Computation, Information, and Intelligence (ENGRI/CS/INFO/COGST 172), Spring 2007
3/5/07: **Lecture 19 aid — More on models of web growth (Lecture 18 was the prelim)**

Topics: Finish the uniform-attachment model of Web evolution, and move on to the preferential-attachment model.

I. Reminders: the general set-up for modeling Web evolution

- On the Web, the number of documents with in-degree x is $x^{-2.1}$ (the "startling" log-log straight line).
- Our template for Web-evolution models: At each time step $t \geq 1$, we add a new document, and probabilistically choose ℓ links from d_j to some of the $n_0 + (j-1)$ pre-existing documents, allowing repeated links to the same document.
- $\text{In}(\text{docID} = j, \text{time} = t)$ denotes an estimate of d_j's in-degree at time t.

II. A heavy tail We find that according to the distribution function described above for the Web, for a given constant D, the number of documents with in-degree at least D is roughly $\frac{1}{D^{1.1}}$.

III. From last time: uniform attachment ("completely random") We suppose that links are chosen *uniformly at random* to the pre-existing documents — this means that at a given time t, each pre-existing page has the same probability, $1/\left(n_0 + (t-1)\right)$, of being selected as the "receiving end" of a given new link.

IV. Preferential attachment ("rich get richer") [1] We suppose that links are chosen so that a pre-existing document is pointed to with probability proportional to the number of links that already point to it. This means that at a given time t, each pre-existing page d_j has probability proportional to $\text{In}(\text{docID} = j, \text{time} = t) + \ell$ of being selected as the "receiving end" of a given new link. (We need the ℓ term (or other positive constant) to get the process off the ground.)

Using the same "rate" assumption as before yields the following for a given document d_j:

$$\frac{d\text{In}(\text{docID} = j, \text{time} = t)}{dt} = \ell \frac{\text{In}(\text{docID} = j, \text{time} = t) + \ell}{\sum_{d_k : k < t} \left[\text{In}(\text{docID} = k, \text{time} = t) + \ell\right]};$$

the denominator ensures that the sum of the linked-to probabilities over all pre-existing documents is 1. After some rewriting, integration yields

$$\text{In}(\text{docID} = j, \text{time} = t) = c'(j)\sqrt{2t + n_0 - 2} - \ell.$$

We solve for $c'(j)$ for $j \geq 1$ as before to determine the dependence on j.

---

[1] This model, proposed by Barabási, Albert, and Jeong (1999), has a "rich get richer" behavior.