6.207/14.15: Networks
Lecture 5: Generalized Random Graphs and Small-World Model

Daron Acemoglu and Asu Ozdaglar
MIT

September 23, 2009

Outline

- Generalized random graph models
- Graphs with prescribed degrees – Configuration model
- Emergence of a giant component in the configuration model
- Small-world model
  - Clustering
  - Average path lengths

Reading:

- Jackson, Sections 4.1.2, 4.1.4-4.1.6, 4.2.1, 4.2.6, 4.2.7.
- EK, Chapter 20.

Configuration Model—1

- We have seen that the Erdös-Renyi model has a Poisson degree distribution, which falls off very fast.
- Our next goal is to generate random networks with a "given degree distribution".
- One of the most widely method used for this purpose is the configuration model developed by Bender and Canfield in 1978.
- The configuration model is specified in terms of a degree sequence, i.e., for a network of n nodes, we have a desired degree sequence $(d_1, \ldots, d_n)$, which specifies the degree d_i of node i, for $i = 1, \ldots, n$.
  - Given a degree distribution $P(d)$, we can generate the degree sequence for n nodes by sampling the degrees independently from the distribution $P(d)$, i.e., $d_i \sim P(d)$.
  - A law of large numbers argument establishes that the frequency of degrees $P^{(n)}(d)$ converges to the degree distribution $P(d)$ as n goes to infinity.

Configuration Model—2

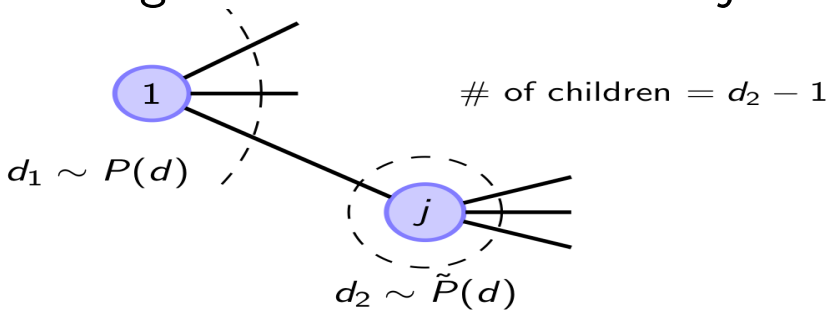
- Given $(d_1, \ldots, d_n)$, we construct a sequence where node 1 is listed d_1 times, node 2 is listed d_2 times, and so on:

$$\underbrace{1, 1, 1, 1, \ldots, 1}_{d_1 \text{ entries}} \quad \underbrace{2, 2, \ldots, 2}_{d_2 \text{ entries}} \quad \cdots \quad \underbrace{n, n, n \ldots, n}_{d_n \text{ entries}}.$$

- We can think of this as giving each node i in the graph d_i "stubs" sticking out of it, which are ends of edges-to-be.
- We randomly pick two elements of the sequence and form a link between the two nodes corresponding to those entries.
- We delete those entries from the sequence and repeat.
- Remarks:
  - The sum of degrees needs to be even (or else an entry will be left out at the end).
  - It is possible to have more than one link between two nodes (thus generating a "multigraph").
  - Self-loops are possible.

Distribution of the Degree of a Neighboring Node—1

- We will use a branching process approximation to study the giant component in the configuration model.
- For this we need to understand the distribution of the degree of a neighboring node, i.e., given some node i with degree d_i, consider a neighbor j. What is the degree distribution of node j?



- Naive intuition: Same distribution as node i. Example: Consider a graph with 4 nodes and links {1,2}, {2,3}, {3,4}.
  - We have $P(1) = P(2) = 1/2$.
  - If we randomly pick a link and then randomly pick an end of it, there is a $2/3$ chance of finding a node with degree 2 and $1/3$ chance of finding a node with degree 1.
  - Reflects the fact that higher degree nodes are involved in a higher percentage of the links.

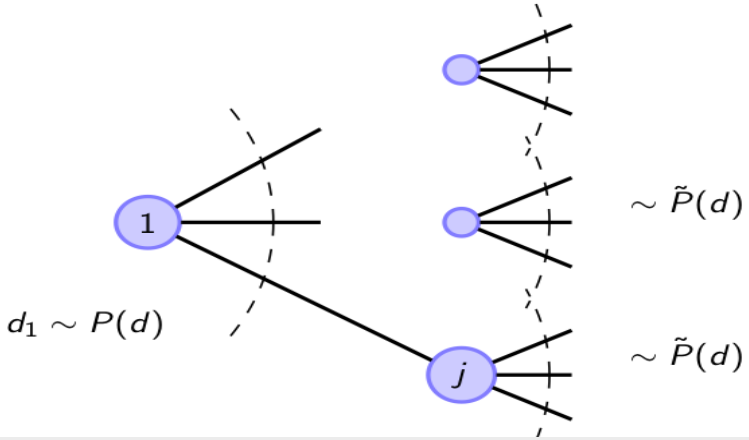
Distribution of the Degree of a Neighboring Node—2

- The degree of a node we reach by following a randomly chosen edge is not given by $P(d)$.
- There are d edges that arrive at a node of degree d, we are d times as likely to arrive at that node than another node that has degree 1.
- Thus, the degree distribution of the neighboring node $\tilde{P}(d)$ is proportional to $dP(d)$,

$$\tilde{P}(d) = \frac{dP(d)}{\sum_k kP(k)} = \frac{dP(d)}{\langle d \rangle}.$$

- Another way to see this is:

$$\tilde{P}(d) = \frac{\text{endpoints attached to degree } d \text{ nodes}}{\text{total number of endpoints}} = \frac{dnP(d)}{n\sum_k kP(k)}.$$



Emergence of a Giant Component in the Configuration Model—1

- We will use a branching process approximation to analyze the emergence of the giant component.
  - We ignore self loops (can be shown to have small probability) and conflicts (do not matter until the graph grows to a substantial size).
- Note that we have

$$\begin{aligned}
\mu &= \tilde{\mathbb{E}}[\text{number of children}] = \tilde{\mathbb{E}}[d-1] \\
&= \sum_d d\tilde{P}(d) - 1 \\
&= \sum_d \frac{d^2 P(d)}{\langle d \rangle} - 1 \\
&= \frac{\langle d^2 \rangle}{\langle d \rangle} - 1.
\end{aligned}$$

Emergence of a Giant Component in the Configuration Model—2

- Using the branching process analysis, this yields the following threshold for the emergence of the giant component:

  Subcritical: $\mu < 1$, or equivalently

  $$\frac{\langle d^2 \rangle}{\langle d \rangle} < 2 \quad \Leftrightarrow \quad \langle d(d-2) \rangle < 0.$$

  Supercritical: $\mu > 1$, or equivalently

  $$\langle d(d-2) \rangle > 0.$$

- In the case of an Erdös-Renyi graph, we have $\langle d^2 \rangle = \langle d \rangle + \langle d \rangle^2$, and so the giant component emerges when

  $$\langle d^2 \rangle > \langle d \rangle \quad \Leftrightarrow \quad \langle d \rangle > 1.$$

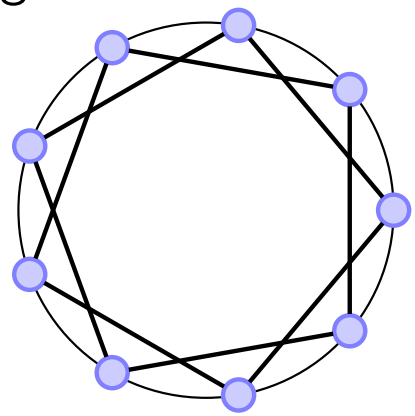
- Since $\langle d \rangle = (n-1)p$ in the Erdös-Renyi graph, this indeed yields the threshold function $t(n) = \frac{1}{n}$ for the emergence of the giant component.

Small-World Model

- Erdös-Renyi model has short path lengths (recall the giant component analysis using branching process approximation). However, they have a Poisson degree distribution and low clustering.
- Generalized random graph models (such as the configuration model) effectively addresses one of the shortcomings of the Erdös-Renyi random graph model, its unrealistic degree distribution.
- However, they fail to capture the common phenomenon of clustering observed in social networks.
- A tractable model that combines high clustering with short path lengths is the small-world model, proposed by Watts and Strogatz in 1998.
- The model follows naturally from combining two basic social network ideas: homophily (the tendency to associate to those similar to ourselves) and weak ties (the links to acquaintances that connect us to parts of the network that would otherwise be far away).
    - Homophily creates high clustering while the weak ties produce the branching structure that reaches many nodes in a few steps.

Small-World Model

- The small-world model posits a network built on a low-dimensional regular lattice (capturing geographic or some other social proximity), and then adding or moving random edges to create a low density of "shortcuts" that join the remote parts of the lattice to one another.
- The best studied case is a one-dimensional lattice with periodic boundary conditions, i.e., a ring.
- We consider a ring with n nodes and join each node to its neighbors k or fewer hops (lattice spacings) away.
  - This creates nk edges.



$k = 2$

Figure: A ring lattice with $k = 2$.

Small-World Model

- The small-world model is then created by taking a small fraction p of the edges in this graph and "rewiring" them.
- The rewiring procedure involves going through each edge in turn, and with probability p, moving one end of that edge to a new location chosen uniformly at random from the lattice.
  - Expected number of total shortcuts is nkp.

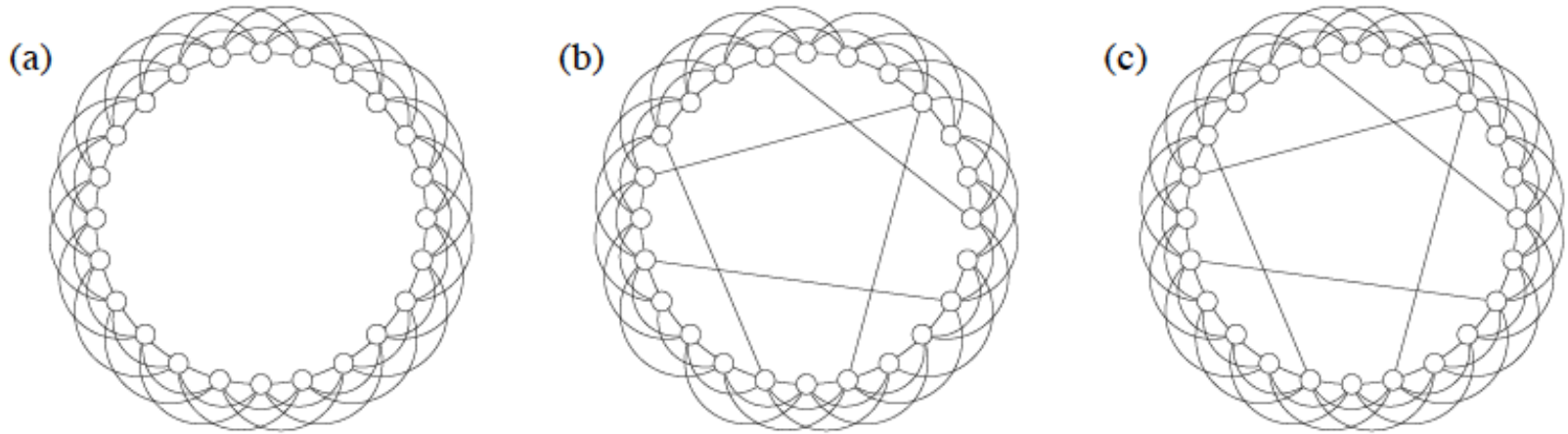


Figure: A small world model with $k = 3$; part (a) illustrates $p = 0$, part (b) illustrates rewiring with probability $p > 0$, part (c) illustrates addition of random links with probability $p > 0$.

Small-World Model

- A more mathematically tractable variant of the model was proposed by Newman and Watts in 1999.
  - No edges are rewired. Instead "shortcuts" joining randomly chosen node pairs are added to the ring lattice.
  - The parameter p is defined as the probability per edge on the underlying lattice of there being a shortcut in the graph (to make it similar to the previous model).
  - Hence, the mean total number of shortcuts is nkp and mean degree is $2k + 2kp$.

Clustering vs Path Lengths in the Small World Model

- Addition of random links allows the small-world model to interpolate between a regular lattice (p=0) and a random graph.
  - Regular lattice has high clustering $Cl(g) = \frac{3k-3}{4k-2}$, long paths $O(\frac{n}{k})$.
  - Random graph has low clustering and short paths.
- Watts and Strogatz showed by numerical simulation that there exists a sizable region in between the two extremes in which the model has both low path lengths and high clustering.

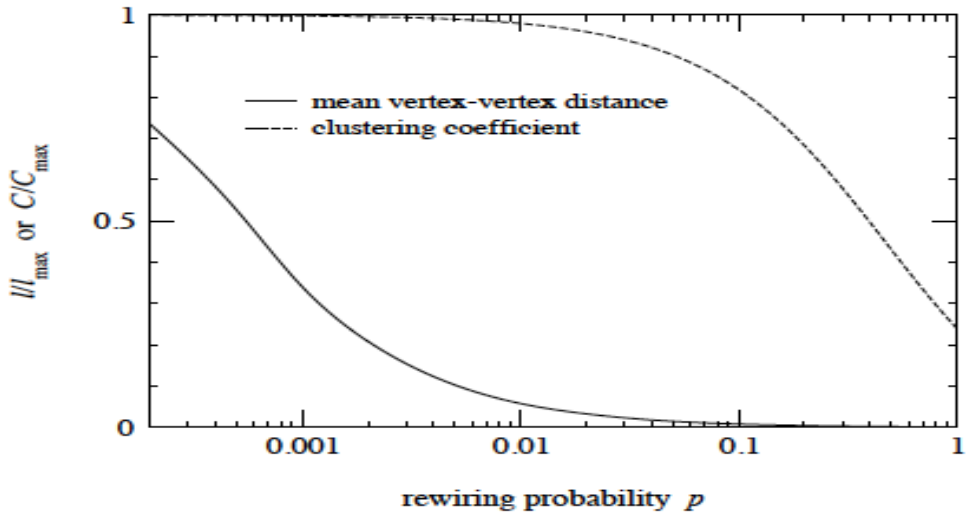
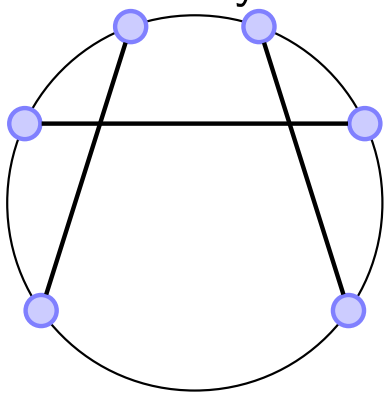


Figure: Clustering coefficient and average path length in the small-world model of Watts and Strogatz.

Average Path Length in the Small-World Model—1

- We next show that the average path length in the small-world model is small [i.e., $O(\log(n))$].
- To simplify the analysis,we consider a continuum approximation:
  - We take a continuum of nodes around a ring with unit circumference.
  - We throw u random shortcuts: we choose u pairs of points independently and uniformly at random, and connect them.



circumference=1

- Let $f(u)$ denote the expected distance along the circle between two random points on this graph (assumes shortcuts have 0 distance).
- Since the unit circumference in the continuous model maps into n arcs in the discrete model, average distance $f(u)$ maps into $\frac{nf(u)}{k}$ arcs.

Average Path Length in the Small-World Model—2

- We next show that for large u, $f(u)$ can be approximated as $f(u) = \frac{\log(u)}{u}$.
- Since $u = npk$, this implies that the average distance in terms of number of arcs satisfies
$$\frac{nf(u)}{k} = \frac{n\log(npk)}{npk^2} \approx \log(n).$$
- We analyze the continuous model by discretizing it into u intervals of length $\delta = 1/u$.
- The shortcuts generated can be represented as an Erdös-Renyi model with the δ-length intervals corresponding to the nodes.
- With this identification, we have
$$\mathbb{E}[\text{number of edges}] = u, \qquad \mathbb{E}[\text{number of end points}] = 2u,$$
$$\mathbb{E}[\text{degree}] = 2.$$

Average Path Length in the Small-World Model—3

- Hence, the link formation probability satisfies $p(n) = \frac{2}{n}$, suggesting that there exists a giant component.
- Any two nodes in the giant component (or intervals in the continuous model) can be connected by a path of $\log(u)$ nodes (or intervals).
- Moreover, it can be shown that any node (interval) which is not in the giant component can be connected to the giant component on average in a constant c number of nodes (intervals).
- Hence the distance between any two intervals satisfies:

$$\text{distance} \leq \log(u) \cdot \frac{1}{u} + \frac{c}{u},$$

showing that $f(u) \approx \frac{\log(u)}{u}$.