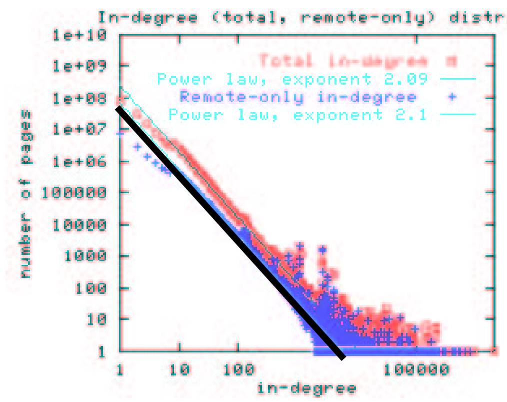
# CS/ENGRI 172, Fall 2002
## 10/28/02: Lecture Twenty-Five Handout

**Topics**: Local web structure; mathematical models of link creation.

### Power laws and Web in-degree distributions

A "power law" is a relationship of the form $y = x^{-\alpha}$, where $\alpha$ is a constant. Observe that if we take the log of both sides, we get the linear relationship $\log(y) = -\alpha \log(x)$.

The *(in-)degree distribution* of a given collection of linked documents gives, for each possible in-degree $x$, the number (or fraction) of documents that have in-degree equal to $x$. Here is Figure 1 of the Broder et al. (2000) reading, which shows the in-degree distribution, on a log-log scale, for their 200M-document crawl from the Web. The line corresponding to $\alpha = 2.1$ is highlighted.




### Conventions and notation

We use the integer-valued variable $t \geq 0$ to stand for time. The constant $n_0 \geq 1$ is the number of documents that exist at time $t = 0$; call them $d_{-1}, d_{-2}, \ldots, d_{-n_0}$. At the $j^{th}$ time step $(t = j \geq 1)$, we add a new document, named $d_j$; hence, positive subscripts indicate when a document was added. We then grant to $d_j$ a constant number $1 \leq \ell \leq n_0$ 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 our estimate of $d_j$'s in-degree at time $t \geq \min(j, 0)$ (there is no point computing the in-degree of a document at a time before it existed).

### Uniform attachment ("random" links) [adapted from Erdös and Rényi (1960)]

We suppose that links are chosen uniformly at random to the pre-existing documents — this means that each pre-existing page has the same probability, $1/(n_0 + t - 1)$, of being selected as the endpoint of a given new link. Roughly speaking, we can then assume

$$\frac{dI_j(t)}{dt} = \ell \frac{1}{n_0 + t - 1}.$$

Integrating with respect to $t$ on both sides gives us that

$$I_j(t) = \ell \cdot \ln(n_0 + t - 1) + c(j)$$

and we can compute $c(j)$ for $j \geq 1$ by observing that $I_j(j) = 0$ (so that $I_j(t)$ is a function of $j$).

(OVER)

**Preferential attachment ("rich get richer") [Barabási, Albert, Jeong 1999]**

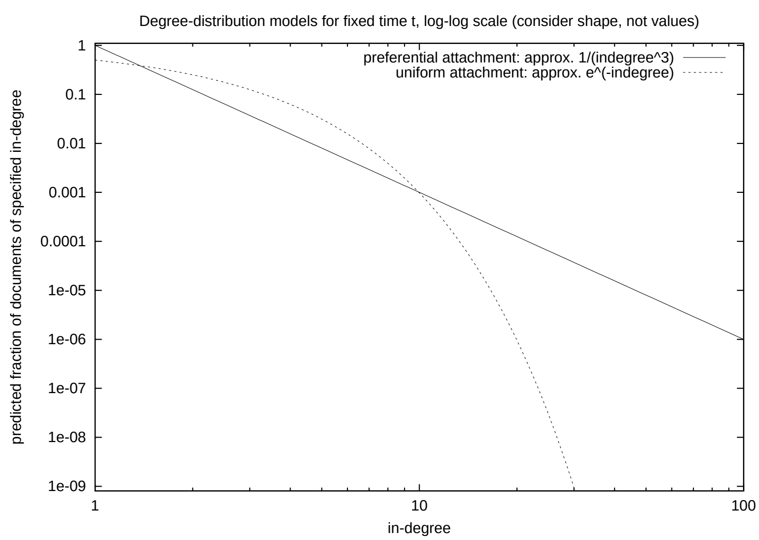
We suppose that we choose to link a new document to document $d_j$ with probability proportional to $I_j(t) + \ell$ (we need the $\ell$ term (or other positive constant) to get the process off the ground). Then we use

$$\frac{dI_j(t)}{dt} = \ell \frac{I_j(t) + \ell}{\sum_{\text{linkable docs } d_k} [I_k(t) + \ell]},$$

to find $I_j(t) = c'(j)\sqrt{2t + n_0 - 2} - \ell$. We solve for $c'(j)$ for $j \geq 1$ in the same way as above to determine the dependence of $I_j(t)$ on $j$.

**Validation of the models**

Here we plot the predicted degree distributions (using calculations based on our computations above) of uniform and preferential attachment at a fixed time, using a log-log scale. We ignored various constants, so focus on the shape of the curves, not the particular values.




But how well do these models do at predicting the large fraction of *communities* that have been observed in the Web [Kumar, Raghavan, Rajagopalan, Tomkins 1999]?

**Copying [Kumar, Raghavan, Rajagopalan, Sivakumar, Tomkins, and Upfal 2000]**

This model involves an extra constant $\beta, 0 < \beta < 1$. Each new document chooses links as follows:

- with probability $\beta$, it chooses $\ell$ pages uniformly at random from the pre-existing documents and links to them.
- with probability $1 - \beta$, it chooses some page $p$ uniformly at random and adds simply copies $p$'s $\ell$ links as its own.