CS 683: Advanced Design & Analysis of Algorithms January 23, 2008

Lecture 2

Lecturer: John Hopcroft *Scribe: Hu Fu, June*

1 Giant components in real world graphs

- **Graph of Protein Interactions**
  Using data from a paper (*Science*, July 30, 1999, 285, pp751–753) that recorded 3602 pairwise interactions among 2735 proteins, a graph was formulated by representing each protein with a vertex and then connecting two vertices with an edge if the corresponding proteins interact with each other. The numbers of the connected components of different sizes are shown in the table blow:

Size	1	2	3	4	5	6	7	8	9	10	11	12	$\cdots$	15	16	$\cdots$	1850	1851
Number	48	179	50	25	14	6	4	6	1	1	1	0	$\cdots$	0	1	$\cdots$	0	1

  It can be seen that a giant component dominates the graph, while all other connected components are of considerably small sizes.

- **Graph of Papers that Share Authors**
  A database of papers was used to construct a graph, where each paper is represented by a vertex, and two vertices are linked if the papers they represent share an author. A count of the connected components in this graph shows that, except some small components of sizes up to 14, the remaining vertices are all part of a giant component of size 27488 — a phenomenon similar to the one found in the graph of protein interactions.

- **Graph of Synonyms**
  Another study was performed on a large number of words, and a graph of synonyms was obtained. Each word is present in the graph as a vertex, and two vertices are linked by an edge if the corresponding words can be used as synonyms in some context. The sizes of the connected components in this graph are as follows: 1, 2, 3, 4, 5, 14, 16, 18, 48, 117, 125, 1128, 30242. This time, there is again a giant component, but unlike the above cases, there are other components of sizes that are not negligible.

  However, if we see the graphs as growing, where the number of edges is steadily increasing, then it can be imagined that the giant component emerges by "swallowing up" smaller ones, and that this last graph of synonyms is on the verge of the appearance of a more dominant giant graph, where the components of intermediate sizes are yet to be merged into the giant one.

  Experiments and theoretical analysis with $G(n,p)$ (introduced in the last lecture, where n is the number of vertices and p is the probability of each possible edge in the graph existing) verifies this view.

2 Behavior of $G(n,p)$ as p goes up

In a random graph $G(n,p)$ where n is arbitrarily large, we increase p from 0, and the following phenomena can be sequentially observed:

- $p = 0$: the graph consists of n isolated vertices.
- $p = \frac{1}{n^2}$: the expected number of edges in the graph is one.
- $p = \frac{d}{n^2}$, $(d > 1)$: The expected number of edges is d, and almost surely, all the components in the graph are of sizes one or two.

- $p = \frac{\log n}{n}$, $p = \frac{1}{n^{3/2}}, \cdots$, as long as $p \leq O(\frac{1}{n})$, there are (almost surely) only trees of size at most $\log n$.
- $p = \frac{d}{n}$, $(d < 1)$: there are a constant number of components with cycles in them, where "constant" means *independent of n*. Almost surely, all components are trees or unicyclic and are of size at most $\log n$.
- $p = \frac{1}{n}$: a component of size $n^{2/3}$ emerges, and it is almost surely a tree, because new edges will morelikely appear in smaller components.
- $p = \frac{d}{n}, (d > 1)$: a giant component of constant fraction of vertices appears, and all other components are of size at most $\log n$ (so there are no two giant ones). This occurs because the probability of an edge connecting two large components is high.
- $p = \frac{1}{4}\frac{\log n}{n}$: the giant component swallows up the components of intermediate sizes, and the graph is left with only the giant component and isolated vertices.
- $p = \frac{\log n}{n}$: all isolated vertices have been swallowed up by the giant component, and the graph becomes connected.
- p is a constant: almost surely the diameter of the graph is two

3 Exemplary properties of $G(n, p)$

3.1 Finding cliques in $G(n, 1/2)$

If we look at $G(n, 1/2)$, it is really a dense graph — the expected number of edges in it is $n^2/2$, while a clique of n vertices has only $\frac{n(n-1)}{2}$ edges. So how large a clique could we find in $G(n, 1/2)$?

Finding a clique of size $\log n$ in $G(n, 1/2)$ is trivial. We arbitrarily pick a vertex v_1, and with high probability, it has at least $n/2$ edges. We arbitrarily pick one of them and the vertex v_2 that it reaches. With high probability, v_2 has at least $n/4$ neighbors that are also neighbors of v_1. We arbitrarily pick a vertex again and continue this process. With high probability we can pick $\log n$ vertices adjacent to each other, and they constitute a clique of size $\log n$.

There is a similar but much harder problem: it can be proved that, given any $\epsilon > 0$, with high probability there is a clique of size $(2 - \epsilon) \log n$, but to find such a clique turns out to be a very hard problem and have implications in P vs. NP..

The phenomena that we observed in Section 2 were obtained by increasing p while taking arbitrarily large n. A simulation that, while increasing p, fixes n as a fairly large number (say, $100,000$), produces results that do not agree with the theoretical analysis on the values of p when certain phenomenon occurs. This is suggested as a problem for a project.

3.2 Expected number of triangles in $G(n, d/n)$

Claim: The expected number of triangles in $G(n, d/n)$ is constant as n goes to infinity.

Two observations give some intuition of the claim: As n increases,

1. the probability that there is an edge between two fixed vertices u and v decreases.
2. the number of triples of vertices increases.

The effects of these two observations counter act each other and result in the constant number of triangles as n increases.

Proof. Given any three vertices in $G(n, d/n)$, the probability that there are edges between each pair of them is $(d/n)^3$. Therefore,

$$\text{The expected number of triangles} = \binom{n}{3}\frac{d}{n}^3 = \frac{n(n-1)(n-2)}{n} \cdot \frac{d^3}{n^3} \to \frac{d^3}{6},\ (n \to \infty)$$

□

Note that this proof is not confined to triangles, but is applicable to any pattern that includes k vertices and k edges (e.g. the rectangles). Additionally, these effects can be seen in graphs of 1000 vertices.

3.3 Diameter of $G(n, p)$ (p a constant)

Claim: When p is a constant and $0 < p < 1$, the diameter of $G(n, p)$ is almost surely two.

Proof. If G has a diameter of at least three, then there exists non-adjacent vertices u and v such that for all $w \in V$, w is not adjacent to both u and v. We call such u, v a bad pair, and show that the expected number of such pairs is zero, which proves the claim.

Let X be the number of bad pairs. We label all pairs of vertices by $1, 2, \cdots, \binom{n}{2}$, and let

$$X_i = \begin{cases} 1, & \text{if the } i^{\text{th}} \text{ pair is bad;} \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$X = \sum_{i=1}^{\binom{n}{2}} X_i.$$

$$E[X_1] = Pr\,(\text{a pair } (u, v) \text{ is bad}) = (1-p)(1-p^2)^{n-2},$$

$$E[X] = \binom{n}{2} E[X_1] = \binom{n}{2}(1-p)(1-p^2)^{n-2} \to 0,\ (n \to \infty)$$

□

Note: the fraction of graphs that fail this is $1/n$, which $((1/n) \to 0)$ as $(n \to \infty)$. Also, $G(n, p)$ have average degrees evenlyspread, real world graphs have clustering.

4 Phase transitions

Definition: If there exists a $P(n)$ such that, for $\lim_{n\to\infty} \frac{P_1(n)}{P(n)} = 0$, then almost surely $G(n, P_1(n))$ does not have a property, and for $\lim_{n\to\infty} \frac{P_2(n)}{P(n)} = \infty$, then almost surely $G(n, P_2(n))$ has the property, we say that $P(n)$ is a threshold for the property.

Definition: If there exists a $P(n)$ such that, $G(n, cp(n))$ for $c < 1$ does not have a property and $G(n, cP(n))$ for $c > 1$ has the property, then we say that $P(n)$ is a sharp threshold for the property.

It is interesting to ask which properties have thresholds and which have sharp thresholds. Is there a necessary and sufficient condition for each? This is suggested for a project, and is perhaps an open problem.

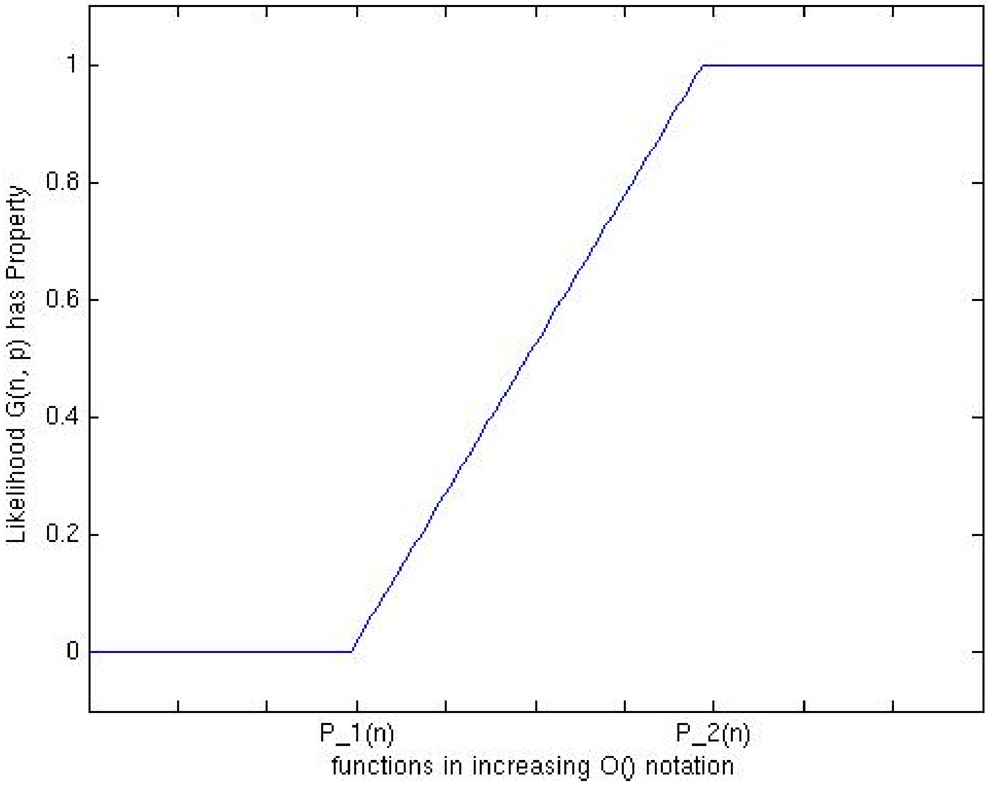

Figure 1: Transition Phase