Lecture 8: Network Formation: Dynamic Models and Preferential Attachment

Alexander Wolitzky

MIT

6.207/14.15: Networks, Spring 2022

Growing Random Networks

So far, have focused on **static** random graph models, where edges among n nodes are formed "all at once" according to some pre-specified probability distribution.

- E.g. ER model, configuration model, small-worlds model

In reality, most networks form **dynamically**, where new nodes are born over time and form attachments to existing nodes when they're born.

- Consider the creation of web pages. When each web page is designed, it includes links to existing pages.
- Also: friendship networks, citations, professional contacts.

Evolution over time introduces a natural heterogeneity to nodes: some nodes are older than others, tend to have higher degrees.

- Correspondingly, these models often end up generating networks with some realistic features.
- Also, dynamics provide an *explanation* for why we get a particular degree distribution.

Power Laws

An important realistic feature that often arises in dynamic models is **extremely imbalanced degrees**.

- E.g. in many different Web snapshots, it has been observed that the distribution over websites of the number of in-links (or out-links) approximates a **power law distribution**, where the fraction of websites with k links is approximately proportional to $k^{-\alpha}$, for α between 2 and 3.

Many social, economic, and biological phenomena are well-approximated by power laws.

- Population of cities (with $\alpha \approx 1$, so NYC$\approx$2xLA, 3xChicago).
- Number of employees of firms (with $\alpha \approx 1$).
- Top incomes/wealth (with $\alpha > 1$, perhaps ≈ 2 for income and ≈ 1.5 for wealth, so income is more equal than wealth, and both are more equal than city or firm sizes).
- Number of copies of a gene in a genome.

Power Laws (cntd.)

Caveat: Sometimes people claim that anything with fatter tails than exponential is a power law, even in cases where distribution is not very close to a true power law.

At end of lecture, return to this and compare power laws with other "heavy-tailed" distributions.

Power Laws (cntd.)

Formally, a nonnegative random variable X has a **power law distribution** if its tail falls polynomially with power α: formally,

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

for constants $c > 0$ and $\alpha > 0$, where $f(x) \sim g(x)$ means $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$. (Also called **"fat-tailed."**)

An example of a commonly used power law distribution is the **Pareto distribution**, given by

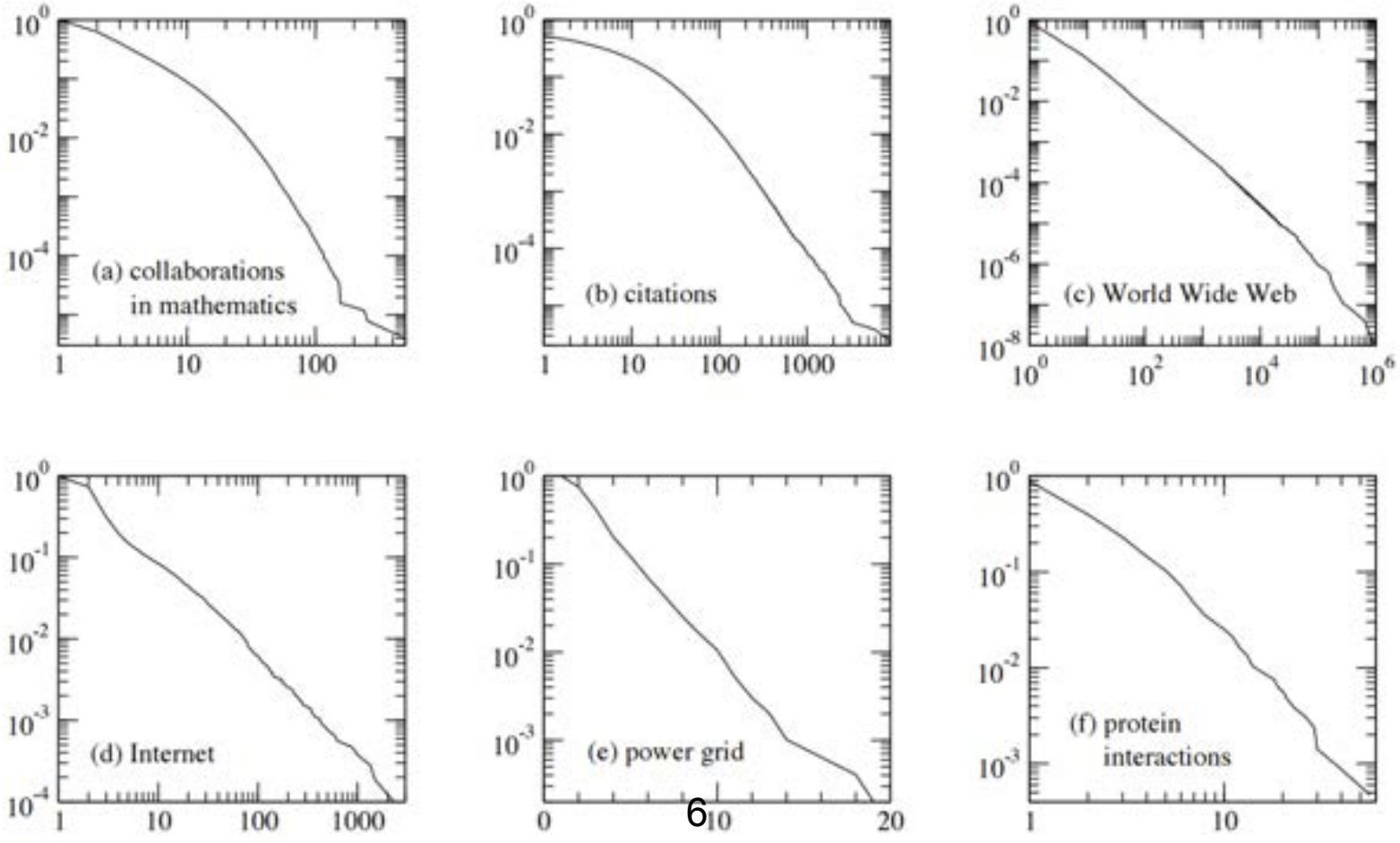
$$\mathbb{P}(X \geq x) = \left(\frac{x}{t}\right)^{-\alpha}$$

for some $\alpha > 0$ and $t > 0$.

- X is supported on $[t, \infty)$.
- The density function is $f(x) = \alpha t^{\alpha} x^{-\alpha-1}$.
- If $\alpha \leq 2$, then X has infinite[5]variance.
- If $\alpha \leq 1$, then X also has infinite mean.

Examples

A simple test for when a data-set exhibits a power-law distribution is to plot the counter-cumulative distribution function or the density function on a log-log scale and see if it looks linear.



History of Power Laws

Power laws have been observed in a variety of fields for some time.

Earliest standard reference is to Pareto in 1897, who introduced the Pareto distribution to describe income distributions.

- Pareto observed that there are many more individuals with large fortunes than would appear in Gaussian or other common distributions.

Another early reference is Zipf 1916, in describing city sizes and English word frequences.

- *Zipf's law:* the frequence of the j^{th} most common word in English (or other common languages) is proportional to j^{-1}.
- "the" =2x "of" =3x "and".

The ideas were further develped by Simon 1955, who showed that power laws arise when *"the rich get richer."*

- As we'll see, power laws can arise when the amount you get depends on the amount you already have.

Rich-Get-Richer

The rich-get-richer mechanism is a plausible explanation for many examples of power laws.

- A city grows in proportion to its current population as people have children.
- Large fortunes may be generated by earning a similar rate of return on a larger initial fortune.
- Gene copies arise in part due to accidental duplications, which occur roughly in proportion to the number of existing copies.

In all of these examples, older nodes will be much "richer" than younger nodes, so the cross-sectional distribution of nodes will have a fat tail.

- Cities that have been growing for longer are bigger.
- Fortunes that have been accruing compound interest for longer are bigger.
- Genomes contain more copies of older genes.

Rich-Get-Richer (cntd.)

Caveat: there's more to "rich-get-richer" than age.

- Big cities have something going for them that attracts people, and then this something together with time for reproduction leads to growth.
- However, a well-established regularity is that the growth rate of cities is independent of their size. So small "initial differences" in population do have a large effect.
- For fortunes and genomes, time can be even more of a dominant factor.

Our point is not that in reality age is always the dominant factor, but rather that a very simple model where age is the only dimension of intrinsic heterogeneity is already one way to generate power laws.

Rich-Get-Richer (cntd.)

The rich-get-richer mechanism implies high sensitivity to initial conditions/fluctuations.

- Salganik, Dodds, and Watts (2006) created a music download site with 48 obscure songs.
- Each visitor to the site can listen to the songs and is also shown the current download count for each song.
- Each visitor is randomly assigned to one of 8 "parallel copies" of the site, which started out identically.
- Final market share of different songs varied widely—more than can be explained by chance if each visitor had decided what to download independent of the download counts.

Cumulative Advantage/Preferential Attachment

Price (1965) applied these ideas to networks, with a particular emphasis on citation networks.

- Found that in-degrees (the number of times a paper has been cited) have power law distributions.
- His explanation was that an article would gain citations over time in a manner proportional to the current number of citations.
- This is exactly what would happen if researchers find articles by reading the references of articles they already know.
- Price called this dynamic link-formation process **cumulative advantage.**

Today this process is called **preferential attachment** after the influential model of Barabasi and Albert (1999) that we'll cover today.

Uniform Attachment Model

Before studying the preferential attachment model, we study a dynamic variation on the ER model, where nodes are born over time and form m links to existing nodes when born, uniformly at random.

This is called the **uniform attachment model**. Features:

- Older nodes have higher expected degrees.
- **But** since links are formed randomly, only very old nodes have degrees much higher than average.
- Specifically, we will see that the fraction of nodes with degree greater than d equals $e^{-\frac{d-m}{m}}$.
- This exponential degree distribution is similar to the Poisson distribution in the usual ER model. In particular, it exhibits "thin tails."
- So the uniform attachment model is a dynamic model of network formation, but it does not generate a power-law degree distribution.

Uniform Attachment Model: Details

Nodes are born over time and randomly form links to existing nodes when born.

- Index the nodes by their birth order, so node i is born at date i, for $i = 0, 1, \ldots$.
- At birth, a node forms m undirected links with existing nodes. Let $d_i(t)$ be the degree of node i at time t.

We consider a convenient version of this model where $d_t(t) = m$ for all $t > m$, and for $t \leq m$ the new node links with everyone.

- Equivalently, the first newborn node is born at time $t = m + 1$, and the pre-existing network at that point is the complete network on m nodes.

Mean-Field Approximation

In this model, analyzing the distribution of realized networks is challenging.

- E.g. The highest degree nodes are very likely to be the oldest ones, so if by chance the old nodes are slow to pick up links the number of high-degree nodes will look very different.

It is much easier to keep track of the evolution of **expected** degrees over time.

- Keeping track of the expected properties of a stochastic process rather than the realized properties is called **mean-field approximation.**
- In this case, we'll also track expected degrees as if the model were in continuous time rather than discrete time.
  - This lets us characterize the evolution of expected degrees according to simple differential equations.
  - This is a second layer of approximation, but the error introduced here is smaller and easier to quantify than that introduced by the mean-field approximation.

Evolution of Expected Degrees

Initial condition: $d_t(t) = m$ for all t.

Starting time at $t = m + 1$, the change in the expected degree of node i at time $t > i$ is given by

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

since each new node links randomly links to m of the t existing nodes.

- Rate of forming new links falls with $1/t$ due to increased competition for links.

This differential equation has solution (for $t \geq i$)

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

- We next use this solution to derive an approximation to the degree distribution.

Distribution of Expected Degrees

Note that the expected degrees of nodes are increasing over time.

- If we ask how many nodes have expected degree ≤ 100 at time t and we know that a node born at time τ has expected degree $= 100$ at time t, then we are equivalently asking how many nodes were born on or after time τ.
- At time t, this fraction is $1 - \frac{\tau}{t}$ (assuming $\tau > m$).

Hence, for any degree d and time t, let $i(d)$ be a node such that $d_{i(d)}(t) = d$. The resulting CDF is $F_t(d) = 1 - \frac{i(d)}{t}$.

We can solve for $i(d)$ according to

$$d = m + m \log \frac{t}{i(d)} \iff \frac{i(d)}{t} = e^{-\frac{d-m}{m}},$$

for $d < m\left(1 + \log \frac{t}{m}\right)$ (the maximum expected degree at time t).

Hence, for $d < m\left(1 + \log \frac{t}{m}\right)$, the fraction of nodes with expected degree less than d is $F_t(d) = 1 - e^{-\frac{d-m}{m}}$.

Distribution of Expected Degrees (cntd.)

We showed that, for $d < m\left(1 + \log \frac{t}{m}\right)$, the fraction of nodes with expected degree greater than d equals $e^{-\frac{d-m}{m}}$.

- For $t = \infty$, this is an exponential distribution with support $[m, \infty)$ and mean $2m$.
  - Exponential, "thin-tailed" distribution, similar to usual ER.
- Interestingly, for $d < m\left(1 + \log \frac{t}{m}\right)$, $F_t(d)$ does not depend on t!
  - How is this possible?
  - The maximum expected degree increases over time, but more and more nodes have smaller expected degrees.
  - Fraction of nodes with d less than any $d_0 < m\left(1 + \log \frac{t}{m}\right)$ is constant over time.
  - Fraction of nodes with $d = m\left(1 + \log \frac{t}{m}\right)$ decreases over time as population grows.
- It can be proved that, as $t \longrightarrow \infty$, $F_t(d)$ is in fact the distribution of *realized* degrees (not just expected degrees), but the proof is beyond our scope.

Preferential Attachment Model

As in the uniform attachment model:

- Nodes are born over time and indexed by their date of birth.
- The system starts with m nodes all connected to each other.
- When born, each new node forms m undirected links with pre-existing nodes.

However, instead of linking randomly, the new node links to pre-existing nodes with **probability proportional to their degrees**.

- Thus, the probability that existing node i gets a link from a new node at time t equals

$$m\frac{d_i(t)}{\sum_{j=1}^{t} d_j(t)} \text{ (instead of } \frac{m}{t} \text{ in uniform model).}$$

This will make a big difference: fraction of nodes with degree greater than d equals $\left(\frac{d}{m}\right)^{-2}$, which for large d is much greater than the corresponding fraction of $e^{-\frac{d-m}{m}}$ in the uniform model.

Evolution of Expected Degrees

We proceed much like in the uniform attachment model.

Starting time at $t = m + 1$, there are tm total links in the system at time t.
Hence, the probability that node i gets a new link at time t equals $m\frac{d_i(t)}{2tm} = \frac{d_i(t)}{2t}$.

This implies that the evolution of expected degrees is given by (under mean-field approximation)

$$\frac{d}{dt} d_i\left(t\right) = \frac{d_i\left(t\right)}{2t},$$

with initial condition $d_t\left(t\right) = m$.

- Rate of forming new links falls with $d_i\left(t\right)/t$, which is slower than $1/t$ rate under uniform attachment.

Evolution of Expected Degrees

We proceed much like in the uniform attachment model.

Starting time at $t = m + 1$, there are tm total links in the system at time t.
Hence, the probability that node i gets a new link at time t equals $m\frac{d_i(t)}{2tm} = \frac{d_i(t)}{2t}$.

This implies that the evolution of expected degrees is given by (under mean-field approximation)

$$\frac{d}{dt} d_i(t) = \frac{d_i(t)}{2t},$$

with initial condition $d_t(t) = m$.

- Rate of forming new links falls with $d_i(t)/t$, which is slower than $1/t$ rate under uniform attachment.

This equation has solution

$$d_i(t) = m\left(\frac{t}{i}\right)^{1/2}.$$

Comparison with Uniform Attachment

With preferential attachment,

$$d_i(t) = m\left(\frac{t}{i}\right)^{1/2}.$$

With uniform attachment,

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

In both cases, older nodes have higher expected degrees, but expected degree increases with age much faster under preferential attachment than under uniform attachment ($\sqrt{t}$ vs. $\log t$).

- Hence, at any point in time, the distribution of expected degrees will be much more dispersed under preferential attachment.

Distribution of Expected Degrees

As in uniform attachment model, nodes' expected degrees are increasing over time.

- To find fraction of nodes with expected degrees below d at time t, suffices to identify which node $i(d)$ has expected degree exactly d at time t.

This is given by

$$d = m\left(\frac{t}{i(d)}\right)^{1/2} \iff \frac{i(d)}{t} = \left(\frac{m}{d}\right)^2,$$

for $d < (mt)^{1/2}$ (the maximum expected degree at time t).

Hence, for $d < (mt)^{1/2}$, the fraction of nodes with degree greater than d is

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

The corresponding density function is

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

Distribution of Expected Degrees (cntd.)

We have shown that, under preferential attachment, the fraction of nodes with degree greater than d at time t equals

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

- Again, for $d < d^{\max}$, the fraction of nodes with degree greater than d does not depend on t.
- This expected degree distribution is a **power law with exponent -3**.
  - Remarkably similar to distribution of web links.
  - Very different from exponential distribution resulting form uniform attachment model.
- As in the uniform attachment model, it can be verified that, as $t \to \infty$, $P_t(d)$ is in fact the distribution of realized degrees. Again, proof is beyond our scope.
- Preferential attachment thus provides an explanation for power-law degree distributions.

Pareto vs. Log-Normal

Another "heavy-tailed" distribution is the **log-normal distribution**: this is the distribution of e^X where X is a normal random variable.

- "Heavy-tailed" means tails falls off slower than exponential. "Fat-tailed" means tails are approximated by a power function $x^{-\alpha}$. So fat-tailed implies heavy-tailed, but not vice versa. The Pareto distribution is fat-tailed (and hence heavy-tailed); the log-normal distribution is heavy-tailed but not fat-tailed.

By the central limit theorem, the geometric (multiplicative) mean of n iid random variables X_i converges to a log-normal distribution.

- Thus, in a population (e.g., of cities, fortunes, or genomes) if growth **rate** is random and independent of current size, the cross-sectional distribution converges to a log-normal distribution.
- This is called the **law of proportionate effect** or **Gibrat's law**, after Robert Gibrat who observed in the 1930s that firm sizes (# employees) followed an approximately log-normal distribution and proposed random growth rates.

Zipf vs. Gibrat

There is some debate over whether things like city population or firm size are better approximated by a Pareto distribution (Zipf's law) or a log-normal distribution (Gibrat's law).

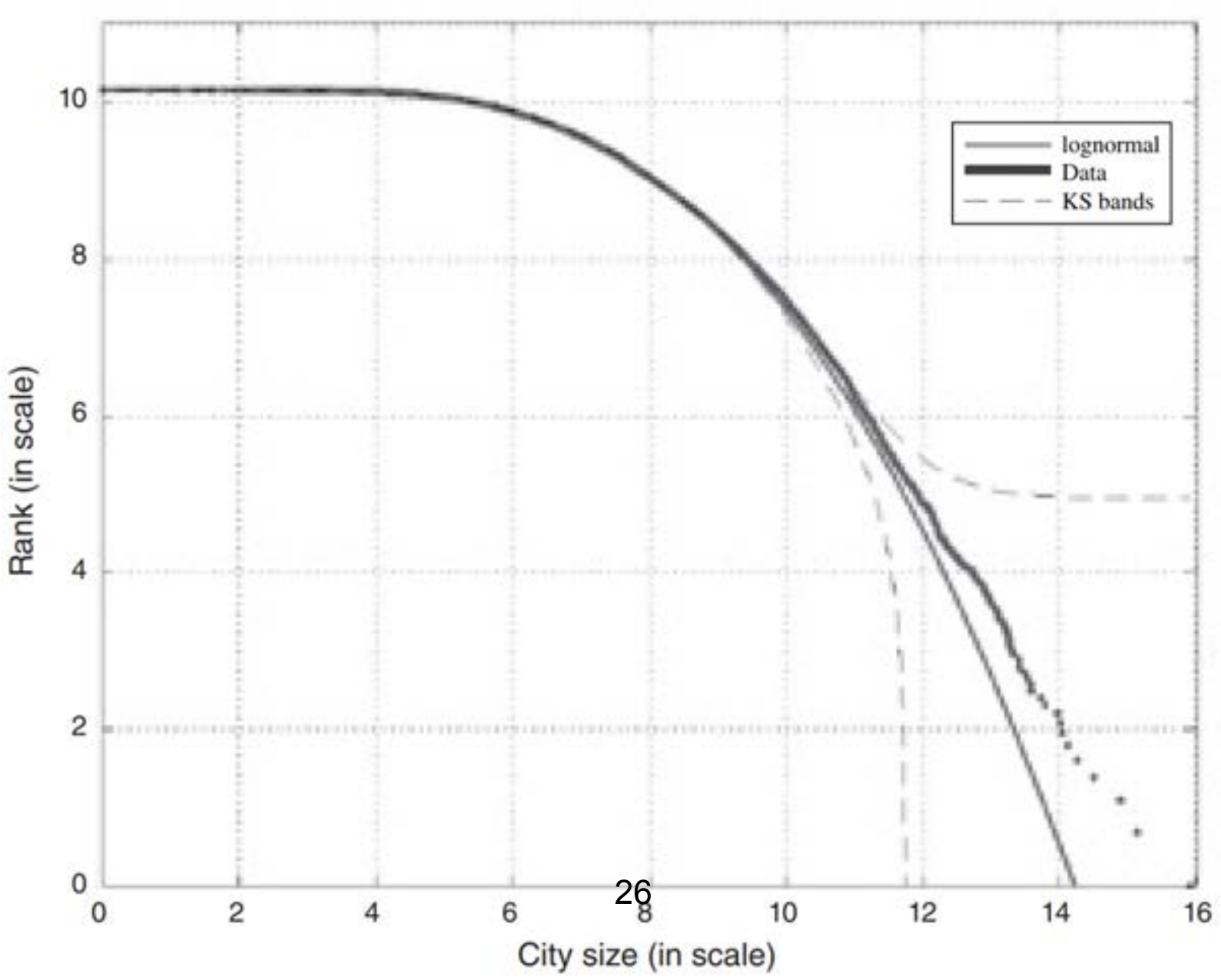
From our analysis today, one important factor determining which is more likely is whether there is a fixed number of growing cities or a growing number of growing cities.

- Fixed number of growing cities with iid growth rates $\Longrightarrow$ log-normal distribution.
- Growing number of growing cities with iid growth rates $\Longrightarrow$ Pareto distribution.
  - This type of process is also called a **Kesten process**, after mathematician Harry Kesten (1973).

At least for cities, the bulk of the distribution is very close to log-normal, but the upper tail may be closer to Pareto.

- The jury is still out.

Zipf vs. Gibrat



Summary

- Dynamic network formation models add realism by modeling how networks form over time.
- If new nodes form links uniformly at random, process generates thin-tailed, ER-like networks.
- If new nodes link to existing nodes in proportion to their degrees, process generates networks with power-law degree distributions.
- Dynamic network formation with prefential attachment is one explanation of power-law degree distributions, as observed among web pages, citation networks, and friendship networks.
- A related type of distribution is a log-normal distribution, which arises from a fixed population of "units" with random growth rates. There is a debate about whether some important empirical distributions are better described as Pareto or log-normal.

14.15 / 6.207 Networks

Spring 2022