6.207/14.15: Networks
Lecture 6: Growing Random Networks and Power Laws

Daron Acemoglu and Asu Ozdaglar
MIT

September 28, 2009

Outline

- Growing random networks
- Power-law degree distributions: Rich-Get-Richer effects
- Models:
  - Uniform attachment model
  - Preferential attachment model

Reading:

- EK, Chapter 18.
- Jackson, Chapter 5, Sections 5.1-5.2.

Growing Random Networks

- So far, we have focused on static random graph models in which edges among "fixed" n nodes are formed via random rules in a static manner.
  - Erdös-Renyi model has small distances, but low clustering and a rapidly falling degree distribution.
  - Configuration model generates arbitrary degree distributions.
  - Small-world model provides a tractable model that has small distances and high clustering.
- Most networks form dynamically whereby new nodes are born over time and form attachments to existing nodes when they are born.
- Example: Consider the creation of web pages.
  - When a new web page is designed, it includes links to existing web pages. Over time, an existing page will be linked to by new web pages.
- The same phenomenon true in many other networks:
  - Networks of friendships, citations, professional relationships.
- Evolution over time introduces a natural heterogeneity to nodes based on their age in a growing network.

Emergence of Degree Distributions

- These considerations motivate dynamic or generative models of networks.
- These models also provide foundations for the emergence of natural linkage structures or degree distributions.
- What degree distributions are observed in real-world networks?
  - In social networks, degree distributions can be viewed as a measure of "popularity" of the nodes.
  - Popularity is a phenomenon characterized by extreme imbalances: while almost everyone goes through life known only to people in their immediate social circles, a few people achieve wide visibility.
- Let us focus on the concrete example of World Wide Web (WWW), i.e., network of web pages.
- In studies over many different Web snapshots taken at different points in time, it has been observed that the degree distribution obeys a power law distribution, i.e., the fraction of web pages with k in-links (or out-links) is approximately proportional to $k^{-2.1}$ (or $k^{-2.7}$).

Power Law Distribution—1

- Many social and biological phenomena also governed by power laws.
  - Population sizes of cities observed to follow a power law distribution.
  - Number of copies of a gene in a genome follows a power law distribution.
- Some physicists think these correspond to some "universal laws", as illustrated by the following quote from Barabasi that appeared in the April 2002 issue of the *Scientist*:
  - "What do proteins in our bodies, the Internet, a cool collection of atoms, and sexual networks have in common? One man thinks he has the answer and it is going to transform the way we view the world."
- A nonnegative random variable X is said to have a power law distribution if

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

for constants $c > 0$ and $\alpha > 0$. (Here $f(x) \sim g(x)$ represents that the limit of the ratios goes to 1 as x grows large.)

- Roughly speaking, in a power law distribution, asymptotically, the tails fall of polynomially with power α.

Power Law Distribution—2

- Such a distribution leads to much heavier tails than other common models, such as Gaussian and exponential distributions.
  - In the context of the WWW, this implies that pages with large numbers of in-links are much more common than we'd expect in a Gaussian distribution.
  - This accords well with our intuitive notion of popularity exhibiting extreme imbalances.
- One specific commonly used power law distribution is the Pareto distribution, which satisfies

$$\mathbb{P}(X \geq x) = \left(\frac{x}{t}\right)^{-\alpha},$$

  for some $\alpha > 0$ and $t > 0$.
- The Pareto distribution requires $X \geq t$.
- The density function for the Pareto distribution is $f(x) = \alpha t^{\alpha} x^{-\alpha-1}$.
- For a power law distribution, usually α falls in the range $0 < \alpha \leq 2$, in which case X has infinite variance. If $\alpha \leq 1$, then X also has infinite mean.

Examples

- A simple method for providing a quick test for whether a data-set exhibits a power-law distribution is to plot the (complementary) cumulative distribution function or the density function on a log-log scale.

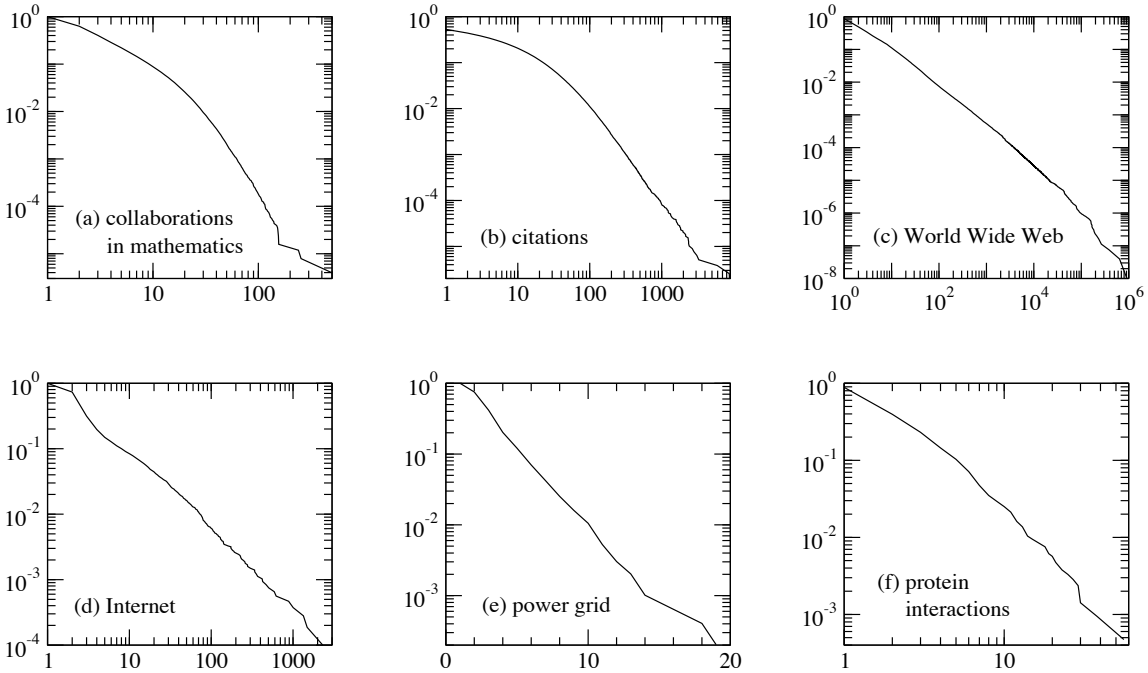

Figure: Cumulative degree distributions for six different networks (degree k vs. the cumulative probability distribution) [Newman 03].

History of Power Laws—1

- Power laws had been observed in a variety of fields for some time.
- The earliest apparent reference is to the work by Pareto in 1897, who introduced the Pareto distribution to describe income distributions.
  - When studying wealth distributions, Pareto observed power law features, where there were many more individuals who had large amounts of wealth than would appear in Gaussian or other distributions.
- Power laws also appeared in the work of Zipf in 1916, in describing word frequencies in documents and city sizes.
  - The empirical principle, known as *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}.
- These ideas were further developed in the work of Simon in 1955, who showed that power laws arise when "the rich get richer", when the amount you get goes up with the amount you already have.

History of Power Laws—2

- Recall the examples:
  - A city grows in proportion to its current size as a result of people having children.
  - Gene copies arise in large part due mutational events in which a random segment of the DNA is accidentally duplicated (a gene which already has many copies more likely to be in a random stretch of DNA)
- All of these examples exhibit rich get richer effects.
- Rich get richer effects quite fragile, there is great sensitivity to unpredictable initial fluctuations.
  - Empirically studied by Salganik, Dodds and Watts (2006): They created a music download site with 48 obscure songs. A visitor to the site can listen to the songs and also is shown the "current" download count for each song.
  - Each visitor at random is assigned to 8 "parallel copies" of the site, which started out identically.
  - Market share of different songs varied considerably across different copies.

History of Power Laws—3

- In 1965, Price applied these ideas to networks, with a particular focus on citation networks.
- Price studied the network of citations between scientific papers and found that the in degrees (number of times a paper has been cited) have power law distributions.
- His idea was that an article would gain citations over time in a manner proportional to the number of citations the paper already had.
- This is 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.

Uniform Attachment Model

- Before studying the preferential attachment model, we discuss a dynamic variation on the Erdös-Renyi model, in which nodes are born over time and form edges to existing nodes at the time of their birth.
- Index the nodes by the order of their birth, i.e., node i is born at date i, $i = 0, 1, \ldots$.
- A node forms undirected edges to existing nodes when it is born. Let $d_i(t)$ be the degree of node i at time t.
- Start the network with $m + 1$ nodes (born at times $0, \ldots, m$) all connected to one another.
  - Thus, the first newborn node is the one born at time $m + 1$.
- Assume that each newborn node uniformly randomly selects m nodes from the existing set of nodes and links to them (ignore repetitions).

Evolution of Expected Degrees

- We will use a continuous-time mean-field analysis to track the evolution of the "expected degrees of nodes".
- We have the initial condition $d_i(i) = m$ for all i, every node has m links at their birth.
- The change at time $t > i$ of the expected degree of node i is given by

$$\frac{d\, d_i(t)}{dt} = \frac{m}{t},$$

since each new node at each time spreads its m new links randomly over the t existing nodes at time t.

- This differential equation has a solution

$$d_i(t) = m + m \log\left(\frac{t}{i}\right).$$

- From this solution, we derive an approximation to the degree distribution.

"Expected" Degree Distribution

- We first note that the expected degrees of nodes are increasing over time.
  - If we ask how many nodes have degree ≤ 100 and we know that a node born at time τ has degree $= 100$ at time t, then we are equivalently asking how many nodes were born on or after time τ.
  - This implies that at time t, the fraction of nodes having degree less than or equal to 100 would be $\frac{t-\tau}{t}$.
- For any d and any time t, let $i(d)$ be a node such that $d_{i(d)}(t) = d$. The resulting cumulative distribution function then is $F_t(d) = 1 - \frac{i(d)}{t}$.
- Applying this technique to the uniform attachment model, we solve for $i(d)$ such that

$$d = m + m \log\left(\frac{t}{i(d)}\right), \qquad \text{which yields} \qquad \frac{i(d)}{t} = e^{-\frac{d-m}{m}},$$

  and therefore the distribution function $F_t(d) = 1 - e^{-\frac{d-m}{m}}$.
- This is an exponential distribution with support from m to infinity and a mean degree of $2m$.

Preferential Attachment Model

- Nodes are born over time and indexed by their date of birth.
- Assume that the system starts with a group of m nodes all connected to one another.
- Each node upon birth forms m (undirected) edges with pre-existing nodes.
- Instead of selecting m nodes uniformly at random, it attaches to nodes with probabilities proportional to their degrees.
  - For example, if an existing node has 3 times as many links as some other existing node, then it is 3 times as likely to be linked to by the newborn node.
- Thus, the probability that an existing node i receives a new link to the newborn node at time t is m times i's degree relative to the overall degree of all existing nodes at time t, or

$$m \frac{d_i(t)}{\sum_{j=1}^{t} d_j(t)}.$$

Preferential Attachment Model

- Since there are tm total links at time t in the system, it follows that $\sum_{j=1}^{t} d_j(t) = 2tm$. Therefore, the probability that node i gets a new link in time t is $\frac{d_i(t)}{2t}$.
- Hence, we can write down the evolution of expected degrees in continuous time as
$$\frac{d\, d_i(t)}{dt} = \frac{d_i(t)}{2t},$$
with initial condition $d_i(i) = m$ (assuming degree is a continuous variable).
- This equation has a solution:
$$d_i(t) = m\left(\frac{t}{i}\right)^{1/2}.$$
- As before, expected degrees of nodes are increasing over time.
- Hence to find the fraction of nodes with degrees below a certain level d at time t, we need to identify which node is exactly at level d at time t.
- Let $i(d)$ be the node that has degree d at time t, or $d_{i(d)}(t) = d$.

Preferential Attachment Degree Distribution

- From the degree expression, this yields

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

leading to the distribution function

$$F(d) = 1 - m^2 d^{-2},$$

with a corresponding density function

$$P(d) = 2m^2 d^{-3}.$$

- Thus, the (expected) degree distribution is a power law with exponent -3.
- This is the argument given by Barabasi and Albert (1999).
- Networks generated by preferential attachment look very different from earlier models with similar average degree.

Master Equation Method—1

- In subsequent work, Dorogovstev, Mendes and Samukhin (2000), took a different approach, using what they call the "master equation" to obtain rigorous asymptotics for the mean degree of the nodes.
- Let p_k denote the fraction of nodes in the network with degree k.
- The probability that a new edge attaches to a node of degree k is

$$\frac{kp_k}{\sum_d dp_d} = \frac{kp_k}{2m},$$

since the mean degree of the network is $2m$ (there are m edges added for each node, and each edge contributes two ends to the degrees of nodes).

- Thus, the mean number of nodes of degree k that gain an edge when a single new node with m edges is added is $m\frac{kp_k}{2m} = \frac{kp_k}{2}$.
- The number of nodes with degree k, given by np_k, thus decreases by this amount (since the nodes that get new edges become nodes with degree $k+1$).

Master Equation Method—2

- The number of nodes with degree k also increases because of influx from nodes of degree $k-1$ that have just acquired a new edge (except for nodes of degree m, which have an influx of exactly equal to 1 due to the addition of the new node with m edges).
- Let $p_{k,n}$ denote the value of p_k when the graph has n nodes.
- Then we can write the dynamics as

$$(n+1)p_{k,n+1} - np_{k,n} = \frac{1}{2}(k-1)p_{k-1,n} - \frac{1}{2}kp_{k,n}, \qquad \text{for } k > m,$$

$$(n+1)p_{m,n+1} - np_{m,n} = 1 - \frac{1}{2}mp_{m,n}, \qquad \text{for } k = m.$$

- Focusing on stationary solutions $p_{k,n+1} = p_{k,n} = p_k$, it follows that

$$p_k = \begin{cases} \frac{1}{2}(k-1)p_{k-1} - \frac{1}{2}kp_k & \text{for } k > m, \\ 1 - \frac{1}{2}mp_m & \text{for } k = m. \end{cases}$$

Master Equation Method—3

- Rearranging for p_k, we find $p_m = 2/(m+2)$ and $p_k = p_{k-1}(k-1)/(k+2)$, or

$$p_k = \frac{(k-1)(k-2)\cdots m}{(k+2)(k+1)\cdots(m+3)} p_m = \frac{2m(m+1)}{(k+2)(k+1)k}.$$

- In the limit of large k, this gives a power law degree distribution $p_k \sim k^{-3}$.