C284r: Social Data Mining **Fall 2015**

Lecture 1: Random Graphs

Lecturer: Charalampos E. Tsourakakis *December 1st, 2015*

1.1 Random Graphs

1.1.1 What is a random graph?

Formally, when we are given a graph G and we say this is a random graph, we are wrong. A given graph is fixed, there is nothing random to it. What we mean though through this term abuse is that this graph was sampled out of a set of graphs according to a probability distribution. For instance, Figure 1.1 shows the three possible graphs on vertex set $[3] = \{1, 2, 3\}$ with 2 edges. The probability distribution is the uniform, namely, each graph has the same probability $\frac{1}{3}$ to be sampled.

1.1.2 $G(n,p), G(n,m)$

- *Random binomial graphs,* $G(n,p)$: This model has two parameters, the number of vertices n and a probability parameter $0 \leq p \leq 1$. Let $\mathcal{G}$ be the family of all possible labelled graphs on the vertex set $[n]$. Notice $|\mathcal{G}| = 2^{\binom{n}{2}}$. The $G(n,p)$ model assigns to a graph $G \in \mathcal{G}$ the following probability

$$\mathbf{Pr}\left[G\right] = p^{|E(G)|}(1-p)^{\binom{n}{2}-|E(G)|}.$$

- *Uniform random graph,* $G(n,m)$: This model has two parameters, the number of vertices n and the number of edges m, where $0 \leq m \leq \binom{n}{2}$. This model assigns to all labelled graphs on the vertex set $[n]$ with exactly m edges equal probability. In other words,

$$\mathbf{Pr}\left[G\right] = \begin{cases} \frac{1}{\binom{\binom{n}{2}}{m}} & \text{if } |E(G)| = m \\ 0 & \text{if } |E(G)| \neq m \end{cases}$$

Notice that in the $G(n,p)$ model we toss a coin independently for each edge, and with probability p we add it to the graph. In expectation there will be $p\binom{n}{2}$ edges. When $p = \frac{m}{\binom{n}{2}}$, then a random binomial graph in expectation has m edges and intuitively the two models should behave similarly. For this p the two models behave similarly in a quantifiable sense. We start with the following simple observation.

Fact 1.1 *A random graph $G(n,p)$ with m edges is equally likely to be any of the $\binom{\binom{n}{2}}{m}$ graphs with m edges.*

Proof:

Consider any graph with m edges, call it G.

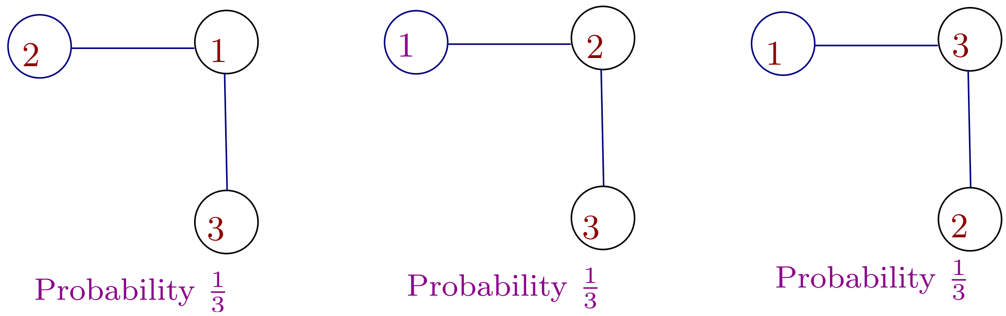

Figure 1.1: A random graph on $\{1,2,3\}$ with 2 edges with the uniform distribution

$$
\begin{aligned}
\mathbf{Pr}\left[G(n,p)=G \,|\, |E(G(n,p))|=m\right] &= \frac{\mathbf{Pr}\left[G(n,p)=G\right]}{\mathbf{Pr}\left[|E(G(n,p))|=m\right]} \\
&= \frac{p^m(1-p)^{\binom{n}{2}-m}}{\binom{\binom{n}{2}}{m}p^m(1-p)^{\binom{n}{2}-m}} \\
&= \frac{1}{\binom{\binom{n}{2}}{m}}
\end{aligned}
$$

■

Definition 1.2 *Define a graph property $\mathcal{P}$ as a subset of all possible labelled graphs. Namely $\mathcal{P} \subseteq 2^{\binom{[n]}{2}}$.*

For instance $\mathcal{P}$ can be the set of planar graphs or the set of graphs that contain a Hamiltonian cycle. We will call a property $\mathcal{P}$ as monotone increasing if $G \in \mathcal{P}$ implies $G+e \in \mathcal{P}$. For instance the Hamiltonian property is monotone increasing. Similarly, we will call a property $\mathcal{P}$ as monotone decreasing if $G \in \mathcal{P}$ implies $G-e \in \mathcal{P}$. For instance the planarity property is monotone decreasing.

Exercise: Think of other monotone increasing and decreasing properties.

Consider any monotone increasing property $\mathcal{P}$. Intuitively, the more edges the graph has, the more likely a random graph has property $\mathcal{P}$[1]. Indeed,

Lemma 1.3 *Suppose $\mathcal{P}$ is a monotone increasing property and $0 \le p_1 < p_2 \le 1$. Let $G_i \sim G(n,p_i), i=1,2$. Then,*

$$\mathbf{Pr}\left[G_1 \in \mathcal{P}\right] \le \mathbf{Pr}\left[G_2 \in \mathcal{P}\right].$$

Proof: We will generate $G_2 \sim G(n,p_2)$ from a graph $G_1 \sim G(n,p_1)$. The idea is called *coupling*. After generating G_1 we will generate a graph $G \sim G(n,p)$ and we will output the union of $G_1 \cup G$ as our G_2. We need to choose p in such way that we respect the probability distributions. To see how to choose p observe the following: an edge in G_2 does not exist with probability $(1-p_2)$. In $G_1 \cup G$ this happens with probability $(1-p)(1-p_1)$. By setting

[1] I will use interchangeably the terms a graph *has* property $\mathcal{P}$ and a graph *belongs* in $\mathcal{P}$.

$$(1-p_2) = (1-p)(1-p_1)$$

and solving for p we have achieved our goal. Given that the property is monotone increasing, we obtain the result. ■

Exercise: Prove an analog lemma for the $G(n,m)$ model.

Now we prove two facts before we give a general statement for the asymptotic equivalence of the two models.

Fact 1.4 *Let $\mathcal{P}$ be any graph property, $p = \frac{m}{\binom{n}{2}}$, where $m = m(n)$, $\binom{n}{2} - m \to +\infty$. Then, asymptotically*

$$\mathbf{Pr}\left[G(n,m) \in \mathcal{P}\right] \le \sqrt{2\pi m}\,\mathbf{Pr}\left[G(n,p) \in \mathcal{P}\right].$$

Proof:

The probability that we obtain a given graph G depends only on the number of its edges. Also notice that there exist $\binom{\binom{n}{2}}{k}$ graphs with k distinct edges, for any $0 \le k \le \binom{n}{2}$. Therefore, from the law of total probability we obtain the following expression:

$$\begin{aligned}
\mathbf{Pr}\left[G(n,p) \in \mathcal{P}\right] &= \sum_{m'=0}^{\binom{n}{2}} \mathbf{Pr}\left[|E(n,p)| = m'\right] \times \mathbf{Pr}\left[G(n,p) \in \mathcal{P} \,\middle|\, |E(n,p)| = m'\right] \\
&\ge \mathbf{Pr}\left[|E(n,p)| = m\right] \times \mathbf{Pr}\left[G(n,p) \in \mathcal{P} \,\middle|\, |E(n,p)| = m\right] \\
&= \mathbf{Pr}\left[|E(n,p)| = m\right] \times \mathbf{Pr}\left[G(n,m) \in \mathcal{P}\right].
\end{aligned}$$

It suffices to prove that

$$\mathbf{Pr}\left[|E(n,p)| = m\right] \ge \frac{1}{\sqrt{2\pi m}}.$$

For this purpose we will use Stirling's formula[2]

$$n! = (1+o(1))\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}.$$

Also, we observe that the random variable $|E(n,p)|$ is a binomial variable, i.e., $|E(n,p)| \sim Bin(\binom{n}{2}, p)$. Therefore,

$$\begin{aligned}
\mathbf{Pr}\left[|E(n,p)| = m\right] &= \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2}-m} \approx \Big(\frac{\binom{n}{2}}{2\pi m(\binom{n}{2} - m)}\Big)^{1/2} \\
&\ge \frac{1}{\sqrt{2\pi m}}.
\end{aligned}$$

■

[2]Check out this post `http://gowers.wordpress.com/2008/02/01/removing-the-magic-from-stirlings-formula/` for a neat proof by Timothy Gowers.

Exercise: The following fact is left as an exercise. You can solve it either by using the central limit theorem or by more tedious computations using appropriate asymptotic approximations.

Fact 1.5 *Let $\mathcal{P}$ be a monotonically increasing (decreasing) graph property, $p = \frac{m}{\binom{n}{2}}$. Then, asymptotically*

$$\mathbf{Pr}\left[G(n,m) \in \mathcal{P}\right] \leq 3\mathbf{Pr}\left[G(n,p) \in \mathcal{P}\right].$$

The following theorem gives precise conditions for the asymptotic equivalence of $G(n,p), G(n,m)$ [**?**], see also [**?**].

Theorem 1.6 *Let $0 \leq p_0 \leq 1, s(n) = n\sqrt{p(1-p)} \to +\infty$, and $\omega(n) \to +\infty$ as $n \to +\infty$. Then,*

(a) if $\mathcal{P}$ is any graph property and for all $m \in \mathbb{N}$ such that $|m - \binom{n}{2}p| < \omega(n)s(n)$, the probability $\mathbf{Pr}\left[G(n,m) \in \mathcal{P}\right] \to p_0$, then $\mathbf{Pr}\left[G(n,p) \in \mathcal{P}\right] \to p_0$ as $n \to +\infty$.

(b) if $\mathcal{P}$ is a monotone graph property and $p_- = p_0 - \frac{\omega(n)s(n)}{n^3}$, $p_+ = p_0 + \frac{\omega(n)s(n)}{n^3}$ then from the facts that $\mathbf{Pr}\left[G(n,p_-) \in \mathcal{P}\right] \to p_0, \mathbf{Pr}\left[G(n,p_+) \in \mathcal{P}\right] \to p_0$, it follows that $\mathbf{Pr}\left[G(n,p\binom{n}{2}) \in \mathcal{P}\right] \to p_0$ as $n \to +\infty$.

1.1.3 History

The theory of random graphs was founded by Paul Erdös and Alfred Rényi in a series of seminal papers. Erdös and Rényi studied originally the $G(n,m)$ model. Gilbert proposed the $G(n,p)$ model. Some people refer to random binomial graphs as Erdös-Rényi or Erdös-Rényi-Gilbert. Nonetheless, it was Erdös and Rényi who set the foundations of modern random graph theory.

Before the series of Erdös-Rényi papers, Erdös had discovered that the probabilistic method could be used to tackle problems whose statements were purely deterministic. For instance, one of the early uses of random graphs was in Ramsey theory. We define the Ramsey number

$$R(k,l) = \min\{n : \forall c : E(K_n) \to \{\text{red,blue}\} \exists \text{ red } K_k \text{ or blue } K_l\}.$$

Example: Prove $R(3,3) = 6$. The next challenge is to show $R(4,4) = 18$.

In one of the next lectures we will study the maximum clique in $G(n,p)$. Specifically, by studying the maximum clique size in $G(n,1/2)$, we will see why $R(k,k) \geq 2^{k/2}$. Now, let's see a proof based on the union bound.

Theorem 1.7 (Erdös, 1947)

$$R(k,k) \geq 2^{k/2}.$$

Proof: Color each edge of the complete graph K_n with red or blue by tossing a fair coin, independently from the other edges. For a fixed subset $S \subseteq [n]$, $|S| = k$ let A_S be the event that S is monochromatic, i.e., all the $\binom{k}{2}$ edges get the same color. Clearly, $\mathbf{Pr}\left[A_S\right] = 2^{1-\binom{k}{2}}$. Notice that if $\mathbf{Pr}\left[\cup_{S \subseteq V, |S|=k} A_S\right] < 1$ then the probability that none of the k-sets is monochromatic is > 0 which means that there exists a 2-coloring which violates the Ramsey property. Hence this would suggest that $R(k,k) > n$.

Based on the union bound

$$\cup_{S \subseteq V, |S|=k} \mathbf{Pr}\left[A_S\right] \leq \binom{n}{k} 2^{1-\binom{k}{2}}$$

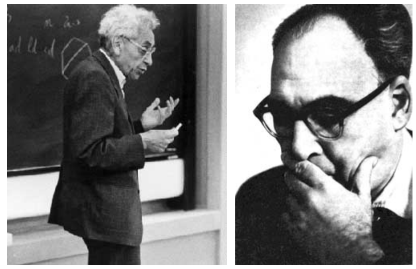

Figure 1.2: Erdös & Rényi, founders of random graph theory

we can deduce that $R(k,k) > n$ if $\binom{n}{k}2^{1-\binom{k}{2}} < 1$. When $n = \lfloor 2^{k/2} \rfloor$ then this condition holds. Let's check it.

$$\binom{n}{k}2^{1-\binom{k}{2}} < \frac{n^k}{k!}2^{1-\binom{k}{2}} < 1.$$

■

1.2 Thresholds

We formalize the notion of a threshold. begin with a formal definition of what we described the previous time.

Definition 1.8 (Threshold) *A function $p^* = p(n)$ is a threshold for a monotone increasing property*[3] $\mathcal{P}$ *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 1.9 *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.

1.3 Basic tools: First and Second moment methods

The next two elementary probabilistic tools are very powerful. Just with these tools, many non-trivial results can be proved.

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

Theorem 1.10 (Markov's Inequality) *Let X a be non-negative integer valued random variable. Then,*

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

Proof:

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

■

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 $\mathbf{Pr}\left[X > 0\right] = o(1)$, where X is a non-negative random variable of interest. It turns out that computing $\mathbb{E}\left[X\right]$ can be much easier than directly computing $\mathbf{Pr}\left[X > 0\right]$ in numerous cases. If $\mathbb{E}\left[X\right] = o(1)$ then by Markov's inequality

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

we obtain that $X = 0$ **whp** . This use is known as the *first moment method.* 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

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

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

Theorem 1.11 (Chebyshev's Inequality) *Let X be any random variable. Then,*

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

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

Corollary 1.12 (Second moment method) *Let X be a non-negative integer valued random variable. Then,*

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

For completeness, here is the proof.

Proof:

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

■

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 1.4 illustrates a use of the first and second moment methods.

1.4 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

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

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 $n \to +\infty$.

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

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

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

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

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

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

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

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

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 1.3. 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

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

Similarly, for case (b), we obtain that

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

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

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

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

1.5 More on Random Graphs

Here is a list of textbooks and other resources on random graphs.

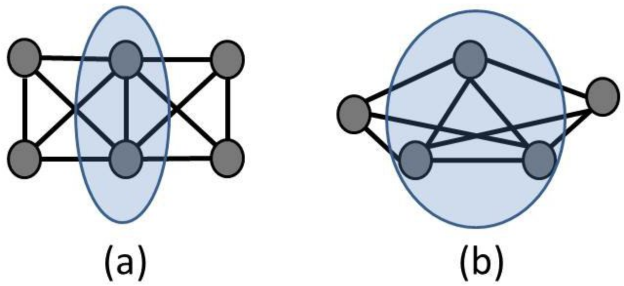

Figure 1.3: The two cases we need to consider in the covariance estimation for K_4s. Intersections of the two copies are highlighted with a shaded blue area.

- Random graphs, by *Béla Bollobás* [?]
- Complex graphs and networks, by *Fan Chung Graham and Linyuan Lu* [?]
- Random graphs, by *Svante Janson, Tomasz Luczak and Andrzej Rucinski* [?]
- Random Graph Dynamics, by *Rick Durrett* [?]
- Random Graphs and Complex Networks, by *Remco Van Der Hofstad* available online at `http://www.win.tue.nl/~rhofstad/NotesRGCN2013.pdf`
- Networks, Crowds, and Markets: Reasoning About a Highly Connected World, by *David Easley and Jon Kleinberg* available online at `http://www.cs.cornell.edu/home/kleinber/networks-book/`
- Alan Frieze's notes, available online at `http://www.math.cmu.edu/~af1p/Teaching/RandomGraphs/RandomGraphs.html`
- The Probabilistic Method by *Noga Alon and Joel Spencer* [?].