Mathematical Toolkit **Autumn 2015**

Lecture 13: November 13, 2015

Lecturer: Madhur Tulsiani

1 Threshold Phenomena in Random Graphs

We consider a model of Random Graphs by Erdős and Rényi [ER60]. To generate a random graph with n vertices, for every pair of vertices $\{i, j\}$, we put an edge independently with probability p. This model is denoted by $\mathcal{G}_{n,p}$.

Let G be a random $\mathcal{G}_{n,p}$ graph and let H be any fixed graph (on some constant number of vertices independent of n). We will be interested in understanding the probability that G contains a copy of H. We start by computing this when H is K_4, the clique on 4 vertices.

Definition 1.1 *We define* k-clique *to be a fully connected graph with* k *vertices.*

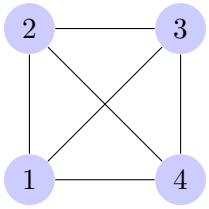

Figure 1: 4-Clique

As a convention, we will count a permutation of a copy of K_4 as the *same* copy. We define the random variable

$$Z \;=\; \text{number of copies of } K_4 \text{ in } G \;=\; \sum_C X_C\,.$$

where C ranges over all subsets of V of size 4 and the random variable X_C is defined as

$$X_C = \begin{cases} 1 & \text{if all pair of vertices in the set } C \text{ have an edge in between them} \\ 0 & \text{otherwise} \end{cases}$$

We have $\mathbb{E}[X_C] = p^6$, since the probability of connecting all 4 vertices (using 6 edges) in the 4-tuple is p^6. So we have the expectation of Z :

$$\mathbb{E}[Z] \;=\; \sum_C \mathbb{E}[X_C] \;=\; \binom{n}{4} \cdot p^6$$

We observe that,

$$\mathbb{E}[Z] \to 0 \;\text{ when } p \ll n^{-2/3} \qquad \text{and} \qquad \mathbb{E}[Z] \to \infty \;\text{ when } p \gg n^{-2/3}\,.$$

Here, by $p \ll n^{-2/3}$, we mean that $\lim_{n\to\infty}(p/n^{-2/3}) = 0$ and $p \gg n^{-2/3}$ is defined similarly. We will prove that there is in fact a threshold phenomenon in the probability that G contains a copy of K_4. When $p \ll n^{-2/3}$, the probability that a random graph G generated according to model $\mathcal{G}_{n,p}$ contains a copy of K_4, goes to 0 as $n \to \infty$. On the other hand, when $p \gg n^{-2/3}$, this probability tends to 1.

Theorem 1.2 *Let G be generated randomly according to the model $\mathcal{G}_{n,p}$ graph. We have that:*

- *If $p \ll n^{-2/3}$, then $\mathbb{P}[G$ contains a copy of $K_4] \to 0$ as $n \to \infty$.*
- *If $p \gg n^{-2/3}$, then $\mathbb{P}[G$ contains a copy of $K_4] \to 1$ as $n \to \infty$.*

Proof: As above, we define the random variable Z,

$$Z \;=\; \text{number of copies of } K_4 \text{ in } G \;=\; \sum_C X_C\,.$$

The case when $p \ll n^{-2/3}$ can be easily handled by Markov's inequality. We get that,

$$\mathbb{P}[Z > 0] \;=\; \mathbb{P}[Z \;\geq\; 1] \;\leq\; \frac{\mathbb{E}[Z]}{1}\,.$$

Since $\mathbb{E}[Z] \to 0$ as $n \to \infty$ when $p \ll n^{-2/3}$, we get that $\mathbb{P}[G$ contains a copy of $K_4] \to 0$.
When $p \gg n^{-2/3}$, we want to show that $\mathbb{P}[Z > 0] \to 1$ i.e., $\mathbb{P}[Z = 0] \to 0$. We use Chebyshev's inequality to prove this. We first compute the variance of Z.

$$\mathsf{Var}[Z] \;=\; \mathsf{Var}\left[\sum_C X_C\right] \;=\; \sum_C \mathsf{Var}[X_C] + \sum_{C\neq D} \mathsf{Cov}[X_C, X_D]$$

Since $\mathbb{E}[X_C] = p^6$, we have $\mathsf{Var}[X_C] = p^6 - p^{12}$. Also, for two distinct sets C and D, we consider four different cases depending on the number of vertices they share.

- **Case 1:** $|C \cap D| = 0$. Since no vertex is shared, X_C and X_D are independent and hence $\mathsf{Cov}[X_C, X_D] = 0$.
- **Case 2:** $|C \cap D| = 1$. Since the variables X_C and X_D depend on *pairs* of vertices in the sets C and D, and the two sets do not share any pair, we still have $\mathsf{Cov}[X_C, X_D] = 0$.
- **Case 3:** $|C \cap D| = 2$. Since C and D share a pair of vertices, there are 11 pairs which must all have edges between them in G, for both X_C and X_D to be 1. Thus, we have $\mathbb{E}[X_C X_D] = p^{11}$ and
$$\mathsf{Cov}[X_C, X_D] \;=\; \mathbb{E}[X_C X_D] - \mathbb{E}[X_C]\cdot\mathbb{E}[X_D] \;=\; p^{11} - p^{12}\,.$$
- **Case 4:** $|C \cap D| = 3$. in this case C and D share 3 pairs and thus there are 9 distinct pairs of vertices which must all have edges between them for both X_C and X_D to be 1. Thus,
$$\mathsf{Cov}[X_C, X_D] \;=\; \mathbb{E}[X_C X_D] - \mathbb{E}[X_C]\cdot\mathbb{E}[X_D] \;=\; p^{9} - p^{12}\,.$$

Also, there are $\binom{n}{6} \cdot \binom{6}{4}$ pairs C and D such that $|C \cap D| = 2$, and $\binom{n}{5} \cdot \binom{5}{4}$ pairs such that $|C \cap D| = 3$. Combining the above cases we have,

$$\begin{aligned}
\mathsf{Var}[Z] &= \sum_C \mathsf{Var}[X_C] + \sum_{C \neq D} \mathsf{Cov}[X_C, X_D] \\
&= \binom{n}{4} \cdot p^6(1-p^6) + \binom{n}{6} \cdot \binom{6}{4} \cdot (p^{11} - p^{12}) + \binom{n}{5} \cdot \binom{5}{4} \cdot (p^9 - p^{12}) \\
&= O(n^4 p^6) + O(n^6 p^{11}) + O(n^5 p^9)\,.
\end{aligned}$$

Applying Chebyshev's inequality gives

$$\begin{aligned}
\mathbb{P}[Z = 0] \;\leq\; \mathbb{P}[|Z - \mathbb{E}[Z]| \geq \mathbb{E}[Z]] \;&\leq\; \frac{\mathsf{Var}[Z]}{(\mathbb{E}[Z])^2} \\
&= \frac{1}{\binom{n}{4}^2 \cdot p^{12}} \cdot \left(O(n^4 p^6) + O(n^6 p^{11}) + O(n^5 p^9)\right) \\
&= O\left(\frac{1}{n^4 p^6}\right) + O\left(\frac{1}{n^2 p}\right) + O\left(\frac{1}{n^3 p^3}\right)\,.
\end{aligned}$$

For $p \gg n^{-2/3}$, all the terms on the right tend to 0 as $n \to \infty$. Hence, $\mathbb{P}[Z = 0] \to 0$ as $n \to \infty$. ■

The above analysis can be extended to any graph H of a fixed size. Let Z_H be the number of copies of H in a random graph G generated according to $G_{n,p}$. Using the previous analysis, we have $\mathbb{E}[Z_H] = \Theta\left(n^{|V(H)|} \cdot p^{|E(H)|}\right)$. Hence, $\mathbb{E}[Z] \to 0$ when $p \ll n^{-|V(H)|/|E(H)|}$ and $\mathbb{E}[Z] \to \infty$ when $p \gg n^{-|V(H)|/|E(H)|}$. Thus, it might be tempting to conclude that $p = n^{-|V(H)|/|E(H)|}$ is the threshold probability for finding a copy of H. However, consider the graph in Figure 2. For this

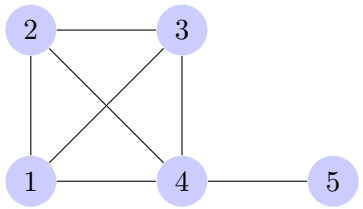

Figure 2: Subgraph H containing K_4

graph, we have $|V(H)|/|E(H)| = 5/7$. But for p such that $p \gg n^{-5/7}$ and $p \ll n^{-2/3}$, a random G is extremely unlikely to contain a copy of K_4 and thus also extremely unlikely to contain a copy of H. For an arbitrary graph H, it was shown by Bollobás [Bol81] that the threshold probability is $n^{-\lambda}$, where

$$\lambda \;=\; \min_{H' \subseteq H} \frac{|V(H')|}{|E(H')|}\,.$$

2 Chernoff/Hoeffding Bounds

We now derive sharper concentration bounds for sums of independent random variables. We start by considering n *independent* boolean random variables $X_1, ..., X_n$, where X_i takes value 1 with probability p_i and 0 otherwise. Let $Z = \sum_{i=1}^n X_i$. We set $\mu = \mathbb{E}[Z] = \sum_{i=1}^n \mathbb{E}[X_i] = \sum_{i=1}^n p_i$.

We will try to derive a bound on the probability $\mathbb{P}[Z \geq t]$ for $t = (1+\delta)\mu$. Using the fact that the function e^x is strictly increasing, we get that for $\lambda > 0$

$$\mathbb{P}[Z \geq (1+\delta)\mu] \;=\; \mathbb{P}\left[e^{\lambda Z} \geq e^{\lambda(1+\delta)\mu}\right] \overset{(Markov)}{\leq} \frac{\mathbb{E}\left[e^{\lambda Z}\right]}{e^{\lambda(1+\delta)\mu}}.$$

We now have:

$$\begin{aligned}\mathbb{E}\left[e^{\lambda Z}\right] \;=\; \mathbb{E}\left[e^{\lambda(X_1+\ldots X_n)}\right] \;=\; \mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda X_i}\right] &\overset{(independence)}{=} \prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda X_i}\right] \\ &= \prod_{i=1}^{n}\left[p_i e^{\lambda} + (1-p_i)\right] \\ &= \prod_{i=1}^{n}\left[1 + p_i(e^{\lambda}-1)\right].\end{aligned}$$

At this point, we utilize the simple but very useful inequality:

$$\forall x \in R,\;\; 1 + x \leq e^x.$$

Since all the quantities in the previous calculation are non-negative, we can plug the above inequality in the previous calculation and we get:

$$\mathbb{E}\left[e^{\lambda Z}\right] \;\leq\; \prod_{i=1}^{n}\exp\Big((e^{\lambda}-1)p_i\Big) \;=\; \exp\Big((e^{\lambda}-1)\mu\Big)$$

Thus, we get

$$\mathbb{P}[Z \geq (1+\delta)\mu] \;\leq\; \exp\Big((e^{\lambda}-1)\mu - \lambda(1+\delta)\mu\Big).$$

We now want to minimize the right hand-side of the above inequality, with respect to λ. Setting the derivative of the exponent to zero, we get

$$e^{\lambda}\mu - (1+\delta)\mu = 0 \quad\Rightarrow\quad \lambda = \ln(1+\delta)\,.$$

Using this value for λ, we get

$$\mathbb{P}[Z \geq (1+\delta)\mu] \;\leq\; \frac{\exp\Big((e^{\lambda}-1)\mu\Big)}{\exp\Big(\lambda(1+\delta)\mu\Big)} \;=\; \frac{e^{\delta\mu}}{(1+\delta)^{(1+\delta)\mu}} \;=\; \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}.$$

Similarly, we can get that

$$\mathbb{P}[Z \leq (1-\delta)\mu] \;\leq\; \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}.$$

(Note that $\mathbb{P}[Z \leq (1-\delta)\mu] = \mathbb{P}\left[e^{-\lambda Z} \geq e^{-\lambda(1-\delta)\mu}\right]$.)

When $\delta \in (0,1)$, the bounds above expressions can be simplified further. It is easy to check that

$$\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} \leq e^{-\delta^2\mu/3},\;\; 0 < \delta < 1\,.$$

So we get:

$$\mathbb{P}\left[Z \geq (1+\delta)\mu\right] \leq e^{-\delta^2\mu/3}, \quad \text{for } 0 < \delta < 1.$$

Similarly:

$$\mathbb{P}\left[Z \leq (1-\delta)\mu\right] \leq e^{-\delta^2\mu/3}, \quad \text{for } 0 < \delta < 1.$$

Combining the two we get

$$\mathbb{P}\left[|Z-\mu| \geq \delta\mu\right] \leq 2e^{-\delta^2\mu/3}, \quad \text{for } 0 < \delta < 1.$$

2.1 Coin tosses one more

We will now compare the above bound with what we can get from Chebyshev's inequality. Let's assume that $X_1, ..., X_n$ are independent coin tosses, with $\mathbb{P}[X_i = 1] = \frac{1}{2}$. We want to get a bound on the value of $Z = \sum_{i=1}^{n} X_i$. Using Chebyshev's inequality as stated in (**??**), we get that

$$\mathbb{P}\left[|Z-\mu| \geq \delta\mu\right] \;\leq\; \frac{\mathsf{Var}\left[Z\right]}{\delta^2\mu^2}.$$

And since in this particular case we have that $\mathsf{Var}[Z] = n/4$ and $\mu = n/2$, we get that

$$\mathbb{P}\left[|Z-\mu| \geq \delta\mu\right] \;\leq\; \frac{1}{\delta^2 n}.$$

The above bound is only inversely polynomial in n, while the Chernoff-Hoeffding bound gives

$$\mathbb{P}\left[|Z-\mu| \geq \delta\mu\right] \;\leq\; 2 \cdot \exp\left(-\delta^2 n/6\right),$$

which is exponentially small in n. This fact will prove very useful when taking a union bound over a large collection of events, each of which may be bounded using a Chernoff-Hoeffding bound.

3 Balanced Allocations

We consider the following problem of allocating jobs to servers: We are given a set of n servers $1, \ldots, n$ and m jobs arrive one by one. We seek to assign each job to one of the servers so that the loads placed on all servers are as balanced as possible.

In developing simple, effective load balancing algorithms, randomization often proves to be a useful tool. We consider two approaches for this problem:

- **Random Choice**: one possible way to distribute the jobs is to simply place each job on a random server, chosen independently of the previous allocations.
- **Two Random Choices**: For each job, we choose two servers independently and uniformly at random and place the job on the server with less load (breaking ties arbitrarily).

We will show that using two random choices significantly reduces the maximum load on any server. For the entire analysis, we will work with the case when $m = n$. The analysis easily extends to

an arbitrary m, but it easier to appreciate the bounds when $m = O(n)$ (and in particular when $m = n$).

It is convenient to think of the above in terms of the so called "balls and bins" model. Each job can be thought of a s ball and each server is a bin. We think of assigning job j to a server i as throwing the j^{th} ball in bin i. The load of a server is the same as the number of balls in the corresponding bin.

3.1 Random choice

Suppose $Z_i =$ number of balls in bin i. We can write

$$Z_i = \sum_j X_{ij}, \quad \text{where} \quad X_{ij} = \begin{cases} 1 & \text{if ball } j \text{ is thrown in bin } i \\ 0 & \text{otherwise} \end{cases} .$$

Then, we have that each Z_i is a sum of $m(= n)$ independent random variables with $\mathbb{E}[Z_i] = 1$. Let $K = \frac{3 \ln n}{\ln \ln n}$. By Chernoff/Hoeffding bounds, we have that for each i,

$$\mathbb{P}[Z_i \geq K] \;\leq\; \left(\frac{e}{K}\right)^K .$$

To bound the maximum load in across all bins, we use a union bound to say that

$$\mathbb{P}[\exists i \in [n].\ Z_i \geq K] \;\leq\; \sum_{i=1}^{n} \mathbb{P}[Z_i \geq K] \;\leq\; n \cdot \left(\frac{e}{K}\right)^K ,$$

which is at most $\frac{1}{n}$ for the above value of K. Hence, with probability at least $1 - \frac{1}{n}$, the maximum number of balls in a bin is at most $\frac{3 \ln n}{\ln \ln n}$.

References

[Bol81] Béla Bollobás, *Threshold functions for small subgraphs*, Mathematical Proceedings of the Cambridge Philosophical Society, vol. 90, Cambridge Univ Press, 1981, pp. 197–206.

[ER60] Paul Erdős and A Rényi, *On the evolution of random graphs*, Publ. Math. Inst. Hungar. Acad. Sci **5** (1960), 17–61.