CS174 **Lecture 9** **John Canny**

Random Graphs

We first need to define what we mean by a random graph. We will do so with an algorithm. To get a random graph $G(n, m)$ with m edges and n vertices, we do the following:

```
Algorithm RandG1(n, m)
  for i = 1 to m do
    pick i,j in {1,...,n} at random until {i,j} is not in E
    add edge {i,j} to E
```

This process also allows us to think of building a random graph one edge at a time. We can ask a variety of questions about random graphs. Is G connected? Does G have a k-clique? A hamiltonian path? etc. In the probabilistic case, the sample space is the set of possible graphs, and an experiment is generating a graph from this space. Each of the questions above is an event (subset of the sample space), and the answer to those questions the probability of that event.

Is G connected?

We now have a full bag of tools to approach this problem. To get the probability that the random graph $G(n, m)$ is connected, we first find the expected value of m that makes the graph connected, and then apply tail bounds to compute the probability of this happening for particular m. We assume an incremental model of adding one edge at a time.

First notice that $m \geq n - 1$ for a connected graph (a tree is a minimally-connected graph). But we use one of the results we already know to get a much stronger bound for random graphs. Hint: Try thinking of choosing random edges as a form of coupon collecting.

As we add edges, we watch the number of connected components of the graph. Initially, the graph has n vertices and no edges, so there are n connected components. The first edge always connects two points, and gives us $n - 1$ connected components. The second edge also reduces the number of connected components to $n - 2$. The third may or may not reduce the number. We use epochs to model the different phases of the process. Let X_k be the number of random edges added while there are k connected components, until there are $k - 1$ connected components. We have shown that $X_n = 1$, and $X_{n-1} = 1$. If we define

$$X = \sum_{k=2}^{n} X_k$$

then X counts the total number of edges that we add until the graph is connected. Our goal then, is to compute $\mathrm{E}(X)$.

Now define p_k to be the probability that an edge added while there are k components reduces the number of components. We cant compute p_k exactly, but we can give a lower bound. Assume

v is one endpoint of the edge we are adding. Then there are at least $k-1$ other vertices to which we can connect v and reduce the number of components (these other vertices lie on the other components). In total there are $n-1$ other vertices to which we can connect v. So the probability that this edge reduces the number of components is $\geq (k-1)/(n-1)$. But this bound holds for any choice of v, so it also bounds p_k:

$$p_k \geq \frac{k-1}{n-1}$$

Now observe that X_k is a geometric random variable with success probability p_k. Its expected value is $1/p_k \leq (n-1)/(k-1)$. So we have

$$\mathrm{E}\left(X\right) = \sum_{k=2}^{n} \mathrm{E}\left(X_k\right) \leq \sum_{k=2}^{n} \frac{n-1}{k-1} = (n-1)H_{n-1}$$

where H_{n-1} is the $(n-1)^{st}$ harmonic number. In other words, an upper bound on $\mathrm{E}\left(X\right)$ is about $n \ln n$.

The next step is to apply tail bounds on the probability of m being much larger than its mean. To get a useful bound, we need to apply Chebyshev. Since X is a sum of independent R.V.'s (strictly speaking they are not independent, but their probability bounds are independent), we can add up their variances. Each X_k is a geometric random variable with success probability p_k. So its variance (lecture 8) is $(1-p_k)/p_k^2$. Then

$$\mathrm{Var}\left[X\right] = \sum_{k=2}^{n} \mathrm{Var}\left[X_k\right] = \sum_{k=2}^{n} \frac{1-p_k}{p_k^2} \leq \sum_{k=2}^{n} \frac{(n-k)(n-1)}{(k-1)^2}$$

and we can split up this sum:

$$\sum_{k=2}^{n} \frac{(n-k)(n-1)}{(k-1)^2} = (n-1)^2 \sum_{k=2}^{n} \frac{1}{(k-1)^2} - (n-1) \sum_{k=2}^{n} \frac{1}{(k-1)}$$

We have seen both kinds of sums on the RHS before (lecture 8), and they can be approximated respectively as:

$$n^2\pi^2/6 - n \ln n$$

and therefore σ_X is at most $\approx n\pi/\sqrt{6}$.

To apply Chebyshev, we set the probability of exceeding the mean at 0.01, then $t = 10$ in the Chebyshev formula:

$$\Pr\left[|X - \overline{X}| \geq t\sigma_X\right] \leq \frac{1}{t^2}$$

which requires that $X - \overline{X} \geq t\sigma_X$ or

$$X \geq n \ln n + 10n\pi/\sqrt{6}$$

So just as we saw for coupon collecting, we have very high probability of connecting up the graph (better than 0.99) when the number of edges m is a linear multiple of n bigger than $\overline{X}$ which is $n \ln n$.

Does G have a k-clique?

Definition: A clique in an undirected graph $G = (V, E)$ is a subset of vertices $U \subset V$ such that every pair of vertices in U is connected by an edge of G (i.e., for all $i \neq j \in U$, we have $\{i, j\} \in E$. If U has k vertices, we call it a k-clique.

Finding cliques in graphs, and in particular large cliques, is an important problem that shows up in many applications. Given G and k, the problem of deciding whether G contains a k-clique is NP-complete. Here we investigate the problem for random graphs. We'll use a different model called the $G_{n,p}$ model for random graphs. Rather than fixing the number of edges, we fix the probability of each edge being included in the graph. To generate a $G_{n,p}$ graph, we do the following:

```
Algorithm RandG2(n, p)
  for every pair {i < j} in {1,...,n}, do
    toss a coin with Pr[Heads] = p
    if heads, add edge {i,j} to E
```

Notice that the expected number of edges in such a random graph is $\binom{n}{2}p$, which is the number of possible edges times p. So by varying p, we get more or less dense graphs. The following question is typical in the fields of random graphs and average-case analysis of algorithms:

- How large does p have to be before a random graph G is very likely to contain a 4-clique?

We approach this problem in the usual way, define indicator random variables for each subset of 4 vertices that indicate the presence of a clique. That is, define X_S for each subset $S \subset V$ of 4 vertices as:

$$X_S = \begin{cases} 1 & \text{if } S \text{ are the vertices of a 4-clique} \\ 0 & \text{otherwise} \end{cases}$$

and then $X = \sum X_S$ is the total number of 4-cliques in the graph. We will first estimate $\mathrm{E}\,(X)$ as a function of p. Then we will compute the variance of X and use the Chebyshev bound to show that p is a "threshold parameter". That is, there is a value p_0 such that for $p > p_0$ there almost certainly is a 4-clique, while for $p < p_0$ there is almost certainly not a 4-clique in G.

First of all, since a 4-clique has 6 edges, it is easy to see that $\Pr[X_S = 1] = p^6$. Since X_S is an indicator r.v., we also have that $\mathrm{E}\,(X_S) = \Pr[X_S = 1] = p^6$. The total number of subsets of 4 vertices is $\binom{n}{4}$, so

$$\mathrm{E}\,(X) = \sum \mathrm{E}\,(X_S) = \binom{n}{4} p^6$$

Its tempting to infer that if $\mathrm{E}\,(X) > 1$ then G is very likely to contain a clique (this happens for $p > 1.7n^{-2/3}$). But that is not necessarily true. Its possible that the distribution of G has a "long tail" of low total probability that inflates the expected value, but has low total probability for $X > 0$. To prove that we have high probability of a clique, we need to compute the variance and apply Chebyshev. Now we know that

$$\mathrm{Var}\,(X) = \mathrm{E}\,(X^2) - \mathrm{E}\,(X)^2 = \sum_S \mathrm{E}\,(X_S^2) - \sum_S \mathrm{E}\,(X_S)^2 + \sum_{S \neq T} \mathrm{E}\,(X_S X_T) - \sum_{S \neq T} \mathrm{E}\,(X_S)\mathrm{E}\,(X_T)$$

Definition: The covariance $\mathrm{Cov}\,(X_S, X_T)$ of two random variables is defined as $\mathrm{E}\,(X_S X_T) - \mathrm{E}\,(X_S)\mathrm{E}\,(X_T)$.

For independent random variables X_S and X_T, the covariance is zero. With this definition, the variance of X can be written:

$$\mathrm{Var}\,(X) = \sum_S \mathrm{Var}\,(X_S) + \sum_{S \neq T} \mathrm{Cov}\,(X_S, X_T)$$

Now the variance of X_S is simple, because it is represents a Bernoulli trial with success probability p^6. From the earlier formula for variance of a Bernoulli trial (lecture 6), we have:

$$\mathrm{Var}\,(X_S) = p^6(1 - p^6)$$

The covariances are tricky, and they vary depending on the degree of similarity between S and T. So we consider three cases:

S **and** T **have 0 or 1 vertices in common.** In this case, there are no edges in common in the (possible) 4-cliques on S and T. Then X_S and X_T are independent, so the covariance $\mathrm{Cov}\,(X_S, X_T)$ is zero.

S **and** T **have 2 vertices in common.** In this case, there is one edge in common in the (possible) 4-cliques on S and T. If $X_S X_T = 1$ then a total of 11 edges (6 each for S and T, less the common edge) must be present. So $\mathrm{E}\,(X_S X_T) = \Pr[X_S X_T = 1] = p^{11}$. Then

$$\mathrm{Cov}\,(X_S, X_T) = \mathrm{E}\,(X_S X_T) - \mathrm{E}\,(X_S)\mathrm{E}\,(X_T) = p^{11} - p^{12}$$

S **and** T **have 3 vertices in common.** In this case, there are 3 edges in common in the (possible) 4-cliques on S and T. If $X_S X_T = 1$ then a total of 9 edges (6 each for S and T, less the 3 common edges) must be present. So $\mathrm{E}\,(X_S X_T) = \Pr[X_S X_T = 1] = p^9$. Then

$$\mathrm{Cov}\,(X_S, X_T) = \mathrm{E}\,(X_S X_T) - \mathrm{E}\,(X_S)\mathrm{E}\,(X_T) = p^9 - p^{12}$$

Since the first case had zero covariance, we only need to consider the last two cases. To evaluate the sums $\sum \mathrm{Cov}\,(X_S, X_T)$ we need to count the number of pairs S, T. In the case of two vertices in common, we can choose S first, then the two vertices in common, then the other two vertices in T. The number of ways of doing that is

$$\binom{n}{4}\binom{4}{2}\binom{n-4}{2} \approx \frac{n^6}{8}$$

In the case of three vertices in common, we can choose S first, then the three in common, then the one other vertex of T. The number of pairs is

$$\binom{n}{4}\binom{4}{3}\binom{n-4}{1} \approx \frac{n^5}{6}$$

The number of sets of S alone is $\binom{n}{4} \approx n^4/24$. Now we can substitute into the formula for variance of X:

$$\mathrm{Var}\,(X) = \sum_S \mathrm{Var}\,(X_S) + \sum_{S \neq T} \mathrm{Cov}\,(X_S, X_T) \approx \frac{n^4}{24} p^6(1 - p^6) + \frac{n^6}{8}(p^{11} - p^{12}) + \frac{n^5}{6}(p^9 - p^{12})$$

Earlier, we noted that $p = 1.7n^{-2/3}$ gives $\mathrm{E}(X)$ of about 1. We introduce a constant c, and plug $p = cn^{-2/3}$ into the variance formula:

$$\mathrm{Var}(X) = \frac{c^6}{24} + O(n^{-1})$$

and the standard deviation is:

$$\sigma_X \approx \frac{c^3}{2\sqrt{3}}$$

Substituting $p = cn^{-2/3}$ into the expected value formula gives:

$$\mathrm{E}(X) = \binom{n}{4} p^6 \approx \frac{c^6}{24}$$

Now we can see that the distribution of X "converges" as c increases. That is, the expected value grows as c^6, while the standard deviation (the width of the distribution) grows as c^3.

To apply Chebyshev, pick say $t = 10$. We choose $\overline{X} = 101$, and solving for c gives $c = 2424^{1/6} \approx 3.665$. Then $\sigma_X \approx 10.0$, and Chebyshev gives

$$\Pr[X < 1] = \Pr[X - \overline{X} < -t\sigma_X] \le \Pr[|X - \overline{X}| > t\sigma_X] \le \frac{1}{t^2} = \frac{1}{100}$$

So there is almost certainly (prob > 0.99) a 4-clique if $p = 3.665n^{-2/3}$.

On the other hand, if we pick $t = 10$ and $\overline{X} = 0.009$, solving for c gives $c = 0.776$. The standard deviation is $\sigma_X \approx 0.095$. Then

$$\Pr[X > 0.959] = \Pr[X - 0.009 > 0.95] \le \Pr[|X - \overline{X}| > t\sigma_X] \le \frac{1}{t^2} = \frac{1}{100}$$

So there is almost certainly not (prob < 0.01) a 4-clique if $p = 0.776n^{-2/3}$.

This is what we mean by $p_0 = 1.667n^{-2/3}$ is a "threshold value". There is almost certainly a clique for p larger than p_0, and almost certainly no clique for values of p less than p_0.