going out from rank sinks are weighted $1/(d + 1)$;
3. finally, for every rank sink x of outdegree $d \geq 0$ and every node y that is not a successor of x, we add a new arc from x to y weighted $1/[(d + 1)(N - d)]$.

Thus, P's rows for nodes that are not rank sinks go unchanged. The only difference is that rank sinks must "donate to the world" $1/(d + 1)$ of their rank. Note that we get back exactly PageRank's patch for terminal nodes, as their rows are filled with $1/(0+1)(N-0) = 1/N$ as usual. In this sense, TruRank is a true generalisation of PageRank.

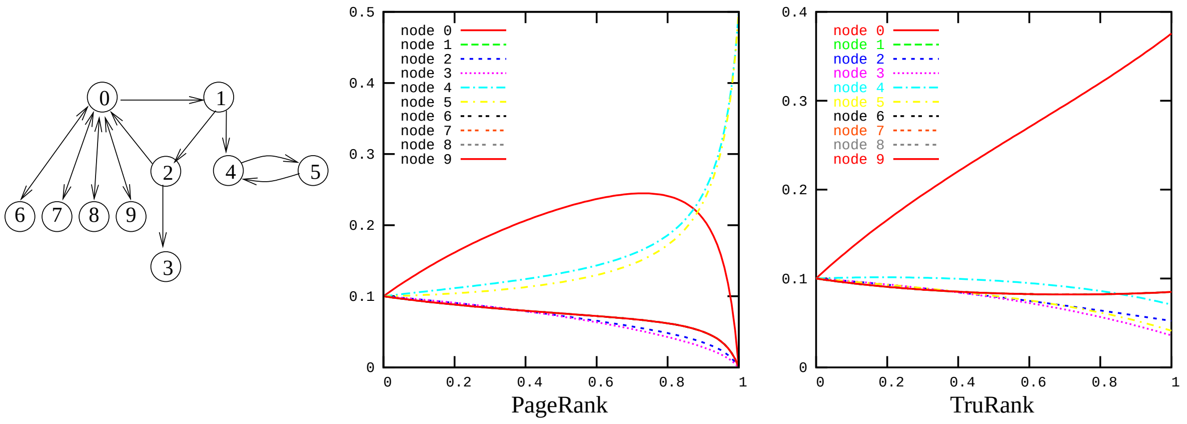

Figure 1: A sample graph, and the relative behaviour of PageRank and TruRank when α changes.

$$\begin{bmatrix}
0 & \frac{1}{5} & 0 & 0 & 0 & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5}\\
0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0\\
\frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0\\
\frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10}\\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0\\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\quad
\begin{bmatrix}
0 & \frac{1}{5} & 0 & 0 & 0 & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5}\\
0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0\\
\frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0\\
\frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10}\\
\frac{1}{18} & \frac{1}{18} & \frac{1}{18} & \frac{1}{18} & \frac{1}{18} & \frac{1}{2} & \frac{1}{18} & \frac{1}{18} & \frac{1}{18} & \frac{1}{18}\\
\frac{1}{18} & \frac{1}{18} & \frac{1}{18} & \frac{1}{18} & \frac{1}{2} & \frac{1}{18} & \frac{1}{18} & \frac{1}{18} & \frac{1}{18} & \frac{1}{18}\\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

Figure 2: PageRank and TruRank iteration matrices for the graph of Figure 1.

Now P can be easily seen to be irreducible, and thus TruRank has a unique stationary distribution, even for $\alpha = 1$ (we assume aperiodicity). Experimentally, TruRank and PageRank largely coincide for $\alpha \ll 1$ (Kendall's τ is larger than .95), but as α approaches 1 we believe that TruRank gives a meaningful "true to the web" distribution.

Intuitively, in TruRank rank sinks are punished by forcing them to release part of their rank. The smaller the number of outlink of a page, the stronger the punishment: in particular, as it was happening in PageRank, a page with no outlinks must donate all of its rank.

3. AN EXAMPLE

In Figure 1 we show a sample graph. The only rank sinks are node 4 and 5, and the reader can easily see the difference with the corresponding rows. The behaviour of the two ranking algorithms when the damping factor α goes from 0 to 1 is also shown in Figure 1: we can clearly see that the few nodes in the only looped terminal component (4 and 5) absorbe all available PageRank. On the contrary, we can see that TruRank makes the ranking of the various nodes converge consistently.

Indeed, node 0 (clearly the most referenced node in the graph) preserve its TruRank when $\alpha \to 1$, absorbing about one third of the ranking of the whole graph.

4. COMPUTING TRURANK

TruRank can be computed using a trivial modification of the standard power method. The only additional information required is one bit for each node, specifying whether it is a rank sink: it can be easily computed using a slightly modified version of Tarjan's algorithm for strongly connected components.

However, the main, and currently unsolved, problem with TruRank is that albeit Perron–Frobenius theory guarantees the converges of the Markov chain even for $\alpha = 1$, it gives us no guarantee on the precision that is necessary to perform the computation, and on the required number of steps. Thus, computing TruRank on a realistic data set (e.g., one billion pages) when $\alpha \approx 1$ is currently beyond our computational capabilities. More research is necessary to speed up the power method to match the requirements of TruRank.

5. REFERENCES

[1] Paolo Boldi, Massimo Santini, and Sebastiano Vigna. PageRank as a function of the damping factor. In *Proc. of the Fourteenth International World Wide Web Conference*, Chiba, Japan, 2005. ACM Press.

[2] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, Stanford University, Stanford, CA, USA, 1998.