CS 590R
Lecture Notes 20: March 27, 2003
Random Graphs & Diameter of the P2P Network

Lecturer: Gopal Pandurangan
Notes taken by: Maleq Khan

1 Random Graph

A random graph is a graph in which properties such as the number of graph vertices, graph edges, and connections between them are determined in some random way. For many monotone-increasing properties of random graphs, graphs of a size slightly less than a certain threshold are very unlikely to have the property, whereas graphs with a few more graph edges are almost certain to have it. This is known as a *phase transition* or *threshold phenomena.*

$G(n,p)$ **model:** $G(n,p)$ is a random graph with n vertices where each pair of vertices have edge between them with probability p. the existence of any two edges are independent event. The number of possible edges is $\binom{n}{2} = \frac{n(n-1)}{2}$. The expected number of edges is $\binom{n}{2}p$.
When $p = \frac{1}{n}$, w.h.p. there is no cycle in the graph. There are many small components and the size of the largest component is $O(\log n)$.
When $p = \frac{2}{n}$, w.h.p. there is a cycle in the graph but the graph is disconnected.
When $p = \frac{\log n}{n}$, w.h.p. the graph is connected.

$G(n,m)$ **model:** This model is very closely related to $G(n,p)$. n is the number of vertices. m is an integer such that $0 \geq m \geq \binom{n}{2}$. m edges are

randomly selected from $\binom{n}{2}$ possible edges.
If we increase number of edges m from 0 to $\binom{n}{2}$, in the beginning, we have only isolated vertices. As long as m is small, $m = o(n)$, the graph is a forest, consisting of a large number of mostly small trees which grow (by merging) as m increases. The first cycle appears when $\epsilon n < m < (1/2 - \epsilon)n$, where ϵ is a positive integer such that $\epsilon(n) \to 0$ as $n \to \infty$. The graph consist of trees and a few unicyclic components until $m \approx n/2$.
At about this number of edges, a phase transition occurs and the structure of the graph changes dramatically. For example, let $\delta > 0$ be fixed. When $m = (1 - \delta)n/2$, there are many trees and possibly some unicyclic components, and the largest component has $O(\log n)$ vertices, but when $m = (1 + \delta)n/2$, there is a unique giant component with $\approx c(\delta)n$ vertices. But all other components are trees or unicyclic, and the second largest component has $O(\log n)$ vertices.
Note that the phase transition occurs when the mean degree of a vertex $2m/n \approx 1$. As more edges are added, the giant grows and swallows other components, but the picture remains qualitatively the same as long as $m = O(n)$. When $m \gg n$, there is no cycle left outside the giant component. There is still some trees outside the giant component, but the last of them disappears when $m \approx \frac{1}{2}n \log n$, more precisely when $m = \frac{1}{2}n \log n + O_p(n)$.

Random Regular Graph $G(n, r)$: $G(n, r)$ is a regular graph with n vertices where every node connects to $r > 3$ random nodes. The diameter of such graph is $O(\log n)$.

Planer Graph: A planer graph is one that can be drawn on a plane in such a way that there are no "edge crossings," i.e. edges intersects only at their common vertices.

Convex Property of a Graph: Let A, B, and C are three graphs with same set of vertices (they differ only in edges) such that $A \subset B \subset C$. If that a property of a graph is held by A and C implies that B also hold the property, then this property is called convex property. For example, if A and C is planer, then B is also planer. Another example, if number of isolated nodes in both A and C is exactly k, then number of isolated nodes in B is also k.

2 Diameter of the P2P network

Although the P2P network built using the given protocol is a constant degree, which is in $[C, D]$, network, its diameter is logarithmic.

Theorem 2.1 *For any t, such that $t/N \to \infty$, w.h.p. the largest connected component of G_t has diameter $O(\log N)$. In particular, if the network is connected (which has probability $1 - O(\frac{\log^2 N}{N})$) then w.h.p. its diameter is $O(\log N)$.*

A d-node is always connected to a c-node. To prove the theorem, it is sufficient to show that distance between any two c-nodes is $O(\log N)$.

The following ideas are used to prove the theorem:
1. reconnect connections are long-range and random.
2. a node that has many reconnect connections is called a good node. there are many such good nodes.
3. due to the existence of many good nodes, distance between any two c-nodes is as small as logarithmic.

The details of the proof is given below. Since we are dealing with the distance between two c-nodes, through out the rest of the discussion, a node refers to a c-node.

For the purpose of the proof, a good cache node and color of the edges are defined as follows.

Good cache node: A cache node is good if during its time in the cache it receives the set of $r \geq f$ (f is a fixed constant) connections such that:
1. they are reconnect connections;
2. they are not preferred connections;
3. they resulted from different nodes leaving the network.

The color of an edge or connection is blue if it satisfy the above three conditions; otherwise, the color is red. A random f blue edges are considered to be B1 and B2 egdes. $f/2$ edges are said to be B1 and the rest to be B2.

Following the proof of Theorem 2 in lecture notes 19, it can easily be shown that at anytime t, the network is connected with probability $1 - O(\frac{\log^2 N}{N})$ using only the red edges, and that if the network is not connected then w.h.p. the red edges define a connected component of size $N(1 - o(1))$.
The random structure of blue edges plays important role in proving the theorem. However, there are two difficulties involving this random structure. First, although the blue edges are random, the occurance of the edges between pairs of nodes are not independent as in the standard $G(n, p)$ model of random graphs. Second, the total number of blue edges is relatively small; thus the proof needs to use both red and blue edges.

Lemma 2.1 *Let node v enter the cache at time t, where $t/N \to \infty$. Then for a sufficiently large choice of the constant C,* $\Pr($ *v leaves the cache as a good node$) \geq 1/2$. *The f blue edges are distributed uniformly at random among the nodes in the current network. Furthermore, the probability that a c-node is good is independent of other c-nodes.*

Proof: Consider the interval of time in which v was a cache node.
1. New nodes join the network according to a Poisson process with rate 1. The expected number of connections to v from a new node is D/K.
2. A node u reconnect with probability $D/d(u)$. There are K cache nodes. The cache node v gets this connection with probability $1/K$. Nodes leave the network according to a Poisson process with rate $1 - o(1)$. Therefore, the expected number of connections to v as a result of a old node leaving the network is

$$\approx \sum_{u \in V} \frac{d(u)}{|V|} \frac{D}{d(u)} \frac{1}{K} \approx \frac{D}{K} < 1$$

Thus each connection to v, while it is in cache, has a constant probability of being a reconnect connection.
3. Also when a node u leaves the network, the expected number of connections to v from u in unit time is $\approx \frac{d(u)}{N} \frac{D}{d(u)} = D/N$.That is, all old nodes have equal probabilities of being connected to v. Since the expected number of connections to v as a result of one node leaving network is ≤ 1, for sufficiently large C, the $C - D$ connections to v include, with probability $\gamma > 1/2$, $r \geq f$ connections that satisfy the requirement for blue edges.
Further using 3 and the fact that each node leaves the network independently

and identically under the same exponential distribution it follows that each node in the network - irrespective of its degree - has an almost equal probability of being connected to v. Finally, it is easy to see the independence of the events for different c-nodes, since a cache node stays in the cache till it accepts C connections irrespective of other cache nodes. $\square$

Expansion Lemma

For a c-node v in G_t, let

- $T_0(v)$ is arbitrary connected cluster of $O(\log N)$ c-nodes including v , using red edges.
- $T_i(v)$ is the set of c-nodes that are connected by blue edges to $T_{i-1}(v)$, but are not in $T_0(v), \ldots, T_{i-1}(v)$.

We show that neighborhood expand exponentially, i.e. $\frac{T_i(v)}{T_{i-1}(v)} \geq c$ for some $c > 1$; which implies that the diameter is $O(\log n)$.
For $i \geq 1$, i odd (resp. even), let $T_i(v)$ be all the c-nodes in G_t that are connected to $T_{i-1}(v)$ and are not in $\cup_{j=0}^{i-1} T_j(v)$ using B1 (resp., B2) edges.
The nodes in $T_i(v)$ use B1 edges to connect to $T_{i-1}(v)$ and B2 edges to connect to $T_{i+1}(v)$.

Lemma 2.2 *If* $|\Gamma_{i-1}(v)| = o(N)$,

$$Pr\{|\Gamma_i(v)| \geq 2|\Gamma_{i-1}(v)|\} \geq 1 - 1/N^5.$$

Let $W = \Gamma_{i-1}(v)$, $w = |W|$, and let $z \notin W \cup (\cup_{j=0}^{i-1} \Gamma_j(v))$. W.l.o.g. assume that $i-1$ is even. Partition W into W_0, consisting of nodes in W that are older than z, and W_1, consisting of nodes in W that arrived after z.
There are $f/2$ B1 edges and the probability that z is a good cache node is $1/2$. Therefore, The probability that z is connected to W_0 using B1 edges is at least $\frac{1}{4}\frac{f|W_0|}{N}(1 - o(1))$.
Similarly, each node in W_1 has probability $\frac{1}{4}\frac{f}{N}(1 - o(1))$ of being connected to z by B1 edges.
Thus the probability that z is connected to W is at least $\frac{1}{4}\frac{fw}{N}(1 - o(1))$.
and $E[|T_i(v)|] = \frac{f}{4}w(1 - o(1)) = \frac{f}{4}|T_{i-1}(v)|(1 - o(1))$. for $f \geq 8$, $E[|T_i(v)|] \geq$

$2|T_{i-1}(v)|$.

We use a martingale technique to prove that $\Gamma_i(v)$ is concentrated around its mean w.h.p. The proof is completed in the next lecture note (Notes 21) after an introduction to martingale.