CS 683: Advanced Design & Analysis of Algorithms February 6, 2008

Lecture 8

Lecturer: John Hopcroft *Scribe: Sucheta Soundarajan, Yookyung Jo*

1 Relationship between Directed and Undirected Graphs

We look into the relationship of the three graph models, $G(n,p)$, $\vec{G}(n,p)$, and $D(n,p)$.
$G(n,p)$ is an undirected random graph where each edge exists with probability p. $\vec{G}(n,p)$ is a directed random graph that is constructed essentially the same way with $G(n,p)$ except that when a pair of vertices v and w are connected with probability p, the directed edges are made in both directions. $D(n,p)$ is a directed random graph where each directed edge has probability p. Thus, if we look at the two directed edges of a pair of vertices v and w, the edges (w,v) and (v,w) exist independent of each other in $D(n,p)$, while in $\vec{G}(n,p)$ either both exist together or none of them exists.

It is noteworthy that if we can prove these 3 graphs have equivalent properties in some aspects then many of the results we have proved for $G(n,p)$ can carry on to the directed graphs. Especially, we show that

Claim The probability distribution for the number of vertices reachable from a vertex v is the same in each model.

Proof Since it is clear that $G(n,p)$ and $\vec{G}(n,p)$ are equivalent, we only show the claim for $G(n,p)$ and $D(n,p)$.
We apply a graph search algorithm DFS (depth first search) starting from a vertex v and show that DFS algorithm discovers the same set of vertices reachable from v in both $G(n,p)$ and $D(n,p)$ with the same probability.
The important trick in the proof is to alter the order of when we apply the algorithm. Instead of generating the graph first and then applying the algorithm, we start applying DFS from the vertex v and generate the random graph $D(n,p)$ on the fly. At v, DFS considers each vertex in turn from the set of vertices not explored yet, and ask the question "Is there an edge from v to that vertex?". At this time, we flip the coin of probability p and decide whether there is such an edge in $D(n,p)$. Once DFS discovers an edge from v to another vertex, DFS recursively starts from that vertex and goes on. Once DFS is finished, we have the set of vertices reachable from v in $D(n,p)$. It is easy to see that the above process applies identically to the case of $G(n,p)$, if our question is slightly altered to "Is there an undirected edge between v and that vertex?". That is, the above process results in the same set of reachable vertices from v with the same probability.
(Other graph search algorithms such as BFS can also be used instead of DFS. The important point is that once the graph search algorithm asks about the existence of an edge $w \to v$ it does not ask about the reversed edge $v \to w$ later because this would result in different probabilities for the two graph models.)

2 Strongly Connected Component of a Directed Graph

Using the claim that the probability distribution for the number of vertices reachable from a vertex v is the same in all 3 random graph models, we can show that if there is a giant component in an undirected graph $G(n,p)$, the directed graph $D(n,p)$ will have a giant strongly connected component.

If we have a giant component of size greater than $\frac{n}{2}$ in $G(n,p)$, with probability greater than $\frac{1}{2}$ we can find a vertex in $G(n,p)$ that reaches more than $\frac{n}{2}$ number of vertices. In the corresponding $D(n,p)$, we can find a vertex v that reaches more than $\frac{n}{2}$ number of vertices by outgoing edges with probability greater than $\frac{1}{2}$,

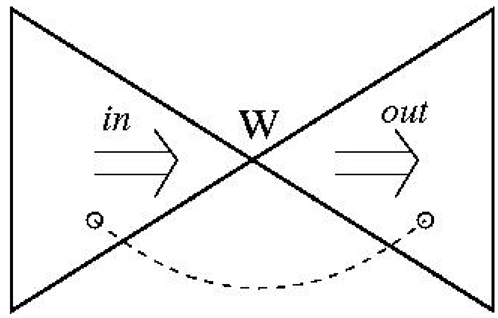


Figure 1: Strongly connected component of a directed graph

because the number of reachable vertices for a vertex in the two graph models has the same distribution. By the same argument if we think in terms of the reversed direction of edges, we can see that with probability greater than $\frac{1}{2}$ we can find a vertex u in $D(n,p)$ that reaches more than $\frac{n}{2}$ number of vertices by incoming edges.

Because we can find a vertex v whose out-component (vertices reachable by v) is greater than size $\frac{n}{2}$ and a vertex u whose in-component (vertices that reach u) is greater than $\frac{n}{2}$ by probability greater than $\frac{1}{2}$ respectively, it is likely that we can find a vertex w whose out-component and in-component are both greater than $\frac{n}{2}$. Also, since w's in and out components are both of size greater than $\frac{n}{2}$, there must be an overlap in the two components.

It is known that if probability p is such that the size of the connected component in $G(n,p)$ is Θn then $\Theta^2 n$ is the size of the strongly connected component in $D(n,p)$.

So far, we have discussed the random graph models whose edges exist with probability p. But, in real life, there are different kinds of random graphs. For example, we might ask the question "When does the power law graph have a giant component?". In the next section, we study random graphs with arbitrary degree distribution.

3 Random Graph of arbitrary degree distribution

How do we generate a random graph with arbitrary degree distribution?
For example, let's say we have a degree distribution shown in the following table.

	out-degree	in-degree
$v1$	3	1
$v2$	0	2
$v3$	1	1
$\vdots$	$\vdots$	$\vdots$

One way to generate a random graph to satisfy the degree distribution is to generate two columns, where in the left column you place each vertex its out-degree number of times and in the right column you place each vertex its in-degree number of times. You do the random permutation on the right column and connect the two vertices in each row from left to right.

$v1$	$v1$		$v1$	$\rightarrow$	$v5$
$v1$	$v2$	after random	$v1$	$\rightarrow$	$v3$
$v1$	$v2$	permulataion	$v1$	$\rightarrow$	$v6$
$v3$	$v3$	$\Rightarrow$	$v3$	$\rightarrow$	$v1$
$\vdots$	$\vdots$		$\vdots$	$\rightarrow$	$\vdots$

The problem with this graph construction is that edges generated this way are not statistically independent with each other. Since we use the property of edge independence many times in proving theorems in random graphs, the violation of edge independence is a serious problem. (The graph could also have self-loop or redundant edges, but when the size of the graph is large, these are negligible.)

4 Giant Component of Random Graphs with Expected Degree Distribution

In this section, we study when a giant component emerges in a random graph with expected degree distribution.

Let λ_i be the fraction of vertices of degree i.
Then, the condition for a giant component emerging is given by

$$\sum_{i=0}^{\infty} i\,(i-2)\,\lambda_i > 0$$

And when $\sum_{i=0}^{\infty} i\,(i-2)\,\lambda_i$ becomes 0, the phase transition occurs.

Let's think about why the condition signifies the emergence of a giant component, by considering a graph search to find a connected component and by looking at how the size of the frontier changes at each time step.

Suppose we have a graph where half of vertices are of degree 1, and the other half of vertices are of degree 2. If we pick a vertex in random, with probability $\frac{1}{2}$ we would pick a vertex with degree 1 and with probability $\frac{1}{2}$ we would pick a vertex with degree 2. But, if we randomly pick an edge and follow the edge to reach a vertex, then, the probability of picking a vertex with degree 1 is $\frac{1}{3}$ and the probability of picking a vertex with degree 2 is $\frac{2}{3}$. The example shows that the probabilibility of picking a vertex of degree i by following a random edge is proportional to its degree i. Thus, the term $i\lambda_i$ in the condition signifies the probability of reaching a vertex with degree i.

Next, let's consider what happens to the size of the frontier in a graph search. At each time step, we consider a vertex v from the frontier and the frontier is modified by exploring the edges from the vertex. If the vertex v has one edge, then that edge must be the edge we followed to reach v, so v has no further edges. This means our frontier has decreased by 1. If the vertex v has two edges then it has only one free edge. By following the free edge, we get a new vertex into the frontier but since we lose v in the frontier, the size of the frontier remains the same. If we continue the argument, we see that if the vertex v has i edges, the size of the frontier would increase by $i-2$.

Thus, we can conclude that the quantity $\sum_{i=0}^{\infty} (i\lambda_i)\,(i-2)$ is the expected change in frontier size at each time step.

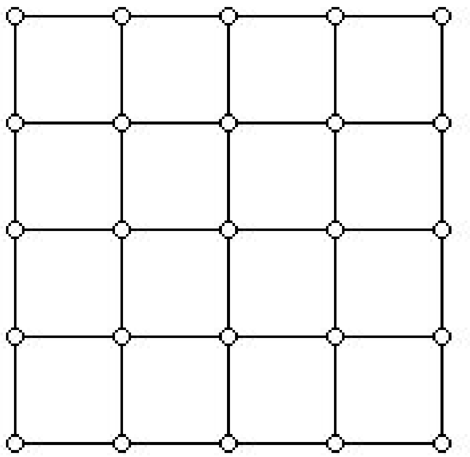
As an example, we could use this condition to check that a random graph $G(n,p)$ goes through a phase transition at $G(n,\frac{1}{n})$. The probability that a vertex has d edges is given as

$$\text{Prob}\,(\text{deg} = d) = \binom{n}{d} p^d\,(1-p)^{n-d}$$

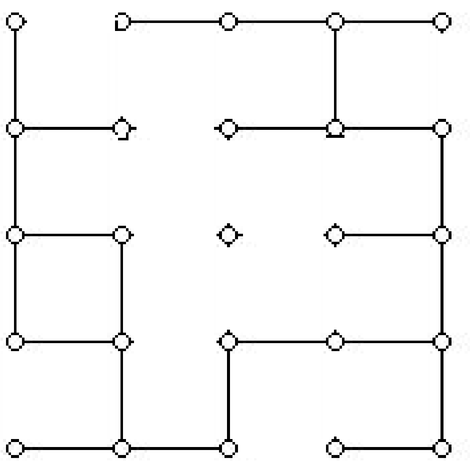
By plugging this into the formula, you will get zero.

5 Grid Graphs

A grid graph is a graph with vertices on the coordinate grid, and edges from the set {(u, v) | u and v are adjacent in the coordinate grid)}. Here is a grid graph in 2 dimensions:



As we did with G(n, p), we can create random grid graphs by specifying the probability that each edge has of existing.



One interesting problem here is finding the critical probability of a giant component, or the critical probability of a path existing from one side to the other.

The critical probability of having a giant component is $\frac{1}{2}$. Intuitively, this makes sense, since each vertex has degree at most 4, and if p = $\frac{1}{2}$, then the expected degree is 2, and so, as discussed in the previous section, the expected size of the frontier will neither increase nor decrease with each step. In the formula given earlier, $\sum_{i=0}^{\infty} i(i-2)\lambda_i$, we find that p = $\frac{1}{2}$ gives us a degree distribution such that the value of the formula is 0.

However, there is a false, but initially appealing, argument that suggests p should be slightly less than $\frac{1}{2}$. As we are searching through the verticies in the graph, we will never reach one with degree 0. Since the probability of a vertex having degree 0 is $\frac{1}{16}$, but we can never reach such a vertex, the other probabilities must all be multiplied by a correction factor of

$\frac{16}{15}$. This suggests that the value p = $\frac{1}{2}$ is a little higher than what we actually need.

Later, we will have a rigorous proof that p = $\frac{1}{2}$.

6 Generating Functions

A sequence a_0, a_1, a_2, ... can be encoded as a power series $\sum_{i=0}^{\infty} a_i x^i$. Such a power series is called a generating function.

Example 1:

For the sequence 1, 1, 1, ..., we have $a_0 = 1, a_1 = 1, a_2 = 1, ...$ and so the corresponding generating function is g(x) = $1 + x + x^2 + ... = \frac{1}{1-x}$.

For the sequence 1, 2, 3, ... we have the generating function $1 + 2x + 3x^2 + ... =$ x $\frac{d}{dx}$ g(x).

Example 2:

Generating functions can be used to solve counting problems. Suppose we have three item types: A, B, and C, and we wish to create a set of five items. Suppose the rules are as follows:

We may select 0 or 1 items of type A.
We may select 0, 1, or 2 items of type B.
We may select 0, 1, 2, or 3 items of type C.

We can represent the A rule with (1+x), the B rule with $(1+x+x^2)$, and the C rule with $(1+x+x^2+x^3)$. Then the generating function for this problem is simply the product of these three terms, or $(1+x)(1+x+x^2)(1+x+x^2+x^3) = 1+3x+5x^2+6x^3+5x^4+3x^5+x^6$. To determine how many ways there are to select 5 items, we find the x^5 term and look at the coefficient. The generating function tells us that there are 3 ways to select 5 items.

Example 3:

Suppose we have an integer valued random variable where p_i is the probability of value i. Then g(x) = $\sum_{i=0}^{\infty} p_i x^i$. Since E(x) = $\sum_{i=0}^{\infty} i p_i$, E(x) is then simply $\frac{d}{dx}$g(x), evaluated at x = 1.

Example 4:

Suppose we have two i.i.d random variables x_1 and x_2, with generating function g(x). What is the generating function for $(x_1 + x_2)$? It is simply $g^2(x) = g_0g_0+(g_0g_1+g_1g_0)x+(g_0g_2+g_1g_1+g_2g_0)x^2+...$

Example 5:

Suppose we have a sequence defined by a recurrence equation:
0, 1, 1, 2, 3, 5, ...

$f_0 = 0$
$f_1 = 1$
$f_i = f_{i-1} + f_{i-2}$, $i \geq 2$. Then we have $f_i x^i = f_{i-1} x^i + f_{i-2} x^i$, so $\sum_{i=2}^{\infty} f_i x^i = \sum_{i=2}^{\infty} f_{i-1} x^i + \sum_{i=2}^{\infty} f_{i-2} x^i$. If we say that $f(x) = f_0 + f_1 x + f_2 x^2 + ...$, then
$f(x) - f_0 - f_1 x = x(f(x) - f_0) + x^2 f(x)$
$f(x) - x = xf(x) + x^2 f(x)$
$f(x) = \frac{x}{(1-x-x^2)}$