

CS599: Structure and Dynamics of Networked Information (Spring 2005)
03/07/2005: Power Law Degree Distributions
Scribes: Viral Shah and Karthik Dantu

1 Introduction

The existence of power law distributions (also known as heavy-tailed distributions) in various natural and man-made scenarios has been demonstrated empirically over the years [6], and attracted a great deal of interest, resulting in models that would naturally predict such distributions. The areas in which power laws have been observed are very diverse, as evidenced by the following, not nearly exhaustive, list:

- Financial Models: price changes of securities in a market [5]
- Biology: lengths of protein sequences in genomes [3]
- Linguistics: word frequencies in the English language [8]
- Internet Topology: node degrees and other parameters of the AS-level topology graph of the Internet [2]
- WWW: Indegrees of web pages
- City populations in the United States
- Number of papers published by scientists [7]
- Distribution of income among people

In particular, the existence of power law degree distributions in the Internet [2] and WWW has caused a flurry of interest within the Computer Science community, resulting in proposals for models that may model the emergence of such distributions.

Definition 1 *A non-negative random variable X is said to have a power law distribution if $\text{Prob}[X \geq x] = cx^{-\alpha}$, for some constants $c, \alpha > 0$.*

If the variable is actually discrete, then the definition implies that $\text{Prob}[X = x] = c'x^{-\alpha'}$, for different values $c', \alpha' > 0$. We will use the definitions interchangeably.

Power law distributions fall into the class of *heavy tailed distributions*: the probability that X assumes a large value is only polynomially small, compared to the exponentially small probabilities for Gaussian, Binomial, or other common distributions.

When given an actual set of data, we can recognize it as a power law most easily in a log-log plot, i.e., in a plot where both axes scale logarithmically. For if $f(x) = c \cdot x^{-\alpha}$ denotes the frequency with which value x was observed, then $\log f(x) = \log(c) - \alpha \cdot \log(x)$. Hence, on a log-log plot, we will observe a straight line with a slope of $-\alpha$ and y-intercept of $\log(c)$.

1.1 Power laws in the WWW

Empirical studies have revealed that the distribution of in(out)-degree d is proportional to $d^{-\alpha}$ for some value $\alpha \in [2.1, 2.2]$. Given the values of α, c , we can calculate the mean of the power law distribution as

$$\mu = \sum_d d \cdot c \cdot d^{-\alpha} = c \sum_d d^{1-\alpha}$$

Notice that the mean is finite iff $\alpha > 2$. In particular, this implies that if the power law indegree distribution were to continue to hold as the WWW grows, the average indegree of pages would remain finite. The fact that the mean indegree of the WWW is finite is not too surprising, given that the average indegree equals the average outdegree, and we don't expect the average web page to have more than a constant number of outlinks, as each requires actual work.

However, there are natural examples of networks that exhibit power laws with $\alpha < 2$, and hence infinite mean:

- The WWW at a site level ($\alpha \approx 1.6$). Here, single sites can include large numbers of web pages, and we may expect the number of pages within a site to grow along with the web.
- Co-authorship in high energy physics ([7], $\alpha \approx 1.2$). High-energy Physics papers often have very large numbers of co-authors (in excess of 1000). It is not clear how to predict the scaling of this graph in the future, but it is not inconceivable that future papers may have even larger author lists. (Notice that by comparison, the mathematics co-authorship graph has a rather large value of $\alpha \approx 2.5$. Many mathematics papers still have a single author.)

2 Preferential Attachment

In trying to explain observed power laws, many models posit a “rich get richer” phenomenon: entities (such as nodes or web pages, plant genera, cities, individuals, etc.) that already have a large value (degree, number of species, populations, wealth etc.) have a tendency to attract more of the same. If this attraction is linear in the current value, then power laws emerge.

In the case of the WWW or other graphs, such a behavior is called *Preferential Attachment*. It is posited that newly arriving nodes link to existing nodes with probability proportional to their current degree. Thus, high-degree nodes are more likely to attract new links. The underlying assumption here is not that nodes make this choice deliberately, but rather that high-degree nodes (e.g., web pages with high indegree) are more likely to be discovered in the first place, and thus linked to. Kumar et al. [4] make this explicit by investigating a copying model, wherein newly arriving nodes randomly select another node, and copy some of that node's outlinks. A mathematically rigorous analysis of such models tends to be quite involved (and is carried out, for instance, in [4]). However, by making some non-rigorous simplifications, we can obtain qualitatively identical results.

Here, we study the following simple model: The graph starts with zero nodes. At each time $t \in \mathbb{N}$, a new node arrives and generates k outlinks; we will label the node with its arrival time t . Each outgoing link is either uniformly random, which happens with probability $1 - \alpha$, or preferential with probability α . In the latter case, the edge links to existing nodes with probabilities proportional to their degrees.

Our analysis will follow the outline from [1]. Let $d_i(t)$ denote the in-degree of node i at time t . Then, because node i arrives at time i with no links yet, we have that $d_i(i) = 0$. At any time t , the total number of inlinks (into all nodes) is kt , and the total number of nodes is t . Hence, the probability that node i is the endpoint of a given new link at time t is $\frac{1-\alpha}{t} + \alpha \frac{d_i(t)}{kt}$. As k new links are created, the expected change in the degree of node i in step t is $\frac{\alpha d_i(t) + k(1-\alpha)}{t}$.

We write $\beta = k(1 - \alpha)$. The difference equation for expected degrees above can be rewritten as a differential equation (here, we are being non-rigorous in our approximation), giving us that

$$\frac{\partial d_i(t)}{\partial t} = \frac{\alpha d_i(t) + \beta}{t}$$

Rearranging and integrating we get,

$$\int \frac{\partial d_i(t)}{\alpha d_i(t) + \beta} = \int \frac{\partial t}{t},$$

which is easily seen to have solution $\frac{1}{\alpha} \ln(\alpha d_i(t) + \beta) = \ln(t) + c$. Solving for $d_i(t)$ now shows that $d_i(t) = \frac{t^\alpha \cdot e^{\alpha c} - \beta}{\alpha}$. We still have to find out the value of the constant c . To determine it, we can use the initial condition that $d_i(i) = 0$. This gives us that $\frac{i^\alpha e^{\alpha c} - \beta}{\alpha} = 0$, which by rearranging yields that $e^{\alpha c} = \beta \cdot i^{-\alpha}$. Substituting this back into the expression for $d_i(t)$ now shows that

$$d_i(t) = \frac{\beta}{\alpha} ((t/i)^\alpha - 1).$$

To find the cumulative function for the number of nodes with degree greater than or equal to d , we first solve the inequality $d_i(t) \leq d$. This yields that the expected degree is at most d whenever $i \geq t \cdot (d \cdot \frac{\beta}{\alpha} + 1)^{-\frac{1}{\alpha}}$. Thus, the fraction of nodes with degree greater than d at time t is

$$\frac{t-i}{t} = \frac{t - t \cdot (d \cdot \frac{\beta}{\alpha} + 1)^{-\frac{1}{\alpha}}}{t} = 1 - (d \cdot \frac{\beta}{\alpha} + 1)^{-\frac{1}{\alpha}}.$$

To obtain the density function from this, we take the derivative with respect to d , which gives us density $\frac{1}{\beta} \cdot \left(d \cdot \frac{\beta}{\alpha} + 1\right)^{-(1+\frac{1}{\alpha})}$.

Thus, we observe a power law with exponent $1 + \frac{1}{\alpha} > 2$. In particular, for $\alpha \approx 0.9$, we obtain the same degree distribution as for the WWW.

Notice that for $\alpha = 0$, the above analysis breaks down (and the result is meaningless). Indeed, for $\alpha = 0$, nodes never attach preferentially, but always choose uniformly at random. Older nodes will still end up with higher degree, as they are around for more rounds, and can thus receive more incoming edges. Notice that each node i , in each round t , receives an expected k/t new incoming links (as there are t competing nodes). Hence, after t rounds, node i will have expected degree $\sum_{j=i+1}^t \frac{k}{j} \approx k \cdot \log(t/i)$ incoming edges. The indices i having degree at most d are thus $i \geq t \cdot e^{-d/k}$, and differentiating the number of vertices of degree at least d with respect to d gives a density of $\frac{t}{k} \cdot e^{-d/k}$. In particular, this implies that the degrees are sharply concentrated. So the age advantage alone does not explain the power law distribution derived above; rather, the preferential attachment was a crucial part of the model.

Notice that the preferential attachment model can be easily extended to include deletion of nodes and edges and rewiring etc., and the same type of analysis based on differential equations can be carried out. Usually, the power law degree distribution is preserved under these modifications, although they usually result in a large number of parameters, whose relative sizes affect the exact value of the exponent.

While preferential attachment gives us the same degree distribution as was observed for the WWW, it fails to obtain several other key properties (among others, all graphs generated are acyclic). For instance, as the links are generated independently at random, the graph will likely not exhibit much community structure (such as triangles or other unusually dense subgraphs). The copying model does slightly better in this respect.

References

- [1] A. Barabasi, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A*, 272:173–187, 1999.
- [2] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proc. ACM SIGCOMM Conference*, pages 251–262, 1999.
- [3] R. Jain and S. Ramakumar. Stochastic dynamics modeling of the protein sequence length distribution in genomes: implications for microbial evolution. *Physica*, 273:476–485, 1999.
- [4] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proc. 41st IEEE Symp. on Foundations of Computer Science*, pages 57–65, 2000.

- [5] B. Mandelbrot. A multi-fractal walk down Wall Street. *Scientific American*, 1999.
- [6] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. In *Allerton Conference*, 2001.
- [7] M. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA*, 98:404–409, 2001.
- [8] G. Zipf. *Selective studies and the principle of relative frequency in language*. Harvard University Press, 1932.