T-79.7003: Graphs and Networks **Fall 2013**

Lecture 2: September 20

Lecturer: Charalampos E. Tsourakakis *Thresholds for subgraphs and connectivity*

2.1 Thresholds

In the previous lecture we discussed the existence of thresholds of monotone properties. Let's begin with a formal definition of what we described the previous time.

Definition 2.1 (Threshold) *A function $p^* = p(n)$ is a threshold for a monotone increasing property*[1] $\mathcal{P}$ *in $G(n,p)$ if*

$$\lim_{n\to+\infty} \mathbf{Pr}\left[G(n,p) \in \mathcal{P}\right] = \begin{cases} 1 & \text{if } p^* = o(p)(p^* \ll p) \\ 0 & \text{if } p = o(p^*)(p \ll p^*) \end{cases}$$

as $n \to +\infty$.

Last time, we discussed the existence of thresholds for various monotone properties. It is natural to ask whether all monotone properties have a threshold. The answer is stated as a theorem without proof.

Theorem 2.2 *Every non-trivial monotone property has a threshold.*

Today, we will discuss two monotone increasing properties, which according to the above theorem have a threshold: the appearance of a K_4 and connectivity. Before we go into the main results of today's class, we will go over some basic tools.

2.2 Basic tools

We will use the following inequalities to bound the binomial coefficient $\binom{n}{k}$.

$$\boxed{\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k.}$$

We will also need to be able to upper- and lower-bound certain expressions. Here are some useful inequalities.

$$\boxed{(1-x)^n \geq 1 - nx, \ \ \forall 0 \leq x \leq 1.}$$

$$\boxed{e^x \geq x + 1, \ \ \forall x.}$$

[1] Of course, in the case of monotone decreasing properties, the two cases above will be flipped.

$$\boxed{e^x \leq x^2 + x + 1, \quad \forall 0 < |x| < 1.}$$

$$\boxed{\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots, \quad \forall 0 < |x| < 1.}$$

$$\boxed{\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \geq \left(\sum_{i=1}^{n} a_i b_i\right)^2.}$$

The above inequality is the Cauchy-Schwartz inequality, which is a special case of Hölder's inequality for $p = q = 2$.

Theorem 2.3 (Hölder's inequality) *For any positive real numbers p, q such that $\frac{1}{p} + \frac{1}{q} = 1$*

$$\boxed{\left(\sum_{i=1}^{n} |a_i b_i|\right) \leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}.}$$

The following inequalities are basic probabilistic tools.

Theorem 2.4 (Markov's Inequality) *Let X a be non-negative integer valued random variable. Then,*

$$\mathbf{Pr}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

Proof:

$$\mathbb{E}[X] = \sum_{k \geq 1} k\mathbf{Pr}[X = k] \geq \sum_{k=t} k\mathbf{Pr}[X = k] \geq t \sum_{k=t} \mathbf{Pr}[X = k] = t\mathbf{Pr}[X \geq t].$$

■

We will use this inequality in two ways in our class. First, it is the basis of the first moment method. In many cases we will need to show that $\mathbf{Pr}[X > 0] = o(1)$, where X is a non-negative random variable of interest. It turns out that computing $\mathbb{E}[X]$ can be much easier than directly computing $\mathbf{Pr}[X > 0]$ in numerous cases. If $\mathbb{E}[X] = o(1)$ then by Markov's inequality

$$\boxed{\mathbf{Pr}[X > 0] \leq \mathbb{E}[X]}$$

we obtain that $X = 0$ **whp** . Furthermore, we will use Markov's inequality to obtain probabilistic inequalities for higher order moments. This is a special case of the following observation. If ϕ is a strictly monotonically increasing function, then

$$\mathbf{Pr}[X \geq t] = \mathbf{Pr}[\phi(X) \geq \phi(t)] \leq \frac{\mathbb{E}[\phi(X)]}{\phi(t)}.$$

For instance, if $\phi(x) = x^2$, then we obtain Chebyshev's inequality.

Theorem 2.5 (Chebyshev's Inequality) *Let X be any random variable. Then,*

$$\mathbf{Pr}\left[|X - \mathbb{E}[X]| \geq t\right] \leq \frac{\mathbb{V}ar[X]}{t^2}.$$

A simple corollary of Chebyshev's inequality is the following:

Corollary 2.6 (Second moment method) *Let X be a non-negative integer valued random variable. Then,*

$$\boxed{\mathbf{Pr}[X = 0] \leq \frac{\mathbb{V}ar[X]}{(\mathbb{E}[X])^2}.}$$

For completeness, here is the proof.

Proof:

$$\mathbf{Pr}[X = 0] \leq \mathbf{Pr}\left[|X - \mathbb{E}[X]| \geq \mathbb{E}[X]\right] \leq \frac{\mathbb{V}ar[X]}{(\mathbb{E}[X])^2}.$$

■

The use of the above corollary is known as the second moment method. Here is how we will typically use it in our class. Let the random variable X of interest be the sum of m indicator random variables $X_1, \ldots, X_m$, where $\mathbf{Pr}[X_i = 1] = p_i$, i.e.,

$$X = X_1 + \ldots + X_m.$$

We will be interested in showing that $X > 0$ **whp** . Even if $\mathbb{E}[X]$ will tend to $+\infty$ this does not suggest that $X > 0$ **whp** . In order to prove this kind of statement, we will use the second moment method. Since $\mathbf{Pr}[X = 0] \leq \frac{\mathbb{V}ar[X]}{(\mathbb{E}[X])^2}$ it will suffice to prove that $\frac{\mathbb{V}ar[X]}{(\mathbb{E}[X])^2} = o(1)$. The problem therefore is reduced to computing or actually *upper-bounding* the variance.

In our typical setting,

$$\mathbb{V}ar[X] = \sum_{i=1}^{m} \mathbb{V}ar[X_i] + \sum_{i \neq j} \mathbb{C}ov[X_i, X_j] \leq \mathbb{E}[X] + \sum_{i \neq j} \mathbb{C}ov[X_i, X_j].$$

To see how we obtained the inequality, notice that $\mathbb{V}ar[X_i] = p_i(1-p_i) \leq p_i = \mathbb{E}[X_i]$. Hence by the linearity of expectation $\sum_i \mathbb{V}ar[X_i] \leq \sum_i \mathbb{E}[X_i] = \mathbb{E}[X]$. The covariance of two random variables A, B is defined as

$$\mathbb{C}ov[A, B] = \mathbb{E}[AB] - \mathbb{E}[A]\,\mathbb{E}[B].$$

In the case of indicator random variables we obtain the following expression:

$$\mathbb{C}ov[X_i, X_j] = \mathbf{Pr}[X_i = X_j = 1] - \mathbf{Pr}[X_i = 1]\mathbf{Pr}[X_j = 1].$$

So, when we apply the second moment, the hard part it to upper bound the sum of covariances. Section 2.3 illustrates a use of the first and second moment methods.

2.3 Emergence of a K_4 in $G(n,p)$

A K_4 is a complete graph on four vertices. Let X be the number of K_4s in $G(n,p)$. We will show that the threshold value p^* is equal to $n^{-2/3}$. The expectation of X

$$\mathbb{E}[X] = \binom{n}{4} p^6.^2$$

Let's see what happens to $\mathbb{E}[X]$ if $p \ll p^*$ or equivalently $p = \frac{p^*}{\omega(n)}$ where $\omega(n)$ is a function that tends to $+\infty$ as $n \to +\infty$.

$$\mathbb{E}[X] = \binom{n}{4} p^6 = \Theta\left(n^4 \left(\frac{n^{-2/3}}{\omega(n)}\right)^6\right) = \Theta\left(\frac{1}{\left(\omega(n)\right)^6}\right) = o(1).$$

Hence by the first moment method we can conclude that when $p \ll n^{-2/3}$

$$\mathbf{Pr}[X > 0] \leq \mathbb{E}[X] = o(1),$$

or equivalently $X = 0$ **whp** . Now, we will prove that $X > 0$ **whp** when $p^* \ll p$ or equivalently $p = p^*\omega(n)$ where $\omega(n)$ is a function that tends to $+\infty$ as $n \to +\infty$. Notice now that the expected value of K4s goes to infinity, namely

$$\mathbb{E}[X] = \binom{n}{4} p^6 = \Theta\left(n^4 \left(n^{-2/3}\omega(n)\right)^6\right) = \Theta\left(\left(\omega(n)\right)^6\right) \to +\infty.$$

However, this does not suggest that $X > 0$ **whp** . We need to apply the second moment method. First, let's define an indicator variable X_i for the i-th labeled copy of K_4 in K_n, $i = 1, \ldots, \binom{n}{4}$. We can write

$$X = X_1 + X_2 + \ldots + X_{\binom{n}{4}}.$$

What is the covariance of two indicator variables here? Well, let's see how dependencies kick in. When two copies of K_4 share no edge then the respective indicator variables are independent. To see why observe that in this case

$$\mathbb{C}ov[X_i, X_j] = \mathbf{Pr}[X_i = X_j = 1] - \mathbf{Pr}[X_i]\mathbf{Pr}[X_j] = p^{12} - p^6 p^6 = 0.$$

Equivalently, for the case of K_4 this happens if two K_4 copies intersect in 0 or 1 vertex. We are left with two cases, which are shown in figure 2.1. Let's consider the covariance for case (a). What is the probability that the two indicator variables are both 1? Since the two copies have two vertices in common, or equivalently 1 edge, the total number of edges is 11. Hence we get that the covariance is

$$\mathbb{C}ov[X_i, X_j] = p^{11} - p^{12}.$$

Similarly, for case (b), we obtain that

[2] The number of edges in K_4 is $\binom{4}{2} = 6$.

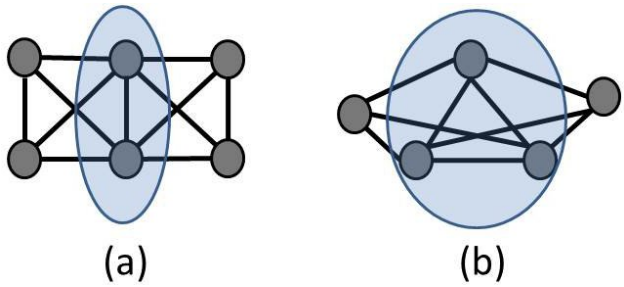

Figure 2.1: The two cases we need to consider in the covariance estimation for K_4s. Intersections of the two copies are highlighted with a shaded blue area.

$$\mathbb{C}ov\left[X_i, X_j\right] = p^9 - p^{12}.$$

Now we have to count how many pairs of indicator variables fall into case (a) and case (b). In case (a) we have $\binom{n}{6}$ ways to choose 6 out of n vertices and $\binom{6}{2,2,2}$ ways to choose the specific labeled configuration. Similarly for case (b), we have $\binom{n}{5}\binom{5}{3,1,1}$ such pairs of indicator variables. Putting everything together gives

$$\mathbb{V}ar\left[X\right] \leq \binom{n}{4}p^6 + \binom{n}{6}\binom{6}{2,2,2}p^{11} + \binom{n}{5}\binom{5}{3,1,1}p^9 = o(n^8p^{12}) = o\big(\left(\mathbb{E}\left[X\right]\right)^2\big).$$

This concludes the proof that $X > 0$ **whp** when $p^* \ll p$.

2.4 Connectivity

In this section we prove that the threshold p^* for connectivity is $\frac{\log n}{n}$. We break the proof of our main result in small incremental steps. In Section 2.4.1 we simulate the transition using Matlab. I think it is a very good practice to simulate certain properties you are interested into. It is not necessary but can help you see what you will need to prove. For instance, in the case of connectivity, the simulation strongly indicates that as we increase p and we get closer and closer to the connectivity threshold, the cause of G being disconnected are isolated vertices.

In Section 2.4.2 we will discuss a crude way of upper bounding the expected number of connected components of order k. Finally in Section 2.4.3 we prove our result. We break our proof into two parts:

- We prove that $\frac{\log n}{n}$ is the threshold for the existence of isolated vertices.
- Let X_k is the number of connected components of size k. We will prove that

$$\mathbf{Pr}\left[X_1 > 0\right] \leq \mathbf{Pr}\left[G \text{ is disconnected }\right] \leq \mathbf{Pr}\left[X_1 > 0\right] + o(1).$$

2.4.1 Matlab simulation

Let's use MATLAB to simulate what we are interested into. We will use David Gleich's MATLAB BGL library. You can find a link in the class Web page. When you download it, add it to your path.

```
addpath('C:\Users\tsourolampis\Libraries\Matlab\matlab_bgl');
addpath(genpath('C:\Users\tsourolampis\Libraries\Matlab\matlab_bgl'));
```

Let's write a routine that creates a random binomial graph on n vertices.

```
function A = Gnp(n,p)

% Generates a random binomial graph on vertex set [n], edge probability p

A = double(rand(n) <= p);
A = triu(A,1);
A = A + A';
```

Let's search the threshold p^*.

```
n = 1000;
for p = 0 : 0.001 : 1
    A = Gnp(n,p);
    A = sparse(A);
    if( max(components(A))== 1)
        pstar = p;
        break
    end
end
```

When you run this little piece of code you will find a very good approximation of p^* which is $\frac{\log n}{n} \approx 0.007$. Let us look how the disconnected graph looks like when we are just below this threshold value.

```
A = Gnp(1000, 0.0068);
A = sparse(A);
[ci sizes] = components(A);
sizes

sizes =

   998
     1
     1
```

This means that we have three connected components. Two vertices are isolated and all the rest are in the same connected component. Let's go a bit above $p = 0.0068$.

```
A = Gnp(1000, 0.00684);
A = sparse(A);
[ci sizes] = components(A);
sizes

sizes =
```

```
999
  1
```

Again, we see that the lack of connectivity is due to one isolated vertex. This is not a coincidence. It turns out that as soon as isolated vertices disappear, the graph is connected **whp** . This is remarkable, as a simple necessary condition for connectivity is sufficient. However at the same time, if you think a bit more about it, it makes sense since the isolated vertices resist most to getting "swallowed" by the giant component (*what is more likely: a small or a large component will get "swallowed" by the giant component first?*).

2.4.2 Preliminaries

We will need to upper bound the expected number of connected components of size k in $G(n,p)$.

Theorem 2.7 *Let X_k be the number of connected components of order exactly k in $G(n,p)$.*

$$\mathbb{E}[X_k] \leq \binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k)}.$$

Proof: Consider a set of k vertices. There are $\binom{n}{k}$ of them. What is the probability they form a connected component? Well, the k vertices have to form a connected component, and at the same time there must be no edge from that component to the rest of the graph. The first probability is upper bounded using a simple union bound by the number of all spanning trees on k vertices which is k^{k-2}[3] times p^{k-1} ($k-1$ is the number of edges of a tree on k vertices). Finally, the probability that there is no edge from that connected component to the rest of the graph is $(1-p)^{k(n-k)}$ since there are $k(n-k)$ possible edges between the connected component and the rest of the graph, and none of them should exist. ■

2.4.3 Main result

Theorem 2.8 *The threshold for the existence of isolated vertices is $p^* = \frac{\log n}{n}$.*

Proof: Let X_1 be the number of isolated vertices, i.e., vertices with degree 0. First, let's write the number of isolated vertices as a sum of indicator variables.

$$X_1 = Z_1 + \ldots + Z_n.$$

Here, Z_i is 1 if vertex i has degree 0. This happens with probability $(1-p)^{n-1}$. Now, we can use the linearity of expectation to compute the expected value of X_1. Specifically,

$$\mathbb{E}[X_1] = \sum_{i=1}^{n} \mathbb{E}[Z_i] = \sum_{i=1}^{n} \mathbf{Pr}[deg(i) = 0] = n(1-p)^{n-1}.$$

When $p^* \ll p$

[3] Cayley's theorem

$$n(1-p)^{n-1} \leq ne^{-p(n-1)} = o(ne^{-\log n}) = o(1).$$

Hence, by the first moment method $\mathbf{Pr}[X_1 > 0] = o(1)$. To prove that when $p \ll p^*$ then $X_1 > 0$ **whp** , we need to apply the second moment method. First, let's try to understand why two indicator variables Z_i, Z_j are correlated always. If someone asks you what is the probability of vertex j being isolated *given* that vertex i is isolated, then you have some information about vertex j. Specifically, you know that edge (i, j) is not there, since if it were, i would not have been isolated. This fact introduces dependencies, but they are weak. So, let's compute the covariance of two indicator random variables.

$$\mathbb{C}ov[Z_i, Z_j] = \mathbf{Pr}[Z_i = Z_j = 1] - \mathbf{Pr}[Z_i = 1]\mathbf{Pr}[Z_j = 1] = (1-p)^{2n-3} - (1-p)^{2n-2} = p(1-p)^{2n-3}.$$

Combined with the fact that all $\binom{n}{2}$ pairs are correlated, it is easy to check (*fill in the details, as you read the notes*) that $\frac{\mathbb{V}ar[X_1]}{(\mathbb{E}[X_1])^2} = o(1)$. Therefore, $X_1 > 0$ **whp** , concluding the proof that p^* is the threshold for the existence of isolated vertices. ■

Theorem 2.9 *Let X_k be the number of connected components of size k in $G(n, p)$. Then,*

$$\sum_{k=2}^{n/2} \mathbf{Pr}[X_k > 0] = o(1).$$

Furthermore,

$$\mathbf{Pr}[X_1 > 0] \leq \mathbf{Pr}[G \textit{ is disconnected }] \leq \mathbf{Pr}[X_1 > 0] + o(1).$$

Proof:

Let $G \sim G(n, p)$. Notice that

$$\mathbf{Pr}[G \text{ is disconnected }] = \mathbf{Pr}\left[\cup_{k=1}^{n/2}\Big(G(n,p) \text{ has a component of order } k\Big)\right] \quad = \mathbf{Pr}\left[\cup_{k=1}^{n/2}\{X_k > 0\}\right].$$

By a union bound we obtain the following upper bound on the probability that G is disconnected.

$$\mathbf{Pr}[G \text{ is disconnected }] \leq \mathbf{Pr}[X_1 > 0] + \sum_{k=2}^{n/2} \mathbf{Pr}[X_k > 0].$$

Also, it is clear that

$$\mathbf{Pr}[G \text{ is disconnected }] \geq \mathbf{Pr}[X_1 > 0].$$

We know how to deal with the $\mathbf{Pr}[X_1 > 0]$ term. Specifically, from our analysis we already know that if $p \ll p^*$, then there exist isolated vertices and therefore G is disconnected **whp** . We also know that if $p^* \ll p$ then **whp** there are no isolated vertices. If we prove that $\sum_{k=2}^{n/2} \mathbf{Pr}[X_k > 0] = o(1)$, then we are done with the proof.

Using Markov's inequality and Theorem 2.7

$$\sum_{k=2}^{n/2} \mathbf{Pr}\left[X_k > 0\right] \leq \sum_{k=2}^{n/2} \mathbb{E}\left[X_k\right] \leq \sum_{k=2}^{n/2} \binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k)} = \sum_{k=2}^{n/2} u_k.$$

We will consider two different ranges for k in order to upper bound the u_k terms that appear in the summation $2 \leq k \leq n/2$. The reason why we do this is because the behavior of the binomial coefficient changes as k varies and we need to take care of these different behaviors carefully enough in order to prove our desired result[4]. Therefore, for $2 \leq k \leq 10$ we can upper bound u_k by a $\Theta\left(\left(\frac{\log n}{n}\right)^{k-1}\right)$ term using the very crude upper bound $\binom{n}{k} \leq n^k$. When $k \geq 10$ we can use the upper bound $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$ to upper bound u_k by a term which is $\Theta\left(n\left(\frac{\log n}{\sqrt{n}}\right)^k\right)$. Therefore, we obtain that $\sum_{k=2}^{n/2} u_k = o(1)$ and

$$\mathbf{Pr}\left[X_1 > 0\right] \leq \mathbf{Pr}\left[G \text{ is disconnected }\right] \leq \mathbf{Pr}\left[X_1 > 0\right] + o(1).$$

■

Therefore, we have obtained as a corollary that the threshold value for connectivity is $p^* = \frac{\log n}{n}$.

Corollary 2.10 *The threshold for the connectivity of $G(n,p)$ is $p^* = \frac{\log n}{n}$.*

Exercise: Use the asymptotic equivalence of $G(n,p)$ and $G(n,m)$ to find the threshold m^* for connectivity of $G(n,m)$.

[4]I am sure most of you are already familiar with this behavior due to the well-known birthday paradox. We are going to see it in detail on the blackboard, but what it says is that $e_{nk} = (1 - \frac{1}{n}) \times \ldots \times (1 - \frac{k-1}{n})$ has constant value when $k = \Theta(\sqrt{n})$ and $e_{nk} = 1 - o(1)$ when $k = o(\sqrt{n})$. Notice now that $\binom{n}{k} = \frac{n^k}{k!} e_{nk}$.