Graphs and Networks Lecture 7

Power-Law Degree Distributions

Daniel A. Spielman September 23, 2010

7.1 Overview

We will examine models of graphs that exhibit degree distributions that satisfy power laws. These are inspired by the models in [BA99, KRR$^+$00]. We will see:

1. Simulations results.
2. The method of heuristic analysis from [BA99].
3. A heuristic treatment, from [Mit03], of the analysis of [KRR$^+$00]. This may be made rigorous using the Martingale technique of Wormald [Wor95].

7.2 Preferential Attachment Models

The preferential attachment models start with a small graph, say with just one vertex and maybe a self-loop. As each time step, a new vertex is added to the graph. The endpoints of edges from this new vertex are biased towards other vertices of higher degree. The model proposed in [BA99] required each new vertex to have a fixed degree, and required that the endpoints of its edges be distributed proportionally to the degrees of existing vertices. If the degree of the new vertex is greater than 1, care needs to be taken to make this model more precise [BR03].

Kumar *et. al.* were interested in the web, so they used directed graphs. They imagined that each new vertex would choose its links uniformly with some probability and as copies of the links of some other node otherwise. For simplicity, we will consider this model in the case that each new vertex has out-degree 1.

I can now state exactly the model we will examine. We begin with one node, which contains a directed edge to itself. This is the only self-loop we will include in the construction. We then choose some probability p. We then add vertices one-by-one, creating one edge leaving each. When we add vertex $t+1$, we do one of two things. With probability p we choose the endpoint of the edge uniformly from the t existing vertices. With probability $1-p$, we choose a random edge already in the graph, and use its endpoint.

7.3 Simulation

I'll show a simulation of this for $p = .5$ and $n = 10^6$. We will see both plots of the number of vertices of each degree and the number of vertices exceeding each degree (the cumulative degree distribution). In a log-log scale these are pretty straight, suggesting that the distribution may satisfy a power law. That is, the number of vertices of degree k will be proportional to $k^{-\alpha}$ for some exponent α.

7.4 A first analysis

We will now present a heuristic justification of this degree distribution. Number the vertices $1, \ldots, n$ according to the order in which they are added to the graph.

Let $X_j(t)$ denote the degree of node j after t vertices have been added. When we add vertex $t+1$, it could increase the degree of node j in one of two ways. With probability p vertex $t+1$ chooses its endpoint uniformly at random, in which case it hits vertex j with probability $1/t$. With probability $1-p$ it chooses vertex j with probability proportional to $X_j(t)$. To see what this probability is, note that t edges already exist in the graph, so the sum of the in-degrees of the verticies is t. This means that vertex j is chosen as an endpoint with probability j/t. In total, the probability that we increment X_j is

$$\frac{p}{t} + \frac{(1-p)X_j(t)}{t}. \tag{7.1}$$

This means that the expected increase in X_j is given by (7.1). In our heuristic analysis, we will try to measure the expectation of X_j by assuming that it receives this expected increase. That is, we set

$$Y_j(t+1) = \begin{cases} 0 & \text{for } j \leq t \\ Y_j(t) + \frac{1}{t}\left(p + (1-p)Y_j(t)\right) & \text{for } j > t. \end{cases}$$

There are two issues with this approximation: it is not clear that $\mathbf{E}\left[X_j(t)\right] = Y_j(t)$ and it is not clear that $X_j(t)$ satisfies any sort of concentration. But, let's carry through and see what we get.

Solving recurrences like the one for $Y_j(t)$ can be tricky. We will use a heuristic to guess the solution. By substitution, one could then confirm that the solution is approximately correct. The heuristic is to treat this as a differential equation. We approximate $Y_j(t)$ by a continuous function of t, which we call $Z_j(t)$. This function should then satisfy

$$\frac{d\, Z_j}{d\, t} = \frac{1}{t}\left(p + (1-p)Z_j\right)$$

We now turn this into a separable equation and solve by integrating:

$$\begin{aligned}\frac{d\ Z_j}{p+(1-p)Z_j} &= \frac{d\ t}{t}\\ \int \frac{d\ Z_j}{p+(1-p)Z_j} &= \int \frac{d\ t}{t}\\ \frac{1}{1-p}\ln\left(p+(1-p)Z_j\right) &= \ln(t)+C\\ \ln\left(p+(1-p)Z_j\right) &= (1-p)\ln(t)+C.\end{aligned}$$

Exponentiating both sides and allowing C to change to some other constant gives

$$\begin{aligned}p+(1-p)Z_j(t) &= t^{1-p}C\\ Z_j(t) &= \frac{1}{1-p}\left(t^{1-p}C-p\right).\end{aligned}$$

So find C, we plug in the boundary condition $Z_j(j)=0$. This gives

$$C=\frac{p}{j^{1-p}}.$$

Substituing in this value for C, we get

$$Z_j(t)=\frac{p}{1-p}\left((t/j)^{1-p}-1\right).$$

If we believe our approximation, then at the end of this process the nodes of degree at least d should be those for which

$$Z_j(n)=\frac{p}{1-p}\left((n/j)^{1-p}-1\right)\geq d.$$

Re-arranging, we see that this happens if

$$\frac{n}{j}\geq\left(\frac{1-p}{p}d+1\right)^{1/(1-p)}\approx\left(\frac{1-p}{p}\right)^{1/(1-p)}d^{1/(1-p)}.$$

This means that node j should have degree at least d if

$$j\leq n\left(\frac{p}{1-p}\right)^{1/(1-p)}d^{-1/(1-p)}.$$

This would be a power law.

If we now compare to our experimental data, we see that it fits it pretty well for d larger than some reasonable number (like 10). Well, it seems to be off by some constant factor. But, the power-law description fits.

On the other hand, the actual predictions of degrees upon which this was based are way off. To see this, look at a plot of the degrees of nodes 1 through 100. You can see that they are all over the place.

This could just be a failure of concentration in the degrees of high degree nodes. If our expectation calculation is correct, then it should probably hold for the low-degree vertices. After all, these are the vertices that show up late in the process. The expected degree of nodes t and $t+1$ should be similar. Moreover, the degrees of these nodes are barely correlated. So, for the degrees that should occur often it is reasonable expect that the number of nodes of those degrees would be concentrated.

7.5 The number of low-degree nodes

We will now go through an analysis that can be made rigorous. This analysis will count the number of nodes of each low in-degree. For low in-degrees, this can be made rigorous. To do it correctly requires more probability that I want to teach right now. But, I'll at least sketch how the argument goes.

First, let's figure out the expected number of nodes of degree 0. A node only goes from degree 0 to degree 1 when it is chosen as a uniform random neighbor of some node that comes along later. If node j has degree 0 when node $t+1$ is added, then the chance that node $t+1$ links to node j is p/t. So, the chance that node j has degree 0 after n nodes have been added is

$$\prod_{t=j+1}^{n} \left(1 - \frac{p}{t-1}\right)$$

For j bigger than $\sqrt{n}$ this is well-approximated by

$$\prod_{t=j+1}^{n} \exp\left(-\frac{p}{t-1}\right) = \exp\left(-p \sum_{t=j}^{n-1} \frac{1}{t}\right).$$

To compute this to first order, we recall a theorem of Euler's which says that

$$\sum_{i=1}^{k} \frac{1}{i} \rightarrow \ln(k) + \gamma,$$

where γ is some absolute constant. So,

$$\sum_{t=j}^{n-1} \frac{1}{t} \approx \ln(n-1) - \ln(j-1) \approx \ln(n/j).$$

This tells us that the probability that node j has degree 0 is approximately

$$\left(\frac{j}{n}\right)^{p}.$$

So, the number of nodes having degree zero should be approximately

$$\sum_{j=1}^{n} \left(\frac{j}{n}\right)^{p} = n^{-p} \sum_{j=1}^{n} (j)^{p} \approx n^{-p} \frac{1}{p+1} n^{1+p} = n \frac{1}{p+1}.$$

This agrees very well with our experimental data.

Indeed, one can show that this quantity is well-concentrated. I know of two ways to do it. The easiest is to show that the events that different nodes have degree 0 are anti-correlated. That is, if one node has degree 0 then it is only less likely that another does. In this case, variables are known to have better concentration then predicted by the Chernoff bounds. The other way uses Martingales, and can be extended to handle the case of other constant degrees.

Warning: the variables in this section have different meanings than in previous sections.

We will again use a differential equation approximation to handle low degrees. Let's see how it would work for the degree 0 nodes. Let $X_0(t)$ be the number of nodes of degree 0 after t have been added. When we add node $t+1$, it will have degree 0. Thus, the number of degree 0 nodes will increase unless that node points to another of degree 0, which happens with probability $pX_0(t)/t$. So,

$$X_0(t+1) = \begin{cases} X_0(t) & \text{with probability } pX_0(t)/t, \text{ and} \\ X_0(t)+1 & \text{with probability } 1 - pX_0(t)/t. \end{cases}$$

As before, define Y_0 to be an approximation of the expectation of X_0 by setting

$$Y_0(t+1) = Y_0(t) + 1 - pX_0(t)/t.$$

We hope that as t becomes larger Y_0 approaches $c_0 t$ for some constant c_0. In this case, c_0 should satisfy

$$c_0 = 1 - pc_0 t/t = 1 - pc_0.$$

This would give

$$c_0 = \frac{1}{1+p},$$

the bound we obtained before.

For higher degrees, let $X_k(t)$ be the number of node of degree k after t have been added. These variables can either increase or decrease. X_k decreases if the edge from node $t+1$ hits a node of degree k, which happens with probability

$$\frac{pX_k}{t} + \frac{(1-p)kX_k}{t}.$$

Similarly, X_k increases if the edge hits a node of degree $k-1$, which happens with probability

$$\frac{pX_{k-1}}{t} + \frac{(1-p)(k-1)X_{k-1}}{t}.$$

So, the expectation of $X_k(t+1) - X_k$ is

$$\frac{1}{t}\left(pX_{k-1} + (1-p)(k-1)X_{k-1} - pX_k - (1-p)kX_k\right).$$

Setting $Y_k(t)$ to be our guess for the expectation of $X_k(t)$, we get

$$Y_k(t+1) = Y_k(t) + \frac{1}{t}\left(pX_{k-1} + (1-p)(k-1)X_{k-1} - pX_k - (1-p)kX_k\right).$$

If we look for a solution to these equations of the form

$$Y_k = c_k t,$$

we get

$$c_k = pc_{k-1} + (1-p)(k-1)c_{k-1} - pc_k - (1-p)kc_k.$$

Re-arranging, we get

$$\frac{c_k}{c_{k-1}} = \frac{p + (1-p)(k-1)}{1 + p + (1-p)k} = 1 - \frac{2-p}{(k+1) - p(k-1)}.$$

We should now check that these estimates agree extremely well with our experimental data.

We can also see that they give a power-law distribution. As k grows large, these give

$$\frac{c_k}{c_{k-1}} \approx 1 - \frac{2-p}{k(1-p)},$$

and so

$$\frac{c_k}{c_j} \approx \prod_{i=j+1}^{k} 1 - \frac{2-p}{1-p}\frac{1}{i} \approx \left(\frac{j}{k}\right)^{\frac{2-p}{1-p}} = \left(\frac{j}{k}\right)^{1+\frac{1}{1-p}}.$$

This is equivalent to saying that for large k,

$$c_k \approx \beta \left(\frac{1}{k}\right)^{1+\frac{1}{1-p}},$$

for some constant β.

7.6 Making That Rigorous

I hope to say a few words about how this can be made rigorous, at least for k a slowly growing function of n.

7.7 Conclusion

This model does have both a good story and its does produce power-law distributions. But, do the graphs it produces resemble the graphs we see in the real world? There are many ways in which they differ, and many ways in which people have corrected the models to fix these differences. The difference that I find most profound is that real-world graphs are very far from being expanders. We now know that most real-world graphs have reasonably large sets of vertices (of sizes 100 to 1000) that are poorly connected to the rest of the graph. Very few models produce this.

References

[BA99] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[BR03] Béla Bollobás and Oliver M. Riordan. Mathematical results on scale-free random graphs. In Stefan Bornholdt and Heinz Georg Schuster, editors, *Handbook of Graphs and Networks: From the Genome to the Internet*, pages 1–34, Weinheim, 2003. Wiley-VCH Verlag.

[KRR+00] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. *Foundations of Computer Science, Annual IEEE Symposium on*, 0:57, 2000.

[Mit03] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2003.

[Wor95] Nicholas C. Wormald. Differential equations for random processes and random graphs. *The Annals of Applied Probability*, 5(4):pp. 1217–1235, 1995.