Computation, Information, and Intelligence (ENGRI/CS/INFO/COGST 172), Spring 2007
3/9/07: **Lecture 21 aid — More on PageRank**

Topics: Motivations for the explicit functional form of PageRank: normalization vs. a model of user browsing behavior.

Announcements:

- Daylight Savings Time ~~strikes~~ starts this Sunday: Spring forward! (Go, Spring!)
- About "showing all work" in Homework Three Part B: you are allowed to say, "computed via calculator" — you are not expected to extract (non-obvious) square roots by hand. But you should delay any use of calculators as long as possible, as the instructions request.

I. Reminder: PageRank Let ϵ be some constant between 0 and 1, exclusive.

- For every d_j in the n-document corpus, set $\mathrm{PR}^{(0)}(d_j)$ to $1/n$.
- Let i be increasing from 1 on, until it's the case that the set of PageRank scores "converges" (in practice, until the change in the set of scores between one value of i and and the next is sufficiently below some small threshold): set

$$\begin{aligned}\mathrm{PR}^{(i)}(d_j) &= (1-\epsilon)\left[\sum_{d\in \mathrm{To}(d_j)} \boxed{\mathrm{PR}^{(i-1)}(d)} \times \frac{1}{\mathrm{outdegree}(d)}\right] \\ &\quad + \epsilon \times \frac{1}{n}.\end{aligned}$$

II. Some facts about probabilities

- The probability of a non-impossible event e_1 happening and then an event e_2 happening is the probability that e_1 happens *times* the probability that e_2 happens *given* that e_1 happened.
- The probability of either (but not both) of two *mutually exclusive alternative events* e_1 and e_2 happening is the probability of e_1 happening *plus* the probability of e_2 happening.
- The sum of the probabilities over all possible mutually exclusive alternatives for a given probabilistic choice must be 1.

III. The "random surfer" model At the very beginning ($i = 0$), the user picks uniformly at random[1] some document to start looking at.

Upon arriving at a document, the user either chooses to follow an existing hyperlink from it, or to randomly jump to any document on the Web. The two cases have probability $(1 - \epsilon)$ and ϵ, respectively (note that these sum to 1), and in either case, the choice among alternatives that then result is made uniformly at random.

We then interpret $\mathrm{PR}^{(i)}(d_j)$ as the probability that the surfer is at document d_j at "time-step" i.

[1] This can be loosened considerably.