Computation, Information, and Intelligence (COMS/ENGRI/INFO/COGST 172), Fall 2005
10/12/05: **Lecture 20 aid — using links to enhance IR**

Agenda: Finish preferential attachment; work towards introducing Google's PageRank algorithm.

I. Reminder: framework for modeling Web evolution

- Start with $d_{-1}, d_{-2}, \ldots, d_{-n_0}$, where there are no links between them. We assume ℓ is an integer between 1 and n_0 inclusive.
- At the j^{th} time step, we add a new document named d_j and grant to d_j ℓ of links to some of the $n_0 + j - 1$ pre-existing documents, allowing repeated links to the same document.

We are interested in computing $I_j(t)$, which is a prediction of d_j's in-degree at time t.

II. Recap

So far, here's what we have discovered about link structures:

- They induce the "bowtie" structure of the Web.
- Link in-degrees follow a highly non-trivial pattern of distribution.

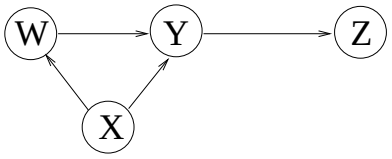
III. Illustrations of potential problems with content analysis

1. "lorry" vs. "truck"
2. The IBM homepage does not contain the word "computer".
3. "candidate X is a felon."

IV. Definitions and conventions

Let d be a document.
To(d): the set of documents that link to d.
From(d): the set of documents that are linked to by d.
For simplicity, we assume that documents have no self-links.
We can write |To(d)| and |From(d)| for the in-degree and out-degree of d, respectively.

V. An example set of Web documents



VI. PageRank, "the" Google algorithm

Introduced by Brin and Page (1998). We give an explicitly iterated version here. Let ϵ be some number between 0 and 1.

- For every d_j in the n-document corpus, set $\text{score}^{(0)}(d_j)$ to $1/n$.
- Repeat until the scores "converge" (the change in scores between one timestep and the next is sufficiently small): set

$$\text{score}^{(t+1)}(d_j) = \frac{\epsilon}{n} + (1-\epsilon) \sum_{d \in \text{To}(d_j)} \frac{\text{score}^{(t)}(d)}{|\text{From}(d)|}.$$