Lecture 43
ECE 664

Random Graphs – An Introduction

Purdue U.
A. C. Kak

- Random graphs are useful for modeling large networks — social networks, computer networks, supply-chain and demand-chain networks for large industrial enterprises, etc.
- When we talk about a random graph of order n, we mean a graph with n vertices.
- For theoretical modeling, the two most commonly used methods for generating a random graph of order n are:

METHOD A:

- Say you want to construct a random graph of order n that has M edges.
- The maximum number of edges in a graph with n vertices is $\boxed{N = \binom{n}{2}} = \frac{n!}{2!\,(n-2)!} = \frac{n(n-1)}{2}$.
- The total number of M-subsets in a set of N objects is $\binom{N}{M} = \frac{N!}{M!\,(N-M)!}$.
- We construct an M-edge random graph by randomly choosing one of these $\binom{N}{M}$ subsets.
- The probability space of graphs thus constructed will have $\binom{N}{M}$ different M-edge graphs in it, each graph occurring with a probability of $\frac{1}{\binom{N}{M}}$.

We will denote this probability space by $G(n, M)$. Note that this probability space exists for particular values of n and M. We are interested in the properties of this space as the order $n \to \infty$.

METHOD B:

- This method, more popular than the previous method, considers a random graph as a structure that evolves one edge at a time, with each edge coming into existence with a probability p.
- Say you want to construct a graph of order n (that is, over n vertices) with edge placement probability p.
- You visit each of the $N = \binom{n}{2}$ potential edge placement sites and, at each site, you flip a loaded coin that shows heads with a probability p. If you see the head, you place an edge between the two vertices being considered. Otherwise, the two vertices

stay unlinked. [Instead of flipping a loaded coin, you can also fire up a random number generator that outputs, say, the integer 1 with probability p.]

- Starting with n isolated vertices, the coin-toss experiment described above will have to be carried out $N = \binom{n}{2}$ times for each graph.
- Each repetition of the above experiment will generate a different graph of order n. The set of all these graphs constitutes a probability space. We will denote this probability space by $G(n,p)$. Obviously, the total number of graphs in $G(n,p)$ is the cardinality of the power set of a set of $N = \binom{n}{2}$ objects. As you will see below, the graphs in $G(n,p)$ do not have the same probability — unlike the graphs in $G(n,M)$.
- The edge-placement coin tosses for constructing a graph G in the probability space $G(n,p)$ constitute $N = \binom{n}{2}$ Bernoulli trials. The probability that G will end up with M edges would then be given by the binomial distribution:

Maximum possible value for M is $N = \binom{n}{2}$

$$\boxed{P(e(G) = M) = \binom{N}{M} p^M q^{N-M}}$$

where $e(G)$ is the number of edges in the graph G and $\boxed{q = 1-p}$. So whereas the random variable $e(G)$ in the probability space $G(n,p)$ possesses a binomial distribution, the probability of a given graph G with M given edges is $p^M q^{N-M}$.

The DeMoivre-Laplace Theorem

- The previous discussion should have given you some idea of the importance of binomial distributions to the theory of random graphs.
- The DeMoivre-Laplace (DL) theorem gives us lower and upper bounds on the values we may expect a binomial random variable, such as $e(G)$ mentioned above, to acquire as the number of trials, N, goes to infinity. These bounds are expressed in terms of the distribution function of a zero-mean unit-variance Gaussian random variable.
- In the rest of our discussion on the DeMoivre-Laplace Theorem, we will denote the binomial random variable by $S_{N,p}$. You may think of $S_{N,p}$ as the number of times the heads show up in

In general, when the mean is μ and the variance σ^2, a Gaussian random variable is characterized by the following density function:

$$N(\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

N Bernoulli trials with a loaded coin that has probability p associated with the heads-up outcome.

- With $S_{N,p}$ defined as above, I'll now present the DL theorem in three different forms:

We will use the notation $\varphi(t)$ and $\Phi(x)$ for the density function and the cumulative probability distribution function of a zero-mean unit-variance Gaussian random variable:

$$\boxed{\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}} \qquad \boxed{\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt}$$

FORM A:

Meaning of '$\approx$'
$h(n) \approx g(n)$ means that $\frac{h(n)}{g(n)} \to 1$ as $n \to \infty$

Meaning of the little 'o' notation
$h(n) = o\{g(n)\}$ means that $\frac{h(n)}{g(n)} \to 0$ as $n \to \infty$

For any fixed value for the probability p, $0 < p < 1$, the probability distribution for $S_{N,p}$ obeys the following bounds as $N \to \infty$:

$$\boxed{P\left(pN + h_1 \leq S_{N,p} \leq pN + h_2\right) \approx \Phi(x_2) - \Phi(x_1)}$$

where $\boxed{x_1 = \frac{h_1}{\sqrt{pqN}}}$, $\boxed{x_2 = \frac{h_2}{\sqrt{pqN}}}$ and with the constraints

$\boxed{h_1 < h_2}$ and $\boxed{|h_1| + |h_2| = o\{(pqN)^{2/3}\}}$. $q = 1-p$

implies that as $N \to \infty$, we can allow for $|h_1|$ and $|h_2|$ to get larger and larger but $|h_1| + |h_2|$ must not rise faster than $(pqN)^{2/3}$

FORM B:

With the same condition on p as before, we have as $N \to \infty$

$$\boxed{P\left(S_{N,p} \geq pN + h\right) \approx 1 - \Phi(x)}$$

where $\boxed{x = \frac{h}{\sqrt{pqN}}}$ and with $\boxed{h = o\{(pqN)^{2/3}\}}$.

implies that we can also allow $h \to \infty$ as $N \to \infty$ but h must NOT rise faster than $(pqN)^{2/3}$

FORM C:

If we also allow $x \to \infty$ as the number of Bernoulli trials $N \to \infty$, Form B simplifies to the following form:

$$\boxed{P\left(S_{N,p} \geq pN + h\right) \approx \frac{1}{x\sqrt{2\pi}} e^{-\frac{x^2}{2}}}$$

- What the DL theorem tells us is that as the number of Bernoulli trials $N \to \infty$, the binomial random variable $S_{N,p}$ begins to look more and more like a Gaussian random variable with mean pN and variance pqN.

- Focusing on Form B, let's choose for h something like $h = \frac{c}{pqN}$. This h satisfies the condition $h = o\{(pqN)^{2/3}\}$ since $\frac{c(pqN)^{-1}}{(pqN)^{2/3}} = c(pqN)^{-5/3} \to 0$ as $N \to \infty$. With this choice for h, since $h \to 0$ as $N \to \infty$, we will get ever closer to the mean pN as N becomes larger and larger. So the limiting form of Form B in this case becomes $P(S_{N,p} \geq pN) \approx 1 - \Phi(0) = 0.5$ since $x \to 0$ as $h \to 0$ and since $\Phi(0) = 0.5$. If the distribution for $S_{N,p}$ looks like a Gaussian centered at pN, this makes sense.

- The bound shown in Form B can be derived from Form A by first choosing a fixed h and then setting $h_1 = h$ and $h_2 = h + (pqN)^{5/8}$. See Bollobas for details.

- As already mentioned, Form C is an algebraic simplification of Form B when we also allow $x \to \infty$ as $N \to \infty$. This algebraic simplification uses the following identity:

$$\boxed{1 - \Phi(x) \approx \frac{\varphi(x)}{x} = \frac{1}{\sqrt{2\pi}\, x} e^{-\frac{x^2}{2}} \quad \text{as } x \to \infty}$$

G(n, M) versus G(n, p) for Random Graph Modeling

- Earlier we talked about two different approaches for modeling random graphs. Our first approach led to the probability space $G(n, M)$ in which every graph comes with M edges and every graph has the same probability of occurrence (which is $\frac{1}{\binom{N}{M}}$). Our second approach led to the probability space $G(n,p)$ in which the different graphs will have different patterns of connectivity, both in terms of which vertices are directly connected and how many edges there are in a graph. For the $G(n,p)$ space, the number of edges is a binomial random variable.

- We will now show that the models $G(n, M)$ and $G(n,p)$ are practically interchangeable provided M is close to pN for large N. This interchangeability is with respect to the various $\boxed{\text{properties}}$ that the graphs might exhibit. So before we can talk about the interchangeability of the probability models, we must explain what we mean by a graph property.

- Examples of graph properties: (1) A graph contains a designated subgraph; (2) A graph contains a Hamiltonian circuit; (3) A graph contains a clique of order 5; etc.

- Formally speaking, $\boxed{\text{a property } Q \text{ is a subset of all the graphs}}$ in a probability space, it being implicit that every graph in the subset must exhibit a designated feature such as possessing a Hamiltonian circuit.

- We say a property Q is monotone whenever a graph $G \in Q$ and $G \subset H$ (meaning that G is a subgraph of H), we also have $H \in Q$. All three example properties listed above are monotonic.

- We call a monotone property Q convex if given three graphs $F \subset G \subset H$ (that is, F is a subgraph of G and G is a subgraph of H) and $F \in Q$ and $H \in Q$ implies $G \in Q$ also.

- Establishing interchangeability between the $G(n, M)$ model and the $G(n,p)$ model requires that we associate probabilities with the graph properties.
- The probability of a property Q, denoted $P(Q)$, means the probability that a randomly selected graph from the set of all graphs exhibits property Q. For example, if half the graphs in the set of all graphs in a probability space exhibit property Q, then $P(Q) = 1/2$.
- If $P(Q) \to 1$ as $n \to \infty$, we say almost every (a.e.) graph has property Q.
- When the set of all graphs is $G(n, M)$, we denote the probability of a property as $\boxed{P_M(Q)}$. And, when the set of all graphs is $G(n,p)$, we use $\boxed{P_p(Q)}$ to denote the probability of a property.

▶ Bollobas has proved that for any monotone property Q, $P_{M_1}(Q) \leq P_{M_2}(Q)$ if $M_1 < M_2$; and $P_{p_1}(Q) \leq P_{p_2}(Q)$ if $p_1 < p_2$. Basically what that says is that any monotone property can <u>always</u> be expected to occur with greater likelihood as the number of edges in a graph increases.

- The following theorem gives a more precise meaning to what we mean by saying that $G(n, M)$ and $G(n,p)$ are interchangeable:

THEOREM:

(i) Let Q be any property and suppose $pqN \to \infty$, then the following two assertions are equivalent:

(a) Almost every graph in $G(n,p)$ has Q.

(b) Given $x > 0$ and $\epsilon > 0$, in n is sufficiently large, then there are $\boxed{\ell > (1-\varepsilon)\, 2x\sqrt{pqN}}$ integers $M_1, M_2, \cdots, M_\ell$ with

$$pN - x\sqrt{pqN} < M_1 < M_2 \cdots < M_\ell < pN + x\sqrt{pqN}$$

such that $P_{M_i}(Q) > 1 - \varepsilon$ for every i, $i = 1, 2, \cdots, \ell$.

(ii) If Q is a convex property and $pqN \to \infty$, then almost every graph in $G(n,p)$ has Q iff for every fixed x almost every graph in $G(n,M)$ has Q where $M = \lfloor pN + x\sqrt{pqN} \rfloor$.

(iii) If Q is any property and $\boxed{0 < p = \frac{M}{N} < 1}$, then

$$P_M(Q) \leq P_p(Q)\, e^{\frac{M}{6}} \sqrt{2\pi pqN} \leq 3\sqrt{M}\, P_p(Q)$$

- Note how the theorem deals with the conceptual difficulties created by the fact that the graphs in the set $G(n,p)$ are allowed to have different number of edges, but every graph in the set $G(n,M)$ must have exactly the same number of edges — M. The theorem sets up an equivalence NOT between one set $G(n,p)$ and one set $G(n,M)$, but between one set $G(n,p)$ and a series of sets $G(n, M_i)$ for different values of i. Each $G(n, M_i)$ is a separate probability space.

- The proof of Part (i) of the theorem involves the following steps: (a) For the probability space $G(n,p)$, we use the DeMoivre-Laplace theorem to estimate the total probability mass that resides in all the graphs whose edge-counts, as given by the binomial random variable $e(G)$, lie between $pN - x\sqrt{pqN}$ and $pN + x\sqrt{pqN}$. (b) We next use another limiting property of binomial distributions which says that as $n \to \infty$, the maximum value of the distribution is upper-bounded by $\frac{1}{\sqrt{2\pi pqN}}$. In conjunction with the probability mass estimated in the previous step, this limit on the maximum value of the distribution is used to estimate a lower bound ℓ on the number of graphs in $G(n,p)$ whose edge counts are in the range $pN - x\sqrt{pqN}$ and $pN + x\sqrt{pqN}$. (c) We next construct a collection of $G(n, M_i)$ sets with M_i taking on values in the range $pN - x\sqrt{pqN}$ and $pN + x\sqrt{pqN}$. In keeping with the nature of the $G(n, M_i)$, the probability associated with an M_i-edge graph is $\frac{1}{\binom{N}{M_i}}$. (d) We denote the smallest of these probabilities by $1 - \varepsilon$. (e) Finally, we show by construction that we can conceive of ℓ probability spaces of type $G(n, M_i)$ with graphs similar to the $G(n,p)$ graphs in step (a).

- The most important conclusion to be drawn from Part (i) of the Theorem is that if we know $P_M(Q)$ with a fair accuracy for every M close to pN, then we also know $P_p(Q)$ with comparable accuracy. This we can do even for those properties that may not be true for all graphs.

- However, the converse of the above is not always true, especially for properties that may not hold for all graphs in $G(n,p)$. To illustrate, let Q be the property that a graph in $G(n,p)$ has an even number of edges. Obviously, $P_p(Q) \approx \frac{1}{2}$ as $n \to \infty$. To construct $G(n,M)$ graphs for which M will be in the vicinity of pN, we note that pN will not be an integer, in general. So we may choose $M = 2\lfloor pN/2 \rfloor$. But now $P_M(Q) = 1$ for every n as $n \to \infty$.

- But we want to be able to go from making assertions in the $G(n,p)$ probability space to making assertions in the $G(n,M)$ probability space since the former probability space is frequently more convenient to work with.

- As it turns out, it is possible to take assertions from $G(n,p)$ to $G(n,M)$ for the following three cases: (1) When the property is convex (this is by virtue of Part (ii) of the Theorem; (2) When the assertion consists of the expected value of a random variable whose value increases with the size of the graph; and (3) When we average the expectations of an arbitrary property with respect to all M for $G(n,M)$ and all p for $G(n,p)$.