Network Science

Lecture 10: 18 August 2009 Lecturer: Samarth Swarap

Scribe: Chris Kuhlman

Power Laws and Distributions

Part of this lecture comes from Chapter 18 of the text.

10.1 Background

This lecture is about characterizing the structure of a network. For example, one central characteristic of a network is its degree distribution. A network may be static, in which case the degree distribution is constant. However, a network may grow, shrink, or otherwise change in time and accordingly the degree distribution will change. It also addresses different probability distributions for representing different kinds of data, and rules of thumb for their use.

10.2 Some Distributions

Flip a coin 100 times. Count the number of times "heads" appears. This is one trial. Plot the number of heads over, say 200 trials, and one gets a normal distribution. See the web page

`http://en.wikipedia.org/wiki/Normal_distribution`

for a nice introduction. The Normal or Guassian probability density function (PDF) is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{-(x-\mu)^2}{2\sigma^2}) \tag{10.1}$$

Informally, the *Central Limit Theorem* (CLT) states that the sum (or average) of many independent random quantities is well-approximated by a Gaussian distribution.

The degree distribution of the "web graph" (where nodes are web pages and there is an edge from page A to page B if page A has a hyperlink to page B) is a power law, not a normal distribution, so it does not conform to the CLT. Why not? The web graph may not have independent links; e.g., two friends probably have links to each others' pages. Also, a multiplicative representation, rather than a sum, may be appropriate.

The distribution that tends to arise in web graphs is the *Heavy Tail Distribution* (HTD). For a description, please see `http://en.wikipedia.org/wiki/Heavy-tailed_distribution`

Also, see Figure 10.1, showing how the tail of an HTD is larger than that for an exponential distribution. Examples of HTDs include power laws and log-normal distributions.

How does one tell whether one has an HTD? Lets look at some examples. These are first-approximations or rough-cuts. A thorough analysis requires more work.

Power law probability density function: $f(k) = ck^{-\alpha}$, where $f(k)$ is the probability of a node having degree k, c is a constant (fit parameter), and α is a fit parameter. (The $f(k)$ and k may represent other phenomena, too.) A power law will plot as a straight line on log-log paper because taking logarithms of the original equation gives a form $y = mx + b$ where $y = \log(f(k))$, $x = \log(k)$, $m = -\alpha$, and $b = \log(c)$.

Experimental studies often indicate that a web graph (based on a part of the web) is often given by $f(k) \approx ck^{-(2+\delta)}$, where δ is a small number $\ll 1$.

Exponential probability density function: $f(k) = c\exp(-\alpha k)$, where we have the variables as above. Taking logarithms again, we obtain $\log(f(k)) = \log(c) - \alpha k$, which is a straight line on semi-log paper: $\log(f(k))$ vs. k.

Lets look at a web graph where $k = 1000$ edges. A power law will give $f(k) \propto k^{-2} = 10^{-6}$. An exponential, in contrast, will give, say $f(k) \propto 2^{-1000} \approx 10^{-300}$. Hence, for reasonable constants c, there is a much fatter tail on the power law than on the exponential.

10.3 Models of Popularity

These models are often of the power law form: $f(k) = ck^{-\alpha}$

The idea is that there is a feedback effect in popularity: if you are popular, you get more popular. This leads to the "rich get richer" model.

10.3.1 Rich-Get-Richer Model

Components of a web graph model follow. Nodes are web pages and edges $\langle k, l\rangle$ are directed, from node k to node l. This edge means that page k contains a hyperlink to page l. In this simple model, we assume that each node has out-degree of exactly 1. A node can have any value of in-degree.

1. pages, which are network nodes, are created in order, and named in order: $1, 2, 3, \ldots, N$.
2. page j links to an earlier web page i (i.e., an edge is formed from j to i) as follows:
   (a) with probability p, j links to i chosen uniformly randomly.
   (b) with probability $(1 - p)$, j chooses i uniformly randomly, and links to the page i points to (i.e., this is a copy decision).

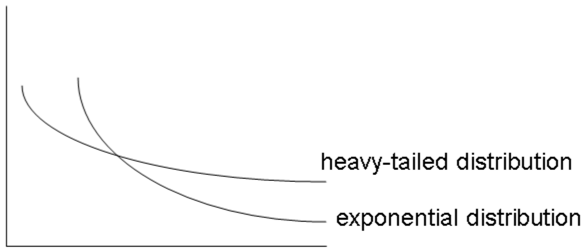

Figure 10.1: Comparison of a heavy tailed distribution and an exponential distribution.

Item 2b above is equivalent to the following: with probability $(1 - p)$, j links to l with probability $\propto$ (l's in-degree)/(total degree of network).

An example of item 2b is the following. Suppose $\langle 17, 14 \rangle$: i.e., page 17 has a hyperlink to page 14. Now we introduce node 18. For node 18, suppose we randomly choose node 17. Since node 17 points to node 14, we form edge $\langle 18, 14 \rangle$.

The model has the following features:

1. If $p = 1$, we get a growing random graph (though not Erdős-Rényi because of the outdegree=1 restriction).
2. If $p = 0$, we get a power law degree distribution. Incoming nodes would never have in-degree> 0 because their in-degree is identically zero initially, and hence has zero chance of increasing. So item 2a above allows the chance for incoming nodes to have in-degree> 0.

Question: what is the expected degree distribution for this web graph model?

Let $X_j(t)$ be the in-degree of node j, $indeg(j)$, at time $t(> j)$. $t > j$ because node j is introduced at time j, per the construction above, and so node j cannot have an edge point to it until time $t \geq (j + 1)$.

1. Initial condition: $X_j(t = j) = 0$

2. Expected change in X_j over time:
   (a) with probability p, node $(t+1)$ chooses j uniformly random.
   (b) with probability $(1-p)$, node $(t+1)$ chooses j with probability $X_j(t)/t$.

We have $Pr[\text{node } (t+1) \text{ links to } j] = p/t + (1-p)X_j(t)/t$, where the 2 contributions are from items 2a and 2b immediately above. Note that this is a discrete time model and that t is time and the total number of nodes.

Now, we turn to a deterministic continuous approximation model to get some insight.

10.3.2 Deterministic Continuous Approximation Model

Approximate $X_j(t)$ by the continuous function $x_j(t)$.

1. Initial condition: $x_j(t = j) = 0$
2. Growth equation (i.e., expected change in x_j over time): $dx_j/dt = p/t + (1-p)X_j(t)/t$, where the RHS contains the same 2 terms as above, consistent with the 2 ways for node j to increase its in-degree.

Look at the growth equation and integrate it. We have

$$\begin{aligned}
\int \frac{dx_j}{p + (1-p)x_j} &= \int \frac{dt}{t} && (10.2)\\
\ln(p + (1-p)x_j) &= (1-p)\ln(t) + \ln(A) && (10.3)\\
p + (1-p)x_j &= At^{(1-p)} && (10.4)\\
x_j(t) &= \frac{At^{(1-p)} - p}{(1-p)} && (10.5)
\end{aligned}$$

Using the initial condition (see above),

$$x_j(j) = \frac{Aj^{(1-p)} - p}{(1-p)} = 0. \tag{10.6}$$

Replacing A gives

$$x_j(t) = \frac{\frac{p}{j^{(1-p)}}t^{(1-p)} - p}{(1-p)} = \frac{p}{(1-p)}[(\frac{t}{j})^{(1-p)} - 1] \tag{10.7}$$

This last equation is the number of links into node j (i.e., $indeg(x_j)$) at time t, where $t > j$.

But we want the degree distribution.

Question: How many nodes have at least k links at time t?

$$x_j(t) = \frac{p}{(1-p)}[(\frac{t}{j})^{(1-p)} - 1] \geq k \tag{10.8}$$

or

$$t(k\frac{1-p}{p}+1)^{\frac{-1}{(1-p)}} \geq j \tag{10.9}$$

where

$$(k\frac{1-p}{p}+1)^{\frac{-1}{(1-p)}} \tag{10.10}$$

is the CDF (cumulative density function), and also the fraction of t nodes that will have at least k links.

Now, we have the CDF, but we want the PDF (probability density function). (Note: $cdf(x) = \int_{-\infty}^{x} pdf(x)$.) So take the derivative with respect to k, and negate the result because Equation (10.8) has $\geq k$ and we want $\leq k$:

$$pdf = \frac{1}{(1-p)}[k\frac{(1-p)}{p}+1]^{(\frac{-1}{(1-p)}-1)}\frac{(1-p)}{p} \tag{10.11}$$

This last equation is the equation of the degree distribution.

Recall the power law density function: $f(k) = ck^{-\alpha}$. So from the preceding equation, $\alpha = 1 + 1/(1-p)$. Since $1-p < 1$, $1/(1-p) > 1$, so $\alpha > 2$. As $p \to 1$, $\alpha \to \infty$, and as $p \to 0$, $\alpha \to 2$. So a web graph has a high level of feedback (i.e., a large exponent in absolute value), and there is some connection back to experimental studies that find the exponent of web graphs to be about 2, as mentioned above.

This model will produce only trees (and forests); think: each node has out-degree of 1.

10.4 Models with Independent Multiplicative Effects

The model above has additive effects but does not satisfy the independence assumption of the CLT, and therefore gives a power law instead of a Normal distribution.

Question: What if a model satisfies the independence assumption but violates the additiveness assumption? In particular, what if the effects are multiplicative?

We (typically) get a log-normal distribution. X has a log-normal distribution if $\ln(X)$ has a normal distribution. The PDF is given by

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}}\exp[\frac{-(\ln(x)-\mu)^2}{2\sigma^2}] \tag{10.12}$$

See wiki web page

`http://en.wikipedia.org/wiki/Log-normal_distribution`

for details of the log-normal distribution.

Just as the CLT has a practical interpretation with respect to normal distributions, the log-normal distribution has an interpretation: the product of many independent random quantities is well-approximated by the log-normal distribution.

Exercise Derive Equation (10.12) from Equation (10.1).

Hints for this exercise are as follows. Let $Y = g(X)$, where g is a monotonically increasing function and X, Y are random variables. Hence $X = g^{-1}(Y)$. Now, the CDF of Y is $F_Y(y) = Pr[Y \leq y]$ and we have

$$\begin{aligned} F_Y(y) &= Pr[Y \leq y] \\ &= Pr[g(X) \leq g(x)] \\ &= Pr[X \leq x] \\ &= F_X(x) \end{aligned} \tag{10.13}$$

where the second line follows from $Y = g(X)$, the third line follows from the monotonicity of g, and the fourth line follows by definition. Also

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}(F_Y(y)) \\ &= \frac{d}{dy}(F_X(x)) \\ &= f_X(x)\frac{dx}{dy} \end{aligned} \tag{10.14}$$

The rest is left to the student.

Back to lecture. Take natural logarithms of each side of Equation (10.12) to get

$$\begin{aligned} \ln(f_X(x)) &= -\ln(x) - \ln(\sigma\sqrt{2\pi}) - \frac{(\ln(x) - \mu)^2}{2\sigma^2} \\ &= -\frac{\ln(x)^2}{2\sigma^2} + (\frac{\mu}{\sigma^2} - 1)\ln(x) - \ln(\sqrt{2\pi}\sigma) - \frac{\mu^2}{2\sigma^2} \end{aligned} \tag{10.15}$$

If σ^2 is sufficiently large, then the distribution will show up as a straight line on a log-log plot for a range of x values. However, eventually the square term will dominate.

Similar effects are also produced by power laws with an exponential cutoff, which are functions of the form: $k^{-\alpha}e^{-k}$. This last equation looks something like Figure 10.2 below. Thus it can be quite hard to tell from looking at data whether a power law or a log-normal will provide the best description. The relative merits of the two distributions have been under debate for many decades now. Where does this log-normal distribution arise?

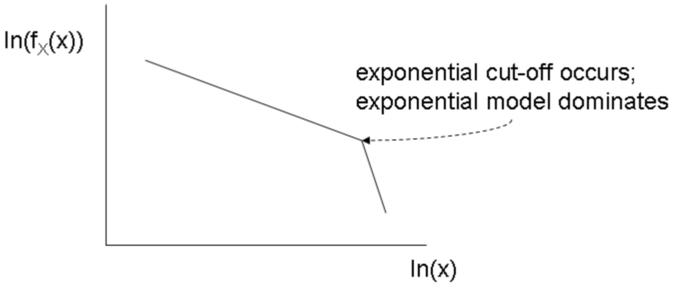

Figure 10.2: Plot of $f_X(x)$ versus $\ln(x)$ showing how a power law with an exponential cutoff exhibits a sharp drop at large $\ln(x)$.

Number of web pages is one example. Let $n_s(t)$ be the number of web pages on day t. We use the model (due to Huberman and Adamic)

$$\begin{aligned} n_s(t+1) &= n_s(t) + g(t+1)n_s(t) \\ &= [1 + g(t+1)]n_s(t) \end{aligned} \tag{10.16}$$

Note this is now a multiplicative model since $n_s(t+1)$ is obtained from a product involving $n_s(t)$. Previously, in contrast, we summed coin flips.

In the previous equation, $g(t)$ is the growth rate; it fluctuates about the mean g_0 in an uncorrelated fashion. So the model is multiplicative and we write $g(t) = g_0 + \xi(t)$, where the fluctuation is the second term. From this, one may derive the following log-normal distribution

$$Pr(n_s) = \frac{1}{n_s\sqrt{t}\sqrt{2\pi\sigma^2}} \exp(\frac{-(\ln(n_s) - g_0 t)^2}{2\sigma^2 t}) \tag{10.17}$$

10.4.1 Pareto Distribution

This distribution has been around for over 100 years, originally developed to look at wealth distribution, but in 1953 Champernowne resuscitated it.

CDF: $Pr[X \geq x] = (x/k)^{-\alpha}$.

PDF: $f_X(x) = \alpha k^{\alpha} x^{-\alpha-1}$.

Notice that the Pareto distribution is a power law. Pareto introduced it to model income distributions but did not specify a generative model. Champernowne suggested the following generative model. We specify a minimum income, $m > 0$. We specify income ranges as a function of some constant γ: $(m, \gamma m), (\gamma m, \gamma^2 m), (\gamma^2 m, \gamma^3 m), \ldots, (\gamma^{j-1} m, \gamma^j m), \ldots$. Define $p_{ij} = f(i, j)$ as the probability of moving from income range i to income range j, i.e. the probability of changing from i to j depends only on i and j.

What is the equilibrium income distribution?

Example:

$$p_{ij} = \begin{cases} 1/3, & \text{when } i = j - 1 \\ 2/3, & \text{when } i = j + 1 \end{cases} \tag{10.18}$$

and

$$p_{11} = 2/3 \tag{10.19}$$

Notice that to change class, income must change multiplicatively. So we would guess a log-normal distribution as the equilibrium distribution, based on our earlier discussion. Champernowne shows that the distribution is not log-normal; rather, it is Pareto. So, why is it not log-normal? The answer is because we have a minimum income, $m > 0$. If we have no minimum income; i.e., no artificial minimum so that $m = 0$, then we get a log-normal distribution.

The takeaways are these:

1. small changes in details of models make a difference.
2. models are closely related.

10.5 The Long Tail

See article by Chris Anderson, "The Long Tail," 2004.

In Figure 10.3, we are looking at the relationship between a book's popularity and sales volume. The question here is how do the ideas of dominance and niche relate with respect to Google's or Amazon's search results? One the one hand, Google reinforces the top sites for each query because many people do not look beyond the first page of 10 hits. This supports the notion that (in this case) dominant books get even more exposure. On the other hand, Google provides many pages for each query, and so even less-popular hits get displayed—just not on the first page. But for those willing to look beyond the first Google page, information on less-popular or niche items can be obtained. The second point—the ability to find obscure results—is also facilitated by web searches. Retailers like Amazon try to shift the curves in Figure 10.3 to the right, to increase the market for non-dominant books.

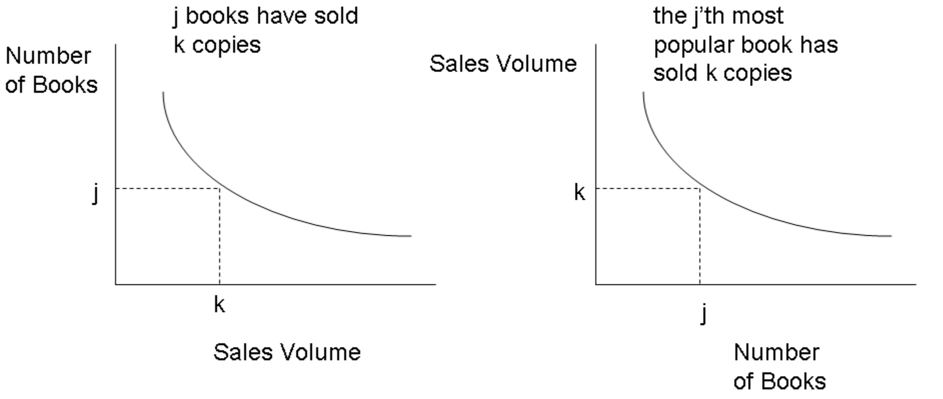

Figure 10.3: (a) Function showing how many books j have sold k copies each; and (b) function showing how many copies k that the j'th most popular book has sold.