Lecture 7 (Nov. 22, 2011)

We begin by introducing the *Erdős-Renyi random graph model.* Actually, it was already introduced by Jeff when he lectured on local coloring two weeks ago, but now here it is for the first time in the actual lecture notes. To begin with, an informal definition :

Definition 16 : Let n be a positive integer and $p \in [0, 1]$. The *random graph* $G(n, p)$ has n vertices and is obtained by choosing each of the $\left(\begin{array}{c} n \\ 2 \end{array} \right)$ possible edges randomly and independently with probability p.

The informal terminology "random graph" is a bit misleading, because strictly speaking a random graph is not a graph at all, it is a probability space. To be able to say what space, we first need to introduce the notion of a *product measure* :

Definiton 17 : Let (Ω, μ) be a discrete probability space and suppose the underlying set Ω is a Cartesian product $\Omega = \Omega_1 \times \cdots \times \Omega_k$ of k sets. Then the measure μ is called a *product measure* if there exist probability measures μ_i on Ω_i, for $i = 1, ..., k$, such that, for any point $(\omega_1, ..., \omega_k) \in \Omega$,

$$\mu[\{(\omega_1, ..., \omega_k)\}] = \prod_{i=1}^{k} \mu_i[\{\omega_i\}].$$

In this case, we write that

$$\mu = \prod_{i=1}^{k} \mu_i.$$

We can now give a more formal definition of a random graph :

Definition 18 : The *random graph* $G(n, p)$ is the probability space (Ω, μ), where $\Omega = \{0, 1\}^{\left(\begin{array}{c} n \\ 2 \end{array} \right)}$, i.e.: Ω is the Cartesian product of $\left(\begin{array}{c} n \\ 2 \end{array} \right)$ copies of the two-element set $\{0, 1\}$, and $\mu = \prod \mu_p$, i.e.: the product of the same number of copies of the measure μ_p on $\{0, 1\}$ given by

$$\mu_p(\{0\}) = 1 - p, \qquad \mu_p(\{1\}) = p.$$

In the study of random graphs, one of the most important concepts is that of a *threshold function.*

Notation : Let A be a graph property and G a graph. We write $G \models A$ to denote the fact that G has the property A. For example, if A is the property "is connected", then $G \models A$ means that G is connected.

Definition 19 : Let A be a graph property and $t : \mathbf{N} \to [0,1]$ a function. Then t is said to be a *threshold function* for the property A if two conditions hold :

(I) If $p(n) = o[t(n)]$ then $\mathbb{P}[G(n,p(n)) \models A] \to 0$ as $n \to \infty$,
(II) If $t(n) = o[p(n)]$ then $\mathbb{P}[G(n,p(n)) \models A] \to 1$ as $n \to \infty$.

Remark If t is a threshold for some property A, then so is $c \cdot t$ for any constant c such that $||c \cdot t||_\infty \leq 1$.

First application : Subgraph threshold

Given a graph property A, there are basically three stages in the analysis of the threshold phenomenon for that property :

(I) Prove that a threshold exists.
(II) Compute the threshold[1]
(III) Investigate more closely what happens as the threshold is crossed.

We will concentrate in this course on stage (II). There are some very general theorems about existence of thresholds, but to do justice to these would require too long a detour toward mathematical logic. Stage (III) obviously is likely to be more technical, and it cannot be undertaken before stage (II) anyway. Note however that, in speaking of "crossing the threshold" we are adopting a *dynamic* model of random graphs $G(n,p)$, where we think of the parameter p as growing, and the edges of the graph "growing" accordingly.

The graph property we have chosen, in order to exhibit how the second moment method can be used to compute thresholds, is that of *subgraph containment*. So let H be any fixed graph. The graph property $\mathcal{A} = \mathcal{A}_H$ under consideration is "contains a copy of H", so that $G \models \mathcal{A}_H$ means that the graph G contains a copy of the graph H. For example, $K_n \models \mathcal{A}_{K_m}$ if and only if $n \geq m$, in which case K_n in fact contains $\left(\begin{array}{c} n \\ m \end{array} \right)$ or $(n)_m$ different copies of K_m, depending on how one counts.

We need some definitions before stating our main result :

Definition 20 : Let H be a graph, with e edges and v vertices. The *density* of H, denoted $\rho(H)$, is the quantity

$$\rho(H) := \frac{e}{v}.$$

[1] Here we are once again deliberately sloppy with our language. Since a threshold function can never be unique (one can always multiply by a constant, for example), one shouldn't speak of "the" threshold. But this is common practice.

The graph H is said to be *balanced* if $\rho(H) \geq \rho(H')$ for every subgraph H' of H.

Theorem 28 *Let H be a balanced graph. Then the function*

$$t(n) := n^{-1/\rho(H)}$$

is a threshold function for the property $\mathcal{A}_H$.

Proof : We need to prove two things, namely :

(I) If $p(n) = o[t(n)]$ then $\mathbb{P}[G(n, p(n)) \models \mathcal{A}_H] = o(1)$.
(II) If $t(n) = o[p(n)]$ then $\mathbb{P}[G(n, p(n)) \models \mathcal{A}_H] = 1 - o(1)$.

Proof of (I) : This part does not require the knowledge that H is balanced. Let e, v denote the number of edges and vertices of H respectively. These quantities are thus constants and do not affect any estimates of orders of magnitude of quantities as $n \to \infty$. Fix an n and $p \in [0, 1]$. For every subset S of the vertices of K_n of size v, let X_S be the indicator variable of the event that, in $G(n, p)$, at least one copy of H appears on the vertices in S. Note that, a priori, many different copies of a single graph may appear on the same set of vertices. But since H is fixed, the number of copies of it which may appear on any set of v vertices is bounded by a function depending only on v. All of this implies that

$$\mathbb{E}[X_S] = \Theta(p^e).$$

Let $X := \sum X_S$, the sum being over all v-element sets of vertices in K_n. Then

$$\mathbb{E}[X] = \sum \mathbb{E}[X_S] = \left(\begin{array}{c} n \\ v \end{array} \right) \cdot \Theta(p^e) = \Theta(n^v p^e). \tag{57}$$

But X just counts the total number of copies of H in $G(n, p)$. So from (57) it is already clear that if $p = p(n) = o[t(n)]$, then $\mathbb{E}[X] = o(1)$, implying that $\mathbb{P}(X = 0) = 1 - o(1)$. This proves part (I).

Proof of (II) : Similarly, (57) implies that if $t(n) = o[p(n)]$ then $\mathbb{E}[X] \to \infty$. All we need to show is that $\mathbb{P}(X > 0) = 1 - o(1)$. We apply the second moment method. To simplify notation, let A_S denote the event indicated by X_S. Clearly, the various conditions introduced in our discussion of the second moment method apply to the present situation, so that it suffices for us to show that, in the notation of the previous lecture, $\Delta^* = o(\mathbb{E}[X])$, where

$$\Delta^* = \sum_{T \sim S} \mathbb{P}(A_T | A_S).$$

Here S is a fixed set of vertices of size v, and the sum runs over all sets T of vertices of size v so that the event A_T is not independent of the event A_S.

Now in the random graph setting, two events are independent if they are defined on disjoint sets of edges. So A_T is dependent on A_S if and only if the edge-sets defined by T and S are not disjoint, which is the case if and only if T and S share at least two vertices. Hence we can write

$$\Delta^* = \sum_{T: 2 \leq |T \cap S| \leq v-1} \mathbb{P}(A_T | A_S) = \sum_{i=2}^{v-1} \sum_{T: |T \cap S| = i} \mathbb{P}(A_T | A_S). \tag{58}$$

In the inner sum, the quantity $\mathbb{P}(A_T|A_S)$ must be the same for every choice of T. The number of such choices is $\begin{pmatrix} v \\ i \end{pmatrix} \begin{pmatrix} n-v \\ v-i \end{pmatrix}$, since T must have i vertices in common with S and $v - i$ other vertices. This number is $\Theta(n^{v-i})$.

Now fix an i and a T. We need an estimate for $\mathbb{P}(A_T|A_S)$. Here it is assumed that at least one copy of H appears on S and want to estimate the probability of at least one copy of H also appearing on T. Up to a constant factor, as before, we may consider a fixed copy of H on S. Let H' be the part of it on $T \cap S$. Once again, up to a constant factor, we may consider a fixed extension of H' to a copy of H on the vertices of T.

At this point we use the fact that H is balanced. It implies that $\rho(H') \leq \rho(H)$, thus H' contains at most ie/v edges and so $H \backslash H'$ contains at least $e - ie/v$ edges. This means that the appearance or otherwise in S of a fixed extension of H' on T is an event which depends on at least $e - ie/v$ independent biased coin tosses, namely the presence or otherwise in $G(n, p)$ of the edges in $H \backslash H'$.

Putting all this together, what we have shown is that, for a fixed i and T,

$$\mathbb{P}(A_T|A_S) = \Theta(p^{e-ie/v}).$$

Substituting this and the estimate for the number of different T:s into (58) we find that

$$\Delta^* = \sum_{i=2}^{v-1} \Theta(n^{v-i}) \cdot \Theta(p^{e-ie/v}) = \sum_{i=2}^{v-1} \Theta \left[(n^v p^e)^{1-i/v} \right].$$

Since $1 - i/v \leq 1 - 2/v < 1$ for every value of i, it is thus clear that the sum is $o(n^v p^e) = o(\mathbb{E}[X])$, which completes the proof of the theorem.

The proof of the following completely general result is more technical.

Theorem 29 *Let H be any graph, not necessarily balanced. Let H' be a subgraph of H of maxmimal density. Then the function $t(n) = n^{-1/\rho(H')}$ is a threshold for the property $\mathcal{A}_H$.*

A full proof is not contained in [AS], but the book contains some technical extensions of Theorem 28 above from which Theorem 29 can be deduced without too much pain. I leave all this as an exercise to the interested reader.

Second application : Concentration of random graph invariants

A graph *invariant* just means any numerical quantity which may be associated to an arbitrary graph. Examples of graph invariants are chromatic number, girth, number of connected components, number of Hamilton cycles etc. Invariants of random graphs $G(n, p)$ are thus (non-negative integer valued) functions of two variables, n and p.

The computation of random graph invariants is a natural counterpart to the problem of computing thresholds. In the former type of problem, one considers a fixed p (the most natural and interesting choice often being $p = 1/2$, as it corresponds to the edges of the graph being chosen by independent tosses of a fair coin) and wants to estimate the value of the invariant as $n \to \infty$. This basically involves estimating the expectation of some random variable X. Of more interest, though, is the degree of predictability of the invariant's value, in other words, how well concentrated the variable X is around its mean. The second moment method sometimes gets us quite strong results, a particularly nice example being the following '*2-values theorem*' :

Theorem 30 *There is an integer-valued function $k(n)$ such that*

$$\mathbb{P}[\omega(G(n, 1/2)) = k(n) \textit{ or } k(n) + 1] \to 1 \textit{ as } n \to \infty.$$

In addition, $k(n) \sim 2\log_2 n$.

Here $\omega(G)$ denotes the *clique number* of a graph G, which is the maxmimal number of vertices in a complete subgraph of G. Note that the theorem does not tell us exactly what two values $\omega(G(n, 1/2))$ is concentrated on, for any given large n, just that there are two such values, and they are of the order of magnitude of $2\log_2 n$. However, it will be clear from the proof of the theorem that the function $k(n)$, and the amount of concentration, is easily[2] computable for any particular n.

Sketch proof of Theorem 30 : I will not go through all the details of the proof, but just give the main ideas. One may consult Chapter 4 of [AS] for the full computations. What is interesting is that, in [AS], they defer a final proof of this theorem to Chapter 10, and there use some more advanced probabilistic machinery, namely the so-called *Janson inequalities.* I think this is unnecessary, however, and that the second moment method suffices to get a full proof. I leave it for yourselves to check this !

Anyway, here is the sketch :

Fix an n and a k and let X be a r.v. which counts the number of cliques of

[2] i.e.: in polynomial time, at least.

size k in $G(n, 1/2)$. We can write (should by now be getting used to this !) $X = \sum X_S$, the sum being taken over all vertex sets S of size k, where X_S indicates that all $\left(\begin{array}{c} k \\ 2 \end{array} \right)$ edges between the vertices of S are, so to speak, "turned on". Thus

$$\mathbb{E}[X] = \left(\begin{array}{c} n \\ k \end{array} \right) \cdot 2^{-\left(\begin{array}{c} k \\ 2 \end{array} \right)}. \tag{59}$$

Denote the quantity on the right hand side of (59) as $f(n, k)$. We have already seen in the very first lecture that $f(n, k)$ becomes less than 1 when k is in the vicinity of $2\log_2(n)$. Since the event $X = 0$ is the same as the event $\omega[G(n, 1/2)] < k$, this is the crucial transition as long as we can show that $\Delta^* = o(f(n, k))$ when the latter is large. The exceptionally high concentration of the clique number comes from the fact that the function $f(n, k)$, which is a decreasing function of k for fixed n, is decreasing very rapidly when k is close to $2\log_2 n$. In fact, direct insertion into the formula for $f(n, k)$ gives that

$$\frac{f(n, k+1)}{f(n, k)} = 2^{-k}\frac{n-k}{k+1},$$

so that when $k \sim 2\log_2 n$, $\frac{f(n,k+1)}{f(n,k)} = n^{-1+o(1)}$.

Some remarks on the estimation of Δ^* : It can be broken up exactly as in (58) above. The analysis is even simpler than in Theorem 28, however, as one doesn't need to worry about those annoying "up to a constant factor" estimates here. One finds (see [AS] for more detail) that

$$\frac{\Delta^*}{\mathbb{E}[X]} = \sum_{i=2}^{k-1} g(i),$$

where

$$g(i) = \frac{\left(\begin{array}{c} k \\ i \end{array} \right) \left(\begin{array}{c} n-k \\ k-i \end{array} \right)}{\left(\begin{array}{c} n \\ k \end{array} \right)} \cdot 2^{\left(\begin{array}{c} i \\ 2 \end{array} \right)}.$$

One needs to show that each term in the sum is $o(1)$, when $\mathbb{E}[X]$ is large and $k \sim 2\log_2 n$. In [AS] they show by direct computation that

$$g(2) \sim \frac{k^4}{n^2} = n^{-2+o(1)},$$

$$g(k-1) \sim \frac{2kn2^{-k}}{\mathbb{E}[X]} = \frac{n^{-1+o(1)}}{\mathbb{E}[X]},$$

and leave the remaining cases to the reader. In fact, one can see (again just by direct insertion into the formula for $g(i)$) that, up to a $1 + o(1)$ factor, the function $g(i)$ starts off by decreasing as i increases, and then starts increasing again as i approaches the order of magnitude of $\log_2 n$. What this implies is that, for every i,

$$g(i) \leq (1 + o(1)) \max\{g(2), g(k-1)\}.$$

It is this estimate which I think allows one to finish off the proof of Theorem 30 without needing to resort to any more advanced techniques. But don't take my word for it : check it yourselves !

Lecture 8 (Nov. 24, 2011)

Devdatt lectured on two CS applications : *quicksort* and *median finding*. Only the latter application involved the second moment method. Photocopies of the relevant sections of [MU], i.e.: sections 2.5 and 3.4 respectively, were handed out in class.