Stat206b Random Graphs	**Spring 2003**

Lecture 3: January 28

Lecturer: David Aldous — *Scribe: Samantha Riesenfeld*

Random graphs with a prescribed degree distribution

A *degree distribution* $(d_0, d_1, d_2, \ldots)$ of a graph specifies for each possible degree the fraction of vertices of that degree, that is

$$d_i = \frac{\#\text{ vertices of degree } i}{\#\text{ vertices}}.$$

Let the random variable D be the degree of a vertex chosen uniformly at random. Then $d_i = \mathbf{Pr}[D = i]$.

For a connected graph on n vertices, clearly $D \geq 1$. The sum of the degrees over all the vertices is at least $2(n-1)$, and therefore the expected value of D is

$$\mathbf{E}[D] \geq \frac{2(n-1)}{n},$$

i.e. in a connected graph, the mean degree is at least 2, asymptotically.

Theme (imprecise): Specify a degree distribution $(d_1, d_2, \ldots)$ with mean degree at least 2. (Note that we assume $d_0 = 0$.) There is a model for a random graph $\mathcal{G}_n$ such that the graph

- has n vertices
- is connected
- the degree D_n of a vertex chosen uniformly at random from the vertex set V_n satisfies $\mathbf{Pr}[D_n = i] \to d_i$ as $n \to \infty$
- $\mathcal{G}_n$ is "completely random" subject to these constraints,

and all such models are essentially equivalent.

The model also has the following properties:

- It is "locally tree-like"; in particular, the cluster coefficient $C_n \to 0$ as $n \to \infty$.
- If $d_i = i^{-(3+\epsilon+\omega(1))}$ as $i \to \infty$, then for the random variable L that is the distance between two vertices chosen uniformly at random, $\mathbf{E}[L] \sim c \log n$ for a constant c depending on (d_i).

Aside: In real-life networks, the cluster coefficient is usually nonzero.

For the theoretical background about this model, there are two Molloy-Reed papers.

Construction: Given nonnegative integers $\Delta_1, \Delta_2, \ldots, \Delta_n$ such that $\sum_i \Delta_i$ is even:

1. Draw n vertices and label them $1, \ldots, n$.
2. For each j, write j on Δ_j cards.
3. Shuffle the cards.
4. Deal the cards 2-at-a-time.
5. When you deal $(j, \hat{j})$, put an edge between the vertices labeled j and $\hat{j}$.

(This experiment was undertaken during class.)

This process gives a random graph in which vertex i has Δ_i edges. For small n, it is rare to get a proper graph—it is easy to get self-loops or two edges between a pair of vertices. Really, this construction gives a distribution on a larger class of graphs (multi-graphs). For large enough n, the process usually gives something acceptable, but to make it work in general, one needs to do some fudging to get things exactly connected or to redefine "proper" graphs.

The model $\mathcal{G}_n$: Given large n, suppose we do the construction with nd_i vertices of degree i. Letting D be the degree of a vertex chosen uniformly at random, then $\mathbf{Pr}[D = i] \to d_i$ as $n \to \infty$.

Fact: As $n \to \infty$, the part of $\mathcal{G}_n$ within distance K (for fixed K) of a randomly chosen vertex v converges to the tree produced in the first K generations of the GWBP with D offspring in the first generation.

Let D^* be the number of offspring in subsequent generations. Then

$$\mathbf{Pr}[D^* = i] = \frac{(i+1)\mathbf{Pr}[D = i+1]}{\mathbf{E}[D]}.$$

To see this, we first pick v. It has D_v edges (D_v has the distribution of D). The neighbors of v are chosen by random picks of cards, which we can assume are basically independent since we are focusing on the early part of the process. Let v_1 be a neighbor of v. The number of children of v_1 is $\deg(v_1) - 1$. The chance that v has an edge to a specific vertex is proportional to that vertex's degree; we can think of the chance that v has an edge to an $i+1$ vertex as the chance that v appears in $i+1$ picks of the total $n\mathbf{E}[D]$ cards. Therefore

$$\mathbf{Pr}[\deg(v_1) = i+1] = \frac{nd_{i+1}(i+1)}{n\mathbf{E}[D]} = \mathbf{Pr}[D^* = i].$$

The analysis of GWBP has heuristic implications for $\mathcal{G}_n$: We will analyze the branching process to compute the distance between two vertices chosen uniformly at random in the graph model.

Easy Case: $D \geq 2$ and $\mathbf{E}[D] > 2$, which implies that $D^* \geq 1$ and $\mathbf{E}[D^*] = \mu > 1$. In this case, GWBP always survives forever. Assume that $\mathbf{E}[(D^*)^{1+\epsilon}] < \infty$ and let Z_k be the number of individuals in the kth generation of GWBP. We know that $Z_k \sim W\mu^k$ for some random variable $W > 0$. (More formally, we know that Z_k/μ^k converges in distribution to some limiting distribution, which is the distribution of a random variable W.) Then

$$\bar{Z}_k = \sum_{j=0}^{k} Z_k \sim \frac{W\mu^k}{1 - 1/\mu}.$$

Consider two random vertices v and $\hat{v}$. For large n, the branching processes that begin growing from v and $\hat{v}$ start off acting very independently. The question is at what stage an edge is likely to appear between the two subtrees; this gives us an estimate of the distance between v and $\hat{v}$.

We want to know the probability of an edge occuring between a vertex in the kth generation of one subtree and a vertex in any of the previous $k-1$ generations of the other subtree. Such an edge has roughly chance $\bar{Z}_{k-1}/n$. The number of vertices in the kth generation is Z_k, and so the expected number of edges is approximately $\mathbf{E}[Z_k \bar{Z}_{k-1}/n] \sim c\mu^{2k}/n$ for some constant c.

The generation k where an edge between subtrees first appears is approximately a solution of

$$\frac{\mu^{2k}}{n} = 1.$$

Let L be the distance from v to $\hat{v}$. Then L should be about $2k \pm O(1) = \log n / \log \mu \pm O(1)$.

Now let's back up for a minute to consider some of the assumptions made. For example, consider the assumption that $\mathbf{E}[(D^*)^{1+\epsilon}] < \infty$. This assumption corresponds to the assumption that $\mathbf{Pr}[D^* = i] = i^{-(2+\epsilon+\omega(1))}$), which in turn corresponds to the assumption that $\mathbf{Pr}[D = i] = i^{-(3+\epsilon+\omega(1))}$). So under these constraints on the power law distribution, the analysis holds.

How might it happen that $\mathcal{G}_n$ is not connected? If nodes i and j have degree 2, then they could end up in a 2-cycle. For fixed i and j, the probability of this is approximately $1/n^2$. There are $d_2 n$ vertices of degree 2. The expected number of 2-cycle components tends to a constant as $n \to \infty$ if d_2 is positive. The expected number of isolated nodes also tends to a constant. It is not necessary to worry too much about bigger size components since the probability of getting such a component is relatively much smaller than the number of possible such components.

Hard Case: $\mathbf{Pr}[D = 1] > 0$, $\mathbf{Pr}[D^* = 0] > 0$, and $\mathbf{E}[D] > 2$. Suppose we want to analyze the graph model given these parameters. Clearly, since the branching process can die, we may have a number of small components, but with very high probability, there will be one giant component. To estimate the number of vertices in this component, we can use that

$$\frac{1}{n}(\#\text{ vertices in giant component}) \to \mathbf{Pr}[\text{GWBP survives forever}] = 1 - \Psi,$$

for some Ψ. In the last class, we showed that $\rho^* = \mathbf{Pr}[\text{GWBP}(D^*) \text{ goes extinct}]$ can be found as a solution of an equation, and that

$$\Psi = \sum_{i=0}^{\infty} \mathbf{Pr}[D = i](\rho^*)^i.$$