# Economics of Information Networks

**Stephen Turnbull**

Division of Policy and Planning Sciences

Lecture 5: December 17, 2020

## Abstract

*In Lecture 5, Part 2 we look at power laws, a kind of statistical distribution that seems to characterize many economic collections: sales of books, videos, and music recordings, number of followers in social networks, and other measures of "popularity."*

# Popularity and power laws

- Consider books, where popularity is measured by number of readers.
- It turns out that a readership of $k$ is achieved by a fraction of publications proportional to $1/k^3$. Kleinberg and Easley report two different measures:
  - the number of *purchasers* of a book, and
  - the total number of *citations* of a scientific paper.
- So the probability of sales of $k$ books is $\mathcal{P}(k) = Ak^{-3}$ where $\sum_{k=1}^{\infty} \mathcal{P}(k) = 1$. So $\mathcal{P}(k)$ satisfies the laws of probability: $\mathcal{P}(k) \leq 1$ and the total probability of all possible $k = 1$.
- So probability of total sales is proportional to a power of total sales, which is why it's called a "power law."A
- A useful trick is to graph a distribution in the *log-log* form, that is, $\log \mathcal{P}(k)$ *vs.* $\log k$, because $\log \mathcal{P}(k) = \log A - a \log k$.

# Why not a power law?

- Since the kinds of networks we're interested in are composed of large numbers of people, the obvious distribution is the *normal distribution* family.
- But popularity distributions are not normal. They're asymmetric, and typically monotonically decreasing: the most frequent class of books (or Twitter accounts) are the ones with very few sales (or followers).
- One alternative is *lognormal distribution*, where the logarithm of sales is normally distributed. The tail of books with few sales is short, and the support of the distribution is positive.
- Another possibility would be a negative exponential distribution, $\mathcal{P}(k) = \alpha e^{-\alpha k}$.
- But popularity distributions have *fat tails*: outliers are much more likely (equivalently outliers at a given percentile rank are much bigger) than these distributions.

# A theory of power law distributions

- Under certain conditions the *Central Limit Theorem ensures* that an aggregate will be normally distributed, so we need a theory of why not.
- We look for violations of the conditions of the theorem: specifically, independence of the quantities aggregated.
- Often, the quantities are actually binary: did a person buy a particular book or not? does a page have a link to a particular page?
- How does one book purchase become (statistically) dependent on another? Or one web page's link targets become dependent on other pages? These are basic questions when investigating if lack of independence might generate non-normality.
- As with information cascades, we will arrange for feedback from other pages into a page's decision where to link.

# A "rich-get-richer" model

1. We are given a parameter $p$, $0 < p < 1$.
2. The first page is $j = 1$. It links to itself.
3. Set $j = j + 1$.
4. Create page $j$, and choose a random Web page $i$ from the existing pages (*i.e.*, one of 1, ..., $j - 1$ with equal probability).
   (a) With probability $p$, page $j$ includes a link *to* this page $i$.
   (b) With probability $1-p$, page $j$ *copies* the link from page $i$.
5. Repeat from Step 3.

# The "rich-get-richer" model, *cont.*

- As the number of pages gets large, the distribution of in-link counts (the number of pages that link to this page) comes to approximate a power law. The limit exponent $a$ on the number of links $k$ in the distribution depends on $p$ turns out to be $a = 1 + \frac{1}{1-p}$.
- This model allows the creation of a single link from page $j$. A fancier model could repeat Step 4 to create multiple, independently generated links from page $j$. This also generates a power law.

# Preferential attachment

- How does this model produce *preferential attachment* (the technical term for "rich get richer")? Consider a page $i$ with $\ell$ in-links. In Step 4b, page $i$ has $\ell$ chances out of $j-1$ to be linked again: its chance of being linked in this case is $\ell$ times that of a page with only one in-link. At this step, each page's chance of being linked is proportional to its number of in-links.
- The probability of being linked in in Step 4b in this model is approximately the expected rate of growth of the number of links of the model. When the rate of growth of a variable is proportional to its current level, it's *exponential growth*.
- The pages with many in-links will accrue more in-links faster than those with a few, so those pages will experience explosive popularity.

# Homework: due 24 December 2020

1. Write a program to simulate the "rich-get-richer" model. Any language is fine (I prefer Python, though). The most important part is the *stopping* rule. *E.g.*, stop when the page with the most links reaches $N$ links, say at $N = 100$. The program should output the distribution of link counts.
2. (Be careful!) What is the stopping rule in the "rich-get-richer" model as presented above?
3. Use a linear regression of the log-log form to determine the value of $a$ that corresponds to various settings of $p$.
4. (Optional) Have the program output a graph of the distribution as an image.
5. (Optional) Have the program output a graph of the distribution in log-log form as an image.
6. (Optional) Enhance the program to allow more than one link per page. Are the distributions produced power laws?

# Homework: due 24 December 2020

1. In an information cascade, we observe preferential attachment. Explain.
2. In the "rich-get-richer" model of web page linking, the preferential attachment is just a mechanical rule. Is preferential attachment in our information cascade examples (the restaurant line and the red/blue urns) "mechanical"? Explain how preferential attachment arises in an information cascade.
3. Describe at least two models of preferential attachment in web page linking based on human psychology. One should explain why pages *copy* links as in the "rich-get-richer" model. Another should give a reason that doesn't involve copying. (*Good* models are an area of active research. An undergraduate student can understand and contribute here, but a *really good* model—and a bit of luck—could get you a Nobel Prize in Economics. I'm not kidding.)

# Path dependence

- Information cascades are useful: we get useful information by watching others' actions. But very bad outcomes may occur if the cascade is *all* the information we get. Which happens depends on the order of arrival.
- In the red/blue urn example, dramatically different outcomes depend on the first three draws. This is an extreme example of *path dependence*.
- A poor technology with a positive network externality can dominate its industry, as long as it starts with a clear lead in users over superior rivals. More path dependence.
- An optimization algorithm in a non-convex problem may converge to a poor local optimum depending on the initialization of the algorithm.
- This is very different from the convergence to a long-run limit that many models (and statistical methods!) depend on.

# Is path dependence real?

- What if the most popular "idols" and "talents" had to start over? Would they again rise to the top or was their success a path-dependent accident of preferential attachment? Such an experiment is impossible.
- The web makes related experiments possible. Salganikik, Dodds, and Watts constructed a web site that offered specific music downloads by unfamiliar artists, of varying quality. Visitors were shown the list of tunes and their download count.
- They duplicated the site 9 times, with independent download counts. (One site had no counts.)
- The result was wide variation in the distributions, although the best songs always ended up in the top half, and the worst in the bottom half.
- The version without counts displayed much lower variation in download counts across the songs.

# The long tail

- In looking at power laws so far, we have emphasized the upper tail of the distribution of sales (links, followers): the blockbusters, the outliers with huge numbers, the influencers.
  - This reflects economies of scale and specialization. Produce one of these, and you're rich.
- But it's not necessarily true that they have a dominant fraction of all sales, especially if there's a preference for variety (or wide dispersion of tastes). An alternative approach is like a supermarket: handle everything. Then the question is what sales volume you can achieve on less popular items.

# Supermarkets *vs.* Internet services

- Note that for a supermarket there's a real problem of shelf life. You have to sell meat and fish and fresh-baked goods on the same day, for example. But not so for books and DVDs!
- There's also a question of storage space for books and DVDs. But this doesn't apply to Internet distribution of e-books and streaming videos! (Or, at least, not in the same way.)

# It's power laws all the way down

- To examine the *long tail* of unpopular items, change viewpoint from the fraction of items selling *exactly* $k$ units to the number selling *at least* $k$ units.
- Changing *fraction* to *number* is just rescales by the number of items. It changes the labels on the vertical axis, but not the curve in the graph.
- The change from *exactly* to *at least* is a real change, although just of "viewpoint" on the underlying data. This distribution is also a power law: $F(k) = \int_k^\infty Ax^{-\alpha}dx = -\frac{A}{1-\alpha}k^{1-\alpha}$ !
- Next, how many units of the $j$-th most popular item are sold? But this is exactly the inverse of the $F$ defined above: if $j = F(k)$ is the number of books that have sold $k$ units or more, then the $j$-th most popular book is exactly the one that sold $k$ units. Draw the curve, then flip the axes.
- The right end of the "flipped" curve is the "long tail": niche items selling few units. A power law!

# Mathematics and meaning

- We've seen three different power laws while examining the "long tail":
  - the original power law distribution of item popularity derived from preferential attachment
  - the power law distribution of "at least this popular" items
  - the inverse power law of percentile rank (the popularity of the $j$-th item).
- They're quantitatively quite different—it is a big mistake to use one of them to answer a question appropriate for the others. *E.g.*, don't try to use the *item popularity* distribution to answer questions about *sales volume of the items* in the tail.
- Each is useful in its own way. For example the *inverse power law* can be used to answer "how much revenue comes from the items with sales rank less than 1000?" (by integrating).
- Mathematical transformations allow us to ask various questions from the same simple data.

# Beyond the math

- While we won't discuss flamewars and social media bullying in detail, clearly these phenomena gain much of their power from preferential attachment. Disinformation is related to information cascades, people believing things because other people believe them.
- Of course there are important psychological aspects, such as *motivated reasoning* (picking facts and relationships that support what you want to believe, and ignoring facts and relationships that contradict your prejudices). But informational effects amplify problems.
- We can also point to potentially opposing aspects of search engines and recommender engines.

# "The Algorithms"

- Search engines and social media *curate* content: they choose content for you, because all available content is an overwhelming (almost unimaginable) quantity. The selection rules are often called "the Algorithm".
- On the one hand, search engines tend to rank candidate links by popularity, not by value of content. This can lead to information cascades.
  - Social media "algorithms" have been shown to encourage "information bubbles" which reinforce prejudice and political division, and to contribute to radicalization.
- To exploit their stocks of "long tail," unpopular items little-known to consumers, vendors like Amazon and Netflix offer *recommender* engines which help consumers to navigate the plethora of little-known items, and enjoy products they would otherwise never hear of.
  - These engines may encourage, rather than undermine, diversity.