

Lecture 4 (October 19, 2005)
prepared by Paul Constantine

1. Modeling the structure of the Internet

By now you all know about the power law relationship in the degrees of the nodes of the Internet. But is that the whole story? What other interesting and relevant questions can we ask regarding the structure of these natural networks? For example, all of our analysis so far has concerned undirected graphs, but the link structure of the World Wide Web is a directed graph. How can we model these directed graphs? Some work has been done on directed scale free graphs by Bollobas, Borgs, and Chase.

Another interesting question arises when considering the persistence of nodes in such networks. In our preferential attachment model, once a node arrives it persists for all of eternity. Clearly this is not the case for any natural graphs we wish to study; web pages leave the WWW all the time. So how can we account for the deletion of nodes in such graphs?

We can also question the random surfer model. Does the real average user of the World Wide Web randomly click on links starting at some randomly selected starting page? A better model might postulate a purpose for information collection, and something like this would invariably consider the results from a search engine. Then the question becomes, how does the search engine affect the popularity of a webpage? Suppose that a particular page always appears among the top results of some query. Then the popularity of that page is self-supportive; a user follows the link to a page because the search engine recommended it and search engine recommends it because users follow the link to that page. So how would a new page with superior content fare among the results? When and how will the search engine recognize the superiority of the new page given that the average user has no knowledge of its existence?

One question that is not addressed by the B-A model is the question of locality or similarity among the nodes. For example, when a new page arrives to connect to the World Wide Web, it will likely link to pages with similar content. This behavior could reasonably create content specific clusters on the web. How might we model this phenomenon? One approach is copying model (*See Durrett's book, Chapter 4*): At each time step, some node is cloned and each of its links are altered with some probability. Clearly this approach will model clustering better than the preferential attachment model. Another appealing feature of the copying model is that it will represent geographic locality among the nodes. When a page arrives at the web, it must connect through some specific ISP, and it may choose its ISP

based on geographic considerations. The copying model represents this behavior nicely. The motivation for the copying model came from biological networks. This sort of “copying” is a natural abstraction of genetic mutations in species; a new species is created by copying the genetic material in an existing species with some probability of modifications.

Another property of these natural scale free graphs that is difficult to model is their self-similarity. Many of the characteristics of the entire graph show up in any smaller subgraph that one may choose to examine: power-law degree distribution, clustering, one large connected component, etc. Self-similarity is rigorously defined in the context of fractals and dynamical systems, but these rigorous definitions have not been precisely applied to networks. We only have inklings of these concepts in networks, so there is still much work to be done here.

Still none of these models consider the role of incentives in the creation of these graphs. When someone creates a new webpage, he or she does not flip a coin. There are many factors to consider when choosing an ISP including sales, customer service, bandwidth, storage space, etc. How can we include these parameters in the model? How do the economic preferences of each of the nodes affect the structure of the resulting network? Some of these parameters may also be conflicting. Suppose you consider two different metrics for choosing an ISP: geographic distance and average distance of the potential parent node to the rest of the nodes in the network (centrality). Papadimitriou and others have showed that these conflicting interests can create heavy tailed distributions in the degrees of the nodes similar to a power law. One might postulate that a power-law distribution is the result of long-term ambivalence.

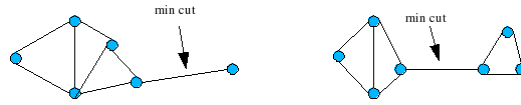
Perhaps a good model for incentives will capture the self-similarity of these networks. For example, a small ISP serving some local geographic region may connect to a larger ISP. Then the subgraph of users connected to the smaller ISP may resemble the larger graph of the ISP’s themselves.

We conclude this section with a well-placed transition question: Why do we care at all about modeling these types of networks? Why do we concern ourselves with accurately simulating the characteristics we observe in these natural graphs? Ultimately our goal is to simulate these networks in order to evaluate the performance of algorithms for tasks such as searching and routing. We also might be interested in questions of e-commerce and online advertising. In order to make predictions about the performance of new algorithms, we need an accurate simulations - a realistic playground.

2. Expander Graphs

Given a network we may ask how reliable it is, or how easy it is to disconnect it. Of course you could examine the size of the minimum cut in graph; this would be one measure of reliability. But there may be two graphs with the same minimum cut in which one cut

disconnects only a single vertex from the rest of the network but in the other network can be split into two pieces of almost equal size.



In many situations the second cut is far more damaging to the performance and functionality of the network.

Therefore we examine a more appropriate measure. Given graph $G(V, E)$ and some nonempty subset $S \subset V$, the cut ratio of S is defined as

$$\rho(S) = \frac{C(S, \bar{S})}{\min(|S|, |\bar{S}|)}$$

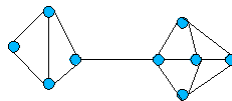
where \bar{S} is the complement of S and $C(S, \bar{S})$ is the number of edges between the sets S and \bar{S} . Then we define the cut ratio of the graph as

$$\rho(G) = \min_{\emptyset \subset S \subset V} \rho(S)$$

For example, suppose K_n is a complete graph on n vertices. Then

$$\max \rho(K_n) = \frac{n}{2}$$

As another example, consider the dumbbell graph on n vertices, D_n :



$$\min \rho(D_n) = \frac{2}{n}$$

Trivially if a graph G is disconnected, $\rho(G) = 0$.

If we restrict the edges of the graph $G(V, E)$ to be linear in n , i.e.

$$|E| = c|V|.$$

for some constant c . Then the average degree of the nodes of G is $2c$. Therefore there is a vertex with degree at most $2c$. In this case there is an S such that $\rho(S) = 2c \Rightarrow \rho(G) \leq 2c$.

Motivated by a quest for a cheap and reliable network, the following question arises: Is it possible to construct a graph with number of edges linear in the number of vertices such that for any subset S , the number of edges going out of S , is $\alpha|S|$? Can we construct a

graph with constant expansion and constant maximum degree? Such a graph can construct a network that is robust against the failure (removal) of the links. We will also show that routing with low-congestion and certain types of search algorithms will perform very well on these graphs.

There is a relatively simple randomized construction for generating constant degree expanders. In fact, we know that for $d \geq 3$, with high probability, a random d -regular graph is a constant expander.

Theorem 4.1 *For all $d \geq 3$ and n sufficiently large, there exists an $\alpha > 0$ such that any random d -regular graph of size n has expansion α .*

Proof: We will prove this theorem for d sufficiently large. Remember how we generate random regular graphs: we take dn minivertices and separate into sets of size d . Create random matchings between these sets of minivertices, then collapse the matchings into a graph of size n .

What does it mean for this graph to have expansion less than α ? We should have a subset S of vertices with $|S| = k < \frac{n}{2}$ such that $|C(S, \bar{S})| < \alpha k$. For that we need $dk - \alpha k$ of the minivertices associated to the vertices in S to be matched to each other. We will show that the probability of this event is very small.

Let's count the possible choices for S :

$$\binom{n}{k}$$

and the minivertices in S and \bar{S} that are connected to each other:

$$\binom{dk}{\alpha k} \binom{dn - dk}{\alpha k}$$

Now we count the ways the $dk - \alpha k$ minivertices of interest may attach to each other. Define $f(m)$ to be the number of matchings between m vertices:

$$f(m) = \frac{\binom{m}{2} \binom{m-2}{2} \cdots \binom{2}{2}}{(m/2)!} = \frac{m!}{(2^{m/2}(m/2)!)}$$

Then the probability of the event that only a certain αk of minivertices match outside of their proper subset is at most

$$\frac{f(dk - \alpha k) f(dn - dk - \alpha k) f(2\alpha k)}{f(dn)}$$

Therefore we have, the probability that there is a subset S with expansion at most α is

$$\sum_{k=1}^{n/2} \alpha k \binom{n}{k} \binom{dk}{\alpha k} \binom{dn-dk}{\alpha k} \frac{f(dk-\alpha k)f(2\alpha k)f(dn-dk-\alpha k)}{f(dn)}$$

In order to bound the above equation, we will use Stirling's inequality

$$c_1 n^{n+(1/2)} e^{-n} \leq n! \leq c_2 n^{n+(1/2)} e^{-n}$$

for some constants c_1 and c_2 .¹ We apply Stirling's inequality to create the following handy inequalities:

$$\binom{n}{k}^k \leq \binom{n}{k} \leq \xi \left(\frac{ne}{k}\right)^k$$

The same bound also shows that:

$$f(m) = c \frac{m^{m+(1/2)} e^{-m}}{2^{m/2} (m/2)^{(m/2)+(1/2)} e^{-(m/2)}} \leq \frac{cm^{m/2}}{e^{m/2}}.$$

By plugging in the above bounds we have:

$$\begin{aligned} & \frac{f(dk-\alpha k)f(2\alpha k)f(dn-dk-\alpha k)}{f(dn)} \\ & \leq \sum_{k=1}^{(n/2)} \alpha k \left(\frac{ne}{k}\right)^k \left(\frac{edn}{dk}\right)^{(2\alpha k)} \frac{(dk-\alpha k)^{(dk-\alpha k/2)} (2\alpha k)^{(\alpha k)} (dn-dk-\alpha k)^{(dn-dk-\alpha k/2)}}{dn^{(dn/2)}} \\ & \leq \sum_{k=1}^{(n/2)} \alpha k \left(\frac{nec}{k}\right)^{(k+2\alpha k)} \left(\frac{dk}{dn}\right)^{((dk-\alpha k)/2)} \\ & \leq \sum_{k=1}^{(n/2)} \alpha k \left(\frac{k}{n}\right)^{((d-2)k-5\alpha k)/2} (ec)^{k+2\alpha k} \quad \text{for } \alpha \text{ sufficiently small} \\ & \leq \sum_{k=1}^{(n/2)} \left(\frac{k}{n}\right)^{((d-3)k)/2} (ec)^{(2k)} \quad \text{for sufficiently large } d \end{aligned}$$

Note that $k/n < 1/2$ so $(k/n)^{2c} \leq (ec)^{-2}$. So d "sufficiently large" means d is large enough to keep a power of 2 so we can bound the above by

$$\leq \sum_{k=1}^{(n/2)} \left(\frac{k}{n}\right)^{2k} \alpha k$$

¹In many situations, the weaker bound of $n^{n/2} \leq n! \leq n^n$ suffices. However, we will need a stronger bound here.

$$\begin{aligned}
&= \alpha \left(\frac{1}{n}\right)^2 + 2\alpha \left(\frac{2}{n}\right)^4 + \dots + \left(\frac{(n/2)}{n}\right)^{(n/2)} \frac{n\alpha}{2} \\
&\leq \alpha \left(\frac{1}{n}\right)^2 + 2\alpha \left(\frac{2}{n}\right)^4 + \frac{n}{2} 3\alpha \left(\frac{3}{n}\right)^6 \\
&\leq \frac{1}{n}
\end{aligned}$$

■

Remarks:

- (i) To get from a “sufficiently large” d to $d \geq 3$, we need a better approximation than Stirling’s formula. Also the constant c must be a function of α . In the proof we need to take derivatives and do some other complicated things, but in the end it works.
- (ii) One can show that a random graph with degree distribution $d_1 \geq d_2 \geq \dots \geq d_n \geq 3$ in the configurational model has constant conductance with high probability. (The configurational model is the procedure we used to generate the regular graph from the groups of minivertices.) For more on this see Durrett’s book, Chapter 3.
- (iii) There are deterministic methods for generating constant degree constant expanders. Both Zig Zag expanders and Ramanujan graphs have these properties.