Based on earlier notes by Alistair Sinclair/Manuel Blum/Douglas Young.

More on Random Graphs: Threshold Phenomena

Definition: A clique in an undirected graph $G = (V, E)$ is a subset of vertices $U \subseteq V$ such that every pair of vertices in U is connected by an edge of G (i.e., for all $i, j \in U$, 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 (for worst-case graphs). Here we investigate the problem for random graphs.

We'll use the so-called $\mathcal{G}_{n,p}$ model for random graphs, which you've already seen in the sample midterm. Recall that the experiment here is as follows:

- start with vertex set $V = \{1, 2, \ldots, n\}$ and edge set $E = \emptyset$
- for each of the $\binom{n}{2}$ possible edges $e = \{i, j\}$ independently, flip a coin with heads probability p; if the coin comes up heads, add e to E
- output $G = (V, E)$

Notice that the expected number of edges in such a random graph is $\binom{n}{2}p$. (Why?) 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 shall see in a moment that the answer is remarkable: there is a sharply defined threshold value of p such that, if p is above this value then G is almost certain to contain a 4-clique, and if p is below it then G is almost certain *not* to contain one. (Note that this is much sharper than what we proved about the number of edges needed for a graph to be connected in Note 4, although in fact a similar threshold result can be proved for that problem with a bit more work.)

The discovery of threshold phenomena is another part of the legacy of Paul Erdös. This calculation will also introduce some important new tricks that are generally useful.

For a random graph G as above, define the r.v. $X = \#$ 4-cliques in G.

Obviously, $X = \sum_S X_S$, where the sum is over all subsets $S \subseteq V$ of four vertices, and

$$X_S = \begin{cases} 1 & \text{if } S \text{ is a clique in } G; \\ 0 & \text{otherwise.} \end{cases}$$

Evidently, $\mathrm{E}(X_S) = \Pr[S \text{ is a clique}] = p^6$. (Why?) Hence $\mathrm{E}(X) = \sum_S \mathrm{E}(X_S) = \binom{n}{4}p^6 \sim \frac{n^4p^6}{24}$.

Now it is clear that $\mathrm{E}(X)$ drops below 1 when $p \approx \text{const} \times n^{-2/3}$. In fact, we can say something stronger:

- if $p \gg n^{-2/3}$ then $\mathrm{E}(X) \to \infty$ as $n \to \infty$;
- if $p \ll n^{-2/3}$ then $\mathrm{E}(X) \to 0$ as $n \to \infty$.

(The notation $p \gg n^{-2/3}$ here means that $\frac{p}{n^{-2/3}} \to \infty$ as $n \to \infty$, and symmetrically for $\ll$.)
So it is tempting for us to conclude from this that $p = n^{-2/3}$ is a threshold value for G to contain a 4-clique. Unfortunately life isn't quite that simple: the above statements only refer to *expectations*, whereas we want a much stronger result about probabilities. Specifically, we want to show:

1. If $p \gg n^{-2/3}$ then $\Pr[G \text{ contains a 4-clique}] \to 1$ as $n \to \infty$.
2. If $p \ll n^{-2/3}$ then $\Pr[G \text{ contains a 4-clique}] \to 0$ as $n \to \infty$.

Now it turns out that Statement 2 follows directly from the expectation result, because the r.v. X is non-negative integer-valued.

Ex: Before reading the proof below, think about this: if X takes on only non-negative integer values and $\mathrm{E}(X) \to 0$, is it possible for X to be > 0 (i.e., ≥ 1) with anything more than vanishingly small probability? $\square$

Proof of Statement 2

By Markov's inequality,

$$\Pr[X > 0] = \Pr[X \geq 1] \leq \mathrm{E}(X) \sim \tfrac{n^4 p^6}{24} \to 0 \qquad \text{if } p \ll n^{-2/3}. \qquad \square$$

Statement 1 presents more of a challenge: in fact, it definitely does <u>not</u> follow just from the fact that $\mathrm{E}(X) \to \infty$. The problem is that we could have $X = 0$ for most graphs but <u>still</u> have $\mathrm{E}(X) \to \infty$ if it happened that, when G does contain a 4-clique, it contains a large number of them. (I.e., X could be 0 with high probability, and much bigger than $\mathrm{E}(X)$ with the remaining small probability. Note that large values of X are possible: X could be as big as $\binom{n}{4}$.) You should make sure you understand this important point.

So, we need some more information about X. Essentially, we need to show that <u>large deviations</u> from the expected value $\mathrm{E}(X)$ are extremely unlikely. This is exactly what the <u>variance</u> measures! This fact is expressed conveniently by Chebyshev's inequality:

Theorem (Chebyshev's Inequality): For any r.v. X,

$$\Pr[\,|X - \mathrm{E}(X)\,| \geq \alpha] \leq \frac{\mathrm{Var}(X)}{\alpha^2}.$$

Proof: Let $\mu = \mathrm{E}(X)$ and define the r.v. $Y = (X - \mu)^2$. Note that Y is non-negative, so we can apply Markov's inequality to it. We get

$$\Pr[\,|X - \mu| \geq \alpha] = \Pr[Y \geq \alpha^2] \leq \frac{\mathrm{E}(Y)}{\alpha^2} = \frac{\mathrm{Var}(X)}{\alpha^2}. \qquad \square$$

Note: Chebyshev's inequality is often written in the equivalent form

$$\Pr[\,|X - \mathrm{E}(X)\,| \geq c\mathrm{E}(X)] \leq \frac{1}{c^2} \frac{\mathrm{Var}(X)}{\mathrm{E}(X)^2}. \qquad \square$$

To use Chebyshev's inequality for our problem, we need to compute $\mathrm{Var}(X)$. We'll do this in a moment, but first let's preview the result.

Claim: For the r.v. $X = \#$ 4-cliques in G, if $p \gg n^{-2/3}$ then $\frac{\mathrm{Var}(X)}{\mathrm{E}(X)^2} \to 0$ as $n \to \infty$.

Proof of Statement 1

Using this fact, together with Chebyshev's inequality, allows us to prove Statement 1 above. For we have

$$\Pr[X = 0] \le \Pr[\,|X - \mathrm{E}(X)\,| \ge \mathrm{E}(X)] \le \frac{\mathrm{Var}(X)}{\mathrm{E}(X)^2} \to 0 \quad \text{as } n \to \infty,$$

assuming that $p \gg n^{-2/3}$. $\square$

It remains only for us to carry out the variance calculation to prove the above Claim.

Proof of Claim: Recall that $X = \sum_S X_S$. In this case, we will expand $\mathrm{Var}(X)$ a bit differently from the way we did in Note 2. The reason for this is that, here, most of the pairs of X_S's are independent, which allows us to ignore their contribution to $\mathrm{Var}(X)$. Specifically, we will write

$$\begin{aligned}\mathrm{Var}(X) = \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2 &= \sum_S \mathrm{E}\left(X_S^2\right) + \sum_{S \ne T} \mathrm{E}(X_S X_T) - \sum_S \mathrm{E}(X_S)^2 - \sum_{S \ne T} \mathrm{E}(X_S)\,\mathrm{E}(X_T) \\ &= \sum_S \mathrm{Var}(X_S) + \sum_{S \ne T} \mathrm{Cov}(X_S, X_T), \qquad (*)\end{aligned}$$

where for two r.v.'s Y, Z, the <u>covariance</u> $\mathrm{Cov}(Y,Z) = \mathrm{E}(YZ) - \mathrm{E}(Y)\,\mathrm{E}(Z)$. (Note that $\mathrm{Cov}(Y,Y) = \mathrm{Var}(Y)$; note also that $\mathrm{Cov}(Y,Z) = 0$ if Y, Z are independent.)

Now since X_S is a 0-1 r.v., we have

$$\mathrm{Var}(X_S) = \mathrm{E}(X_S) - \mathrm{E}(X_S)^2 = p^6 - p^{12} \le p^6.$$

Since there are $\binom{n}{4} = O(n^4)$ sets S, the total contribution of the $\mathrm{Var}(X_S)$ terms in $(*)$ is $O(n^4 p^6)$.

What about the covariance terms $\mathrm{Cov}(X_S, X_T)$ for two distinct sets S, T? We consider three cases.

Case 1: S, T have zero or one points in common.

In this case, a moment's thought should convince you that X_S, X_T are <u>independent</u> (why???), so $\mathrm{Cov}(X_S, X_T) = 0$. So these terms contribute 0 to $(*)$.

Case 2: S, T have two points in common.

In this case $\mathrm{E}(X_S X_T) = p^{11}$, so $\mathrm{Cov}(X_S, X_T) = p^{11} - p^{12} \le p^{11}$.
Moreover, the number of pairs S, T of this form is $\binom{n}{6}\binom{6}{2}\binom{4}{2} = O(n^6)$.
So the total contribution of these covariance terms to $(*)$ is $O(n^6 p^{11})$.

Case 3: S, T have three points in common.

In this case $\mathrm{E}(X_S X_T) = p^9$, so $\mathrm{Cov}(X_S, X_T) = p^9 - p^{12} \le p^9$.
And the number of pairs S, T of this form is $\binom{n}{5}\binom{5}{3}\binom{2}{1} = O(n^5)$.

So the total contribution of these covariance terms to $(*)$ is $O(n^5p^9)$.

Putting all this together in $(*)$ gives us

$$\mathrm{Var}\,(X) = O(n^4p^6) + O(n^6p^{11}) + O(n^5p^9) = n^4p^6 O(1 + n^2p^5 + np^3).$$

Now notice that, since $p \gg n^{-2/3}$, each of the three terms inside this last O-expression is $o(n^4p^6)$ (i.e., its ratio with n^4p^6 tends to 0 as $n \to \infty$). (You should check this.) Hence $\mathrm{Var}\,(X) = o((n^4p^6)^2) = o(\mathrm{E}\,(X)^2)$, which is exactly the statement of the Claim. $\square$

The size of the largest clique

- Given a random graph G in the model $\mathcal{G}_{n,1/2}$ (i.e., where each edge is equally likely to be present or absent), how large is the largest clique in G likely to be?

This question has a surprisingly precise answer: essentially, almost all such graphs have a largest clique of size about $2\lg n$. (In fact the answer is even more precise: for all sufficiently large n, there is an integer m (depending on n) such that the largest clique in G almost certainly has size either m or $m+1$.)

How would we go about answering this question? Well, we would start by defining the r.v. $X_k = \#$ k-cliques in G.

Then $X_k = \sum_S X_S$, where $X_S = \begin{cases} 1 & \text{if } S \text{ is a } k\text{-clique;} \\ 0 & \text{otherwise.} \end{cases}$

Here the sum is over all subsets S of k vertices.

So we get $\mathrm{E}(X_k) = \sum_S \mathrm{E}(X_S) = \binom{n}{k} 2^{-\binom{k}{2}}$. (Why?)

We are interested in the value of k at which $\mathrm{E}(X_k)$ drops below 1. Let's do a bit of algebra:

$$\mathrm{E}(X_k) = \binom{n}{k} 2^{-\binom{k}{2}} \le \left(\tfrac{n\mathrm{e}}{k}\right)^k 2^{-\binom{k}{2}} = \left(\tfrac{n\mathrm{e}}{k2^{(k-1)/2}}\right)^k.$$

For this to be less than 1 we require $n\mathrm{e} < k2^{(k-1)/2}$, or, taking logs, $\lg n + \lg \mathrm{e} < \lg k + \frac{k-1}{2}$. Tidying this up leaves us with the condition

$$k > 2\lg n - 2\lg k + O(1),$$

where $O(1)$ denotes a term that is bounded by a constant as $n \to \infty$.

Now it is easy to check that setting $k = 2\lg n - 2\lg\lg n + O(1)$ satisfies this inequality. Call this value k_0.

In other words, if $k \gg k_0$ then $\mathrm{E}(X_k) \to 0$ as $n \to \infty$.

But now, just as in the proof of Statement 2 in the 4-cliques example, this immediately implies that

$$\Pr[G \text{ contains a } k\text{-clique}] \to 0 \quad \text{as } n \to \infty.$$

What we'd now like to do is to argue, as in the 4-cliques example, that if $k \ll k_0$ then

$$\Pr[G \text{ contains a } k\text{-clique}] \to 1 \quad \text{as } n \to \infty.$$

This turns out to be true, and once again it relies on the fact that the variance of X_k is small. This calculation is quite messy, so we won't do it here. (Look in Chapter 4, Section 5 of the Alon & Spencer book if you are interested.)

The bottom line is:

The size of the largest clique in a random graph G in the $\mathcal{G}_{n,1/2}$ model is within an additive constant of $2\lg n - 2\lg\lg n$ with probability tending to 1 as $n \to \infty$. □

Amazingly, although a clique of size about $2\lg n$ is almost certain to exist in a random graph, nobody has been able to design an algorithm that runs in reasonable time and provably (or even in practice) finds a clique of size even $(1+\epsilon)\lg n$ with reasonable probability! This is an intriguing algorithmic challenge.