I. Discussion: Why’s Page Ranking Hard?

1. **Keywords are a very limited way to express a complex information need:** For a long time, up through the 1980s, information retrieval was the province of reference librarians, patent attorneys, and other people whose jobs consisted of searching collections of documents; such people were trained in how to formulate effective queries, and the documents they were searching tended to be written by professionals, using a controlled style and vocabulary. With the arrival of the Web, the problems surrounding information retrieval exploded in scale and complexity.

2. **The diversity in authoring styles makes it much harder to rank documents according to a common criterion:** on a single topic, one can easily find pages written by experts, novices, children, conspiracy theorists and not being able to tell which is from whom.

3. **There is a correspondingly rich diversity in the set of people issuing queries, and the problem of multiple meanings becomes particularly severe.**

II. Hubs and Authorities

- In response to the one-word query “Cornell,” what are the clues that suggest Cornell’s home page, www.cornell.edu, is a good answer?

- **Links are essential to ranking:** we can use them to assess the authority of a page on a topic, through the implicit endorsements that other pages on the topic confer through their links to it.

- **How do we achieve this?**
  First, collect a large sample of pages that are relevant to the query; Then let pages in this sample “vote” through their links; Now try draw a graph counting in-links for the query “newspapers”
It should look something like this:

- **A List-Finding technique**
  Now look at your Newspaper graph, which page ranks the highest? Usually, there is not necessarily a single, intuitively “best” answer here; there are a number of prominent newspapers on the Web, and an ideal answer would consist of a list of the most prominent among them.
The page's value as a list: $6 = 3(\text{Amazon}) + 3(\text{Yahoo})$

- **Hubs and Authorities**
  - **Authorities for the query**: the prominent, highly endorsed answers to the queries which we were originally seeking.
  - **Hubs for the query**: high-value lists

Now, for each page $p$, we’re trying to estimate its value as a potential authority and as a potential hub, and so we assign it two numerical scores: $\text{auth}(p)$ and $\text{hub}(p)$, both start out as 1.

**Authority Update Rule**: For each page $p$, update $\text{auth}(p)$ to be the sum of the hub scores of all pages that point to it.

**Hub Update Rule**: For each page $p$, update $\text{hub}(p)$ to be the sum of the authority scores of all pages that it points to.

Each update works as follows: – First apply the Authority Update Rule to the current set of scores. – Then apply the Hub Update Rule to the resulting set of scores. Repeat the procedures till you have a sequence of $k$ updates.

Now try the method on your “newspaper” graph. Remember to normalize the scores (:
Advanced Question: How big should k be?
III. PageRank

- Endorsement is best viewed as passing directly from one prominent page to another — in other words, a page is important if it is cited by other important pages.

In this case, we can easily tell that node A has the highest PageRank.

- **Computing PageRank**
  - In a network with n nodes, we assign all nodes the same initial PageRank, set to be 1/n.
  - We choose a number of steps k.
  - Basic PageRank Update Rule: Each page divides its current PageRank equally across its out-going links, and passes these equal shares to the pages it points to.

**Exercise:** Find the Equilibrium PageRank for each page in the above network
Extra Reading: How to appear #1 on Google?
https://moz.com/beginners-guide-to-seo (you don’t have to read them all, in fact, this is a pretty comprehensive guide and you are free to select whichever topic that interests you the most. Enjoy😊)