Elliot Smith
Research 2 - PageRank

1. The key need for the PageRank algorithm is to optimize the results returned to end-users when they perform web searches using search engines. Due to the explosion in the number of web pages (estimated at 150 million in 1998), it is crucial to make sure that the most relative results are returned to the users, in order of importance, so that they will receive the web pages that are most pertinent to their request and with as little need of results parsing as possible to facilitate ease of use.

   The current approach has some glaring shortcomings, most notably:

   a. Due to a lack of holistic web curation, websites may use strategies to artificially inflate the number of citations that they receive, thus making them look more popular or relevant when they are not so.
   b. The diversity of web pages makes it very difficult to set a standard by which to judge all web pages. The comparison between academic papers, where a defined structure is followed, and a web page, which has no structure definitions, makes it hard to compare several web pages together fairly.

2. The key object that is the solution to this need is the PageRank. PageRank is a methodology by which we may apply a ranking for each web page, and thus, have a scale of relative importance to define web request results. Broadly, PageRank will give a higher weighting, or rank, to websites that have a larger number of other web pages that have links pointing to it. Roughly, if web page A has three links pointing to it and web page B has seven links pointing to it, web page B will have a higher PageRank and will be a higher ranked result from the query. The inference drawn here is that if a web page has larger number of links pointing to it, it is believed that it is more important than one with less links pointing to it.

3. One notable weakness or limitation that the author identified of the PageRank algorithm related to manipulation by commercial interests. While the algorithm is virtually immune to manipulation, based on the fact that web pages are judged by how many links are pointing to it, it gives a significant advantage to sites that have more capital to spend on advertising. For example, a certain web page could pay other sites to put a pointer to its web page, while smaller companies with smaller advertising budgets would not be able to do this. The author did not see this as a serious concern, but as the web has evolved over the last twenty years, I believe it would be perceived as such as companies have looked to query result optimization to monetize.

   Another problem (and next step) proposed by the author was how to handle the E component, which is a vector over the web pages used in the rank to account for web

pages that cycle amongst themselves with pointers; the E component allows for random deviations out of this “Rank Sink”. There are several options for how to define the E component, for example, whether E should give uniform weighting to all web pages or with any other method. The downside, for example, with the methodology of giving equal weighting to all web pages is that they are given value simply because they exist, which may not be a very elegant solution.

4. Something interesting I learned from this paper is how simple the PageRank algorithm really is. Often we hear of these great advancements in the world of Computer Science and perceive them as these concepts or programs of monumental size and scope, but often we come to learn that the premise (and the implementation) are much more straightforward than we originally believe. The definition of the algorithm is very simple and makes a lot of sense: if other web pages think web page A is really important, than it probably is! While the idea may seem very intuitive to us now, having developed as computer scientists in the Internet age, to those that were the pioneers of exploration into this world at the time, it must have appeared very revolutionary indeed!

5. Maslov, Sergei; Redner, Sidney. Promise and Pitfalls of Extending Google’s PageRank Algorithm to Citation Networks. Journal of Neuroscience, 29 October 2008.

   This paper finds flaws in Google’s PageRank methodology when applies to scientific publications, and thus, questions its application in this realm. The key issues that the authors raise are:

   1. The importance of the citing paper is ignored; a citation from a highly cited groundbreaking paper will be treated equally as from an obscure paper
   2. Different scientific fields have different standards and practices when it comes to numbers of citations resulting in difficulty in cross-discipline importance comparison (for example: an average paper is cited ~6 times in life sciences but <1 in mathematics)
   3. The evolution in scientific community sizes may give greater or lesser importance to later written papers as the size of those communities expand or contract

   This paper relates to our paper because it talks about a case where perhaps the PageRank solution is not the best way to determine web-based information’s importance. I learned that while the PageRank algorithm is a great tool for generic web page importance, there is still some work to be done to apply it broadly to more esoteric cases, such as the one presented here. While a general web PageRank algorithm is certainly appropriate, perhaps a more domain-specific could be developed in situations like this to improve the importance ranking of results.