Graphs and Networks Lecture 2

Growth and Erdös-Rényi Graphs

Daniel A. Spielman September 3, 2013

2.1 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.

2.2 Announcement

The class is moving to WLH 119.

2.3 Introduction

We will begin class by examining the rate of growth of neighborhoods of vertices in some of our graphs. We will then see that graphs chosen from a natural distribution also have fast growth. In the lecture, we will encounter some inequalities that occur frequently in probabilistic analysis. These appear in boxes. I suggest that you learn them all.

I remark that there are three approaches to the analysis of random graphs. The first, favored by our textbook, is to give an intuitive, but not mathematically rigorous, explanation of what should happen. The second, which I take, is to give a rigorous exposition of weak bounds on what should happen. The third, which would take too long, would be to give a rigorous exposition of sharp bounds on what should happen.

2.4 Growth in Graphs

We are going to examine the growth of BFS (bread-first-search) balls around nodes in graphs. These are the sets of vertices within a given distance of some one vertex. For a vertex a and a number r, we define the *ball of radius r around a* to be the set of vertices of distance at most r from a. Here, the distance from a to b is the least number of edges in a path from a to b. We define

the *sphere of radius r around a* to be the set of vertices of distance exactly r from a. Symbolically, these are

$$B(r,a) \stackrel{\text{def}}{=} \{b \in V : \text{dist}(a,b) \leq r\},$$

and

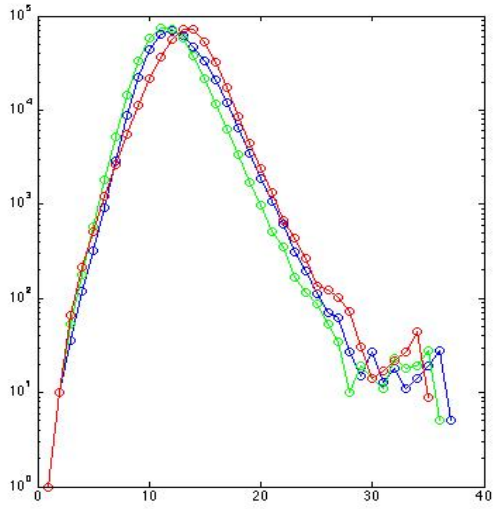
$$S(r,a) = B(r,a) \setminus B(r-1,a).$$

We will now look at the sizes of these spheres in some of our graphs by using the matlab code `randGrowth`. It picks a random vertex, and then returns the numbers of vertices in each sphere around that vertex.

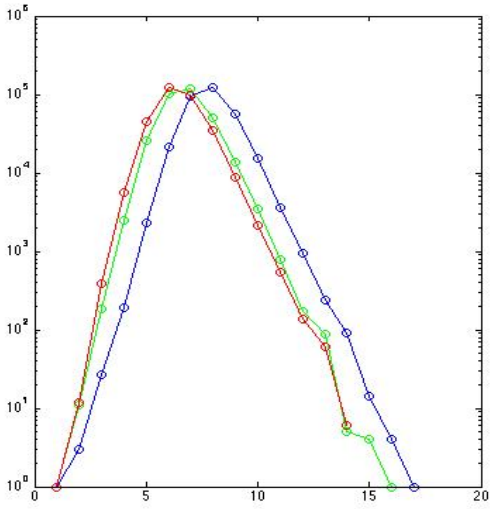
```
>> load Bg_S_cerevisiae
>> sizes = randGrowth(a); sizes'

ans =

           1
           6
         974
        5203
         358
           3

>> sizes = randGrowth(a); sizes'

ans =

           1
         153
        5207
        1177
           7

>> sizes = randGrowth(a); sizes'

ans =

           1
          70
        5079
        1380
          15
```

That was very rapid growth.

```
>> load amazon0601
>> sizes = randGrowth(a); sizes

sizes =

  Columns 1 through 6

           1          10          36         118         319         931

  Columns 7 through 12

        2899        8721       22312       43637       63822       69655

  Columns 13 through 18

       61634       47526       33607       21122       11993        6442

  Columns 19 through 24

        3510        1873        1075         621         313         194

  Columns 25 through 30

         112          71          62          27          15          27

  Columns 31 through 36

          13          18          11          14          19          28

  Column 37

           5

% This has a long tail.  Let's plot the sizes of the spheres.

>> clf;
>> semilogy(sizes); hold on; semilogy(sizes,'o')
>> hold on
>> sizes = randGrowth(a); sizes;
>> semilogy(sizes,'g'); hold on; semilogy(sizes,'go')
>> sizes = randGrowth(a); sizes;
>> semilogy(sizes,'r'); hold on; semilogy(sizes,'ro')
```



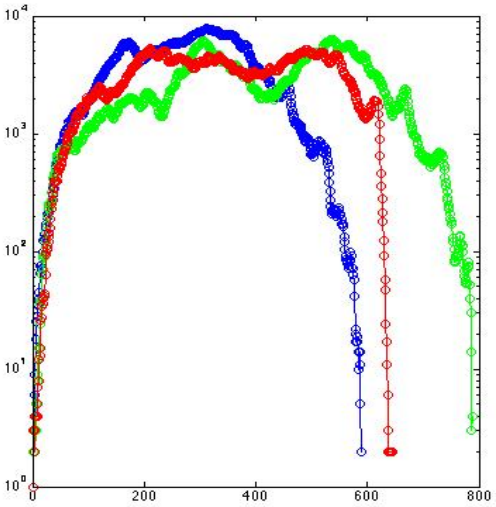

```
>> load dblp
>> sizes = randGrowth(a); sizes

sizes =

  Columns 1 through 6

           1           3          27         191        2285       21806

  Columns 7 through 12

       93641      123185       55657       15323        3667         946

  Columns 13 through 17

         239          90          14           4           1

>> clf
>> semilogy(sizes); hold on; semilogy(sizes,'o')
>> sizes = randGrowth(a); sizes;
>> hold on
>> semilogy(sizes,'g'); hold on; semilogy(sizes,'go')
>> sizes = randGrowth(a); sizes;
>> semilogy(sizes,'r'); hold on; semilogy(sizes,'ro')
```



I don't want you to get the idea that all graphs exhibit such growth. Graphs that are inherently low dimensional do not. For example, let's look at a road network.

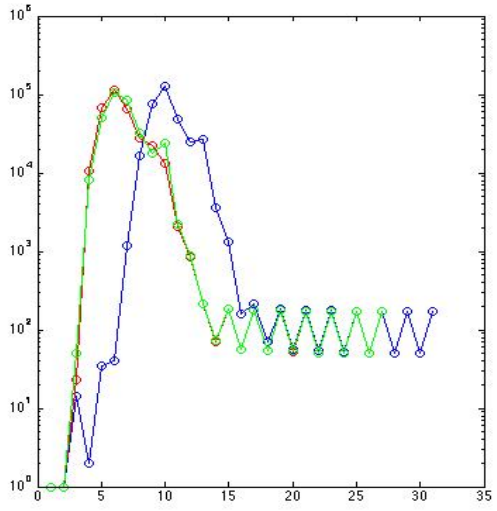
```
>> load roadNet-CA
>> sizes = randGrowth(a); sizes

sizes =

  Columns 1 through 6

          1           2           2           2           2           3

  Columns 7 through 12

          5           8           9          12          12          15

  Columns 13 through 18

         24          36          39          40          64          69

  Columns 19 through 24

         88          91          92          89         109         123

  Columns 25 through 30

        148         165         200         229         237         251
```

```
  Columns 31 through 36

         243          250          263          272          320          366

...

>> clf
>> semilogy(sizes); hold on; semilogy(sizes,'o')
>> hold on
sizes = randGrowth(a); sizes
>> semilogy(sizes,'g'); hold on; semilogy(sizes,'go')
>> sizes = randGrowth(a); sizes;
>> semilogy(sizes,'r'); hold on; semilogy(sizes,'ro')
```



Some graphs exhibit stranger behavior. For example, the following web graph has rapid initial growth, but a long tail.

```
>> load web-NotreDame
>> clf
>> sizes = randGrowth(a);
>> semilogy(sizes); hold on; semilogy(sizes,'o')
>> sizes = randGrowth(a);
>> semilogy(sizes,'r'); hold on; semilogy(sizes,'ro')
>> sizes = randGrowth(a);
>> semilogy(sizes,'g'); hold on; semilogy(sizes,'go')
```



2.5 Random Graphs

We will now introduce the Erdös-Rényi model of random graphs, and prove that it exhibits rapid growth. This is not a particularly good model of "real-world" graphs. We study this model because it is simple and because the analysis of this model is similiar to the analysis of many better models. And, we have to start somewhere.

In the Erdös-Rényi model, each edge is chosen to appear in the graph independently with the same probability. For an integer n and a probability $p \in (0, 1)$, the distribution $\mathcal{G}(n, p)$ produces a random graph on n vertices by choosing each possible edge to appear in the graph independently with probability p.

The expected number of edges in a graph chosen from this distribution will be

$$\binom{n}{2} p,$$

and that the expected degree of each vertex is

$$(n-1)p.$$

As long as p is not too small, it is unlikely that the degree of any vertex deviates too much from its expectation. One proves such statements using "large-deviation bounds". We will see some of these in a few minutes. But first, we will examine whether the graph is connected.

2.6 Connectivity

It turns out that the graph is probably connected if $p > \ln n/n$, and that it is probably not connected if $p < \ln n/n$. This is called a "threshold phenomenon", and it is the mathematical inspiration for

the concept of a "tipping point". We will see others in the next few weeks.

We begin by showing that if $p = (1+\epsilon)\ln n/(n-1)$, for any $\epsilon > 0$, then a graph sampled from $\mathcal{G}(n,p)$ is probably connected. To make that statement clear, we set ϵ to a fixed constant and then look at what happens when n grows big. In our analysis, we will not concern ourselves with small n.

To begin, let a be any vertex, and let I_a be the event that vertex a is *isolated*. That is, that a has no edges. We begin by showing that it is unlikely that there is any isolated vertex. To begin, observe that

$$\mathbf{Pr}[I_a] = (1-p)^{n-1}.$$

To prove an upper bound on this probability, we use the inequality

$$\boxed{1-p \leq \exp(-p)}.$$

We will use this inequality often.

We now compute

$$\mathbf{Pr}[I_a] = (1-p)^{n-1} \leq \exp(-p(n-1)) = \exp(-(1+\epsilon)\ln n) = n^{-(1+\epsilon)}.$$

So, the probability that there exists an isolated vertex satisfies[1]

$$\mathbf{Pr}[\exists a : I_a] \leq \sum_a \mathbf{Pr}[I_a] \leq n \times n^{-(1+\epsilon)} = n^{-\epsilon}.$$

So, it is unlikely that there is any isolated vertex.

It turns out that when isolated verties are unlikely, the graph is probably connected. In order for the graph to be disonnected, there would have to be a subset S of the vertices such that there are no edges leaving S. We call such a set of vertices S a *cut*. If $\sigma = |S|$, then the probability S is a cut is $(1-p)$ raised to the number of possible edges leaving S, which is

$$(1-p)^{\sigma(n-\sigma)} \leq \exp(-p\sigma(n-\sigma))$$

So, the probability that there exists a cut S of size σ is at most

$$\binom{n}{\sigma} \exp(-p\sigma(n-\sigma)).$$

To show that this probability is small, we will divide it into two cases. When σ is small, we will use the bound

$$\boxed{\binom{n}{\sigma} \leq n^{\sigma}}.$$

[1]We are using the "Union Bound" here. The union bound is an extension of the fact that for any two events A and B, $\mathbf{Pr}[A \text{ or } B] \leq \mathbf{Pr}[A] + \mathbf{Pr}[B]$. While this is simple, it is given a name to help us remember it and to help us remember to try using it.

In particular, for $\sigma < \epsilon n/3$, we have

$$\exp\left(-p\sigma(n-\sigma)\right) = \exp\left(-(1+\epsilon)\frac{\sigma(n-\sigma)}{n-1}\ln n\right) \leq \exp\left(-(1+\epsilon/3)\sigma\ln n\right) \leq n^{-(1+\epsilon/3)\sigma}.$$

So, the probability that there is a cut S of size σ is at most

$$n^{\sigma} n^{-(1+\epsilon/3)\sigma} = n^{-(\epsilon/3)\sigma}.$$

Summing over the sizes between 1 and $\epsilon n/3$, we find that the probability that there is such a cut with size in this range is at most

$$\sum_{\sigma=1}^{\epsilon n/3} n^{-(\epsilon/3)\sigma} = n^{-(\epsilon/3)}/(1-n^{-(\epsilon/3)}).$$

This goes to zero as n grows large.

To handle the larger sets, we use the inequality

$$\boxed{\binom{n}{\sigma} \leq \left(\frac{ne}{\sigma}\right)^{\sigma}}.$$

Now, for $\epsilon n/3 < \sigma \leq n/2$, the probability that there is a cut S of size σ is at most

$$\begin{aligned}
\binom{n}{\sigma}\exp\left(-p\sigma(n-\sigma)\right) &\leq \left(\frac{en}{\sigma}\right)^{\sigma}\exp\left(-p\sigma(n-\sigma)\right) \\
&\leq \left(\frac{3e}{\epsilon}\right)^{\sigma}\exp\left(-p\sigma(n/2)\right) \\
&\leq \left(\frac{3e}{\epsilon}\right)^{\sigma}\exp\left(-(1+\epsilon)\ln n\sigma/2\right) \\
&= \left(\frac{3e}{\epsilon n^{(1+\epsilon)/2}}\right)^{\sigma}.
\end{aligned}$$

For $n^{(1+\epsilon)/2} > 3e/\epsilon$, these terms decrease geometrically as n grows. So, their sum is at most

$$\frac{\left(\frac{3e}{\epsilon n^{(1+\epsilon)/2}}\right)}{1-\left(\frac{3e}{\epsilon n^{(1+\epsilon)/2}}\right)},$$

which also goes to zero as n grows large.

2.7 Isolated Vertices

We now observe that if $p = (1-\epsilon)\ln/(n-1)$, then there are likely to be isolated vertices. For this analysis, we use the inequality

$$\boxed{1-p \geq \exp\left(-p(1+p)\right)},$$

which holds for $0 \leq p \leq 1/2$.

We have

$$\mathbf{Pr}\left[I_a\right] = (1-p)^{n-1} \geq \exp\left(-p(1+p)(n-1)\right) = \left(n^{1+p}\right)^{-(1-\epsilon)}.$$

Let's first get rid of that annoying $1+p$. We have

$$n^{1+p} = \exp\left(\ln n + (1-\epsilon)(\ln n)^2/n\right) \leq n\exp\left((\ln n)^2/n\right) \leq 2n,$$

for n sufficiently large.

When this holds, we have

$$\left(n^{1+p}\right)^{-(1-\epsilon)} \geq (2n)^{-(1-\epsilon)}.$$

The expected number of isolated vertices is

$$\sum_a \mathbf{Pr}\left[I_a\right],$$

which is at least

$$\frac{n}{(2n)^{1-\epsilon}} \geq n^{\epsilon}/2.$$

Note that we can improve this bound to $n^{\epsilon}(1-\alpha)$ for any $\alpha > 0$.

Actually, proving that the expected number of isolated vertices is large does not imply that there is probably an isolated vertex: it is conceivable that there is rarely an isolated vertex, but that when there are isolated vertices there are many. You will prove that this is not the case in Problem Set 1.

2.8 Large Deviations

There are many types of large deviation bounds. The ones that we will usually use are the Chernoff-Hoeffding bounds. These are not always the sharpest in all regimes, but they have the advantages that they will always be true and that they will usually allows us to obtain results that are qualitatively correct. They have many forms. We will use[2] the following.

Theorem 2.8.1. *Let $X_1, \ldots, X_n$ be independent Bernoulli (that is, 0/1 valued) random variables with $\boldsymbol{Pr}\left[X_i = 1\right] = p_i$. Let $X = \sum X_i$ and let $\mu = \sum p_i$ be the expectation of X. Then,*

a. for all $0 < \delta < 1$,

$$\boldsymbol{Pr}[X \leq (1-\delta)\mu] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} \leq \exp\left(-\mu\delta^2/2\right),$$

[2]Many forms of Chernoff bounds may be found. It is often convenient to prove one's own. The form we use here appears in [MU05]. Other useful forms and derivations may be found in [AS00, MR95, DP09].

b. for all $\delta > 0$,

$$\boldsymbol{Pr}[X \geq (1+\delta)\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu},$$

c. for $0 < \delta < 1$

$$\boldsymbol{Pr}[X \geq (1+\delta)\mu] \leq \exp\left(-\mu\delta^2/3\right).$$

For now, consider $p = 2\ln n/(n-1)$. For any particular vertex, its expected degree is $\mu = 2\ln n$. We will now show that it is unlikely that there is any vertex with degree larger than $5\ln n$. Setting $\delta = 3/2$ so that

$$(1+\delta)2\ln n = 5\ln n,$$

part b of Theorem 2.8.1 tells us that the probability that any particular vertex has degree larger than $5\ln n$ is at most

$$\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}, \leq (0.46)^{2\ln n} < \exp\left(-(3/2)\ln n\right) = n^{-3/2}.$$

As there are n vertices in the graph, the probability that there is one with degree larger than $5\ln n$ is at most $n^{-1/2}$.

We can also show that it is unlikely that there is a vertex of degree less than $(1/5)\ln n$. From the first Chernoff bound, with $\delta = 9/10$, we learn that the probability that any particular vertex has degree less than $(1/5)\ln n$ is at most

$$\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} = \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{2\ln n} \leq \exp\left(-(4/3)\ln n\right) = n^{-4/3}.$$

So, the probability that there is any vertex with degree less than $(1/5)\ln n$ is at most $n^{-1/3}$.

To see how loose these bounds are, I generated a number of graph from this distribution with $n = 10,000$. The expected degree was just slightly above 23. The maximum degree was usually around 47, and the minimum degree was usually around 5.

2.9 Rapid Growth of BFS Balls

We will now show that BFS balls in Erdös-Rényi graphs with $p = 2\ln n/(n-1)$ grow rapidly. You could do the same analysis with $p = (1+\epsilon)\ln n/(n-1)$ for any $\epsilon > 0$, with a little more work.

Our analysis will have two limiting cases. As when we proved connectivity, singleton sets will be one limiting case. The other will be large sets. Very large spheres cannot grow too much, because they can run out of space to grow. We will show that, until this happens, each sphere probably grows by a factor of at least $(1/5)\ln n$.

Theorem 2.9.1. *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*

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

Proof. Let $b = |B(r,a)|$. Let C be the set of nodes that are not in $B(r,a)$. Each of them has a chance of being a neighbor of a node in $S(r,a)$. The probability that one of them is not a neighbor of a node in $S(r,a)$ is

$$(1-p)^s \leq 1 - ps + \binom{s}{2}p^2 \quad \text{(by inclusion-exclusion)}.$$

So, the chance that each is a neighbor of a node in $S(r,a)$, in which case it is in $S(r+1,a)$, is

$$1 - (1-p)^s \geq ps(1 - ps/2).$$

We pause to simplify this expression. We have

$$ps \leq \frac{2\ln n}{n}\frac{n}{12\ln n} = \frac{1}{6}.$$

So,

$$ps(1 - ps/2) \geq (11/12)ps.$$

So, the expected number of nodes in $S(r+1,a)$ is at least

$$(11/12)ps(n-b) \geq (11/12)^2 psn \geq (10/12)psn = s(20/12)\ln n.$$

We can apply Theorem 2.8.1 to upper bound the probability that the size of this set is less than $(1/5)s\ln n$. To do this, we set $\delta = 46/50$, and find that the probability is at most

$$\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{s(20/12)\ln n} \leq (0.3022)^{s\ln n} \leq n^{-1.2s}.$$

□

There is one subtlety in this proof that I should mention: I used the fact that the edges between $B(a,r)$ and the rest of the graph are random, *even though I am making assumptions about* $B(a,r)$. The reason I can do this is that those edges do not depent on $B(a,r)$. That is, we could have written the statement of the theorem as

$$\mathbf{Pr}\left[|S(r+1,a)| \leq (1/5)\ln n\,|S(r,a)| \,\middle|\, |B(r,a)| \sum n/12\ln n\right] \leq n^{-1.2|S(r,a)|}.$$

One way of thinking of this is to consider the BFS procedure. At the point where it has determined the ball $B(r,a)$, it has not examined any of the edges leaving the set of vertices in this ball.

2.10 Diameter

In the next lecture, we will extend this argument to show that this graph probably has logarithmic diameter.

References

[AS00] Noga Alon and Joel Spencer. *The Probabilistic Method*. John Wiley & Sons, 2000.

[DP09] Devdatt Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, New York, NY, USA, 2009.

[MR95] Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Cambridge University Press, New York, NY, USA, 1995.

[MU05] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2005.