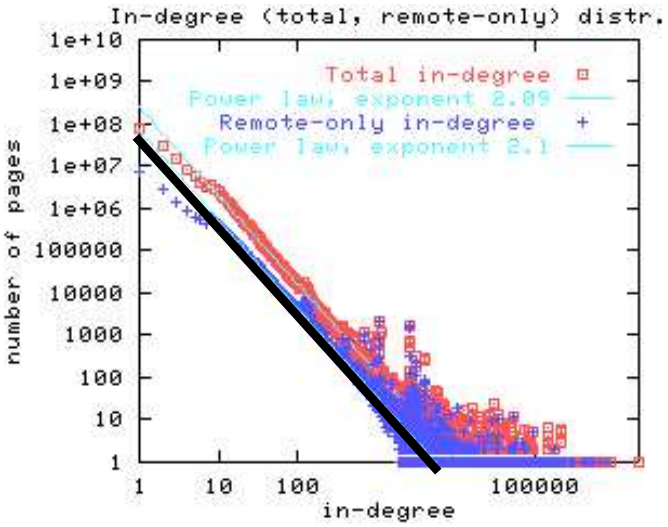
Computation, Information, and Intelligence (ENGRI/CS/INFO/COGST 172), Spring 2007
2/28/07: **Lecture 17 aid — Models of web growth**

Topics: Two mathematical models for the evolution of the Web's hyperlink structure.

I. The Web's in-degree distribution Here is Figure 1 of Broder et al. (2000), which shows the in-degree distribution, on a log-log scale, for their 200M-document Web crawl. The line corresponding to $\alpha = 2.1$ is shown in black.



This means that the ((mostly?) human-authored) data is following the relationship

$$\log(\text{number of documents with in-degree } x) \approx -2.1 \log(x)$$

or that the number of documents with in-degree x is $x^{-2.1}$. Also, for reasonable choices of number D, the number of documents with in-degree at least D is non-negligible: the distribution is said to be "heavy-tailed".

(OVER)

II. Web-evolution models: Template and conventions We consider $t = 0, 1, 2, \ldots$. We pre-determine constants $n_0 \geq 1$ and $\ell > 0$.

- We assume n_0 "original" documents exist at time $t = 0$; they have no links between them.
- At each time step $t \geq 1$, we add a new document, identified by the time. For example, at time $t = 100$, we create a new document d_{100}.

  When a given document d_j is created (that would be time-step j), we probabilistically choose ℓ links fromt d_j to some of the $n_0 + (j - 1)$ pre-existing documents, allowing repeated links to the same document.

III. Estimates for in-degree We are interested in computing $\mathrm{In}(\mathrm{docID} = j, \mathrm{time} = t)$, which is our estimate of d_j's in-degree at time t (where $t \geq j$: there's no point computing the in-degree of a document at a time when the document didn't exist). This estimate is used as a stepping stone to working out whether the in-degree distribution under a given model matches the data observed in the Web.

IV. Uniform attachment ("completely random") [1] 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. We then assume that

$$\frac{d\mathrm{In}(\mathrm{docID} = j, \mathrm{time} = t)}{dt} = \ell \frac{1}{n_0 + t - 1}.$$

Integrating with respect to t on both sides gives us that

$$\mathrm{In}(\mathrm{docID} = j, \mathrm{time} = t) = \ell \times \ln(n_0 + t - 1) + c(j)$$

and we can compute $c(j)$ for $j \geq 1$ by observing that $\mathrm{In}(\mathrm{docID} = j, \mathrm{time} = j) = 0$. (It's actually important to compute the constant: since it depends on j, it is what differentiates the in-degree estimate for one document from that of another.)

[1] This model is adapted from Erdös and Rényi (1960), who considered the properties of "random graphs".