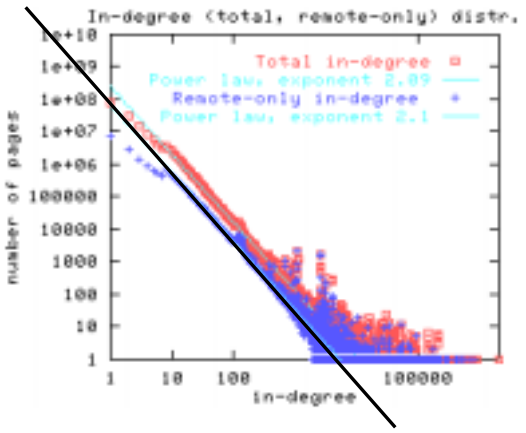
CS/ENGRI 172, Fall 2003: Computation, Information, and Intelligence
10/29/03: Mathematical Models of Link Creation

Power Laws and Web In-degree Distributions

A "power law" is a relationship of the form $y = x^{-\alpha}$, where α 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.

We reproduce here Figure 1 from the Broder et at. (2000) reading, which shows the in-degree distribution, on a log-log scale, for their 200 million document crawl from the Web. The line corresponding to $\alpha = 2.1$ is highlighted.[1]



Setting for Models of Link Creation

- We will 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 $t = 0$.
- At the jth time step ($t = j \geq 1$), we add a new document, named d_j; hence, the subscript indicates when the document was added. We then grant to d_j a constant number l links ($1 \leq l \leq n_0$) to some of the $n_0 + j - 1$ pre-existing documents.
- We will allow repeated links to the same document.

In modeling link creation on the web, we are interested in computing $I_j(t)$, which is our estimate of d_j's in-degree at time t.

Uniform Attachment ("random links") [adapted from Erdös and Rényi (1960)]
The uniform attachment model assumes that when a new page is created, the l links within that page are chosen uniformly at random to the $n_0 + t - 1$ pre-existing documents.
Roughly speaking, we can then assume:

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

[1] The original figure, available in the full copy of the paper at www.www9.org/w9cdrom/160/160.html, is in color, making some of the details easier to view.

so by integration with respect to t we know:

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

and we can compute c(j) by observing that $I_j(j) = 0$ (so that $I_j(t)$ will depend on j).

Preferential Attachment ("rich get richer") [Barabási, Albert, Jeong 1999]
The preferential attachment model assumes that when we create a new page we choose to link to a pre-existing document d_j with probability proportional to $I_j(t) + l$. (We need the l term, or some other positive constant, to get the process off the ground.) We can show that:

$$\frac{dI_j(t)}{dt} = l \frac{I_j(t) + l}{\sum_{\text{all linkable } d_k} I_k(t) + l}$$

Integration on both sides and some calculation will show us that:

$$I_j(t) = c'(j)\sqrt{n_0 + 2t - 2} - l$$

and we can solve for $c'(j)$ as we did above to determine the dependence of $I_j(t)$ on j for $j \geq 1$.

Copying Model [Kumar, Raghavan, Rajagopalan, Sivakumar, Tomkins, and Upfal 2000]
The copying model introduces an extra constant β, $0 < \beta < 1$. Each new document chooses links to pre-existing documents as follows:

- With probability β, it chooses l 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 simply copies p's l links as its own.

Validation of the models
We validate these models by plotting 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 ignore the various constants and focus on the shape of the curves, not the particular values. Recall that we empirically observed a power law relation of in-degrees to number of pages on the web.