1.022 Introduction to Network Models

Amir Ajorlou

Laboratory for Information and Decision Systems
Institute for Data, Systems, and Society
Massachusetts Institute of Technology

Lecture 12

Growing random networks

- Static random graph models:
  - edges among "fixed" n nodes are formed via random rules in a static manner.

– Erdös-Renyi model: small distances, but low clustering and a rapidly falling degree distribution.

– Small-world model: small distances, high clustering.

- Most networks form dynamically:
  - new nodes are born over time
  - attach to existing nodes when they are born.

– Examples: creation of web pages, citations, professional relationships.

– Evolution over time introduces a natural heterogeneity to nodes based on their age in a growing network.

– https://www.youtube.com/watch?v=4GDqJVtPEGg

A bit of history on power laws

- In a power law distribution, the tails fall off polynomially with power α.

$$\mathbb{P}(X \geq x) \sim cx^{-\alpha},$$

  for constants $c > 0$ and $\alpha > 0$.
- Power law degree distributions are observed frequently in real-world networks.
- The earliest apparent reference is the work by *Pareto* in 1897:
  - studying wealth distributions
  - Pareto observed that there were many more individuals who had large amounts of wealth than would appear in Gaussian or exponential distributions.
- Power laws also appeared in the work of *Zipf* in 1916, in describing word frequencies in documents.
  - *Zipf's Law* states that the frequency of the j^{th} most common word in English (or other common languages) is proportional to j^{-1}.

Power laws in preferential attachment model

- In 1965, *Price* studied the network of citations between scientific papers
- Found that the in-degrees (number of times a paper has been cited) have power law distributions.
- An article would gain citations over time proportional to the number of citations the paper already had.
- Consistent with the idea that researchers find some article (e.g. via searching for keywords on the Internet), and then search for additional papers by tracing through the references of the first article.
- The more citations an article has, the higher the likelihood that it will be found and cited again.
- Price called this dynamic link formation process cumulative advantage.
- Today it is known under the name preferential attachment after the influential work of *Barabasi and Albert* in 1999.

Preferential attachment model

- Nodes are born over time and indexed by their date of birth, i.e., node i is born at date i, $i = 1, \ldots$.
- Start the network with $M = 2m + 1$ nodes (born at times $1, \ldots, M$) all connected to one another.
  - Thus, the first newborn node is the one born at time $M + 1$.
- Each node upon birth forms m (undirected) edges with pre-existing nodes.
- It attaches to nodes with probabilities proportional to their degrees.
- Let $k_i(t)$ be the degree of node i at time t.
- The expected number of edges that an existing node i receives at time $t + 1$ is:

$$m \frac{k_i(t)}{\sum_{j=1}^{t} k_j(t)}.$$

- There are tm total links at time t ($t \geq M$), hence
$$\sum_{j=1}^{t} k_j(t) = 2tm.$$
- The expected number of new edges that node i received at time $t+1$ is $\frac{k_i(t)}{2t}$.
- We can write down the evolution of expected degrees in continuous-time as
$$\frac{d\, k_i(t)}{dt} = \frac{k_i(t)}{2t},$$
with initial condition $k_i(i) = m$ (for $i > M$).
- This equation has a solution:
$$k_i(t) = m\sqrt{\frac{t}{i}}$$
- Expected degrees of nodes are increasing over time.
- How to find the fraction of nodes with degrees above a certain level d at time t? identify which node is exactly at level d at time t.

Preferential attachment degree distribution

- Let $i(d)$ be the node that has degree d at time t, or $k_{i(d)}(t) = d$.
- From the degree expression, this yields

$$i(d) = t\left(\frac{m}{d}\right)^2,$$

- All nodes $1, \ldots, i(d)$ have expected degrees $\geq d$ at time t.
- There are a total of t nodes at time t. Therefore,

$$\mathbb{P}\left(k_i(t) \geq d\right) = \frac{i(d)}{t} = \left(\frac{m}{d}\right)^2.$$

- The (expected) degree distribution is power law.
- This is the argument given by Barabasi and Albert (1999).

Configuration model

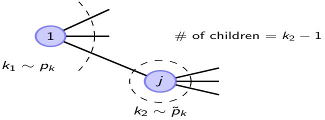
- Goal is to generate random networks with a "given degree distribution".
- One of the most widely method used for this purpose is the configuration model developed by *Bender and Canfield* in 1978.
- Specified in terms of a degree sequence:
  - for a network of n nodes, we have a desired degree sequence $(k_1, \ldots, k_n)$, which specifies the degree k_i of node i, for $i = 1, \ldots, n$.
- Given a degree distribution p_k, we can generate the degree sequence for n nodes by sampling the degrees independently from the distribution p_k, i.e., $k_i \sim p_k$.
- Frequency of degrees $p_k^{(n)}$. $p_k^{(n)} \to p_k$ as $n \to \infty$.

Configuration model

- Given the degree k_i for node i for all $i = 1, \ldots, n$, we create a random network with these degrees as follows:
- Give each node i, k_i "stubs" sticking out of it, which are ends of edges-to-be (there are a total of $\sum_i k_i = 2m$ stubs, where m is the number of edges).
- Choose two stubs uniformly at random and create an edge between the corresponding nodes.
- Choose another pair from the remaining $2m - 2$ stubs, connect those and continue until all the stubs are used up.
- Remarks:
  - This process generates each possible matching of stubs with equal probability.
  - The sum of degrees needs to be even (or else an entry will be left out at the end).
  - It is possible to have self-edges and multiedges.

Degree distribution of a neighboring node

- Needed in studying the giant component in the configuration model.
- 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 $p_1 = p_2 = 1/2$. Pick a link at random, then randomly pick an end of it. What is the degree distribution of this node?
  - there is a 2/3 chance of finding a node with degree 2 and 1/3 chance of finding a node with degree 1.
  - higher degree nodes are involved in a higher percentage of the links.

Degree distribution of a neighboring node

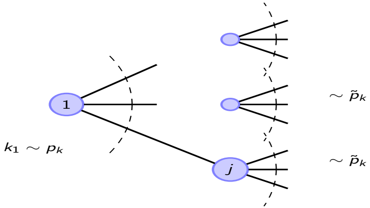
- The degree of a node we reach by following a randomly chosen edge is not given by p_k.
- In the configuration model, an edge emerging from a node has equal chance of terminating at any of the stubs.
- Since there are $2m$ stubs in total, the probability of this edge ending at any particular node of degree k is $k/2m$.
- Since the total number of nodes with degree k is given by np_k, the probability of the edge attaching to a node with degree k is given by

$$\frac{k}{2m} np_k = \frac{kp_k}{\langle k \rangle},$$

where $\langle k \rangle$ is the expected degree in the network and the equality follows from the relation $2m = n\langle k \rangle$.

- Intuitively, there are k edges that arrive at a node of degree k, we are k times as likely to arrive at that node than another node that has degree 1.
- Thus, the degree distribution of the neighboring node $\tilde{p}_k$ is proportional to kp_k,

$$\tilde{p}_k = \frac{kp_k}{\sum_j jp_j} = \frac{kp_k}{\langle k \rangle}.$$



Emergence of a giant component in the configuration model

- 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

$$\begin{aligned}
\mu &= \tilde{\mathbb{E}}[\text{number of children}] = \tilde{\mathbb{E}}[k-1] \\
&= \sum_k k\tilde{p}_k - 1 \\
&= \sum_k \frac{k^2 p_k}{\langle k \rangle} - 1 \\
&= \frac{\langle k^2 \rangle}{\langle k \rangle} - 1.
\end{aligned}$$

Emergence of a giant component in the configuration model

- Using the branching process analysis, this yields the following threshold for the emergence of the giant component:
  Subcritical: $\mu < 1$, or equivalently

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

  Supercritical: $\mu > 1$, or equivalently

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

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

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

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