Graphs and Networks Lecture 3

Erdös-Rényi Graphs, Part 2

Daniel A. Spielman September 5, 2013

3.1 Disclaimer

These notes are not necessarily an accurate representation of what happened in class. They are a combination of what I intended to say with what I think I said. They have not been carefully edited.

You should be able to find a diary of my Matlab session from today's class. It may reveal computations that do not appear in these notes.

3.2 Overview

My plan for today's lecture is to:

- Prove a good bound on the diameter of Erdös-Rényi graphs.
- Observe that it depends on a property of these graphs that is surprising, and which most graphs in our library do not have: every pair of sets of $n/14$ vertices in the graph share an edge. We assume that natural social network graphs probably do not have this property. For example, we could probably find a set of half a billion people in India and half a billion people in China so that no person in one set knows any person in the other.
- Talk about better models that fix some of this.
- Prove that most vertices in Erdös-Rényi graphs are not involved in triangles, or short cycles for that matter.
- Observe that most real world graphs have many, many, triangles and short cycles.

I will put many constants in the bounds that I prove in this lecture. I am not too attached to them. We could tighten them. The key point is that they are constants.

3.3 First Lesson from Chernoff

I begin by recalling the most important less from the Chernoff bounds: if X is a sum of independent $\{0,1\}$ random variables and if the expectation of X is large, then it is very unlikely that X differs

too much from its expectation. The probability it differs by more than δ times it expectation is at most inverse exponential in the expectation, with a constant in the exponent depending on δ.

3.4 Rapid Growth

We begin by recalling the theorem that we proved at the end of last class. Recall that we have defined $B(r, a)$ to be the ball of radius r around the vertex a, and $S(r, a)$ to be the shell (or sphere) consisting of vertices at distance exactly r from a.

Theorem 3.4.1. *Let G be a graph chosen from $\mathcal{G}(n, p)$ with $p = 2 \ln n/n$. Let a be a vertex of G and let r be an integer. If $|B(r, a)| \leq n/12 \ln n$, and $s = |S(r, a)|$, then*

$$\boldsymbol{Pr}\left[|S(r+1, a)| \leq (1/5)s \ln n\right] \leq n^{-1.2s}.$$

Since the term $(1/5) \ln n$ is going to come up a lot, let me call it γ, for *growth*. I will assume that n is large enough that $\gamma = (1/5) \ln n > 2$.

Let's examine a little more what this theorem tells us. Let a be any vertex, let r_a be the largest radius such that $|B(r_a, a)| \leq n/12 \ln n$, and assume that for all $r \leq r_a$,

$$|S(r+1, a)| \geq \gamma \, |S(r, a)| \, .$$

We will now show that this implies that most of the vertices of $B(r_a, a)$ are in its outer shell.

As

$$|B(r, a)| = \sum_{t \leq r} |S(t, a)| \, ,$$

we have

$$|B(r, a)| \leq |S(r, a)| \, (1 + \sum_{t=1}^{r} \gamma^{-t}) \leq \frac{\gamma}{\gamma - 1} \, |S(r, a)| \, .$$

So,

$$|S(r, a)| \geq \frac{\gamma - 1}{\gamma} \, |B(r, a)| \, .$$

As we have assumed $\gamma > 2$, $S(r, a)$ has most of the vertices of $B(r, a)$, for $r \leq r_a + 1$.

This is likely to be the case for all vertices a. As $\gamma \geq 2$, we know that $r_a \leq \log_2 n$. So, we merely need the conditions of the theorem to hold for all vertices a and all r between 1 and $\log_2 n$. As this is $n \log_2 n$ events that need to hold, and each fails with probability at most $n^{-1.2}$, the probability that they all hold is at least $1 - (\log n)/n^{1/5}$, which goes to 1 as n goes to infinity.

3.5 Diameter

Under the assumption of the previous section, we now show that such a graph is likely to have diameter at most $2 \log_2 n + 3$. The assumptions imply that for every vertex a,

$$|S(r_a + 1, a)| \geq \frac{1}{2} \, |B(r_a + 1, a)| \geq \frac{n}{24 \ln n}.$$

We would now like to show that for every pair of vertices a and b, there is probably a short path between $S(r_a + 1, a)$ and $S(r_b + 1, b)$. One way to do this would be to repeat the argument that we used to prove Theorem 3.4.1 for sets of this large size. We would begin by observing that the probability that any vertex (not in $B(r_a + 1, a)$) is not a neighbor of $S(r_a + 1, a)$ is at most

$$(1-p)^{|S(r_a+1,a)|} \leq \exp\left(-p\,|S(r_a+1,a)|\right) \leq \exp\left(-\frac{2\ln n}{n}\frac{n}{24\ln n}\right) = \exp\left(-1/12\right) \leq 1 - 1/13.$$

So, every vertex not in $B(r_a + 1, a)$ has at least a $1/13$ chance of being in $S(r_a + 2, a)$. So,

$$\mathbf{E}\left[|B(r_a+2,a)|\right] \geq n/13.$$

The Chernoff bounds tell us that with probability extremely close to 1, at least $1/14$ of the vertices not in $B(r_a + 1, a)$ are in $B(r_a + 2, a)$. That is, with very high probability,

$$|B(r_a+2,a)| \geq n/14.$$

This follows from the summary of the Chernoff bounds that I gave at the beginning of the lecture . . . I thought of assigning this as a homework problem, but it was too easy.

If we pursued this argument for a few more steps, we would show that $S(r_a + 11, a) > n/2$, for all a. At this point, we would know that $S(r_a + 11, a)$ and $S(r_b + 11, b)$ must overlap. But, we can do one better: we can show that for sufficiently large n, with very high probability, there must be an edge between $S(r_a + 2, a)$ and $S(r_b + 2, b)$.

If $|B(r_a + 2, a)| \geq n/14$ and $|B(r_b + 2, a)| \geq n/14$, and the two sets do not overlap, then there are at least $(n/14)^2$ possible edges between them. In the next section, we show that it is very unlikely that there are *any* two sets of this size that do not have an edge between them. In fact, we show this for p of the form c/n, for a some constant c. It holds in our case as, for n sufficiently large, $2\ln n > c$.

3.6 Lack of Community Structure

We think that most real-world graphs are not random. We expect them to have some sort of community structure. And, if they do, then we would think that it would be possible to find two sets of $n/14$ vertices without any edges between them. But, this doesn't happen in an Erdös-Rényi graph even with $p = c/n$ for a sufficiently large constant c. Let's prove that.

Theorem 3.6.1. *Let $c > 1$ and $\alpha > 0$ be constants such that*

$$c > 2\ln(e/\alpha)/\alpha.$$

Let G be sampled from $\mathcal{G}(n, p)$ with $p = c/n$. Then, the probability that there exist two sets of αn vertices with no edges between them goes to zero as n grows large.

For example, if $\alpha = 1/4$, this would work with $c = 19.1$.

Proof. Let S and T be any two sets of αn vertices. The probability that they have no edges between them is

$$(1-p)^{|S||T|} \leq \exp\left(-\frac{c}{n}(\alpha n)^2\right) = \exp\left(-c\alpha^2 n\right).$$

Taking a union bound over all sets S and T of this size, we find that the probability that there exist any S and T with αn vertices and no edges between them is at most

$$\begin{aligned}\binom{n}{\alpha n}^2 \exp\left(-c\alpha^2 n\right) &\leq \left(\frac{en}{\alpha n}\right)^{2(\alpha n)} \exp\left(-c\alpha^2 n\right) \\ &\leq \exp\left(-c\alpha^2 n + (2\alpha n)\ln(e/\alpha)\right) \\ &\leq \exp\left(n(2\alpha \ln(e/\alpha) - c\alpha^2)\right).\end{aligned}$$

This goes to zero if

$$c\alpha^2 > 2\alpha \ln(e/\alpha) \quad \Longleftrightarrow \quad c > 2\ln(e/\alpha)/\alpha.$$

□

This implies something else shocking: under these conditions, *every* set of αn vertices has at least $(1-\alpha)n$ neighbors.

3.7 Demo

I was planning on doing a demonstration to show that this does not usually happen in the graphs we've been considering. But, it was too easy. The reason is that almost all of the graphs we are considering have a very wide range of degrees. There are many vertices whose degrees are much smaller than the average degree. These tended to be the ones that I found, which is much less impressive.

Let's see this for a graph or two.

```
>> %% let's look at one of our graphs
>> load Bg_S_cerevisiae
>> %% it turns out that it needs some cleaning: it has diagonal entries
>> sum(diag(a))

ans =

   (1,1)           5426

>> %% this will get rid of them
>> a = a - diag(diag(a));
>>
>> %% now, let's check if it is symmetric
>> sum(sum(abs(a-a')))
```

```
ans =

   All zero sparse: 1-by-1

>> %% it is.  But, if it wasn't I could make it symmetric by typing
>> a = double(a+a'>0);
>> %% this also converts all edges weights that aren't 1 to 1.
>> %% Some of our graphs have those.
>> format short g
>> %% now, let's look at the degrees
>> degs = full(sum(a)); % the full is not necessary, but it will make display prettier
>> [min(degs), mean(degs), max(degs)]

ans =

            1       67.095         2873

>> counts = zeros(1,2873);
>> n = length(a)

n =

        6548

>> for i = 1:n, d = degs(i); counts(d) = counts(d)+1; end
>> [1:10;degs(1:10)]

ans =

     1     2     3     4     5     6     7     8     9    10
   678   171   274   140   187   321   168   226   258   275

>>
```

It also makes the analysis we just did irrelevent, as Erdös-Rényi graphs don't look like this. The degrees of their vertices are tightly concentrated around the mean. In the next lecture we will examine models that do look like this.

The one graph that we have built in which the degrees of almost all nodes are close to the average are the k-nearest neighbor graphs. Let's try this with the graph `mnist_knn3`. I will use a heuristic in the package Metis for partitioning the graph into two pieces with as few edges as possible in between. I call this from matlab code that I've written.

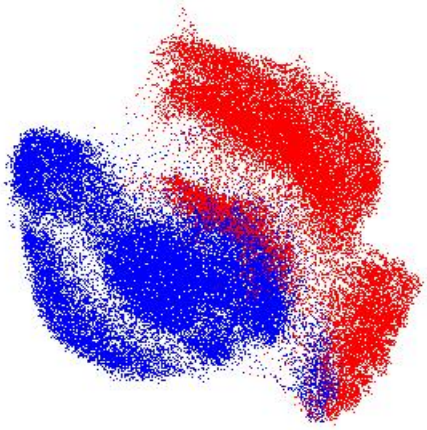
```
>> load mnist_knn3
>> n = length(a)
```

```
n =

       60000

>> part = callMetis(a,2);
>> sum(part==0)

ans =

       29928

>> sum(part==1)

ans =

       30072

>> nnz(a(part==0,part==1)) % count number of non-zero entries

ans =

        1029
```

We have divided the graph into two parts of approximately 30,000 vertices each, and there are only 1029 edges between them. So, if we delete the endpoints of those 1029 edges, we will have at least (approximately) 29,000 vertices on either side with no edges between them.

I'd now like to observe that the code that we used to draw this graph put the vertices from the different sides in different parts of the drawing. To show this, I will recover the coordinates, and then draw the vertices from one side in red and the vertices from the other in blue. I will now draw the edges in between.

```
load ../www/resources/mnist.mat
s = find(part==0);
t = find(part==1);
plot(pos(s,1),pos(s,2),'.');
hold on
plot(pos(t,1),pos(t,2),'r.');
```



3.8 Fixing The Models

One way of fixing the model of Erdös-Rényi graphs is to enforce a community structure. For example, before choosing the graph, one could divide the vertices into two sets A and B. One could then choose probabilities $p > q$, and then add edges between vertices inside the same set with probability p and between different sets with probability q. One can then take this further, beginning with even more sets and a list of different probabilities.

Some peole also like models with overlapping communities.

Of course, this does lead to the problem of deciding how to assign each person to which communities, and how they should overlap. The bottom line is that you can make very complicated models.

3.9 Clustering Coefficients

Another property that most "real-world" graphs have is that they have many triangles. That is, there are many nodes such that many of their neighbors are also neighbors of each other. I assigned reading about this for the first lecture of the class.

This should be obvious for collaboration graphs: if three people collaborate on one thing, then they will be a triangle in the graph. It is more surprising and interesting for other graphs.

In contrast, random Erdös-Rényi graphs are unlikely to have many triangles. Let's count the expected number of triangles in graph chosen from $\mathcal{G}(n, p)$.

Each triangle is specified by its three vertices. So, there are $\binom{n}{3}$ potential triangles. The probability

that all three edges are present is p^3. So, the expected number of triangles in such a graph is

$$\binom{n}{3} p^3 \leq (np)^3/6.$$

For $p = c \ln n/n$, this is $(c \ln n)^3/6$ triangles, which is vanishingly small compared to n.

Let's now count the triangles in some actual graphs. One way to do this is to capture the neighbors of every vertex, and then count how many edges are in it. This will count each triangle 3 times (once for each endpoint).

I wrote code called `countTriangles` to do this for you.

```
>> load mnist_knn3
>> countTriangles(a)

ans =

       89541

>> load Bg_S_cerevisiae
>> countTriangles(a)

ans =

     5524611

>> load soc-Epinions1
>> countTriangles(a)

ans =

     4873443
```

That's a lot of triangles.

I want to quickly point out that the speed of doing this depends quite a bit on how you represent a graph. When Matlab stores a sparse matrix, it stores a list of the non-zero entries in each column. So, it is much faster to get them from columns than from rows. The code `a(:,i)` returns the entries of a in column i, and `find(a(:,i))` returns the list of their positions. To see that this is much faster doing it by columns than by rows, I'll time each on one of our smaller graphs. A bigger graph would take too long. The command `toc` outputs the amount of time that has passed since the last `tic`. This code doesn't do anything with the answers it gets.

```
>> tic; for i = 1:n, ind = find(a(:,i)); end; toc
Elapsed time is 0.190525 seconds.
>> tic; for i = 1:n, ind = find(a(i,:)); end; toc
```

ctrl-c

```
>> toc
Elapsed time is 48.828321 seconds.
```

I couldn't wait any longer. Let's see how far it got.

```
>> i

i =

       16210
```

3.10 Local Tree Structure

We can say something stronger than that there are very few triangles in an Erdös-Rényi graph: there are very few short cycles, and most vertices are unlikely to be in a short cycle.

Let's compute the probability that any vertex a is in a cycle of length at most k. Such a cycle is determined by the $k-1$ other vertices in the cycle, and we should choose them in order. So, there are less than n^{k-1} choices for these other vertices (note that we are over-counting by a factor of 2), and the cycle appears with probability p^k. So, the probability that a is involved in a cycle of length k is at most

$$n^{k-1}p^k \leq= c^k/n.$$

So, most vertices are unlikely to be involved in cycles of length less than k if

$$c^k < n \quad \iff \quad c < n^{1/k}.$$

If a vertex is not involved in any cycles of length less than k, then its first $k/2$ neighbohoods are tree-like. This is also something that we do not expect to find in real-world graphs. But, it is common in many random graph models.

In fact, many algorithms for analyzing graphs, such as those using Belief Propogation or message passing algorithms, are based on the assumption that neighborhoods are tree-like. Sometimes they even work well on graphs that are not tree-like.