Lectures 6–7: Network Formation: Static Models

Alexander Wolitzky

MIT

6.207/14.15: Networks, Spring 2022

Plan

Next unit: probabilistic models of network formation

- There are many different probability models describing what networks are likely to form/be observed.
- We study some of the most important ones.
  - Some more important as mathematically elegant/tractable benchmark models.
  - Others more important for generating "realistic" networks.
  - Need different models because what's "realistic" depends on context/application.

Lectures 6–7: **Static** models of network formation.

- Network forms "all at once."
- Tends to include simplest/most canonical models.

Lecture 8: **Dynamic** models of network formation.

- Network forms "over time."
- Important for understanding power law distributions, homophily, and other realistic features.

Why Study Random Graphs?

Random graph models are especially important for understanding large networks.

- Once go beyond a small handful of nodes, often not very tractable or meaningful to analyze the exact structure of a particular network.
- E.g. the Facebook friend graph within each country is a completely different graph. But presumably these graphs were formed by similar processes and are thus likely to have some important common features.
- We'll see that random graph models are helpful for understanding properties like connectivity (e.g. "small worlds"), clustering/homophily, robustness/fragility, and diffusion processes on large networks.

Why Study Random Graphs? (cntd.)

At the same time, real social and economic networks usually form as the result of deliberate choices, **not** (purely) randomly.
This matters too and we'll come back to it later in the course.

- Network properties that would be very unlikely from a random graph perspective can be prevalent in real-world networks, precisely because real-world networks are formed strategically.
- E.g., star networks are unlikely to form purely by chance, but we often see star networks ("hub-and-spokes") in transportation and information networks, precisely because they create many short paths while economizing on links.

So real-world social and economic networks have both random and strategic aspects, and we'll study both.

The Erdös-Renyi Model

Most of Lectures 6–7 studies the simplest and best-known random graph model: Erdos-Renyi (ER) random graphs.

The ER random graph model is simply that, given n nodes, each possible (undirected) link forms with **independent** probability $p \in (0, 1)$.

- Natural and important mathematical benchmark or starting point for analyzing network formation.
- Can also be realistic in settings where independence assumption is not too implausible (we'll see examples).
- In most social and economic networks, independence is not a good assumption.
  - E.g. clustering, homophily, degree distribution.
- But sometimes simple generalizations of ER do much better and involve similar math.
  - E.g. there could be different types of individuals, where some types form links with higher probability.

Basic Properties of ER Random Graphs

We start by consider some basic properties of ER random graphs.

- Number of Links
- Degree Distribution
- Clustering
- Diameter and Average Path Length.

Basic Properties: # of Links

Let $l_{ij} \in \{0, 1\}$ be Bernoulli random variable indicating presence of link $\{i, j\}$.

Since the number of possible links is $\frac{n(n-1)}{2}$, we have

$$\mathbb{E}\left[\# \text{ of links}\right] = \mathbb{E}\left[\sum_{i \neq j} l_{ij}\right] \Bigg(= \frac{n(n-1)}{2} p.$$

Moreover, by the weak law of large numbers, for all $\alpha > 0$,

$$\lim_{n \to \infty} \mathbb{P} \quad \sum_{i \neq j} l_{ij} - \frac{n(n-1)}{2} p \; \geq \alpha \frac{n(n-1)}{2} \Bigg) \Bigg(= 0.$$

Hence, the number of links is a random variable, but for large n it is **tightly concentrated around its mean**.

ER vs. Gilbert

Technically speaking, ER's original model assumed a fixed number of links m: out of the $n\left(n-1\right)/2$ possible links, randomly form m of them.

The model where each link forms with independent probability p was introduced at the same time by another mathematician named Edgar Gilbert.

However, since the number of realized links is tightly concentrated around its mean, the ER model with $m = pn\left(n-1\right)/2$ links behaves very similarly to the Gilbert model with link probability p.

- This is true for properties that are usually unaffected by adding or removing a very small number of random links, which are most properties we care about.
- For other properties it might make a difference, for example if we ask what is the probability that the number of links is even.

In this class we follow the standard modern practice of focusing on the independent link probability case and calling this the ER model.

Basic Properties: Degree Distribution

Let D be a random variable that represents degree of a node.

D is a binomial random variable with $\mathbb{E}\left[D\right] = \left(n-1\right)p$. That is,

$$\mathbb{P}\left(D=d\right) = \binom{n-1}{d} p^{d} \left(1-p\right)^{n-1-d}.$$

If we let $p = \lambda/\left(n-1\right)$ (i.e. keep the expected degree constant at λ as $n \to \infty$), then the **Poisson limit theorem** says, as $n \to \infty$ D converges to a Poisson random variable, with

$$\mathbb{P}\left(D=d\right) = \frac{e^{-\lambda}\lambda^{d}}{d!}.$$

For this reason, the ER model is also called the **Poisson random graph model**.

This degree distributions falls off (faster than) **exponentially** in d. (E.g., **not** a power-law distribution).

Basic Properties: Conditional Degree Distribution

What is the expected degree of a given node (say, node number 17) *conditional* on the event that it is linked to another given node (say, node number 1)?

- This equals 1 plus the expected number of neighbors of node 17 other than node 1, conditional on the event that 17 and 1 are linked.
- Since links are independent, the latter conditional expectation is simply $(n-2)\,p$, so the overall expected degree is $1 + (n-2)\,p$.
- If we let $p = \lambda / (n-1)$ and take n large, the overall expected degree converges to $1 + \lambda$.
- That is, in a large network, λ is the expected number of "friends" of a given node, and is also the expected number of **other** friends of a friend of mine.
- This is a simple example of the "friendship paradox": on average, your friends have more friends than you do!
- (We'll see that the friendship paradox can be even more severe in other random graph models.)

Basic Properties: Clustering

In expectation, both overall and individual clustering equal p.

- Probability that any "potential triangle" becomes an actual triangle is p.

We will usually consider the case where $p \longrightarrow 0$ as $n \longrightarrow \infty$, so expected degree is finite (or at least much less than n).

In this case, expected clustering goes to 0 as $n \longrightarrow \infty$.

This is an important **unrealistic** feature of ER graphs: most social and economic networks have significant clustering, but ER graphs do not.

Basic Properties: Diameter/Average Path Length

Within each component of size k, by branching process approximation (as in Lecture 1), average path length and diameter are both approximately $\log k / \log \lambda$.

- Branching process approximation usually works well in ER graphs with finite expected degree, because cycles are "rare" in these graphs.

This is a small number within each component.

- The distribution of component sizes is a crucial issues that we will discuss at length.

Small diameter/average path length within each component is realistic for many social and economic networks.

- ER random graphs exhibit "small worlds" (at least within each component).

Asymptotic Properties

Most analysis of ER graphs focuses on **asymptotic properties.**

- This means we let p be a function of n, and ask whether the probability that the realized network has a certain property goes to 0 or 1 as $n \to \infty$.
- This approach allows a clearer mathematical analysis than focusing on finite n: for any fixed n every realized network arises with some positive probability, so assessing the probability that a certain feature arises is difficult.
- Due to law of large numbers-type considerations, asymptotic analysis is usually a good guide for reasonably large networks.

Example: We will see that, if $p(n)$ is below $\log n/n$ then the probability that the network is connected goes to 0 as $n \to \infty$; while if $p(n)$ is above $\log n/n$ then this probability goes to 1.

- But exactly calculating $\Pr(\text{connected})$ for fixed n is very hard.

Threshold Functions

For a given property A (e.g. “the network is connected”, “the network contains at least one cycle”, “the network contains at least one edge”), we say a function $t(n)$ is a **threshold function** if

$$\begin{aligned} \mathbb{P}(A) &\rightarrow 0 \qquad \text{if } \lim_{n\to\infty} \frac{p(n)}{t(n)} = 0, \qquad \text{and} \\ \mathbb{P}(A) &\rightarrow 1 \qquad \text{if } \lim_{n\to\infty} \frac{p(n)}{t(n)} = \infty. \end{aligned}$$

“Property A almost never holds if the link probability is significantly less than $t(n)$, and property A almost always holds if the link probability is significantly greater than $t(n)$.”

- We sometimes write $p(n) \ll t(n)$ for $p(n)/t(n) \rightarrow 0$, and write $p(n) \gg t(n)$ for $p(n)/t(n) \rightarrow \infty$.

This definition makes sense for **monotone properties**: properties such that if a given network (N, E) satisfies it, then so does any network (N, E') with $E \subseteq E'$.

Phase Transitions

If a threshold function exists, we say that a **phase transition** occurs at that threshold.

Analyzing phase transitions let us get clear qualitative insights from models that at first glance seem very complicated.

- E.g., "Large random graphs are connected if the expected degree grows fast than $\log n$" vs. "Simulating network formation on 1000 nodes 1000 times tells me that the network is connected with probability approximately .978."

Finding phase transitions (like the $\log n / n$ threshold for connectivity) was one of Erdos and Renyi's main contributions.

- This was a landmark in graph theory and discrete math. The most cited of Erdos's 1500+ papers.

Simple Example of a Phase Transition: Edges

Define property A as $A = \{\text{number of edges} > 0\}$.

We thus seek a threshold for the emergence of the first edge.

We claim that $t(n) = 1/n^2$ is a threshold function for this property.

- "If $p(n) \ll 1/n^2$ then for large n the network is very likely to have no edges; but if $p(n) \gg 1/n^2$ then for large n the network is very likely to have at least one edge."

Let's prove it. We must prove two things:

1. If $n^2 p(n) \to 0$ then $\mathbb{P}(\#\text{edges} > 0) \to 0$.
2. If $n^2 p(n) \to \infty$ then $\mathbb{P}(\#\text{edges} > 0) \to 1$.

Proof

Recall that $\mathbb{E}\left[\#\text{edges}\right] = \frac{n(n-1)}{2} p\left(n\right) \approx \frac{n^2}{2} p\left(n\right)$.

- First, suppose $n^2 p\left(n\right) \to 0$ as $n \to \infty$.
  Then $\mathbb{E}\left[\#\text{edges}\right] \to 0$.
  This implies that $\mathbb{P}\left(\#\text{edges} > 0\right) \to 0$. (Why?)
  This is the first thing we needed to prove.
- Second, suppose $n^2 p\left(n\right) \to \infty$ as $n \to \infty$.
  Then $\mathbb{E}\left[\#\text{edges}\right] \to \infty$.
  If this implied that $\mathbb{P}\left(\#\text{edges} > 0\right) \to 1$, we'd be done.
  Is this implication valid? That is, are there networks where $\mathbb{E}\left[\#\text{edges}\right] \to \infty$ but $\mathbb{P}\left(\#\text{edges} > 0\right) \not\to 1$?

Proof (cntd.)

- To complete the proof, recall that edge formation is independent, so by the law of large numbers the distribution of edges is tightly concentrated around its mean.
- Formally, number of edges follows a binomial distribution, and hence converges to a Poisson distribution (with mean $\approx \frac{n^2}{2} p(n)$). So

$$
\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\left(\#\text{edges} = 0\right) &= \lim_{n\to\infty} \frac{e^{-\frac{n^2}{2}p(n)} \left(\frac{n^2}{2} p(n)\right)^k}{k!}{}_{k=0} \\
&= \lim_{n\to\infty} e^{-\frac{n^2}{2}p(n)}.
\end{aligned}
$$

- If $n^2 p(n) \to \infty$, this goes to 0.
- This completes the proof that $t(n) = 1/n^2$ is a threshold function for the emergence of links.

A Similar Example: Trees

What is a threshold function for the emergence of **connected triples**?

- Note that

$$\mathbb{E}\left[\#\text{connected triples}\right] = \left(\#\text{triples}\right) p^2 \approx n^3 p^2 .$$

- By a similar analysis as for pairs, a threshold function for the emergence of connected triples is $t(n)$ such that $n^3 p^2$ is constant: that is, $t(n) = 1/n^{3/2}$.

What is a threshold function for emergence of **trees with k nodes**?

- $\mathbb{E}\left[\#\text{k-node trees}\right] \approx n^k p^{k-1}$, so a threshold function is $t(n) = 1/n^{k/k-1}$.

As $p(n)$ increases from $1/n^2$ to $1/n$, we expected to find bigger and bigger trees in the network.

Cycles

What's a threshold function for emergence of a **cycle with k nodes?**

- As we'll see, this turns out to be $t(n) = 1/n$.
- Arbitrarily big trees emerge "before" *any* cycles!

Cycles

What's a threshold function for emergence of a **cycle with k nodes?**

- As we'll see, this turns out to be $t(n) = 1/n$.
- Arbitrarily big trees emerge "before" *any* cycles!

Intuition: Think about growing a tree vs. growing a cycle.

- A tree grows to size k if $k - 1$ links form from nodes in the tree to the infinitely many nodes outside the tree.
- For the tree to grow into a cycle, a link must form from a node already in the tree to one of the finitely many other nodes in the tree. This is much less likely.
- This logic also suggests (correctly) that the threshold for the emergence of a cycle is the same as that for the emergence of a connected component that contains a **positive fraction** of all the nodes in the network.
- Such a component is called a **giant component**.
  This is an important concept in ER graphs.

Component Structure

- Below the threshold of $1/n$, the largest component of the graph includes no more than $c \log(n)$ nodes for some constant c.
- Above the threshold of $1/n$, a **giant component** emerges, which contains a positive fraction of the nodes: that is, at least cn nodes for some constant c.

  (There can't be two giant components, because once $p(n) \geq 1/n$ and there are two components with cn nodes, the probability that there are no links between the components goes to 0.)
- As $p(n)$ increases further, the giant component grows in size, until $p(n)$ reaches the threshold of $\log n/n$, at which point the network becomes connected.
- See Figures 4.4–4.7 in Jackson's book for pictures.

Key questions:

- Why $1/n$ threshold for emergence of giant component?
- Why $\log n/n$ threshold for emergence of connectivity?

Connectivity

Theorem (Erdös and Renyi)

In the ER model, a threshold function for connectivity is $t(n) = \log n / n$.

Let $p(n) = r\frac{\log n}{n}$. We'll show that

$$\begin{aligned} \text{If } r &< 1 \text{ then } \mathbb{P}(\mathit{connected}) \to 0, \\ \text{If } r &> 1 \text{ then } \mathbb{P}(\mathit{connected}) \to 1. \end{aligned}$$

- This implies the theorem, because if $p(n) / t(n) \to 0$ then $\mathbb{P}(\mathit{connected})$ is smaller than it is when $p(n) = r\frac{\log n}{n}$; and if $p(n) / t(n) \to \infty$ then $\mathbb{P}(\mathit{connected})$ is larger than it is when $p(n) = r\frac{\log n}{n}$.

We'll give a heuristic argument.

- See Jackson's book for the proof if you're curious [optional].

Intuition for r<1 Case

To show $r < 1 \implies \mathbb{P}\left(connected\right) \to 0$, we show that the probability that there exists at least one **isolated node** goes to 1.

- If there's an isolated node, the network is disconnected. (Is the converse true?)
- Focusing on the distribution of the number of isolated nodes is the key idea in the proof.

The probability that any given node is isolated is

$$(1-p)^{n-1} \approx e^{-pn} = e^{-r\log(n)} = n^{-r},$$

where the approximation comes from $\lim_{n\to\infty}\left(1-\frac{a}{n}\right)^n = e^{-a}$.

Note that the events that distinct nodes are isolated are **not** independent.

- However, it is intuitive that the correlation between such events is not very large.
- We will therefore proceed **as if** these events are independent, without rigorously justifying this approximation.

Intuition for r<1 Case (cntd.)

Let $X = \#$isolated nodes.

Our approximation gives $\mathbb{E}[X] \approx n \cdot n^{-r}$.

- If $r < 1$ then $\mathbb{E}[X] \to \infty$.
- If $r > 1$ then $\mathbb{E}[X] \to 0$.

This is strongly suggestive of what we're trying to prove.

- If $r < 1$ then on average there are many isolated nodes.
- If $r > 1$ then on average there are 0 isolated nodes.

Intuition for r<1 Case (cntd.)

$\mathbb{E}[X] \to \infty$ does **not** immediately imply that $\Pr(X = 0) \to 0$.

However, under our approximation that the events that distinct nodes are isolated are "almost independent," the number of isolated nodes will be close to $\mathbb{E}[X]$ with high probability.

Therefore, $\mathbb{P}(X \geq 1) \to 1$, and hence $\mathbb{P}(\textit{connected}) \to 0$, completing the proof.

Intuition for r>1 Case

We have seen that $r > 1$ implies that $\mathbb{E}[X] \to 0$ (under our approximation).

- This implies $\mathbb{P}(X \geq 1) \to 0$.
- Does this imply connectivity?
- **No.**
- The event "the graph is disconnected" is equivalent to the existence of a set of k nodes without an edge to the remaining nodes, for some $k \leq n/2$ (but not necessarily $k = 1$).
- To complete the proof, can show that the prob of having any "isolated set" of k nodes (for any $k \leq n/2$) is not much higher than that of having an isolated individual node.
- Intuitively, increasing k can make it a little more likely that there is an isolated set of k nodes, because there are **more** different sets of k nodes than there are individual nodes.
- However, a larger set of nodes is much less likely to have no links to the remaining nodes, and this effect quickly swamps the effect of having more sets to check.

The Giant Component

We just argued that, when $p(n) \ll \log n / n$, the network is disconnected with high probability.

What does the component structure look like in this case?

- In this regime, we've seen that the expected number of isolated nodes (and hence the number of components) goes to infinity.
- **Question:** when does *every* component contain a vanishing fraction of the nodes, and when is there instead a "giant component" with a constant fraction of nodes?
- We'll see that $t(n) = 1/n$ is a threshold function for the existence of a giant component.

The Giant Component (cntd.)

The basic idea for why the threshold function is $1/n$ comes from **branching process approximation**.

- Suppose we knew there were no cycles, i.e. the network is a tree. (Not necessarily true, but useful thought experiment.)
- Then, starting from a given node, what's the expected number of nodes in the tree rooted at this node?
- If $p(n) = \lambda/n$, the root has on average λ distance-1 neighbors, λ^2 distance-2 neighbors, etc..
- (Why is expected number of distance-2 neighbors λ^2 and not $(\lambda - 1)^2$?)
- If $\lambda < 1$, the root's component contains on average $\frac{1}{1-\lambda} < \infty$ nodes.
  - In fact, this is an overcount, since in reality there can be cycles.
  - This implies $\mathbb{P}(\text{giant component}) = 0$, because if $\mathbb{P}(\text{giant component}) = \alpha$ and $\mathbb{E}[\text{size of giant component}|\text{it exists}] = qn$, then $\mathbb{E}[\text{root node's component size}] \geq \alpha \times q \times qn = \infty$.

The Giant Component (cntd.)

- If $\lambda > 1$, the series $\lambda + \lambda^2 + \ldots$ diverges, so in expectation the root's component contains infinitely many nodes.
    - This implies the network will have at least one infinitely large component; we'll see that one of them will actually be "giant," in the sense of having at least *cn* nodes.

More Details: lambda<1

Theorem

Let $p(n) = \lambda / n$ with $\lambda < 1$. For all sufficiently large $c > 0$, we have

$$\mathbb{P}\left(\max_{1 \leq i \leq n} |S_i| \geq c \log n\right) \Bigg(\!\longrightarrow 0,$$

where $|S_i|$ is the size of the component that contains node i.

- That is, if $\lambda < 1$ then with high probability not only are all components of size $\ll n$, they're all of size at most $c \log n$.
- We won't cover the proof.

More Details: lambda>1

A somewhat more detailed argument for why a giant components exists when $\lambda > 1$ (in the actual ER model, not just the branching process approximation):

- Fix a root node.
- Let N_k^{ER} denote the (random) number of distance-k neighbors in the ER model.
- Let N_k^{ER} denote the (random) number of distance-k neighbors in the branching process with the same degree distribution.
- Clearly, $N_k^{ER} \leq N_k^{B}$.
- We want to show it's not "much less."
  That is, the expected number of "overlaps" is small.

More Details (cntd.)

When $p\left(n\right) = \lambda / n$, the expected number of “overlaps” in the branching process is very small until we reach a distance k from the root node such that $N_k^{ER} \geq \sqrt{n}$.

Hence, $N_k^{ER} \approx N_k^{B}$ whenever $\lambda^k \leq c\sqrt{n}$.

When $\lambda > 1$, this implies $\mathbb{P}\left(\exists \text{component with size } > c\sqrt{n}\right) \to 1$.

Moreover, between any two components of size $\sqrt{n}$, the probability of having a link is

$$1 - \left(1 - \frac{\lambda}{n}\right)^{\sqrt{n} \times \sqrt{n}} \approx 1 - e^{-\lambda},$$

which is a positive constant independent of n.

So (intuitively) components of size $> \sqrt{n}$ connect to each other with high probability, forming a connected component of size cn for some $c > 0$.

Size of the Giant Component

When $\lambda > 1$, it's also relatively straightforward to compute the **fraction of nodes** in the giant component.

- As we'll see, this is an important quantity for understanding contagion and diffusion in networks.

Let q be the expected fraction of nodes in the giant component of an n-node network.

- Assume that, for large n, q is also approximately the fraction of nodes in the giant component of an $n+1$-node network (a safe assumption).
- The probability that node $n+1$ is not in the giant component is given by $1-q$.
- The probability that node $n+1$ is not in the giant component is equal to the probability that none of its neighbors are in the (n-node) giant component, so

$$1 - q = \sum_{d} P\left(d\right)\left(1-q\right)^{d}.$$

- This equation has a fixed point $q^{*} \in (0,1)$.

Size of the Giant Component (cntd.)

We can simplify the fixed point equation for q^* as follows:

We have

$$\begin{aligned} 1 - q &= \sum_d P(d)(1-q)^d \\ &\approx \sum_d \frac{e^{-\lambda}\lambda^d}{d!}(1-q)^d . \end{aligned}$$

Recall the math fact that

$$\sum_d \frac{(\lambda(1-q))^d}{d!} \approx e^{\lambda(1-q)}.$$

Therefore, we have

$$q \approx 1 - e^{-\lambda q}.$$

- **Important equation:** this gives the size of giant component as a function of expected degree.

Size of the Giant Component (cntd.)

The size of the giant component is given by

$$q \approx 1 - e^{-\lambda q}.$$

This does not give a simple closed-form solution for q as a function of λ, but we can still plot q as a function of λ:

- First plot λ as a function of q, given by
  $$\lambda = -\frac{\log(1-q)}{q}.$$
- Then swap the axes.

If do this, find that

- $q = 0$ until λ reaches 1.
- Then q increases as a concave function of λ.
  - $q \approx .8$ when $\lambda = 2$. $q \approx .94$ when $\lambda = 3$.
- $q \to 1$ as $\lambda \to \infty$.

An Application: Contagion and Diffusion

- Consider a society of n individuals.
- A randomly chosen individual is infected with a contagious virus.
    - E.g. actual disease, new idea/technology, new fad/fashion.
- Assume the network of interactions in the society is described by an ER graph w/ link prob p.
- Assume any individual is immune (e.g. vaccinated) w/ prob π.
- Question: on average, what fraction of society gets infected?

An Application: Contagion and Diffusion

- Consider a society of n individuals.
- A randomly chosen individual is infected with a contagious virus.
    - E.g. actual disease, new idea/technology, new fad/fashion.
- Assume the network of interactions in the society is described by an ER graph w/ link prob p.
- Assume any individual is immune (e.g. vaccinated) w/ prob π.
- Question: on average, what fraction of society gets infected?
- Can analyze by considering ER graph on $(1-\pi)\,n$ nodes (i.e. removing immune nodes) w/ link prob p, and then determining size of component containing the initially infected node.
- Let $\lambda = p\,(1-\pi)\,n$ denote expected degree after removing immune nodes. (This is the expected number of others that each each infected individual infects.)

Contagion and Diffusion (cntd.)

Three cases:

- $\lambda < 1$:

$$\mathbb{E}\left[\text{fraction infected}\right] \leq \frac{\log n}{n} \approx 0.$$

- $1 < \lambda < \log\left(\left(1-\pi\right)n\right)$:

$$\mathbb{E}\left[\text{fraction infected}\right] = \frac{qq\left(1-\pi\right)n + \left(1-q\right)o\left(n\right)}{n} \approx q^2\left(1-\pi\right),$$

  where q denotes fraction of nodes in the giant component of graph with $\left(1-\pi\right)n$ nodes, i.e. $q = 1 - e^{-\lambda q}$.

- $\lambda > \log\left(\left(1-\pi\right)n\right)$:

$$\mathbb{E}\left[\text{fraction infected}\right] \approx 1 - \pi.$$

Thus, the size of the giant component plays a key role in analyzing diffusion in ER graphs.

- We will return to this when we study diffusion on networks in more detail next week.

Summary: Erdos-Renyi Model

- The ER model is the simplest random graph model.
  Key assumption: independent link formation.
- The ER model is tractable but has some unrealistic features such as very low clustering. However, it's an important benchmark/starting point.
- A key feature of the ER model is threshold phenomena, such as thresholds for the emergence of a giant component or the emergence of connectivity.
- In the ER model, there are either only small components, or one giant component and many small components.
  Understanding the size and structure of the giant component (when it exists) is often important, e.g. for studying diffusion or average path length.

Extensions of the Erdos-Renyi Model

We next consider a couple well-known extensions of the Erdos-Renyi model that can accommodate some more realistic features.

The **configuration model** (Bendor and Canfield, 1978) generalizes ER by allowing an arbitrary degree distribution, rather than being restricted to the Poisson distribution.

- This is a general model for generating “ER-like” networks but with any desired degree distribution.

The **small world model** (Watts and Strogatz, 1998) starts with a “lattice” network with high clustering, and then adds random links as in ER to recover the desirable small worlds properties of ER graphs while maintaining clustering.

- This is a specific simple model for generating clustering while maintaining some nice features of ER networks.

The Configuration Model

Recall that the ER model leads to a Poisson degree distribution, which falls off very fast.

- This may not be realistic. E.g., degree distribution in Facebook friend graph has a long right tail.

The idea of the configuration model is to specify a desired degree distribution in advance and then generate a random network with (approximately) this degree distribution.

The configuration model behaves similarity to the ER model in many ways, but it is flexible enough to accommodate any desired degree distribution, rather than being restricted to the Poisson distribution.

Configuration Model

Start with a **degree sequence** $(d_1, \dots, d_n)$ which specifies the desired degree for each node $i \in N$.

- Alternatively, start with a degree distribution $P(d)$ and generate the degree sequence by sampling iid from $P(d)$.

Given $(d_1, \dots, d_n)$, construct a sequence where node 1 is listed d_1 times, node 2 is listed d_2 times, and so on:

$$\underbrace{1, 1, 1, \dots, 1}_{d_1 \text{ times}} \underbrace{2, 2, \dots, 2}_{d_2 \text{ times}} \cdots \underbrace{n, n, n \dots, n}_{d_n \text{ times}}$$

- Imagine giving each node d_i "stubs" sticking out of it, which are the ends of edges-to-be.

Then, randomly pick two elements of the sequence and form a link between the the corresponding nodes.

- Delete those entries from the sequence and repeat, until no entries remain.

Configuration Model (cntd.)

Remarks:

- The sum of the degrees must be even (or else an entry will be left over at the end).
- It is possible to have more than one link betwen two nodes (so we technically generate a "multigraph").
- It is also possible to have self-loops.

However, if n is large and degrees are bounded, then the proportion of multi-links and self-links will be small.

- If we simply delete all multi-links and self-links at the end, we will be left with a standard graph, and with high probability the degree sequence will be close to the original one.

If n is large and the degree sequence is formed by iid draws from a Poisson degree distribution, the distribution over networks generated by the configuration model and the ER model will be almost the same.

- Thus, the configuration model is essentially a generalization of the ER model.

Basic Properties

Let's start by asking about the same basic properties of the configuration model as we did for ER: degree distribution, conditional degree distribution, clustering.

- Degree distribution is an input in the configuration model, so nothing to investigate here.
- As in the ER model, in the typical case where $\mathbb{E}\left[d\right]/n \to 0$ as $n \to \infty$, clustering goes to 0 as well.

Conditional degree distribution is more subtle and is very important for understanding the component structure (e.g. when there's a giant component).

- For any nodes $i \neq j$, what is the degree distribution of node j conditional on the event that node j is linked to node i?

Conditional Degree Distribution

- Suppose the degree sequence is generated by iid sampling from distribution $P(d)$.
- Consider some node i and then pick a random neighbor j.
- What's the distribution of d_j?

Naive guess: $P(d)$. But this is wrong.

- For example, in the ER model, the expected degree of node j in this situation is $\lambda + 1$, not λ as under the unconditional degree distribution.
- Also, there is no way to reach a node of degree 0 by this method!
- **Explanation:** Higher-degree nodes are involved in a higher percentage of links.
- If we follow a random link, we're likely to end up at a higher-degree node.
- This is the **friendship paradox** again.

The Friendship Paradox (PSet 1 Review)

Fix any network (N, E) with $|E| = m$ (not necessary generated by the configuration model or any other random graph model).

What is the degree of a **randomly chosen node** (i.e. average degree over nodes)?

$$\mathbb{E}\left[d\right] = \frac{\sum_{i\in N} d_i}{n} := \left\langle d\right\rangle .$$

What is the degree of the **node at the randomly chosen end of a randomly chosen link** (i.e. average degree over link-ends)?

- Each node i is selected with probability $d_i/2m$.
  - There are $2m$ link-ends, and d_i of them belong to node i.
- Therefore, the answer is

$$\sum_{i\in N}\left(\frac{d_i}{2m}\right)\Bigg(d_i = \frac{\sum_{i\in N} d_i^2}{2m} = \frac{n\mathbb{E}\left[d^2\right]}{n\mathbb{E}\left[d\right]}\Bigg(= \frac{d^2}{\left\langle d\right\rangle} = \left\langle d\right\rangle + \frac{\mathrm{var}\left(d\right)}{\left\langle d\right\rangle}.$$

- **Important.** Mean-square degree shows up when nodes are sampled with probability proportional to their degree.

Conditional Degree Distribution (cntd.)

For an arbitrary network, the degree distribution of a randomly chosen neighbor of a randomly chosen node is **not** the same as the degree distribution of a node at the randomly chosen end of a randomly chosen link.

- You computed both of these in PSet 1.
  They weren't the same.

For example, in a star with n nodes, if randomly choose node and then random choose neighbor, mean degree equals

$$\frac{n-1}{n} \times (n-1) + \frac{1}{n} \times 1 = n - 2 + \frac{2}{n} \approx n,$$

while if random choose link-end, mean degree equals

$$\frac{1}{2} \times (n-1) + \frac{1}{2} \times 1 = \frac{n}{2}.$$

Conditional Degree Distribution (cntd.)

However, in the configuration model, each stub connects to each of the $2m - 1$ other stubs with equal probability.

- Intuitively, picking a random neighbor is the same as following a random link.
- More carefully, each stub connects to a given degree-d node with probability
$$\frac{d}{2m-1} \approx \frac{d}{2m}.$$
- Each node is "oversampled" in proportion to its degree, exactly as when we choose a node at the end of a random link.
- Therefore, in the configuration model, the degree distribution of a randomly chosen neighbor of a given node (randomly chosen or not) **is** the same as the degree distribution of a node at the randomly chosen end of a randomly chosen link.

Conditional Degree Distribution (cntd.)

We just saw that, in the configuration model, each stub connects to a given node with degree d with probability

$$\frac{d}{2m-1} \approx \frac{d}{2m}.$$

Since the share of nodes with degree d is $P(d)$, each stub connects to some node with degree d with probability

$$\frac{d}{2m} \times nP(d) = \frac{dP(d)}{\langle d \rangle}$$

(since $\langle d \rangle = 2m/n$).

- This is the **conditional degree distribution**.

In particular, the expected degree of a neighboring node equals

$$\sum_d \left(\frac{dP(d)}{\langle d \rangle} \right) \Bigg(d = \frac{d^2}{\langle d \rangle} = \langle d \rangle + \frac{\mathrm{var}(d)}{\langle d \rangle}.$$

Examples

In the ER model, $P\left(d\right)$ is Poisson, and hence $\operatorname{var}\left(d\right) = \langle d\rangle$, so $d^2 \;\; = \langle d\rangle + \langle d\rangle^2$. Therefore,

$$\frac{d^2}{\langle d\rangle} = 1 + \langle d\rangle = 1 + \lambda.$$

- This matches what we found earlier.

For regular networks with degree k, we have

$$\frac{d^2}{\langle d\rangle} = \frac{k^2}{k} = k.$$

- For regular networks, conditioning on being linked to a given node does not affect the degree distribution.

For scale-free networks with $P\left(d\right) = cd^{-\gamma}$, then $d^2 \;\; = \sum_d cd^{-\gamma}\left(d^2\right)$. (This equals ∞ if $\gamma < 3$.

- For scale-free networks, the expected degree of a neighboring node is often **infinite**!

Existence of the Giant Component

We can easily use our formula for the expected degree of a neighboring node to determine when a giant component exists in the configuration model.

For any given starting node, the expected number of distance-2 neighbors (under branching process approximation) is

$$\langle d \rangle \underbrace{\left(\frac{d^2}{\langle d \rangle} - 1 \right)}_{\text{“reproduction number”}} = d^2 - \langle d \rangle .$$

Similarly, expected number of distance-3 neighbors is

$$\left(d^2 - \langle d \rangle \right) \left(\frac{d^2}{\langle d \rangle} - 1 \right) = \langle d \rangle \left(\frac{d^2}{\langle d \rangle} - 1 \right)^2 .$$

Expected number of distance-k neighbors is

$$\langle d \rangle \left(\frac{d^2}{\langle d \rangle} - 1 \right)^{k-1} .$$

Existence of the Giant Component (cntd.)

Expected number of distance-k neighbors is

$$\langle d \rangle \quad \frac{d^2}{\langle d \rangle} - 1 \Bigg)^{k-1} .$$

As in the ER model, a giant component exists if and only if this diverges as $k \to \infty$.

This is the case if and only if

$$\frac{d^2}{\langle d \rangle} - 1 \geq 1 \iff \frac{d^2}{\langle d \rangle} \geq 2.$$

That is, giant component exists iff, on average, each of your neighbors has at least one neighbor other than you.

- Reproduction number ≥ 1.

Examples

In the ER model, $d^2 \;\; / \langle d \rangle = 1 + \lambda$.

- Hence, a giant component exists in the ER model if and only if $\lambda > 1$.
- This matches what we found earlier: threshold function $t(n) = 1/n$.

For regular networks with degree k, $d^2 \;\; / \langle d \rangle = k$.

- Hence, a giant component exists if $k > 2$.
  (Also $k = 2$, but this case is tricky since it's on the boundary.)

For scale-free networks with $P(d) = cd^{-\gamma}$, $d^2 \;\; / \langle d \rangle = \infty$ if $0 < \gamma < 3$.

- Hence, a giant component exists if $0 < \gamma < 3$.

Application to Contagion and Diffusion

If immunize fraction π of the population, reproduction number falls to

$$R = \quad \frac{d^2}{\langle d \rangle} - 1 \Bigg) \Bigg(1 - \pi) \, .$$

Therefore, $R > 1$ (and hence a giant component emerges) iff

$$\frac{d^2}{\langle d \rangle} > 1 + \frac{1}{1-\pi},$$

or equivalently

$$\pi < \pi^* := 1 - \frac{\langle d \rangle}{\langle d^2 \rangle - \langle d \rangle}.$$

Note: Setting $\pi = 0$ recovers the usual condition $\quad d^2 \quad / \langle d \rangle > 2$ for existence of a giant component.

Here π^* is called the **contagion threshold** (or **percolation threshold**): to prevent infection of a positive fraction of nodes by removing random nodes, must remove at least π^* of them.

Examples

ER model: $d^2 \quad = \langle d \rangle^2 + \langle d \rangle$.
This yields threshold $\pi^* = 1 - \frac{1}{\langle d \rangle}$.

For a regular graph with degree k, threshold is

$$\pi^* = 1 - \frac{1}{k-1}.$$

- If $k = 1$ or 2, giant component never emerges.
- If $k = 3$, giant component emerges whenever less than half the population is immune.

Examples (cntd.)

Scale-free graph: $P(d) \sim d^{-\gamma}$, $\gamma < 3$.

- Then d^2 diverges, so contagion threshold is $\pi^* = 1$: unless *everyone* is immune, a nontrivial fraction will get infected.
- **Intuition:** in scale-free networks, there are many nodes with very degree, and these nodes serve as hubs that cause a giant component to exist even if many nodes are eliminated at random.

However, there's an important flip side to this: while the presence of many high-degree node makes eliminating **random** nodes ineffective for preventing contagion, it makes eliminating **targeted** nodes effective.

- Randomly eliminating 99% of the nodes in a scale-free network does not disconnect it, but can show that eliminating the 3% highest-degree nodes does!
- We'll discuss targeting and other "strategic" aspects of diffusion in Lecture 9.

Size of the Giant Component/Infected Population

When the initially infected node is in the giant component, the size of the giant component is the size of the infected population.

We can again approximate this via a branching process, just like for the ER network.

Let $\tilde{P}$ be the conditional degree distribution, and let $\tilde{q}$ be the probability that the branching process does not die out, starting from a neighboring node:

$$1 - \tilde{q} = \pi + (1 - \pi) \sum_{d=1}^{\infty} \tilde{P}(d) (1 - \tilde{q})^{d-1} .$$

Let q be the probability that the branching process does not die out, starting from a random node:

$$1 - q = \sum_{d=0}^{\infty} P(d) (1 - \tilde{q})^{d}$$

The fraction of nodes in the giant component is precisely q.

Small-World Models

The ER model has an unrealistically concentrated degree distribution and unrealistically low clustering.

Generalized random graph models (like the configuration model) can address the unrealistic degree distribution but not the unrealistically low clustering.

A "cheap" way to address low clustering would be to divide society into groups and specify that everyone only links within their own group.

- But this would create a new unrealistic feature, namely that the network is disconnected, or possibly involves high diameter/average path length.
- In contrast, the ER model looks good in terms of diameter/average path length: by branching process approximation, within the giant component, diameter is approximately $\log n / \log \lambda$.

Small-World Models (cntd.)

These considerations have led to interest in developing “small-world models” that are simple to generate and analyze while still having **all** of:

- Realistic degree distribution.
- Small diameter and average path-length.
- High clustering.

The Watts-Strogatz Small-World Model

The most famous such model (despite being very special) is due to Watts and Strogatz (1998).

Watts-Strogatz start with a ring with n nodes, where each node is linked to its $2k$ closest neighbors.

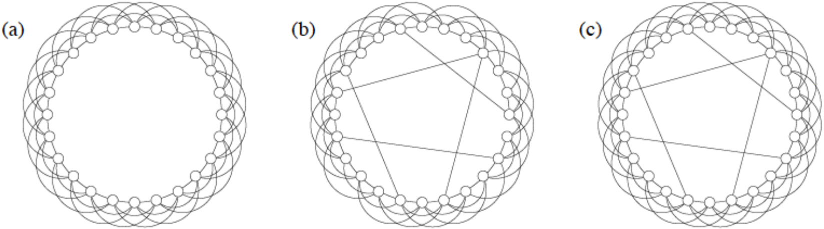
- This network has high clustering but also high diameter/average path-length.

Then, randomly "rewire" a small fraction p of the edges.

- This creates nkp "shortcuts" in the graph.
- For small p, clustering is hardly affected, but diameter/average path-length decrease dramatically: for any $p > 0$, for lage n, diameter/average path-length is proportional to $\log n$ (rather than n, as in the case without rewiring).
- As $p \to 1$, we recover an ER random graph with kn edges.

Can also consider a variant where add fraction p random edges, rather than rewiring.

The Watts-Strogatz Small-World Model



- (a) before rewiring
- (b) after rewiring
- (c) adding edges instead of rewiring

Average Path-Length

Heuristic argument that average path-length is proportional to $\log n$:

- Suppose we add $r = nkp$ random edges.
- Imagine dividing the ring into r **intervals** of length $n/r = 1/kp$.
- Now consider the network where each **interval** is seen as a **node**, and these nodes are connected by the random edges.
- This is an **ER network** with r nodes and r links, so expected degree equals 2, so there is a giant component, which consists of a fraction of nodes that is independent of n.
- Average path-length within the giant component is approximately $\log r$.
- Since interval-length $1/kp$ is constant, nodes not in the giant component can reach the giant component in a constant number of steps c/kp along the ring.
- Hence, diameter is bounded by $\frac{1}{kp}(\log r + c) \approx \log r \approx \log n$.

Summary: Configuration Model and Small-World Model

- The configuration model is a generalization of the ER model that allows for arbitrary degree distributions.
- The configuration model is analyzed similarly to the ER model, but must be careful to distinguish the conditional and unconditional degree distributions.
- Neither the ER model nor the configuration model has realistic clustering. A simple model that obtains realistic clustering while maintaining realistic average path length/diameter is the Watts-Strogatz small world model.