Social and Information Networks

CSCC46H, Fall 2022

Lecture 6

Prof. Ashton Anderson
ashton@cs.toronto.edu

Logistics

Blog posts K-R due Friday

Today

Power laws
Inequality
Unpredictability

How is popularity distributed?

A deeper look at one of our central questions: how connected are people? How many people do people tend to know?

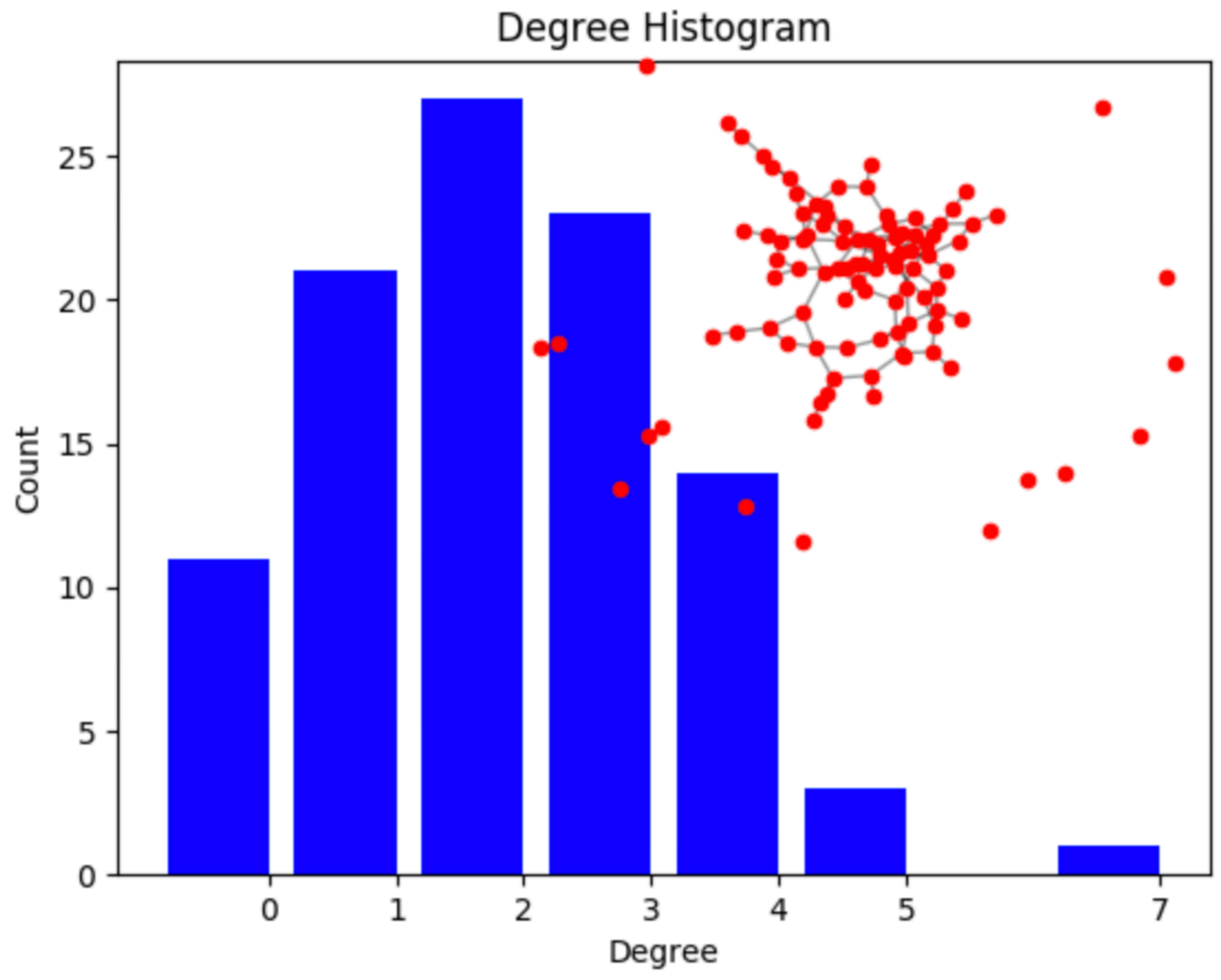
Most know some, and some know a ton

How is popularity *distributed* in the population?

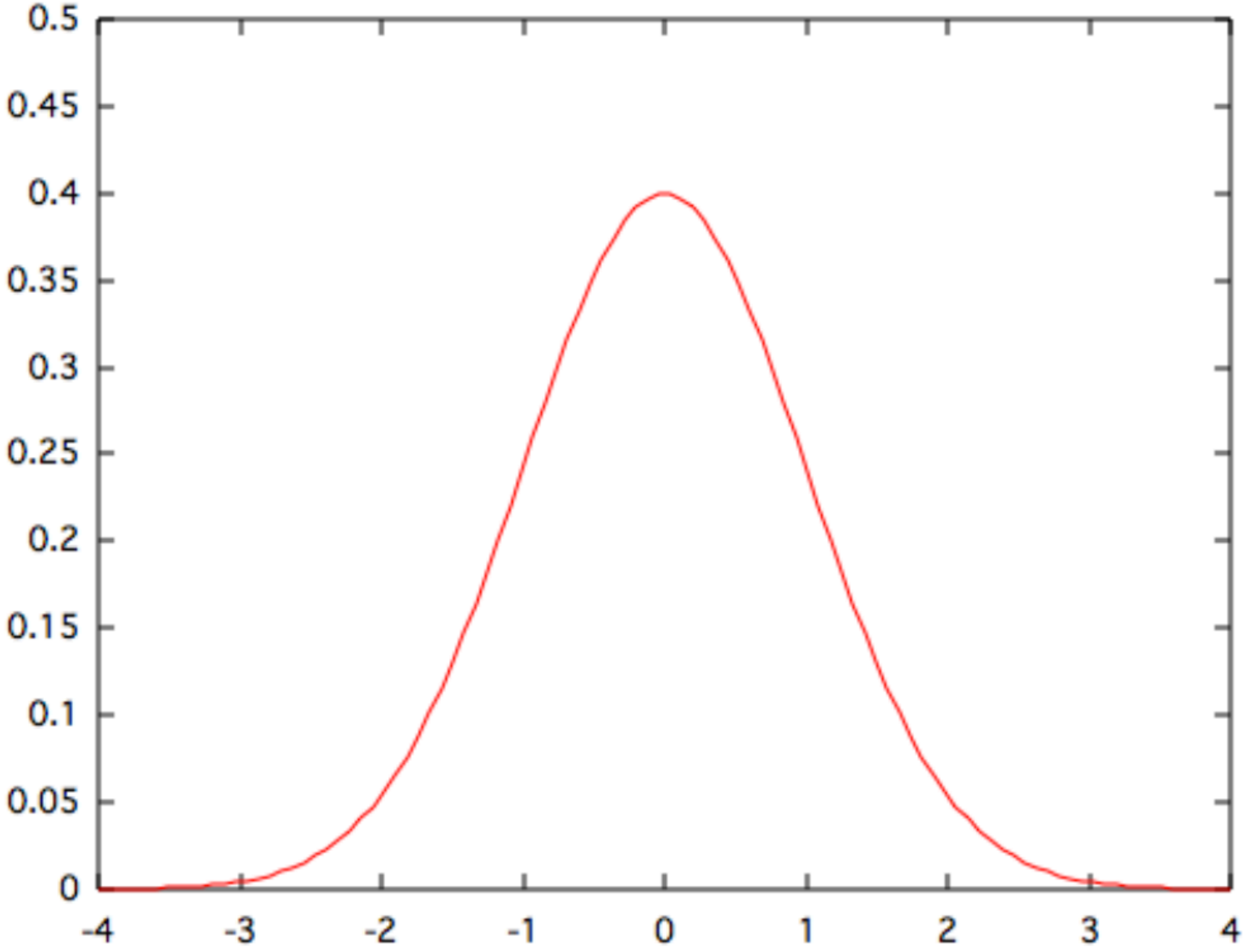
Recall: Degree Distributions

Every node has some number of neighbours, which is their degree

The degree distribution is just the histogram of degrees in the network

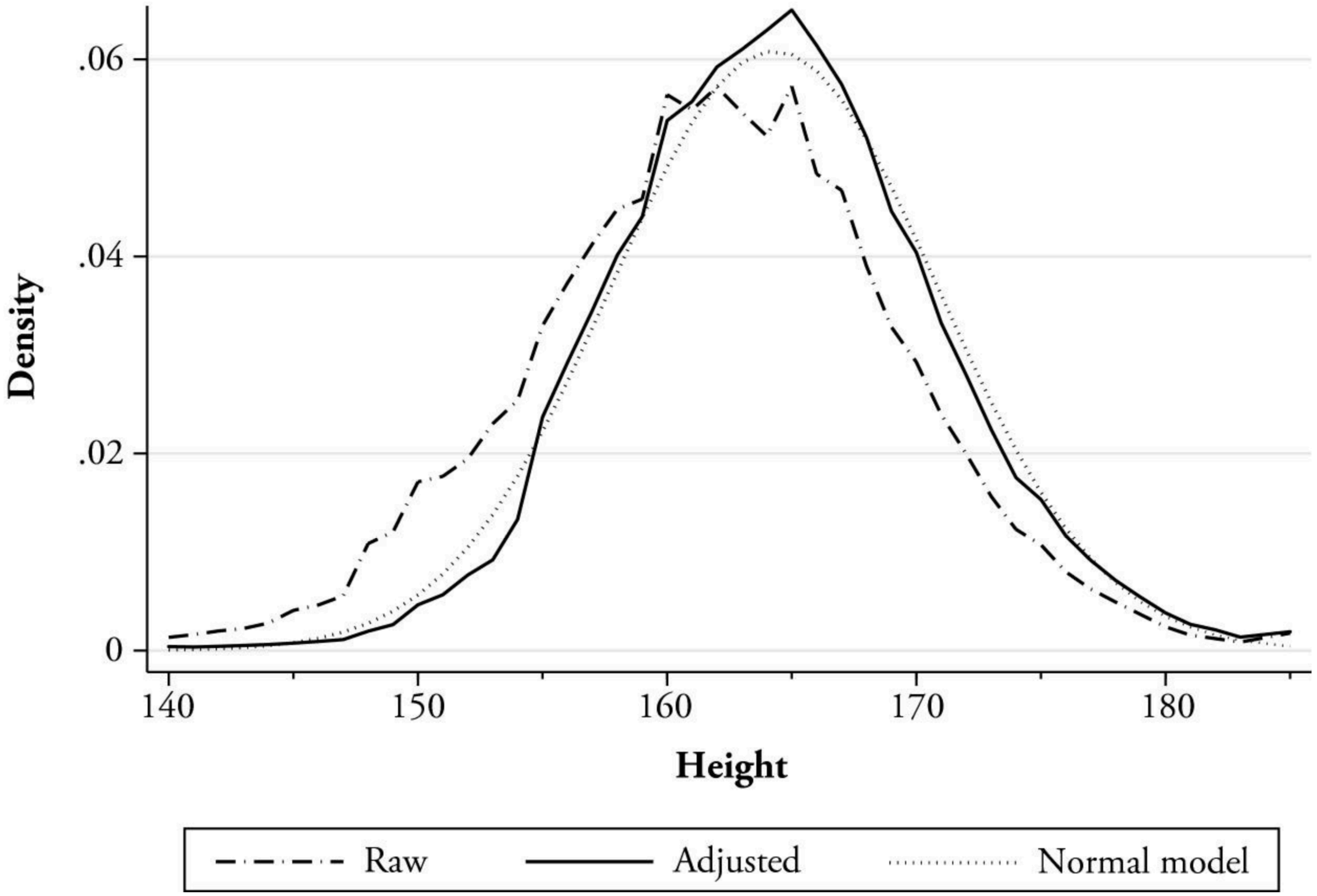


A guess



The normal/Gaussian distribution
Most values are clustered around a typical value

A guess



From "Height and the Normal Distribution: Evidence from Italian Military Data"

Heights of males in the Italian army
Most values are clustered around a typical value

MSN: Degree Distribution

Count, P(k)*n

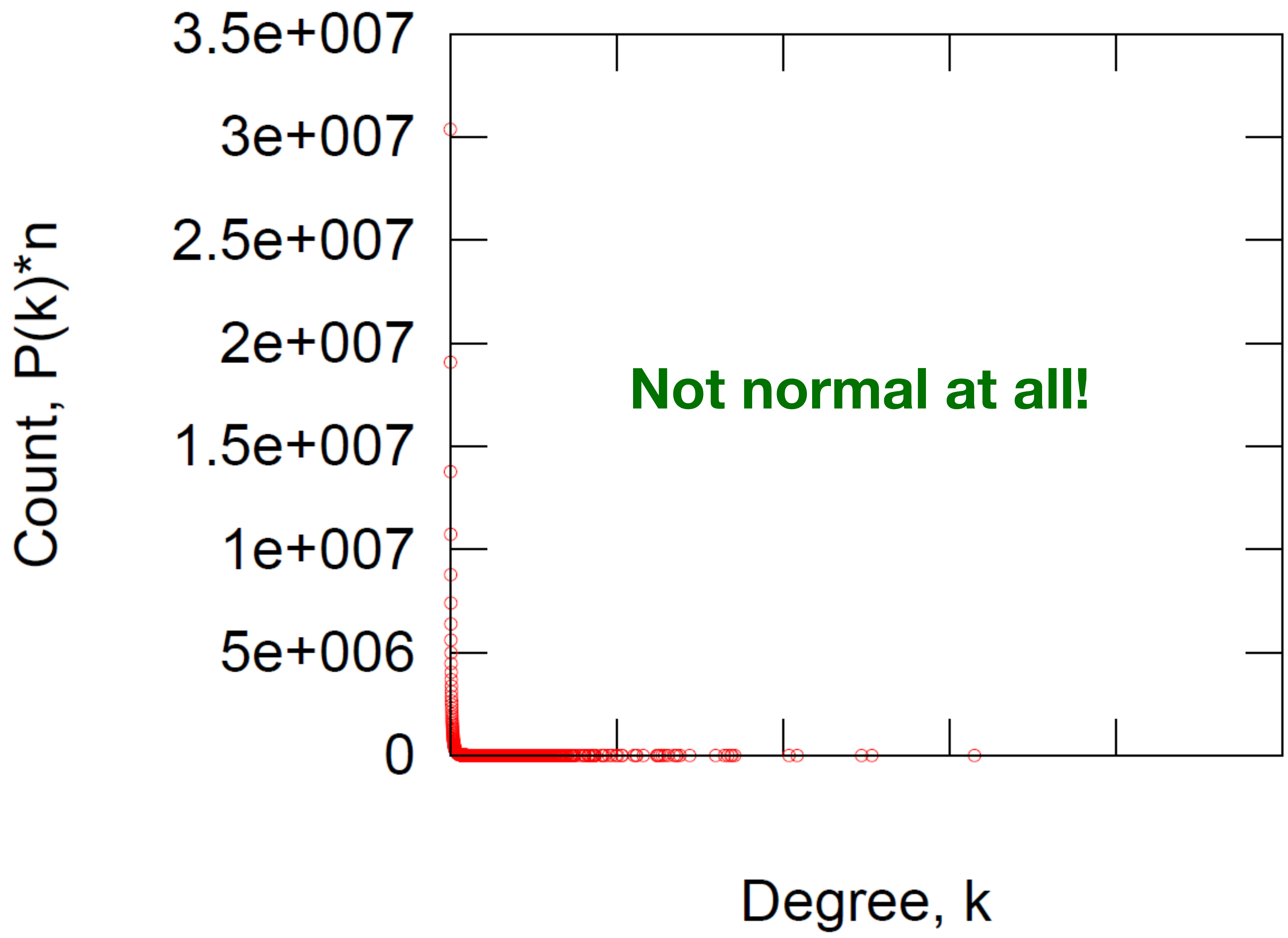
?

Degree, k

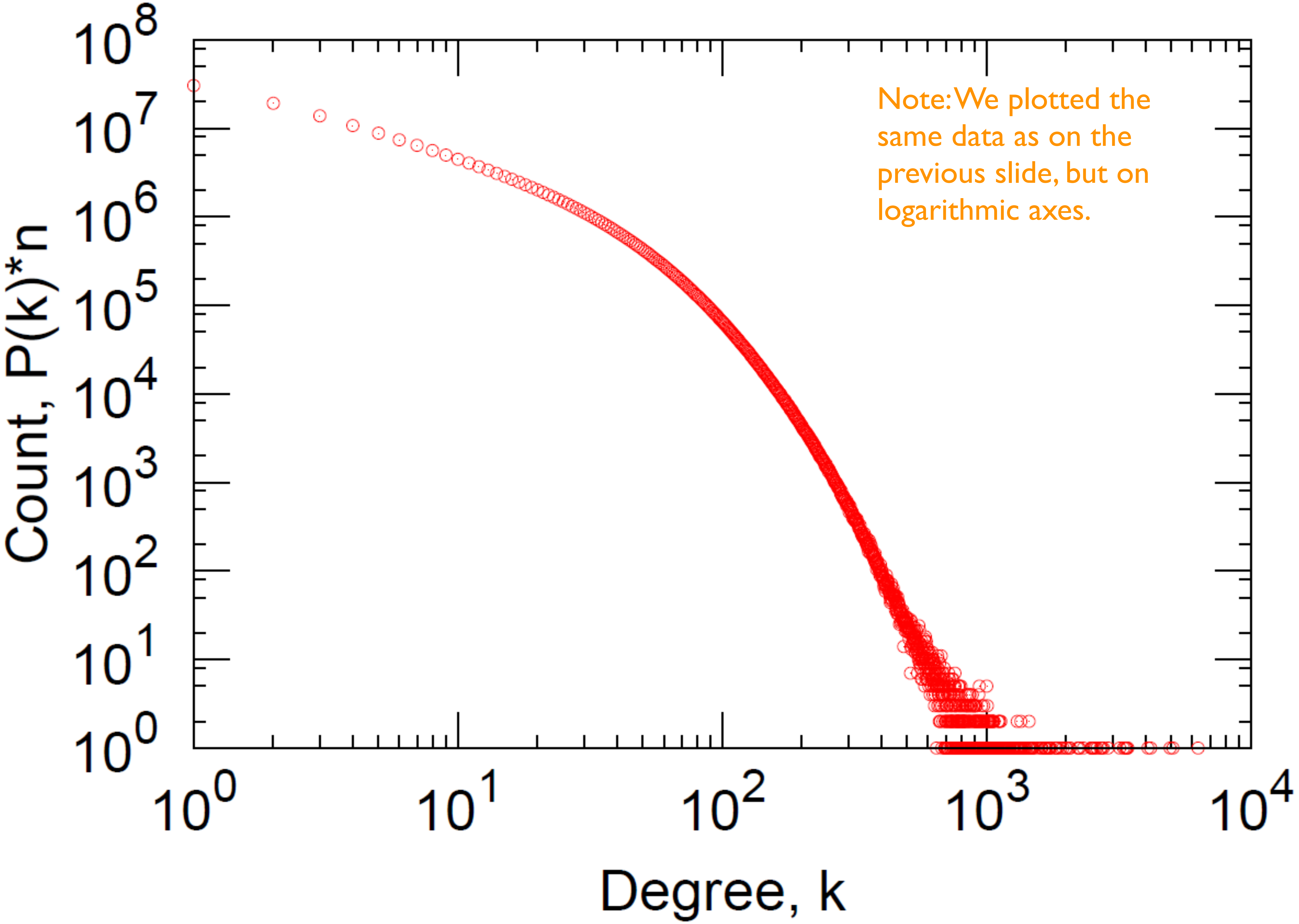
Plot: fraction of nodes with degree k:

$$p(k) = \frac{|\{u|d_u = k\}|}{N}$$

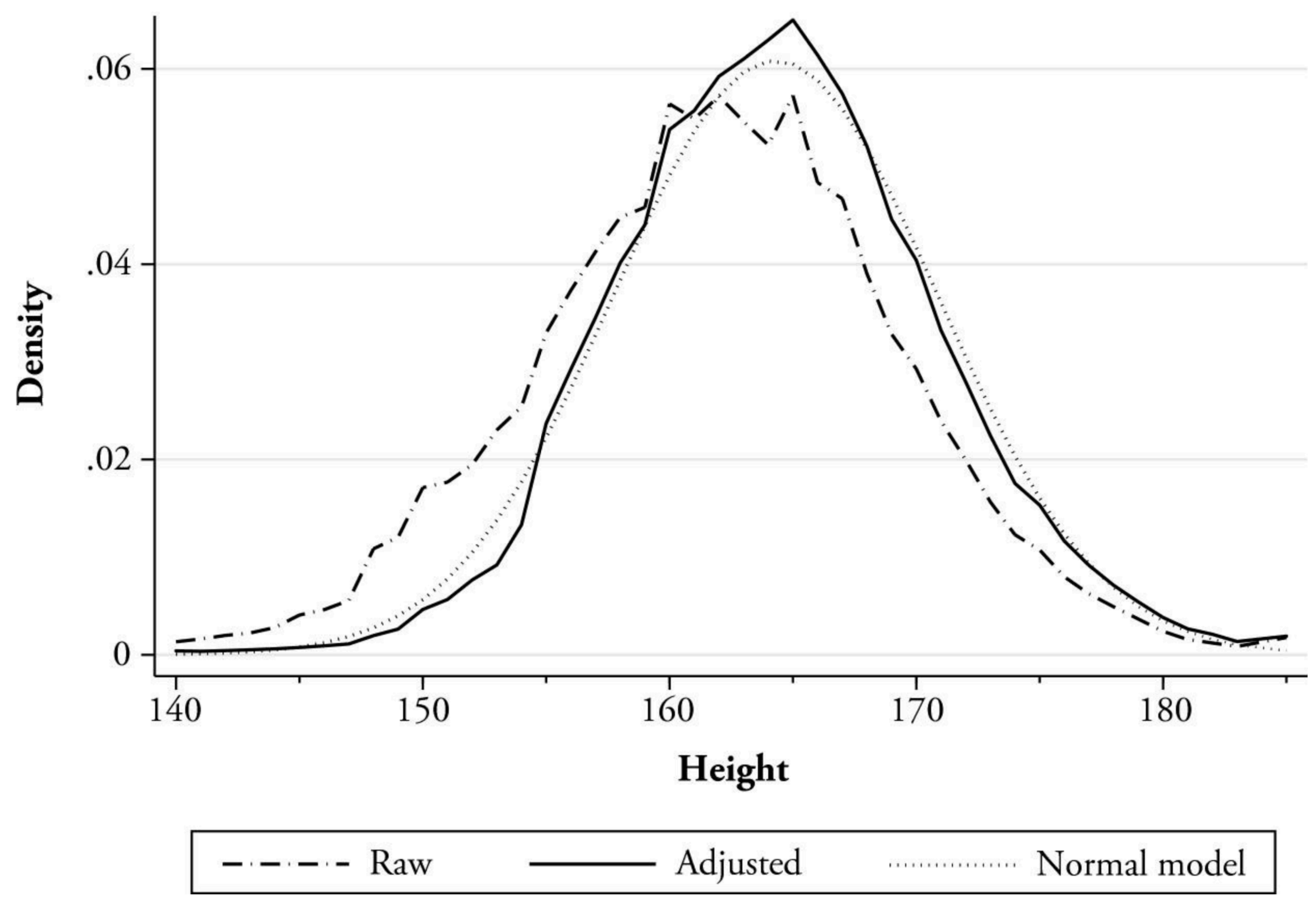
MSN: Degree Distribution

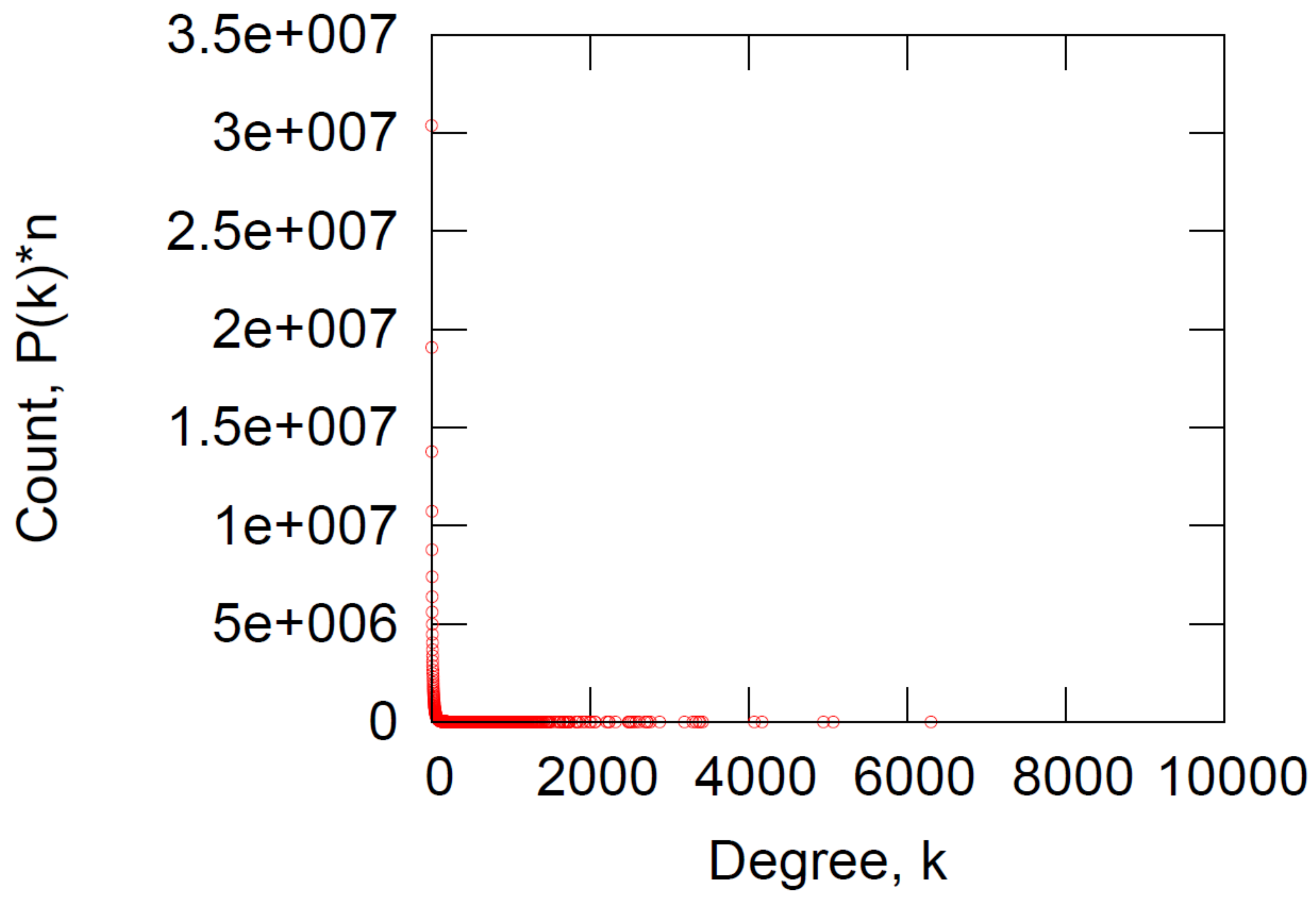


MSN: Log-Log Degree Distribution



Degree distributions in networks





Degree distributions are **heavy-tailed**

Gaussians, which have exponentially decreasing tails, have almost no mass far from their mean

The same is **not true** of heavy-tailed distributions

The Power Law Distribution

The main heavy-tailed distribution we will consider is the **power law**:

$$p(x) \propto x^{-\alpha}$$

For example, Newton's law of universal gravitation follows an "inverse-square law", e.g. a power law:

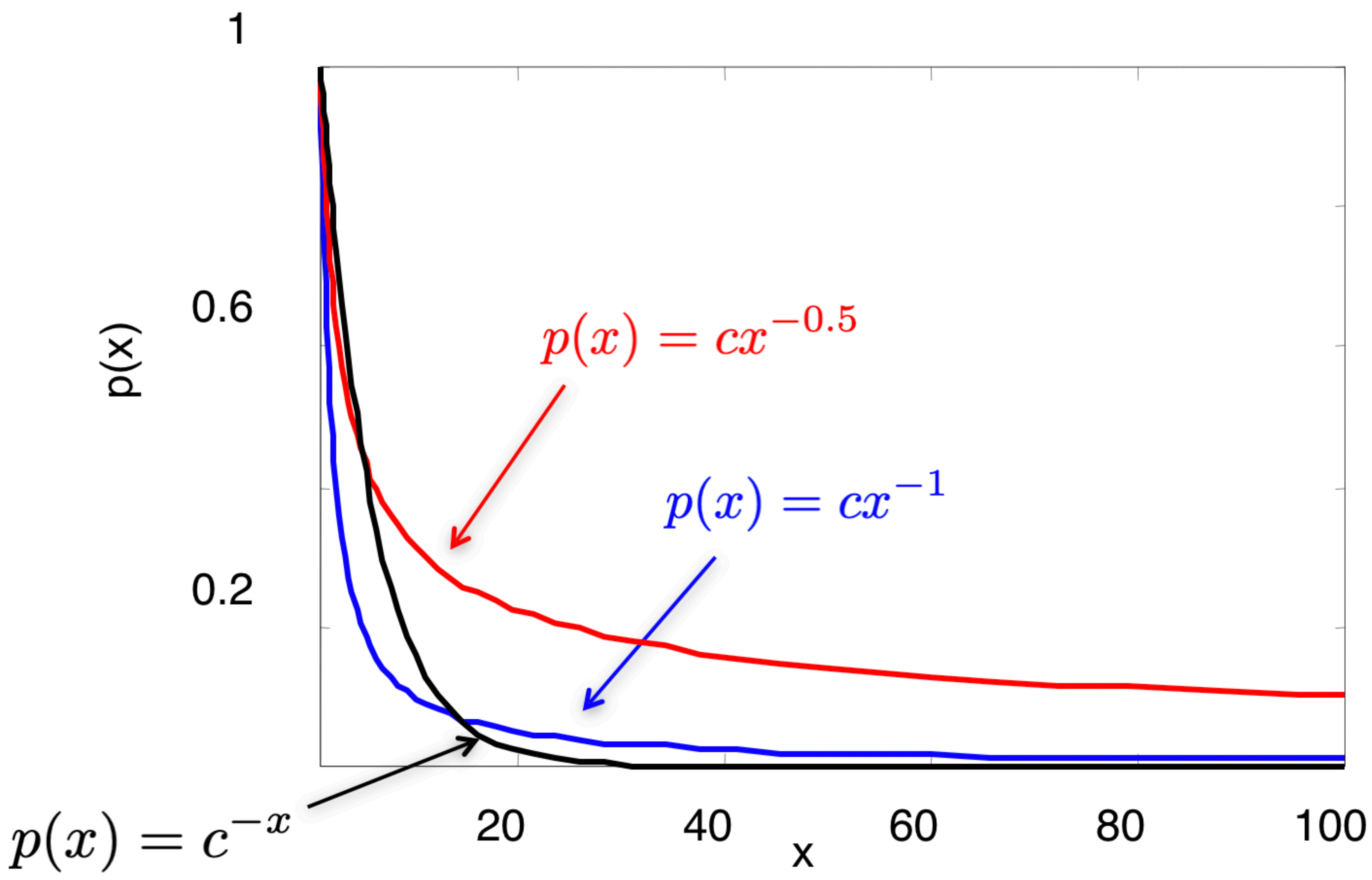
$$F(r) = G\frac{m_1 m_2}{r^2}$$

Where the distance r is the quantity that is changing

To make it an actual distribution, include a normalizing constant c

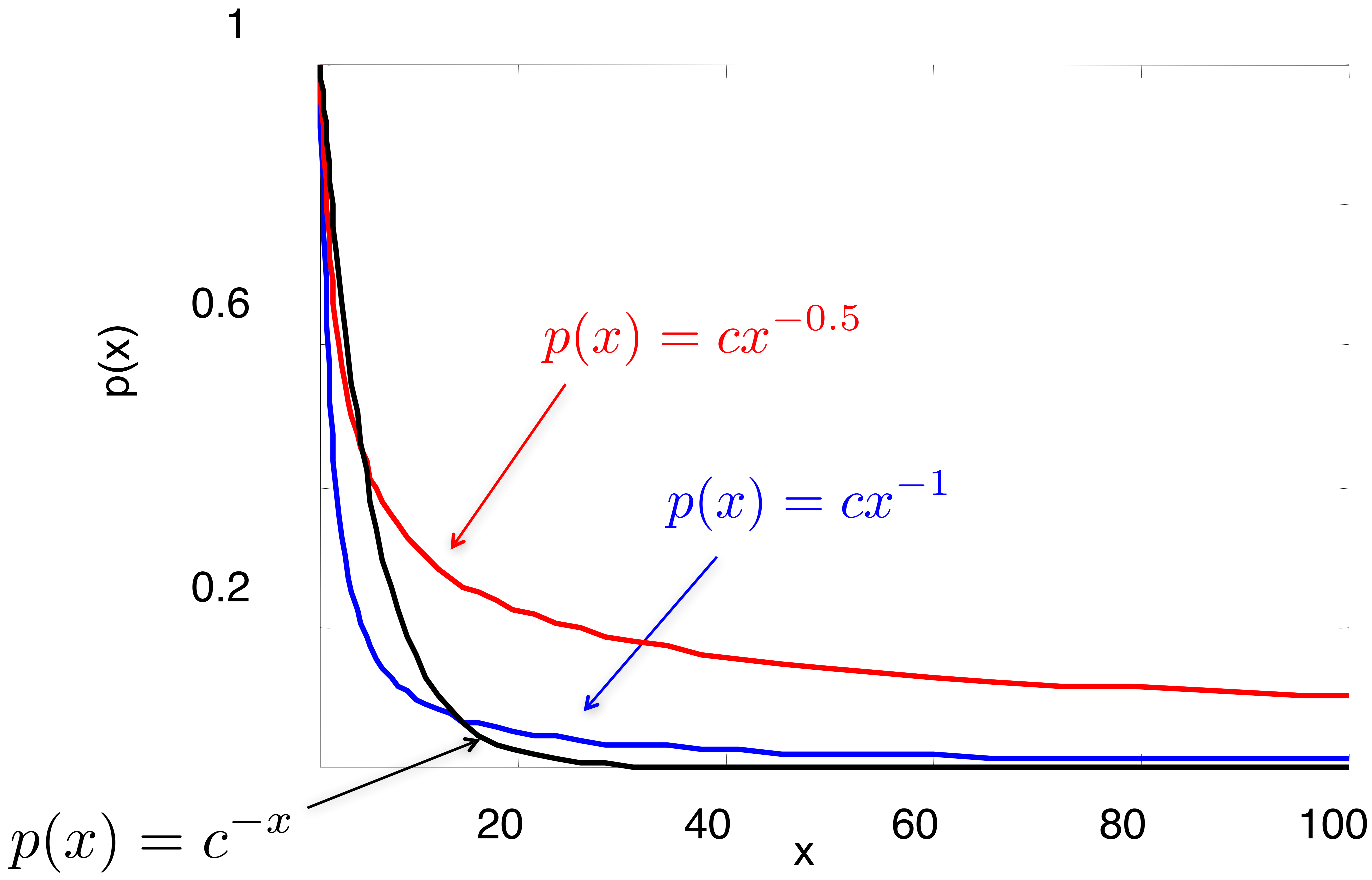
$$p(x) = cx^{-\alpha}$$

Exponential vs. Power-Law



Above a certain x value, the power law is **always** higher than the exponential

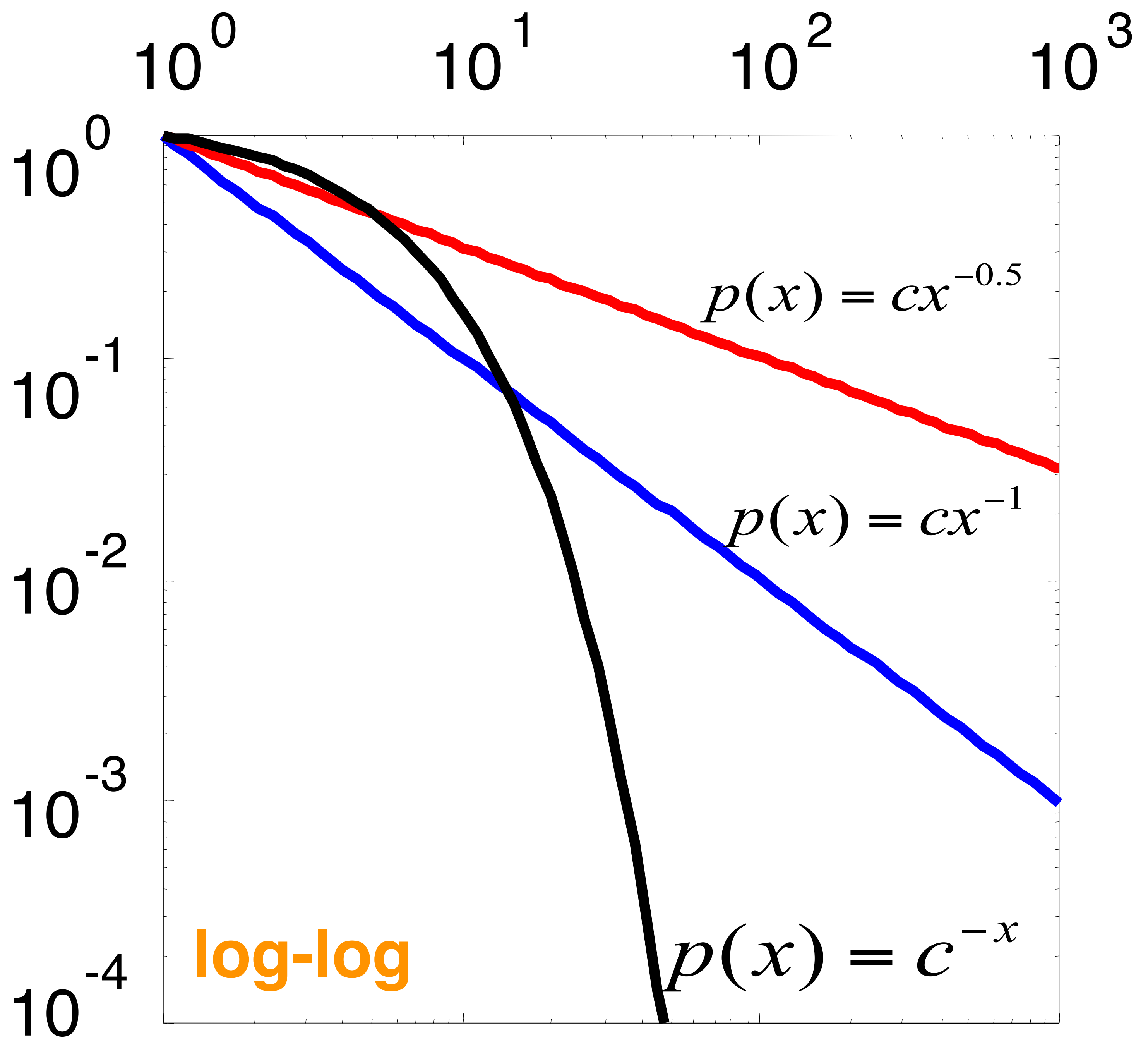
Exponential vs. Power-Law



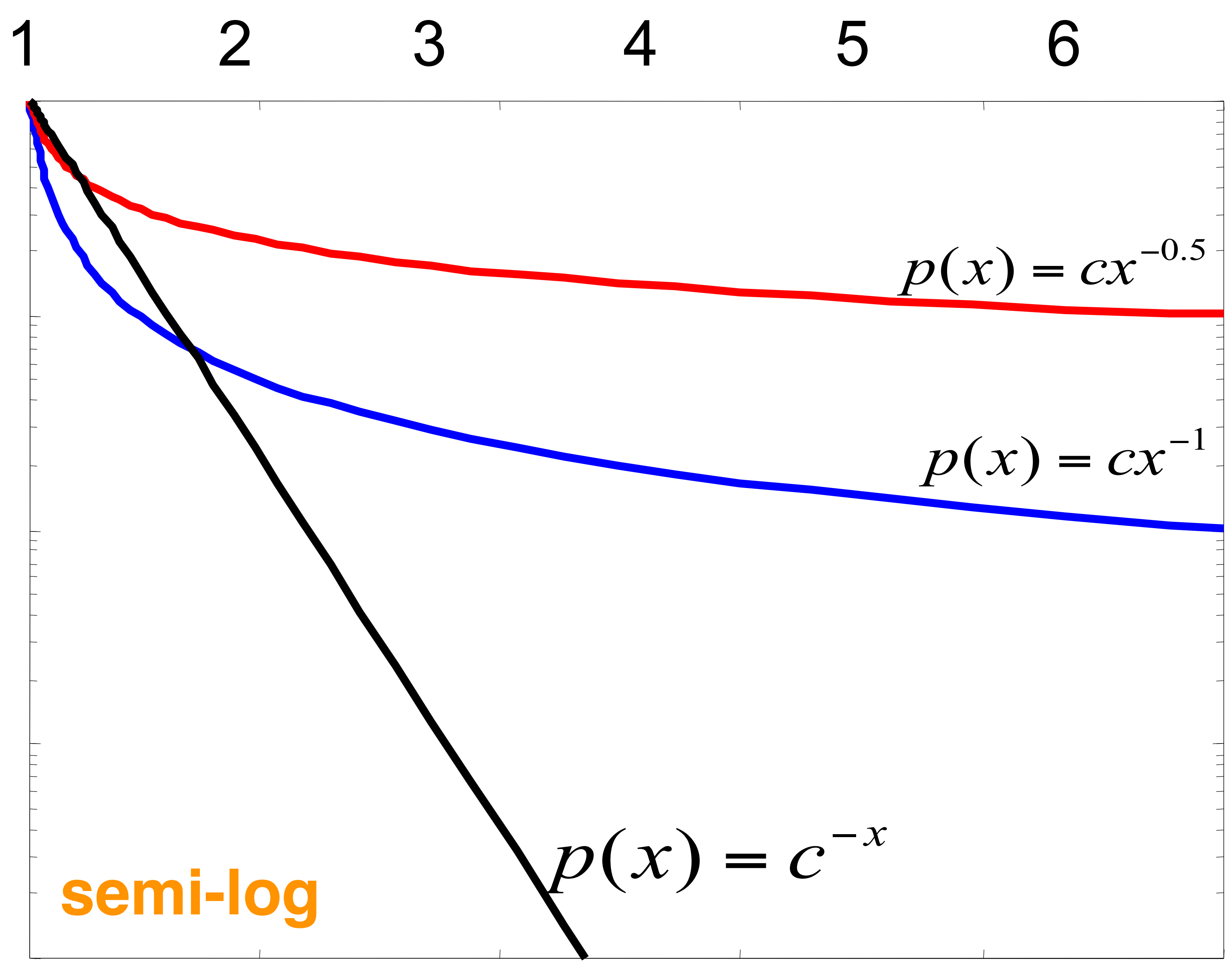
Think: 2^{-1000} is unimaginably tiny, but $1/1000^2$ is only one in a million ($\sim 10^{-302}$ vs. 10^{-6})

Exponential vs. Power-Law

Power-law vs. Exponential on log-log and semi-log (log-lin) scales



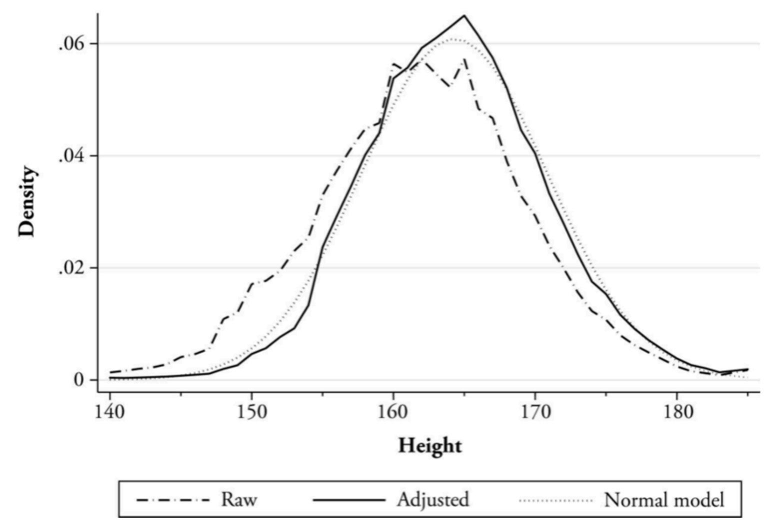
x … logarithmic axis
y … logarithmic axis



x … linear
y … logarithmic

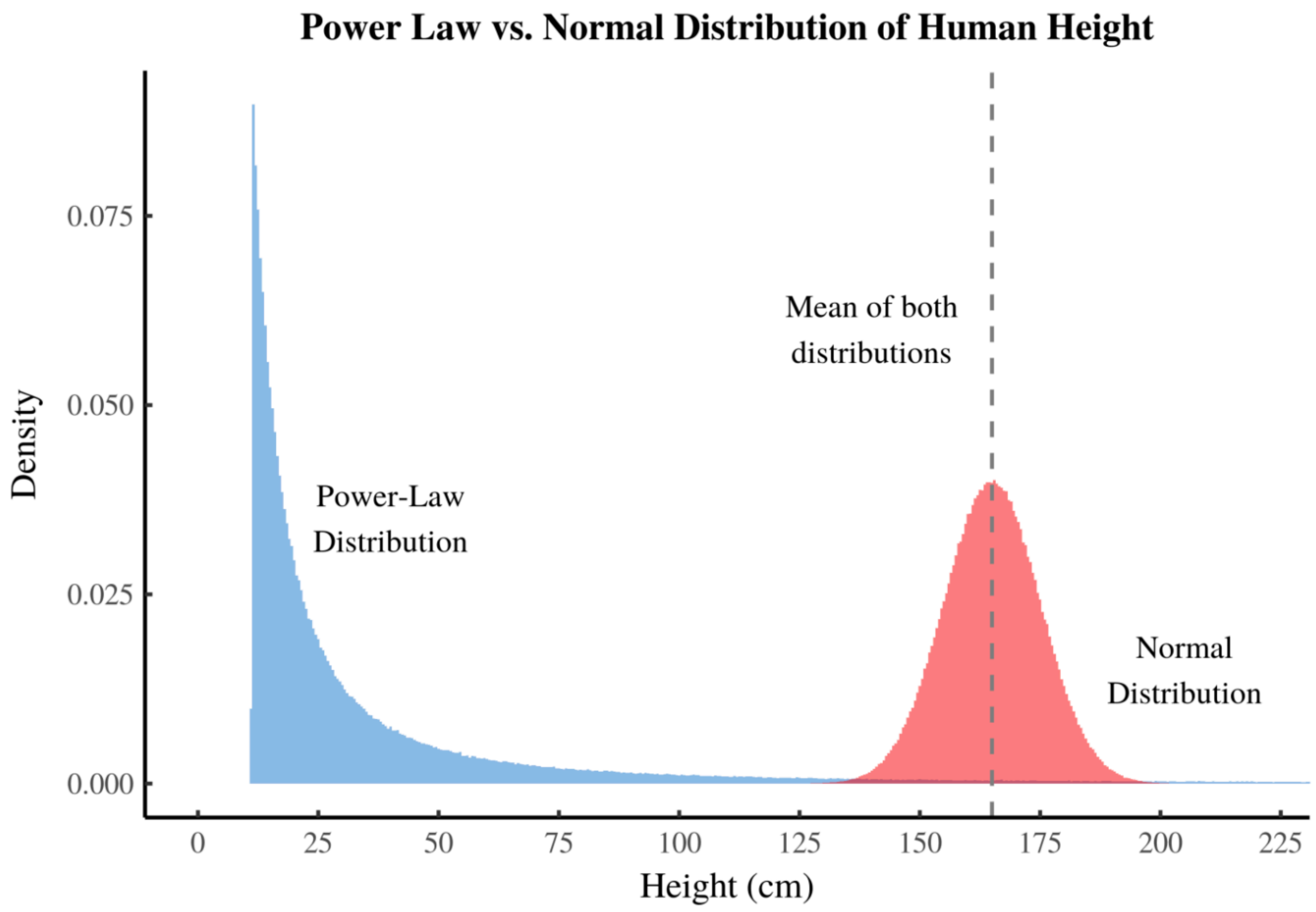
Height as a Power Law

We know that height is distributed *normally* (Gaussian)



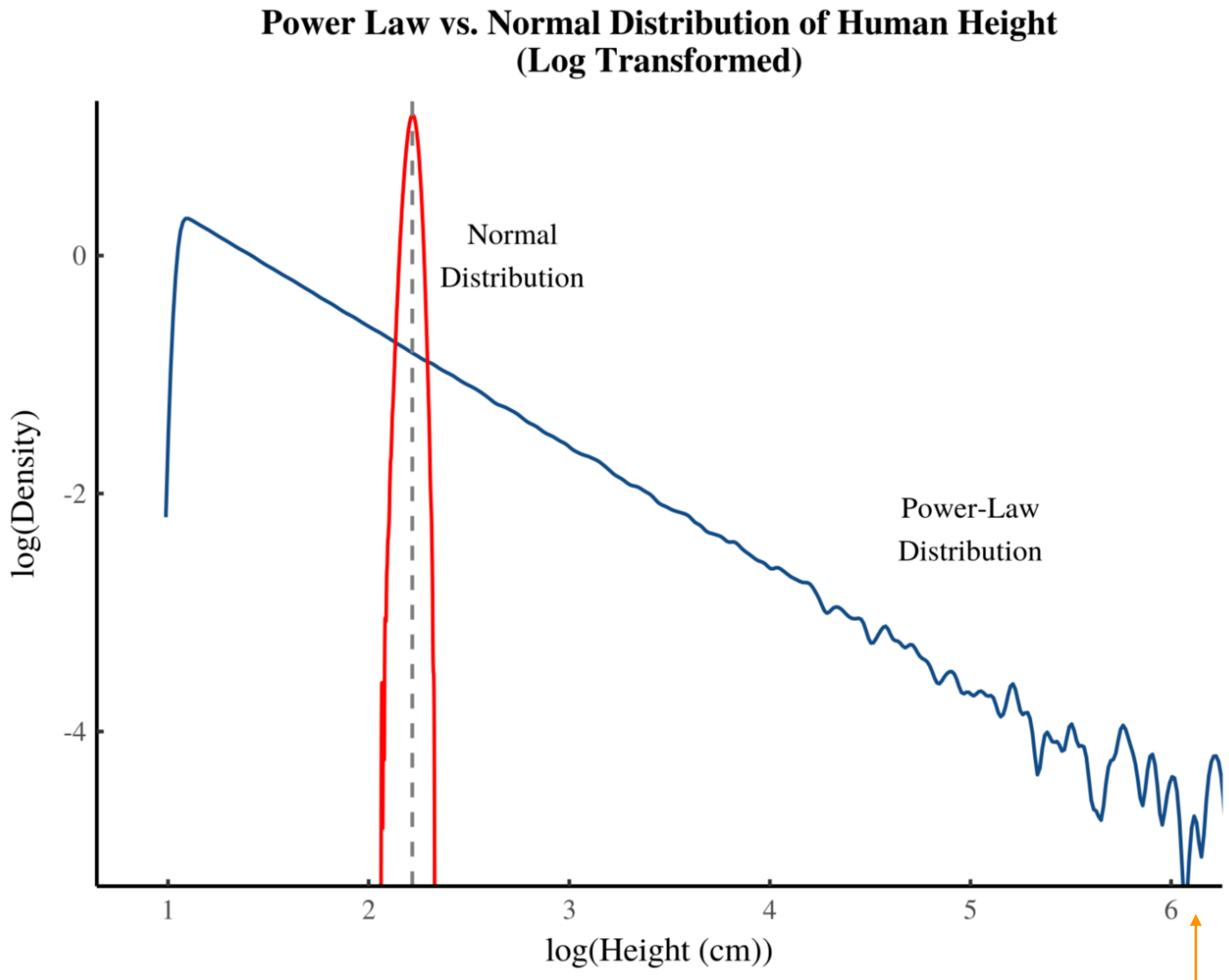
But what if it were a power law?

Height as a Power Law



Why is the mean of the power law so far out?

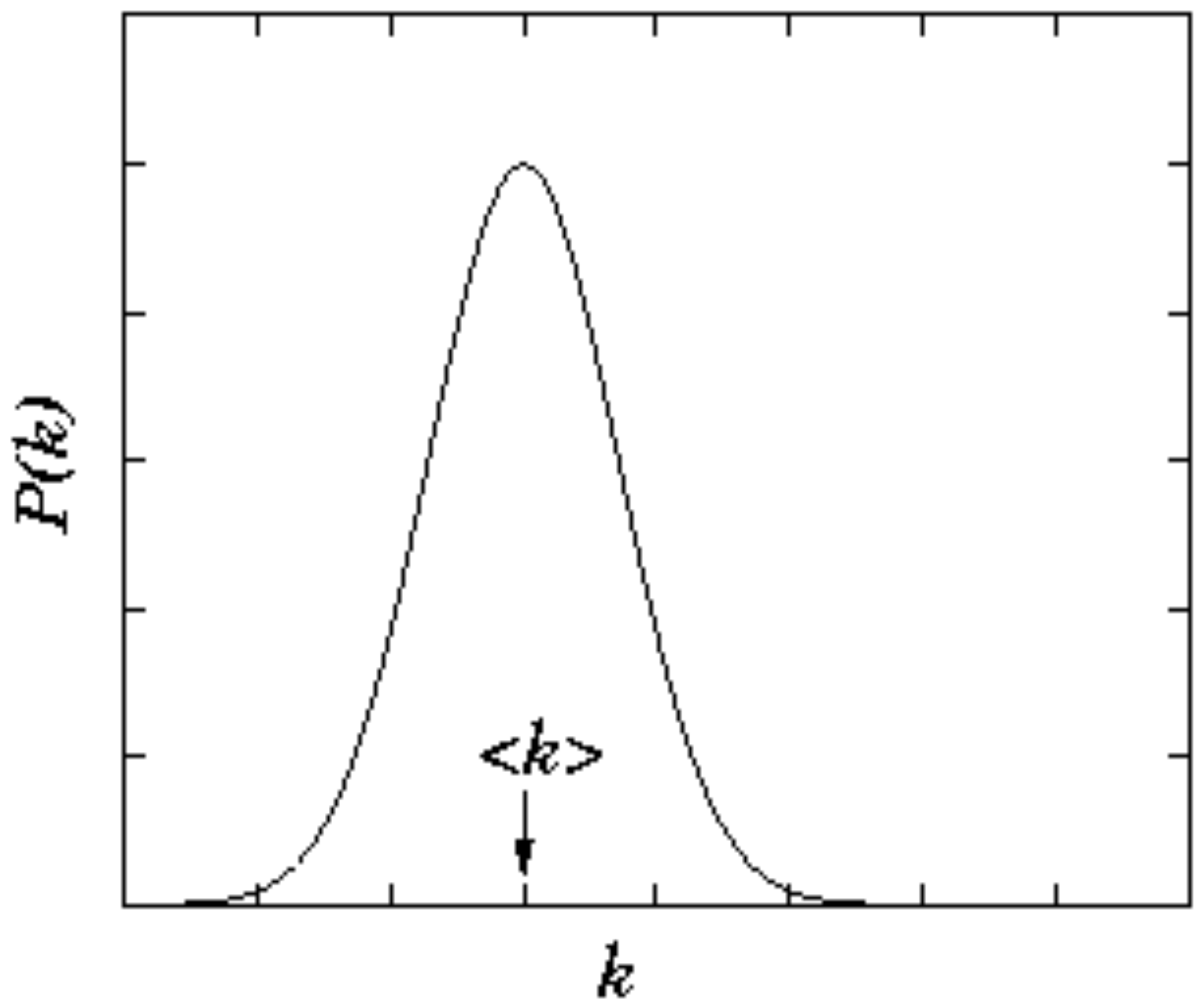
Height as a Power Law



10km tall people

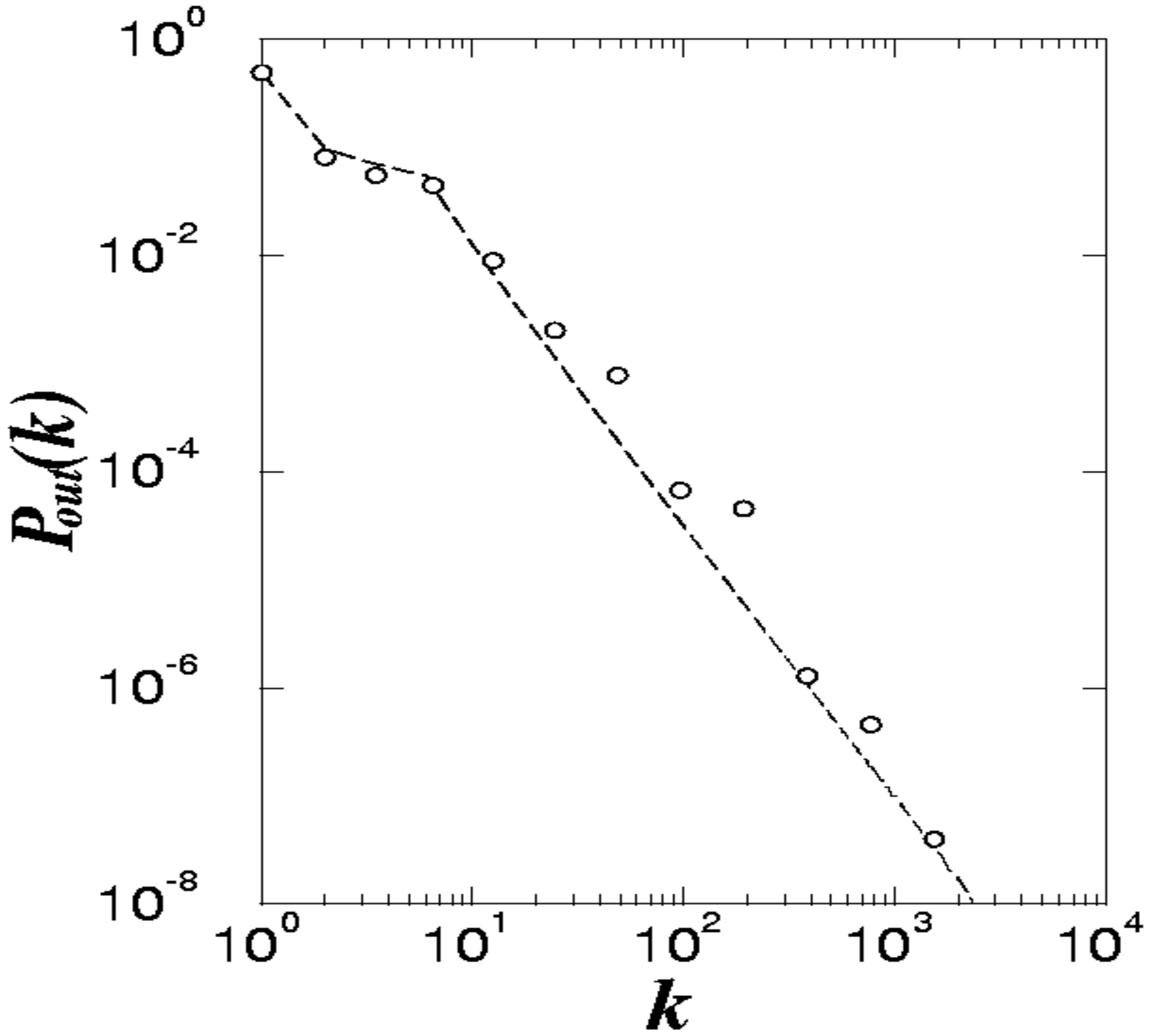
Power Laws in Networks

Expected based on G_{np}



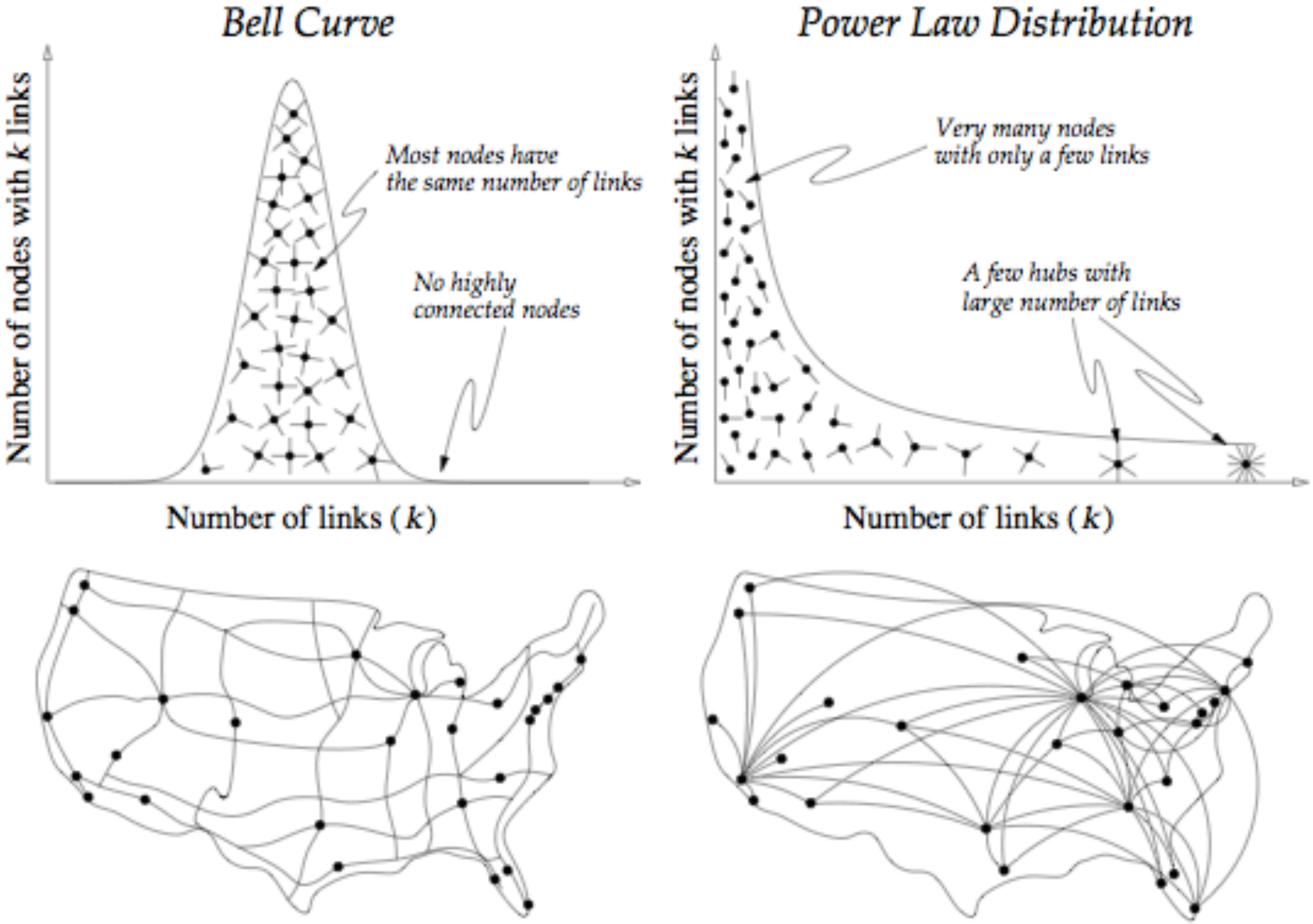
$$P(E) = \binom{E_{\max}}{E} p^{E} (1-p)^{E_{\max}-E}$$

Found in data



$$P(k) \propto k^{-\alpha}$$

Exponential vs. Power-Law



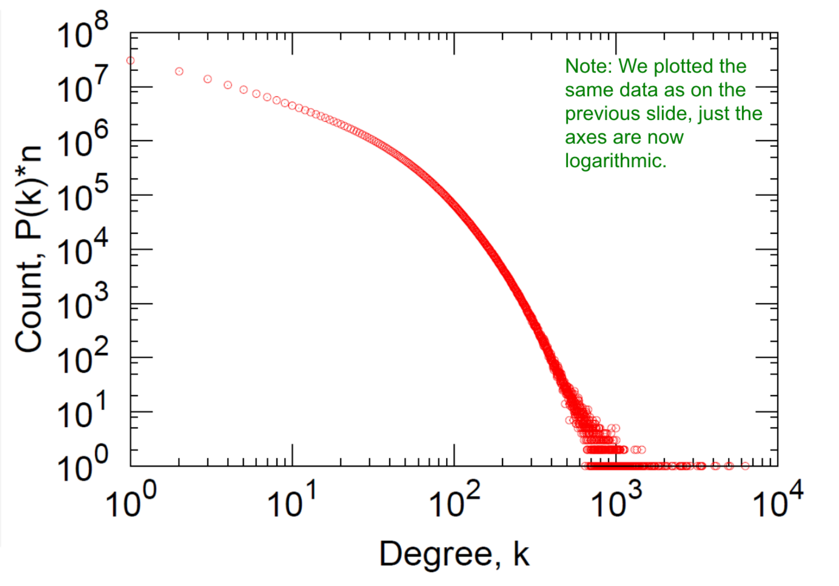
Test for a power law

How can you tell if empirical data follows a power law?

Let f(x) be the fraction of items that have value x

Question: does f(x) = c/x^{α} approximately hold? [for some exponent α and constant c]

$$f(x) = cx^{-\alpha}$$
$$\log f(x) = \log cx^{-\alpha}$$
$$\log f(x) = \log c - \alpha \log x$$

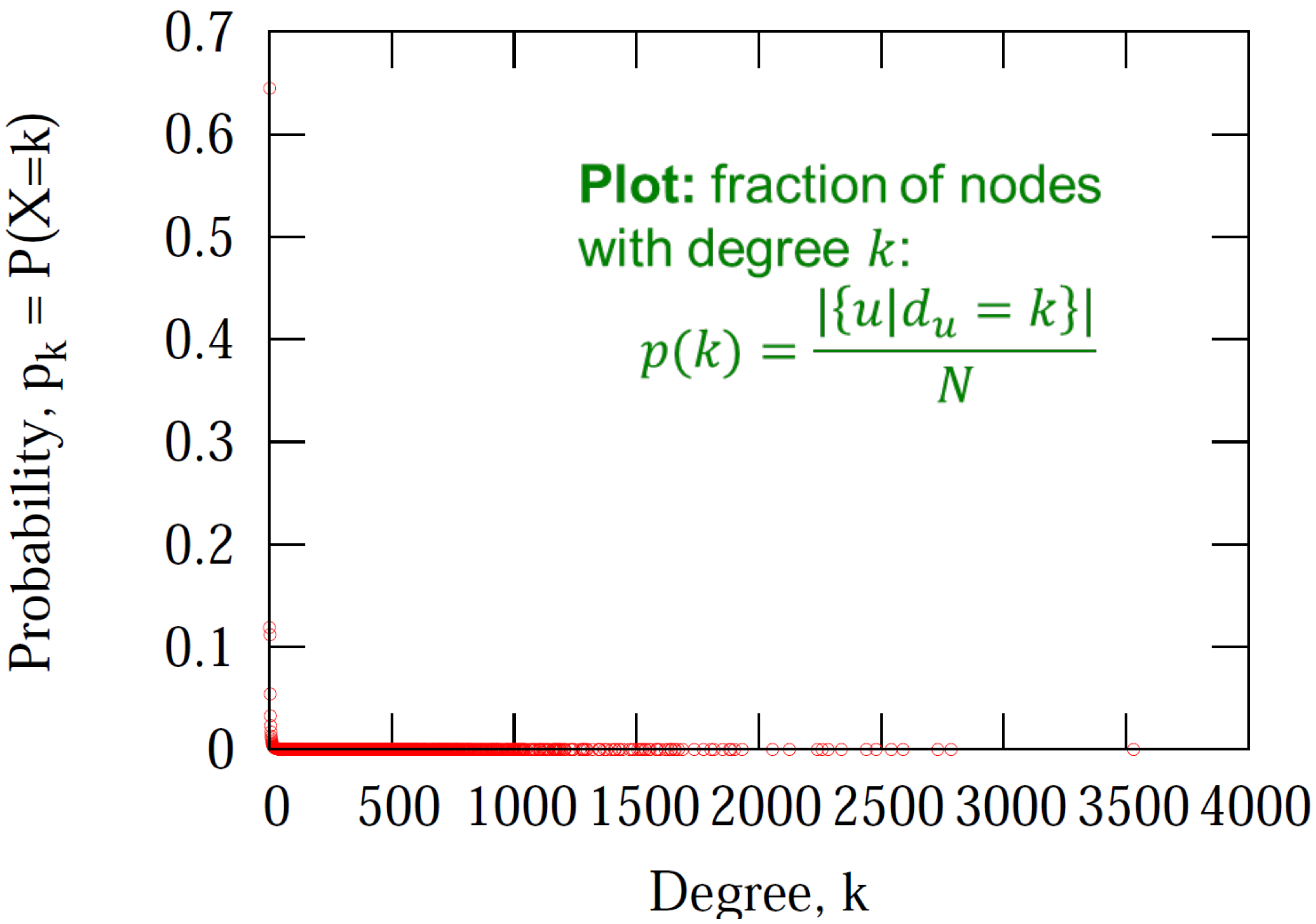


Plot log f(x) as a function of log x

Straight line with slope $-\alpha$!

Node Degrees in Networks

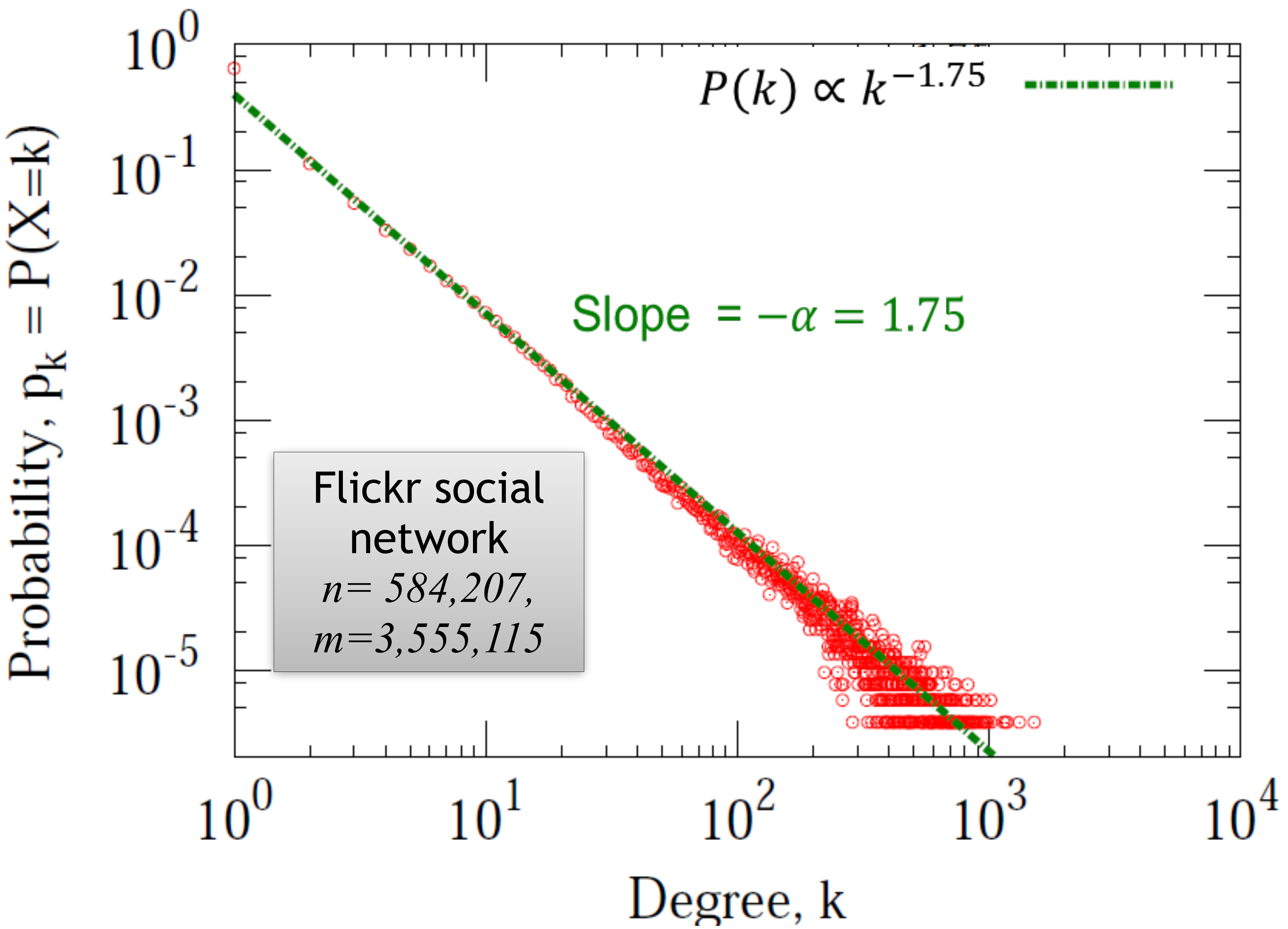
Take a network, plot a histogram of P(k) vs. k



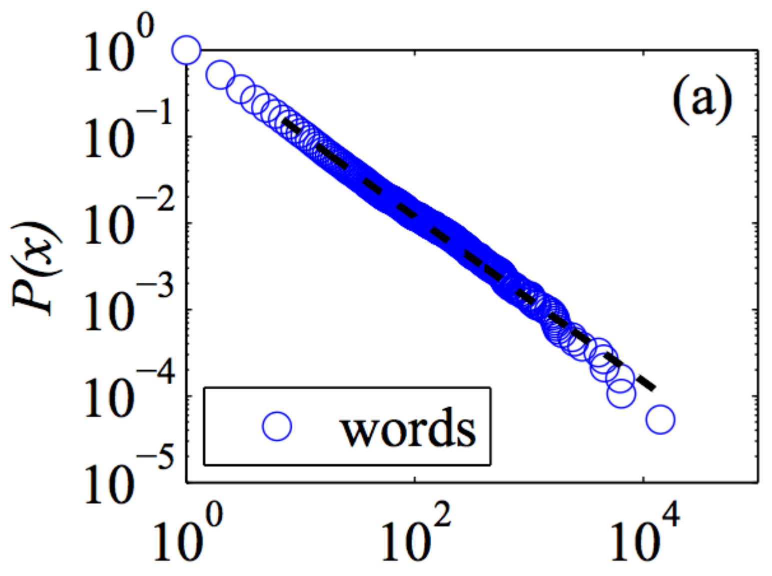
Flickr social network
n= 584,207, m=3,555,115

Node Degrees in Networks

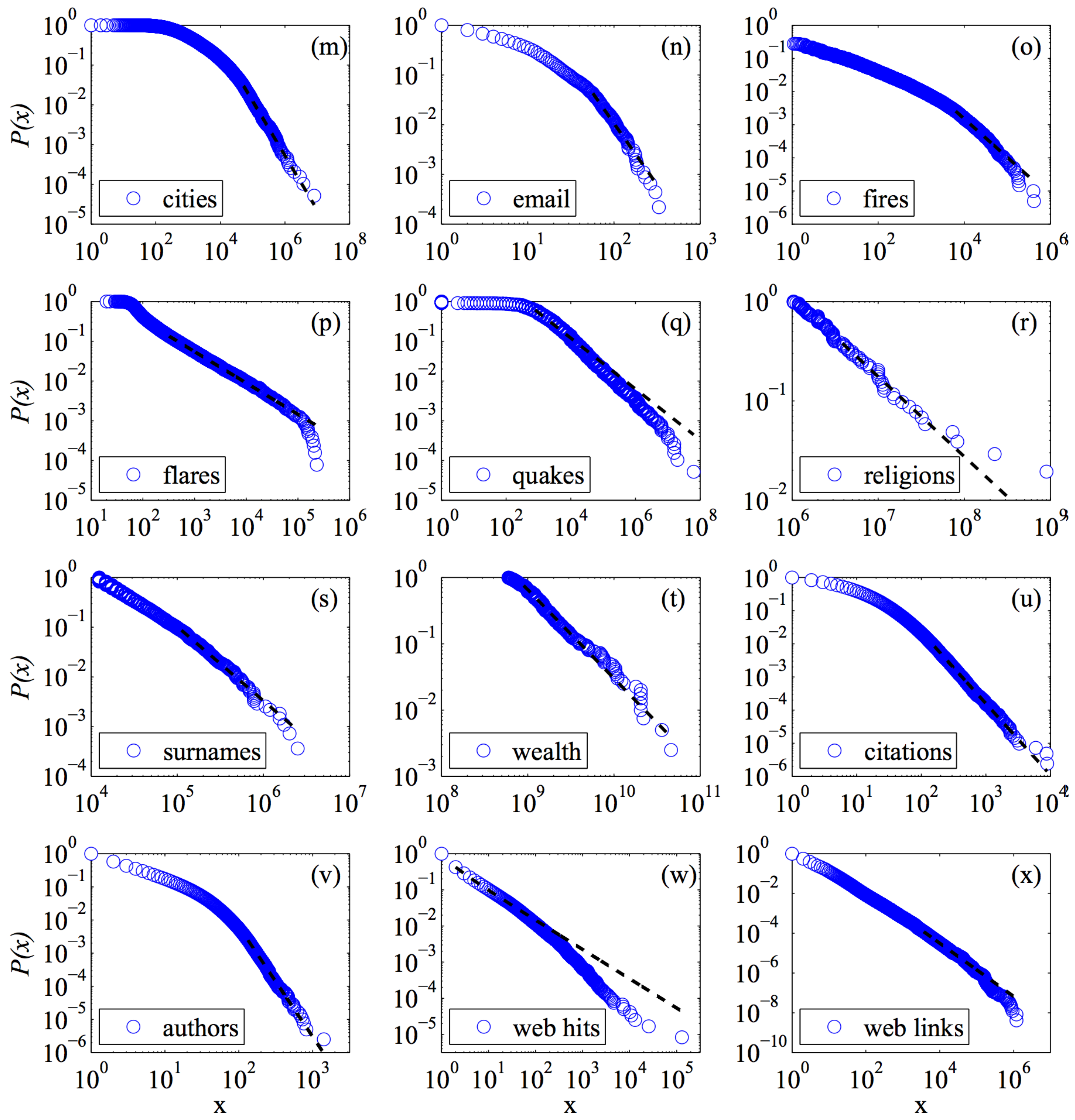
Plot the same data on *log-log* scale:



Power laws are *everywhere*



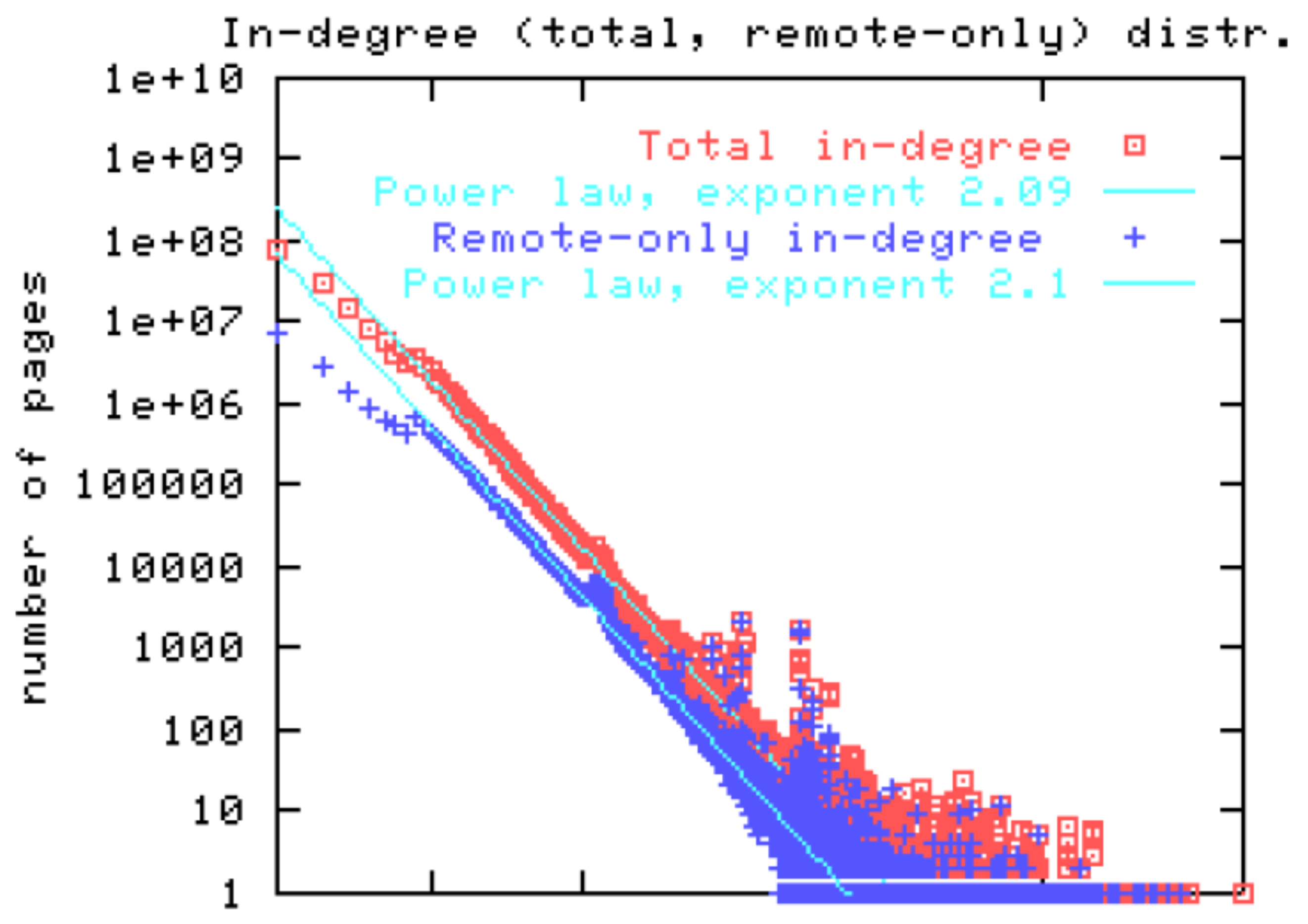
Power laws are *everywhere*

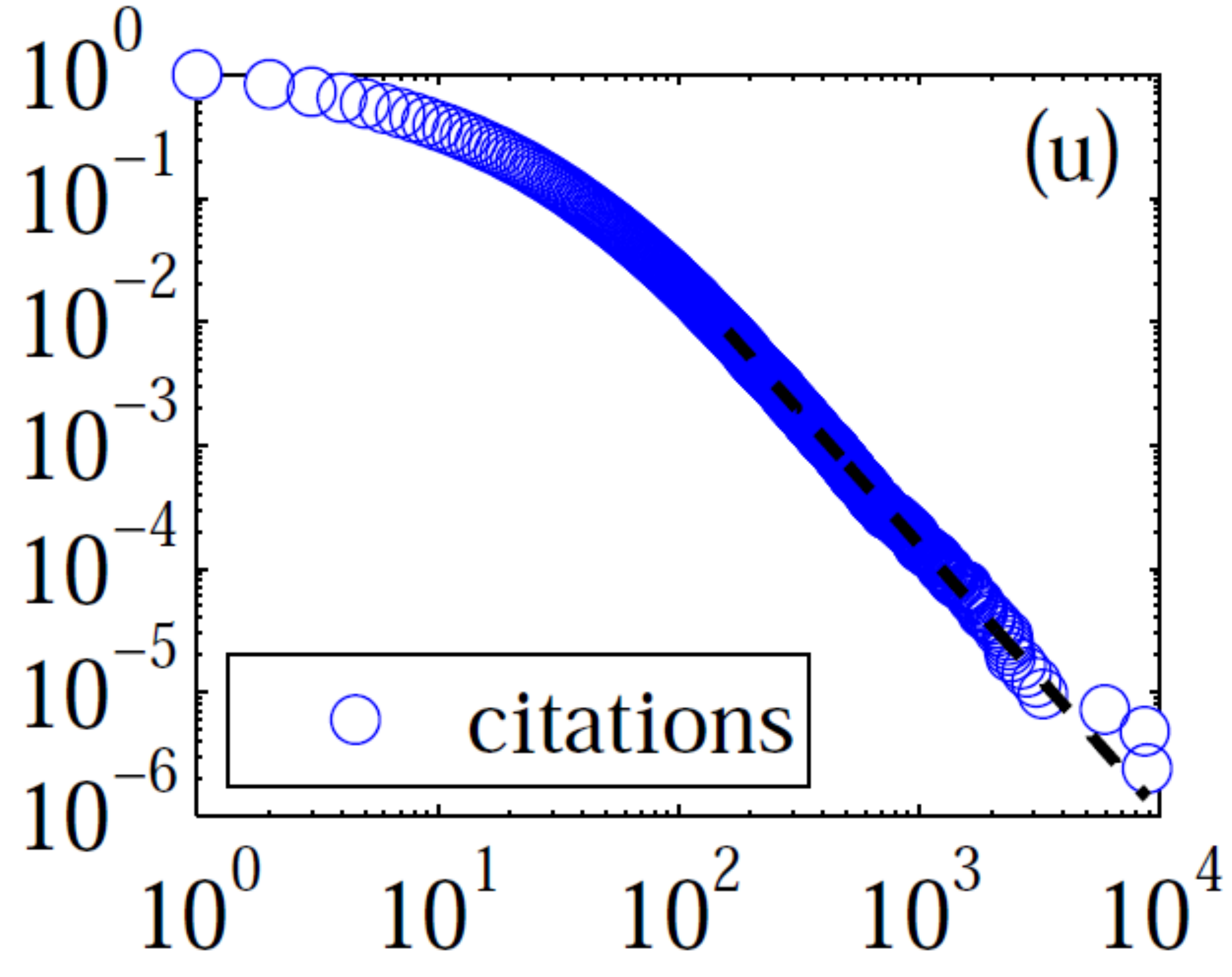


Power-Law Degree Exponents

Power-law degree exponent is typically $2 < \alpha < 3$

- Web graph:
  - $\alpha_{in} = 2.1$, $\alpha_{out} = 2.4$ [Broder et al. 00]
- Autonomous systems:
  - $\alpha = 2.4$ [Faloutsos^3, 99]
- Actor-collaborations:
  - $\alpha = 2.3$ [Barabasi-Albert 00]
- Citations to papers:
  - $\alpha \approx 3$ [Redner 98]
- Online social networks:
  - $\alpha \approx 2$ [Leskovec et al. 07]





Scale-Free Networks

- **Definition:**
  Networks with a power-law tail in their degree distribution are called "scale-free networks"
- Where does the name come from?
  - **Scale invariance:** There is no characteristic scale
  - **Scale-free function:** $\boldsymbol{f(ax) = a^{\lambda} f(x)}$
    - Power-law function: $\boldsymbol{f(ax) = a^{\lambda} x^{\lambda} = a^{\lambda} f(x)}$

The power law is the unique function with this property!

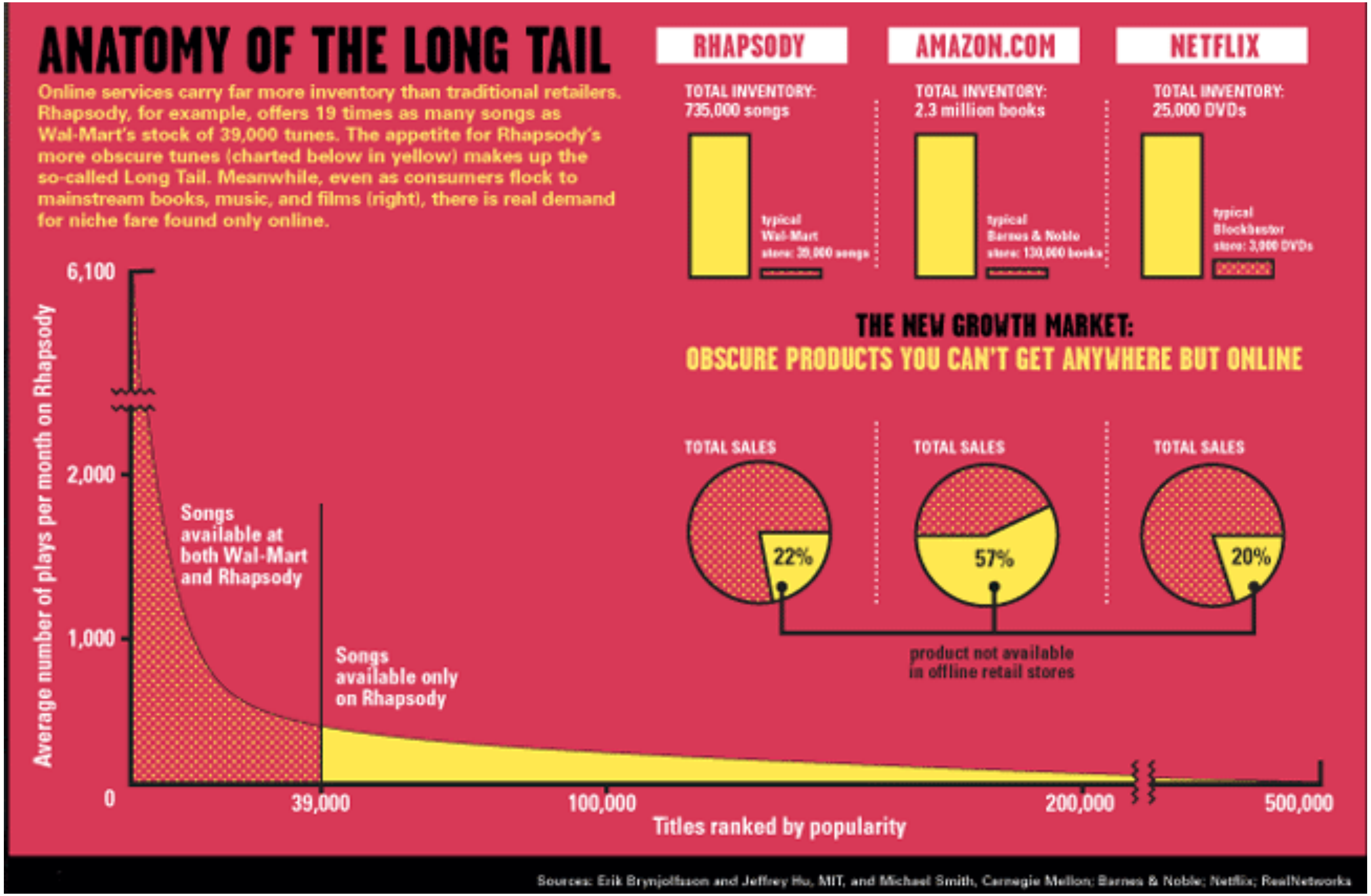
$$f(x) = ax^{-\alpha}$$

$$f(cx) = a(cx)^{-\alpha} = c^{-\alpha} \cdot ax^{-\alpha} = c^{-\alpha} f(x) \propto f(x)$$

Log() or Exp() are not scale free!

$f(ax) = \log(ax) = \log(a) + \log(x) = \log(a) + f(x)$
$f(ax) = \exp(ax) = \exp(x)^a = f(x)^a$

Anatomy of the Long Tail



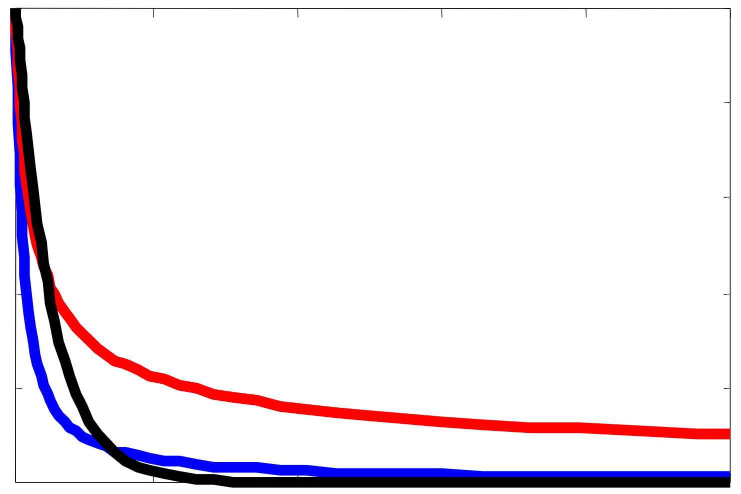
Mathematics of Power-Laws

Heavy-Tailed Distributions

- **Degrees are heavily skewed:**

  Distribution $P(X > x)$ is **heavy tailed if**:

$$\lim_{x\to\infty} \frac{P(X > x)}{e^{-\lambda x}} = \infty$$



- **Note:**
  - **Normal PDF:** $p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - **Exponential PDF:** $p(x) = \lambda e^{-\lambda x}$
    - **then** $P(X > x) = 1 - P(X \leq x) = e^{-\lambda x}$

are not heavy tailed!

Heavy-Tailed Distributions

Various names, kinds and forms:

Long tail, Heavy tail, Zipf's law, Pareto's law

Heavy tailed distributions:

P(x) is proportional to:

power law	$x^{-\alpha}$
power law with cutoff	$x^{-\alpha}\mathrm{e}^{-\lambda x}$
stretched exponential	$x^{\beta-1}\mathrm{e}^{-\lambda x^{\beta}}$
log-normal	$\frac{1}{x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$

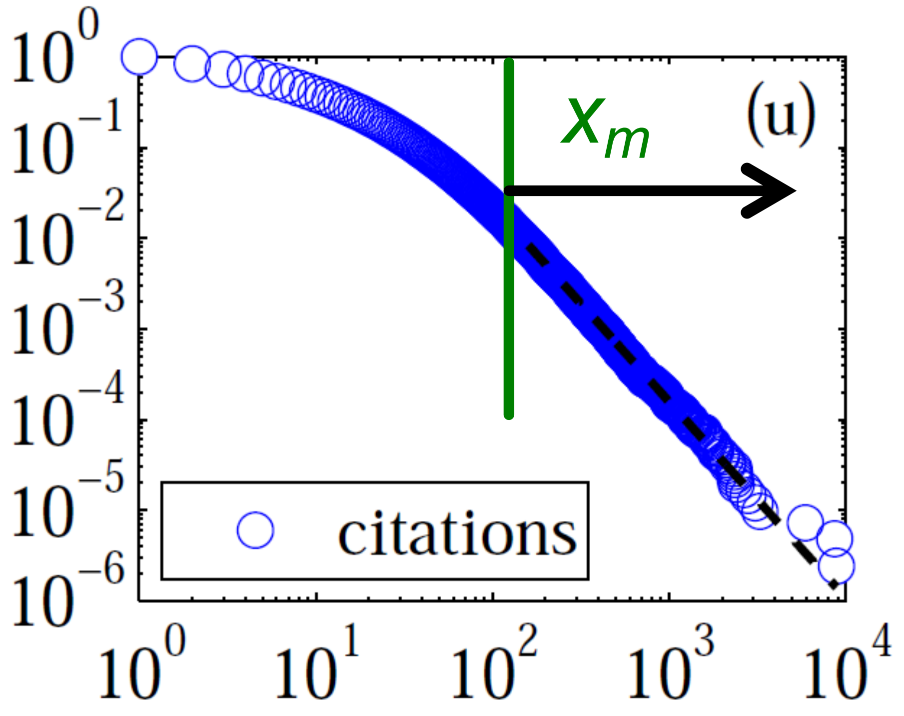
Mathematics of Power-laws

- **What is the normalizing constant?**

  $p(x) = Z\,x^{-\alpha}$ $\qquad Z = ?$

  - $\boldsymbol{p(x)}$ is a distribution: $\int \boldsymbol{p(x)dx = 1}$
  - $1 = \int_{x_m}^{\infty} p(x)dx = Z \int_{x_m}^{\infty} x^{-\alpha} dx$
  - $= -\frac{Z}{\alpha - 1}\left[x^{-\alpha+1}\right]_{x_m}^{\infty} = -\frac{Z}{\alpha - 1}\left[\infty^{1-\alpha} - x_m{}^{1-\alpha}\right]$
  - $\Rightarrow Z = (\alpha - 1)x_m^{\alpha - 1}$

Need: α **> 1 !**



p(x) diverges as $x \to 0$
so x_m is the minimum value
of the power-law distribution
$x \in [x_m, \infty]$

$$\boldsymbol{p(x) = \frac{\alpha - 1}{x_m}\left(\frac{x}{x_m}\right)^{-\alpha}}$$

Integral:

$$\int (\boldsymbol{ax})^{\boldsymbol{n}} = \frac{(\boldsymbol{ax})^{\boldsymbol{n+1}}}{\boldsymbol{a(n+1)}}$$

Mathematics of Power-laws

- **What's the expected value of a power-law random variable *X*?**
- $E[X] = \int_{x_m}^{\infty} x\, p(x) dx = Z \int_{x_m}^{\infty} x^{-\alpha+1} dx$
- $= \frac{Z}{2-\alpha} [x^{2-\alpha}]_{x_m}^{\infty} = \frac{(\alpha-1)x_m^{\alpha-1}}{-(\alpha-2)} [\infty^{2-\alpha} - x_m^{2-\alpha}]$

$$\Rightarrow E[X] = \frac{\alpha - 1}{\alpha - 2} x_m$$

Need: α **> 2 !**

Power-law density:

$$p(x) = \frac{\alpha - 1}{x_m} \left(\frac{x}{x_m}\right)^{-\alpha}$$

$$Z = \frac{\alpha - 1}{x_m^{1-\alpha}}$$

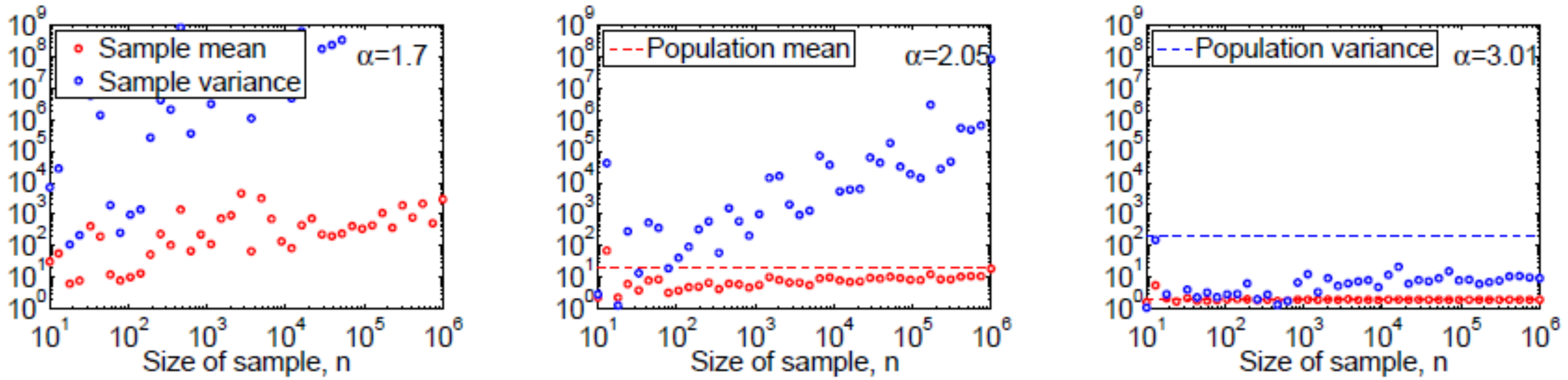
Mathematics of Power-Laws

- **Power-laws have infinite moments!**

$$E[X] = \frac{\alpha - 1}{\alpha - 2} x_m$$

  - If $\alpha \le 2 : E[X] = \infty$
  - If $\alpha \le 3 : Var[X] = \infty$
    - **Average is meaningless, as the variance is too high!**
- **Consequence: Sample average of *n* samples from a power-law with exponent α**

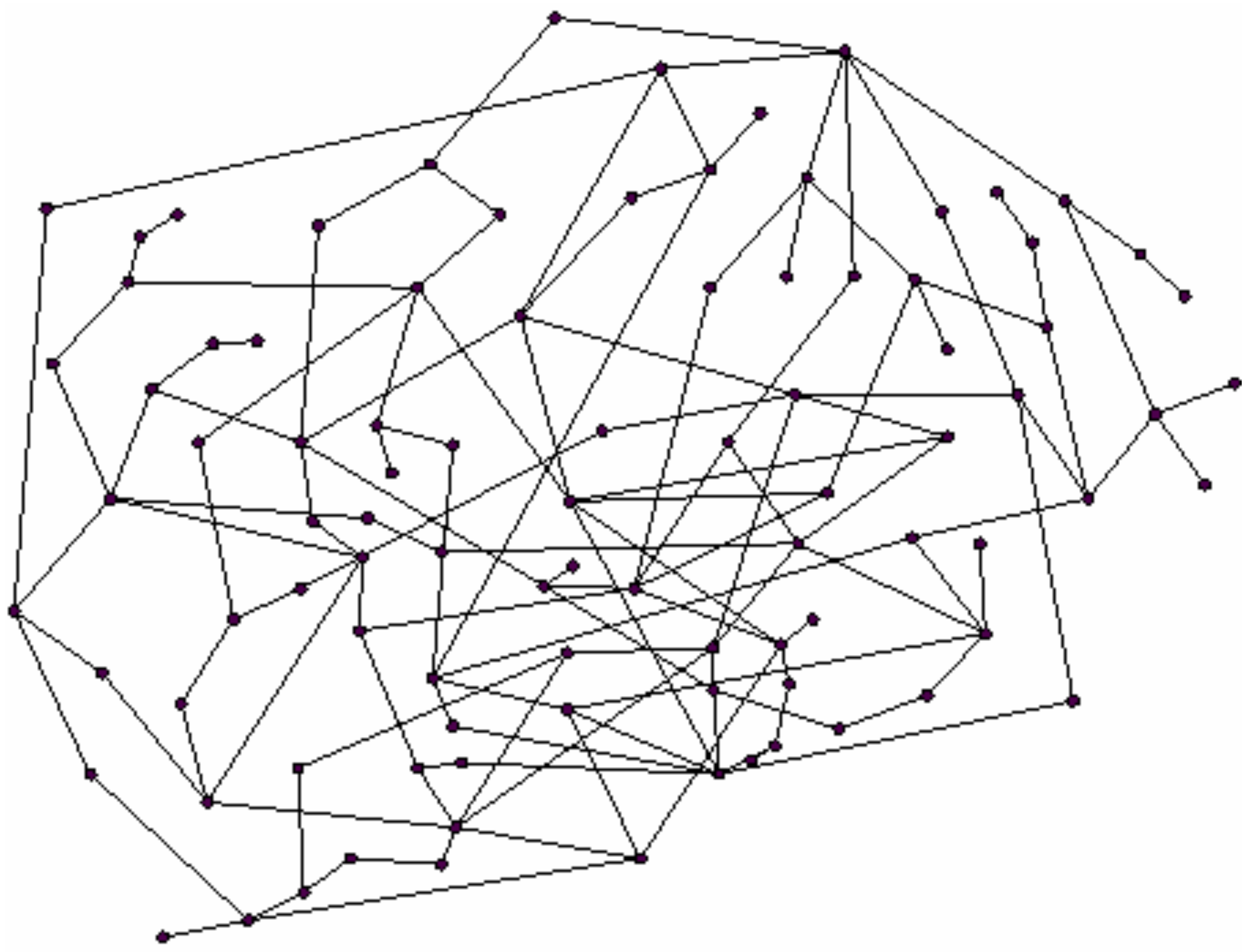
In real networks
2 < α < 3 so:
E[*X*] = const
Var[*X*] = ∞

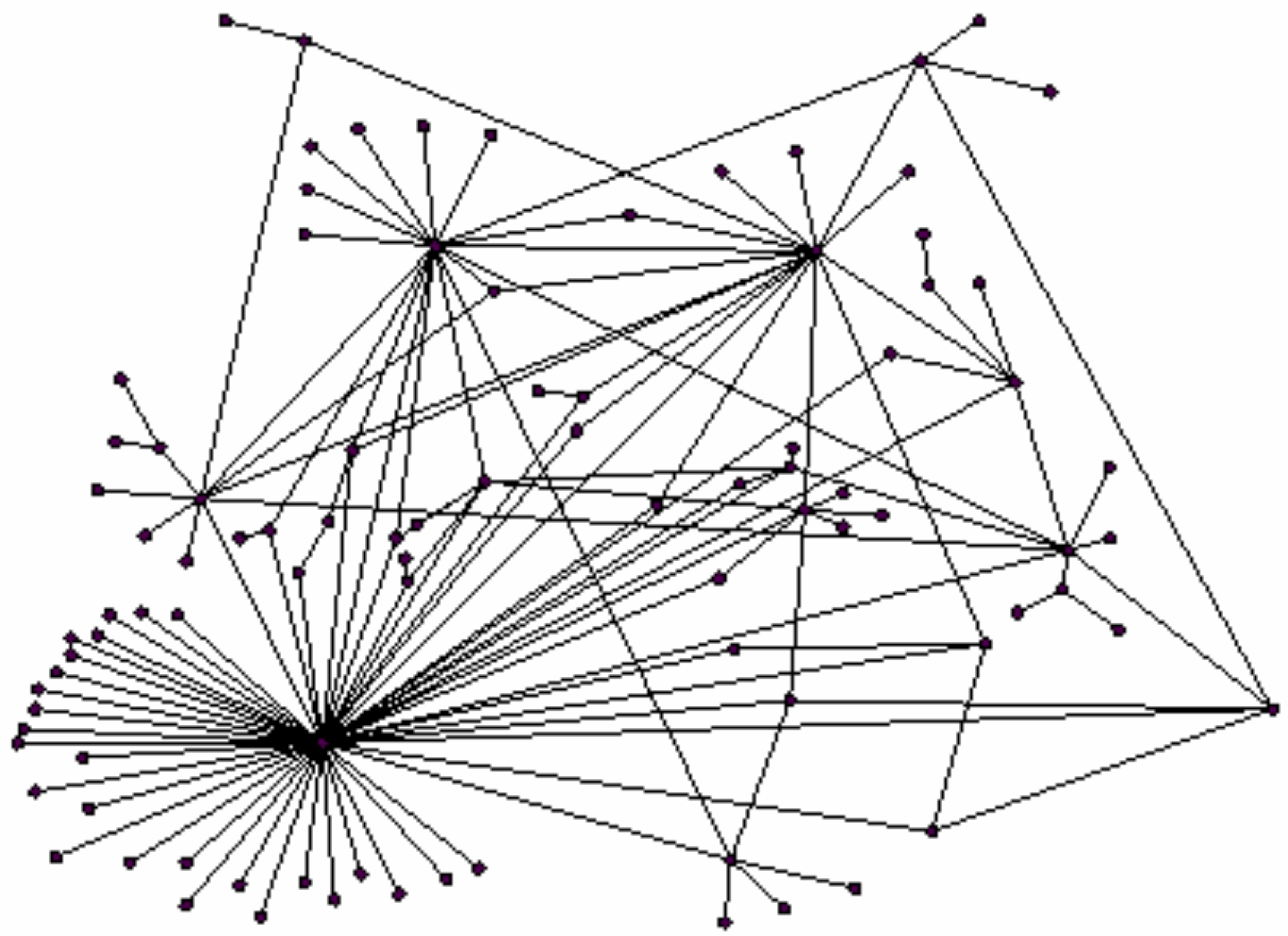


Why are Power-Laws Surprising

- <u>**Can not arise from sums of independent events!**</u>
  - **Recall:** in $\boldsymbol{G_{np}}$ each pair of nodes in connected independently with prob. $\boldsymbol{p}$
    - $\boldsymbol{X}$… degree of node $\boldsymbol{v}$
    - $\boldsymbol{X_w}$ … event that $\boldsymbol{w}$ links to $\boldsymbol{v}$
    - $\boldsymbol{X = \sum_w X_w}$
    - $\boldsymbol{E[X] = \sum_w E[X_w] = (n-1)p}$
  - **Now, what is $\boldsymbol{P(X=k)}$? Central limit theorem!**
    - $\boldsymbol{X_1, \dots, X_n}$: random vars with mean μ, variance σ^2
    - $\boldsymbol{S_n = \sum_i X_i}$: $E[S_n] = \boldsymbol{n\mu}$, $\mathrm{Var}[S_n] = \boldsymbol{n\sigma^2}$, $\mathrm{SD}[S_n] = \boldsymbol{\sigma\sqrt{n}}$
    - $\boldsymbol{P(S_n = E[S_n] + x \cdot \mathrm{SD}[S_n]) \sim \frac{1}{2\pi} e^{-\frac{x^2}{2}}}$

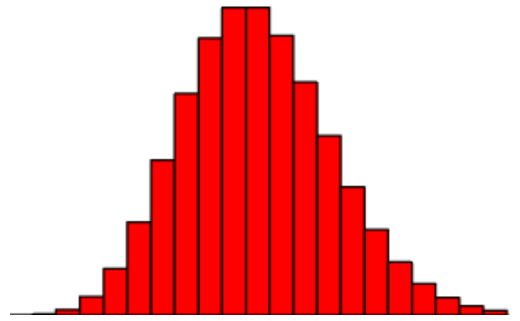
Random vs. Scale-free network





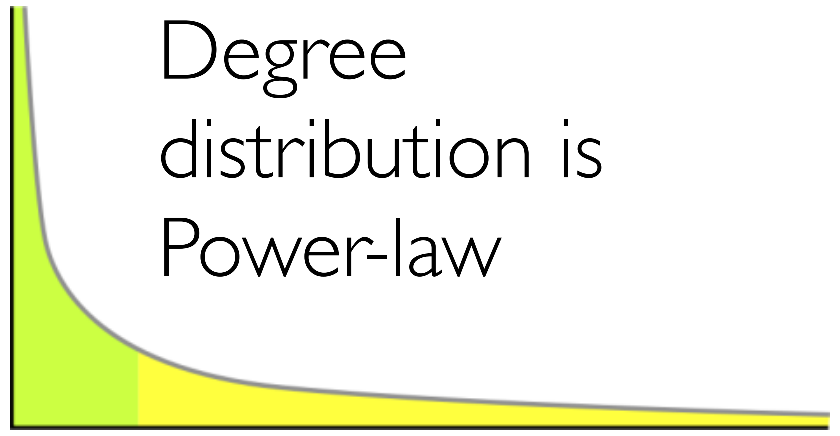
Random network

(Erdos-Renyi random graph)



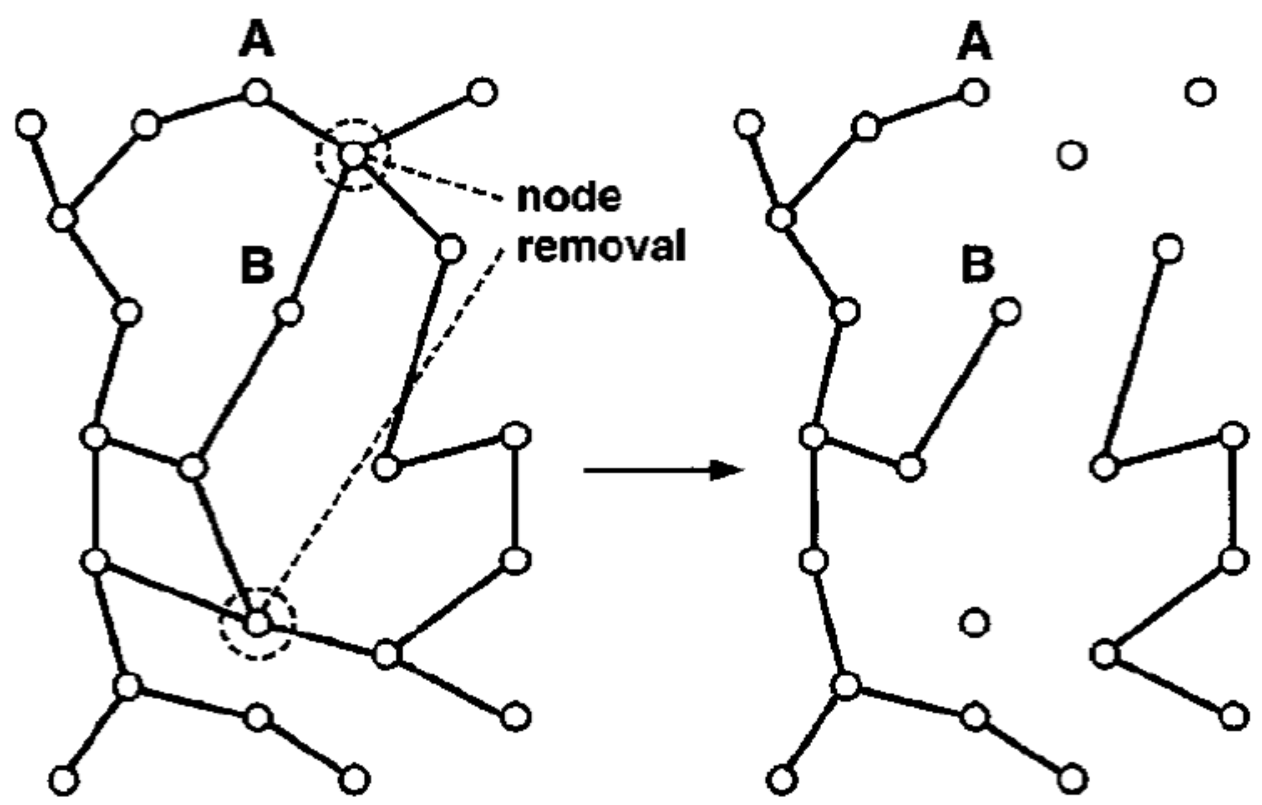
Degree distribution is Binomial

Scale-free (power-law) network



Degree distribution is Power-law

Consequence: Network Resilience



How does network connectivity change as nodes get removed?
[Albert et al. 00; Palmer et al. 01]

Nodes can be removed in two main ways:

Random failure:

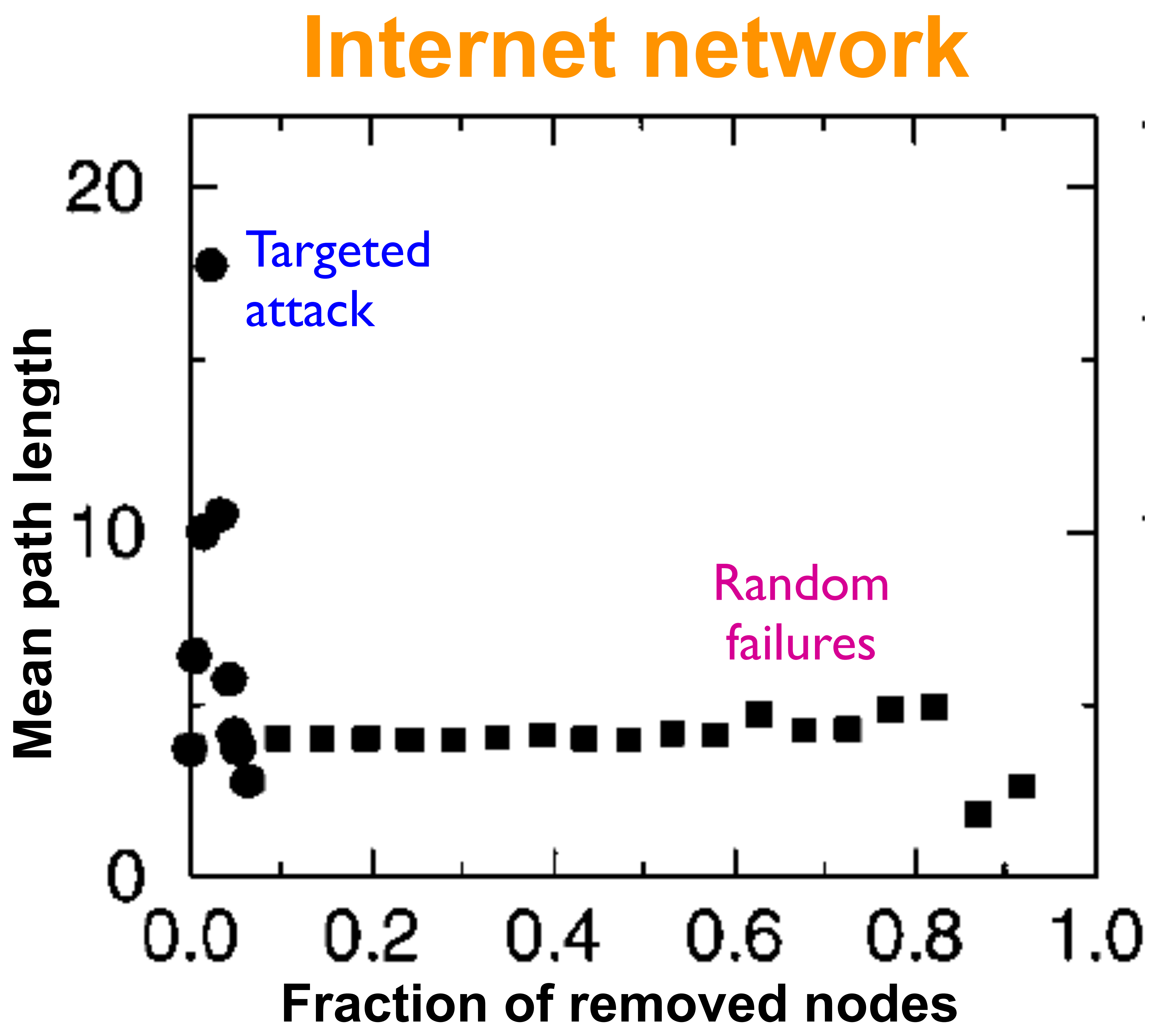
Remove nodes uniformly at random

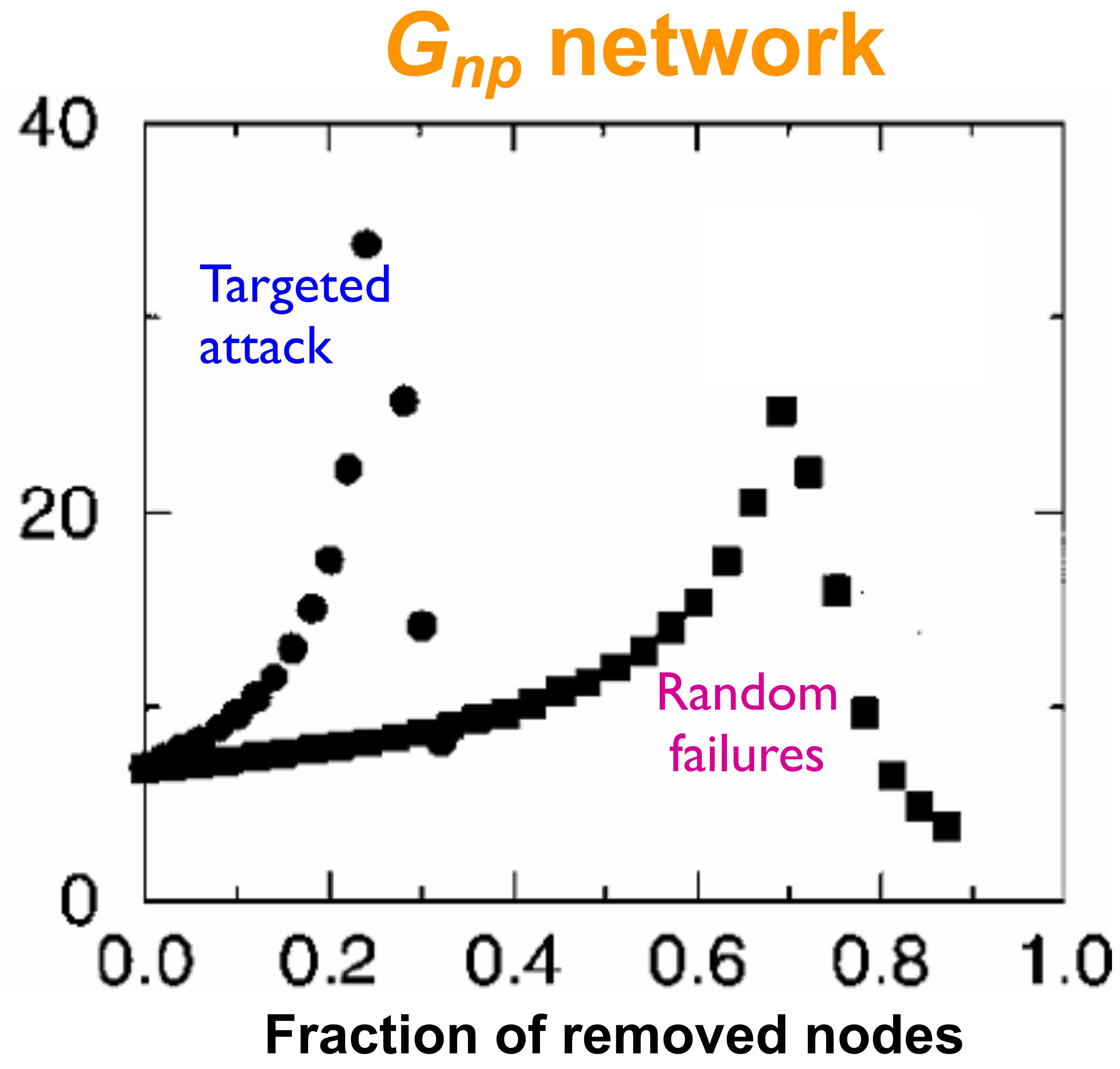
Targeted attack:

Remove nodes in order of decreasing degree

This is important for **robustness of the internet** as well as **epidemiology**

Network Resilience





Real networks are resilient to random failures

G_{np} has better resilience to targeted attacks

- Need to remove all pages of degree **>5** to disconnect the Web
- But this is a very small fraction of all web pages

Inequality

A Thought Experiment

One of the crucial properties of heavy-tailed distributions is **inequality** (in some sense this follows from the *definition* of a heavy-tailed distribution)

Some nodes have millions of connections, some have one

A Thought Experiment

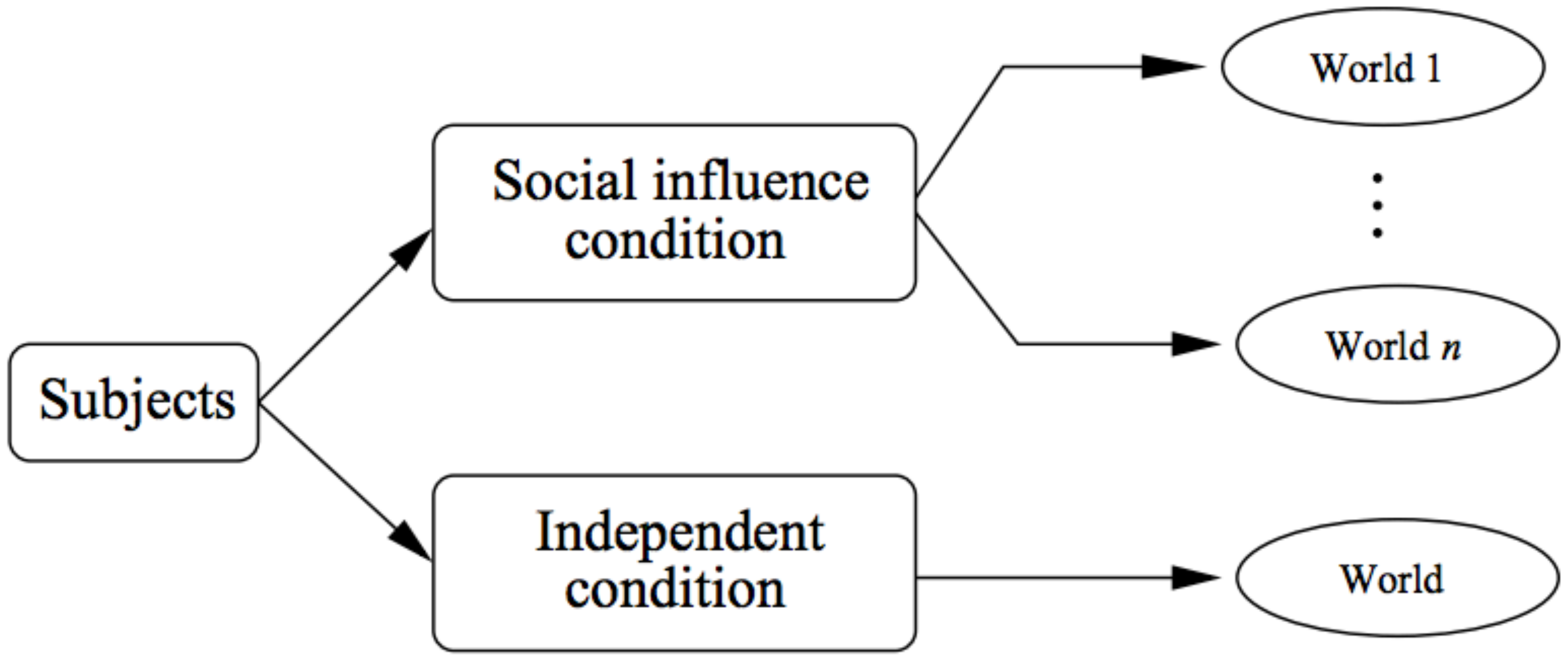
Do Drake/Ariana Grande/The Beatles “deserve” their fame?

If you ran the world over again, would they still have been as big?

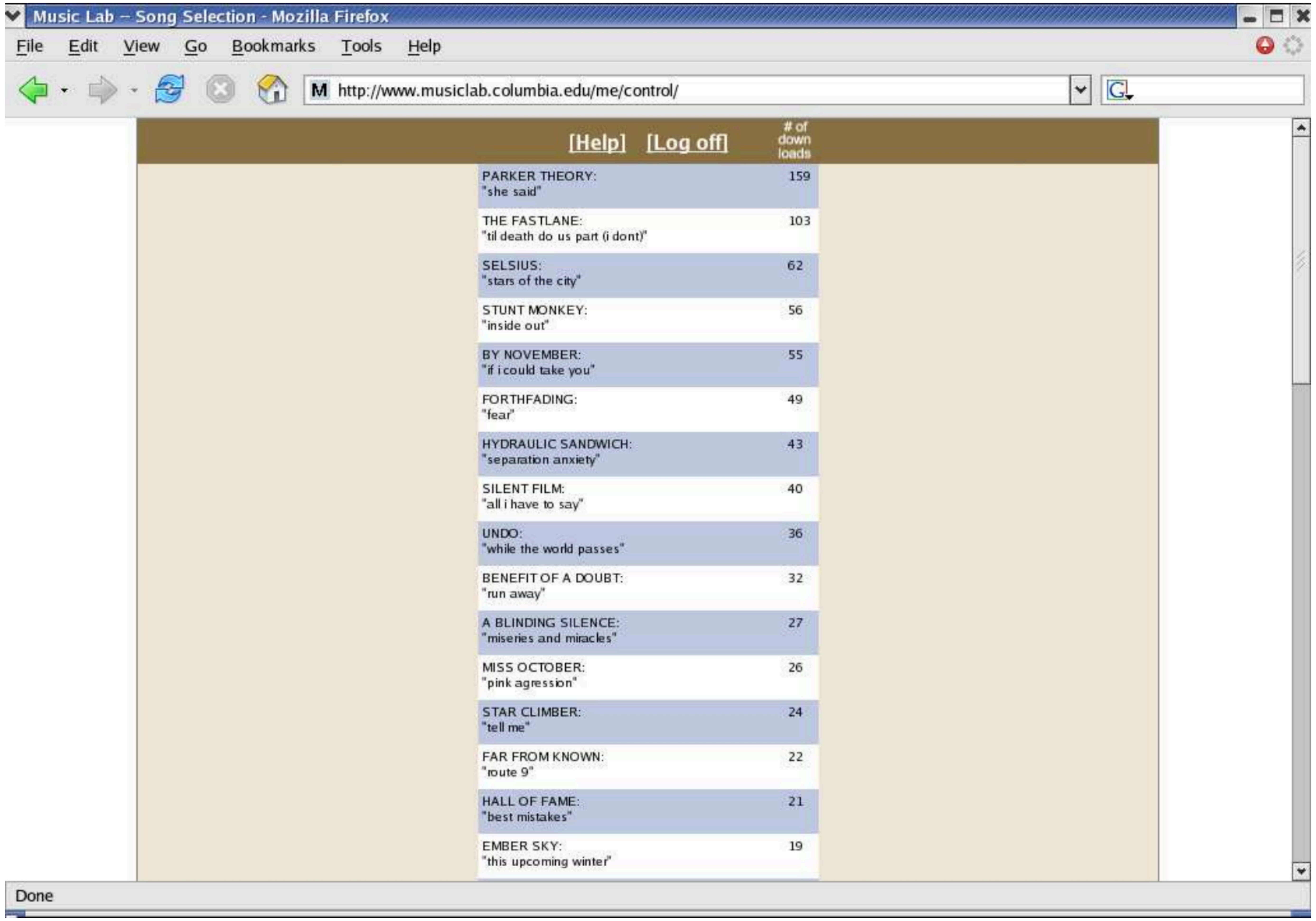
Run the experiment!

Salganik, Dodds, and Watts '06 ran an experiment called MusicLab

Got ~2,000 people to come to their music download site (never-before-heard music)



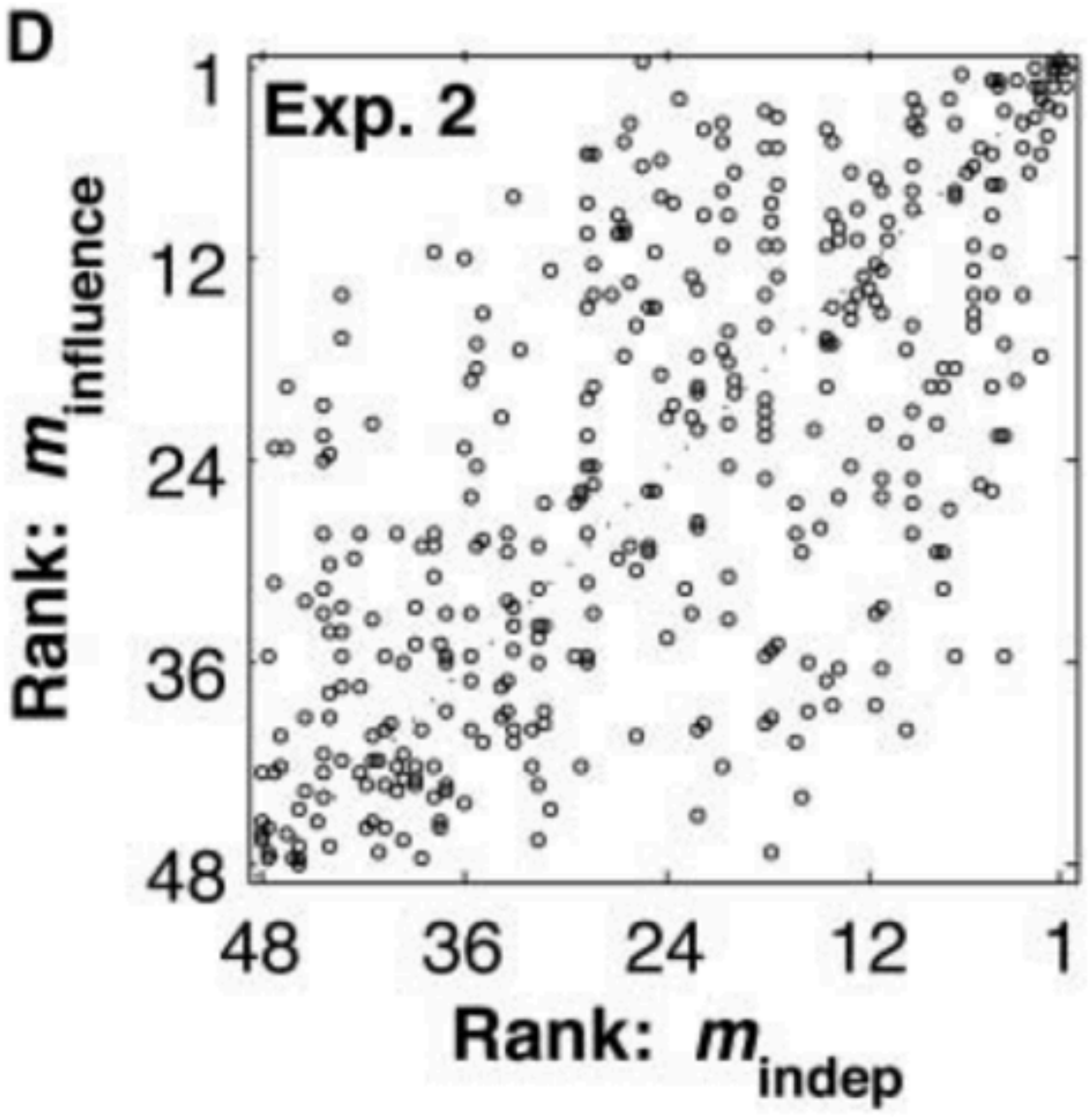
Run the experiment!



Download counts shown in social influence world, not shown in control world

MusicLab:

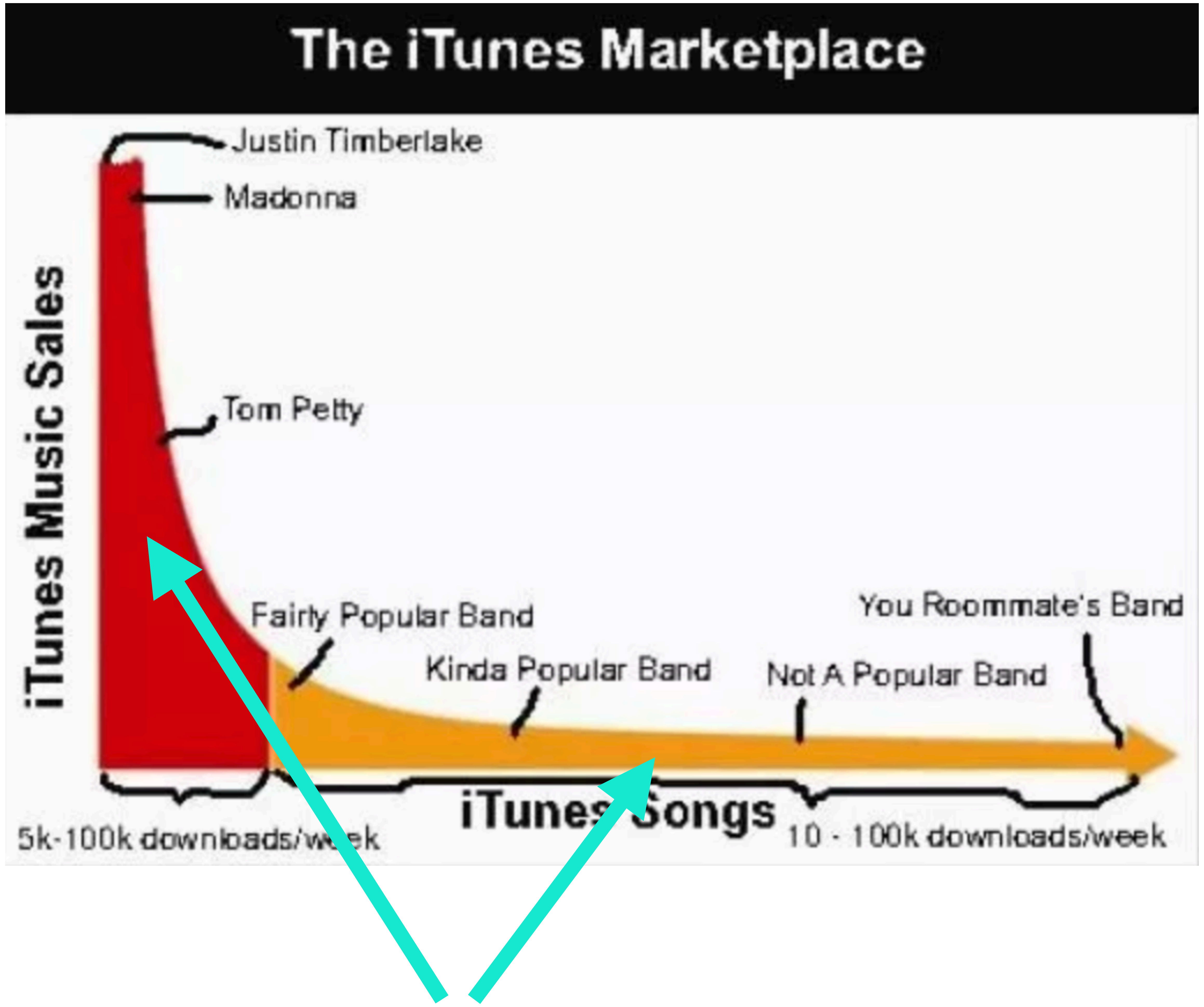
success



"quality"

Success is inherently unpredictable from quality

MusicLab:

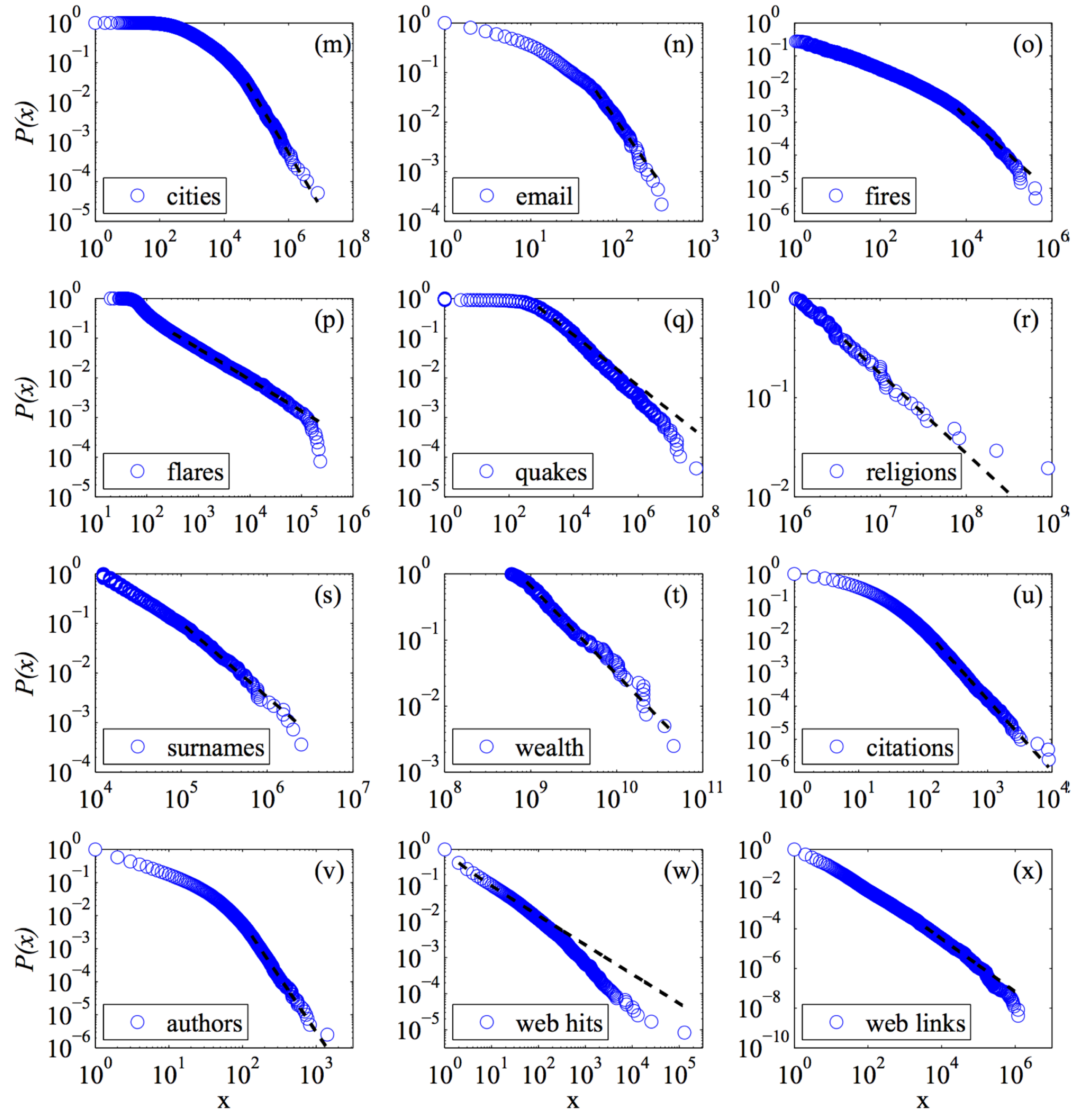


Who ends up here is pretty **random**!

What causes power laws?

What underlying process is keeping the line so straight?

And in such a variety of settings?



Central Limit Theorem : Gaussian :: _____ : Power Laws?

Preferential Attachment Model

Key idea: rich get richer

Normal distributions can come from many independent random variables averaging out

Power laws can arise from the **rich getting richer**

Another way to put it: from the **feedback** introduced by correlated events

Rich Get Richer

Example in networks: new nodes are more likely to link to nodes that already have high degree

Herbert Simon's result:

- Power-laws arise from "**Rich get richer**" (**cumulative advantage**)

Examples [Price '65]

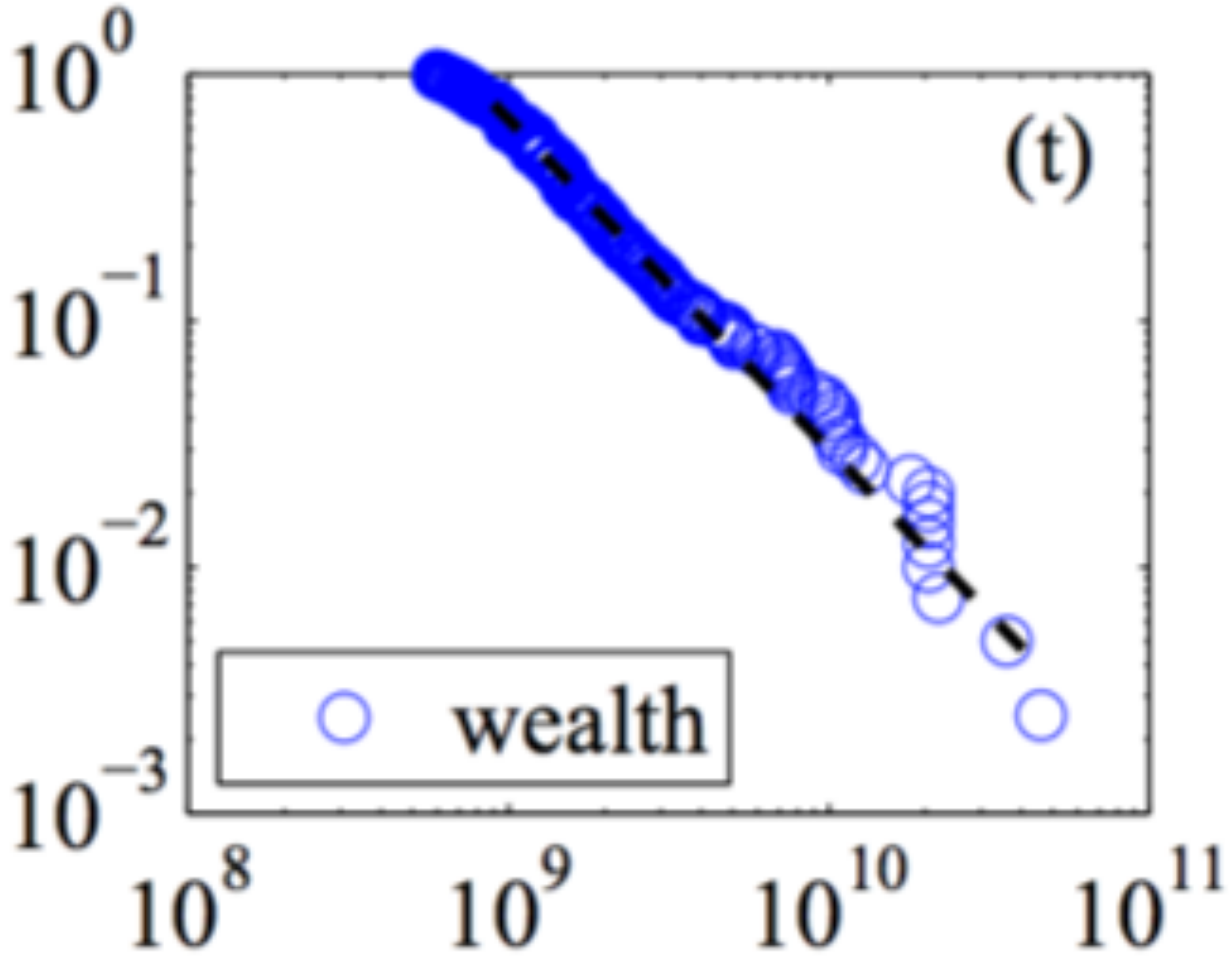
- Citations: New citations to a paper are proportional to the number it already has
- Herding: If a lot of people cite a paper, then it must be good, and therefore I should cite it too

Think back to wealth

People with different amounts of money

All put it in the bank and get compound interest

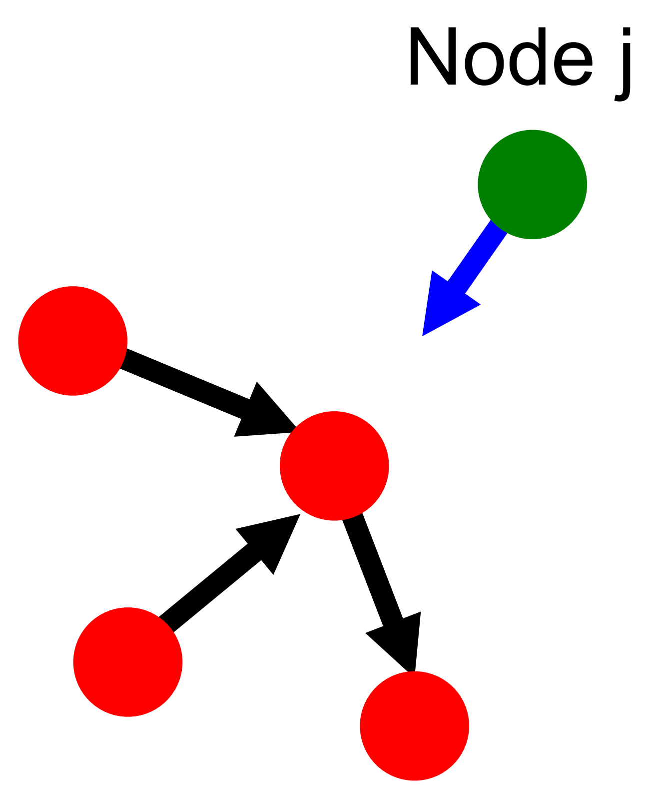
Rich get richer (literally)



The Exact Model

We will analyze the following model:

- Nodes arrive in order $1, 2, 3, \ldots, n$
- When node $\boldsymbol{j}$ is created it makes a **single out-link** to an earlier node $\boldsymbol{i}$ chosen:
  - **1)** With prob. $\boldsymbol{p}$, $\boldsymbol{j}$ links to $\boldsymbol{i}$ chosen **uniformly at random** (from among all earlier nodes)
  - **2)** With prob. $\mathbf{1} - \boldsymbol{p}$, node $\boldsymbol{j}$ chooses $\boldsymbol{i}$ uniformly at random and links **to node *l that i* points to**
    - **This is same as saying:** With prob. $\mathbf{1} - \boldsymbol{p}$, node $\boldsymbol{j}$ links to node $\boldsymbol{l}$ with prob. proportional to $\boldsymbol{d_l}$ (the in-degree of $\boldsymbol{l}$)
  - **Our graph is directed:** Every node has out-degree **1**



The Model Gives Power-Laws

<u>Claim:</u> The described model generates networks where the fraction of nodes with in-degree *k* scales as:

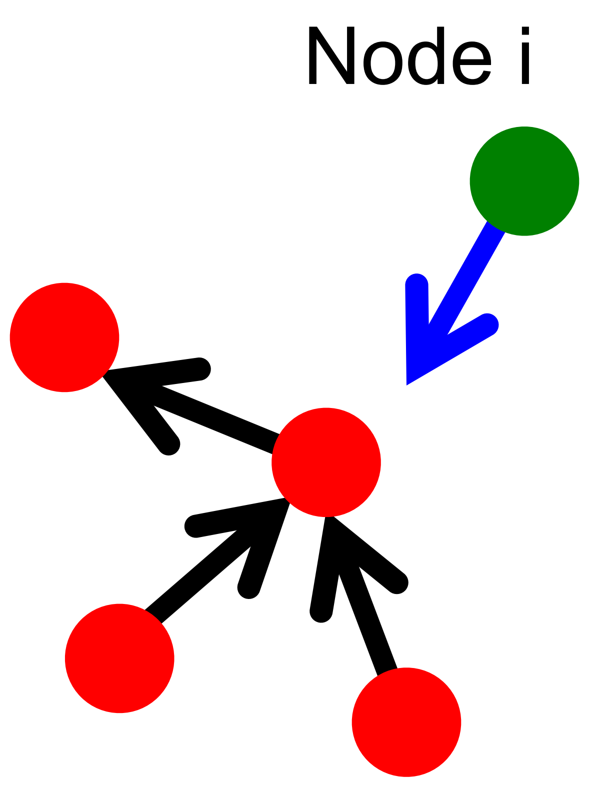
$$P(d_i = k) \propto k^{-(1+\frac{1}{q})}$$

where q=1-p

So we get power-law degree distribution with exponent:

$$\alpha = 1 + \frac{1}{q} = 1 + \frac{1}{1-p}$$

Degrees Over Time: What We Know



- **Initial condition:**
  - $d_i(t) = 0$, when $t = i$ (node i just arrived)
- **Expected change of $d_i(t)$ over time:**
  - Node i gains an in-link at step $t+1$ only if a link from a newly created node $t+1$ points to it.
  - **What's the probability of this event?**
    - With prob. p node $t+1$ links **randomly**:
      - Links to our node i with prob. $1/t$
    - With prob. $1-p$ node $t+1$ links **preferentially**:
      - Links to our node i with prob. $d_i(t)/t$
  - **Prob. node $t+1$ links to i is:** $p\frac{1}{t} + (1-p)\frac{d_i(t)}{t}$

Continuous Approximation

Analyzing this probabilistic discrete process is too involved

- **Consider deterministic and continuous approximation** to the degree of node $\boldsymbol{i}$ as a function of time $\boldsymbol{t}$
  - $\boldsymbol{t}$ is the number of nodes that have arrived so far
  - In-Degree $\boldsymbol{d_i(t)}$ of node $\boldsymbol{i}$ ($i = 1,2, \dots, n$) is a **continuous quantity** and it **grows deterministically** as a function of time $\boldsymbol{t}$
- **Plan: Analyze $\boldsymbol{d_i(t)}$ – continuous in-degree** of node $\boldsymbol{i}$ at time $\boldsymbol{t} > \boldsymbol{i}$
  - **Note: Node $\boldsymbol{i}$ arrives to the graph at time $\boldsymbol{t}$**

Continuous Degree

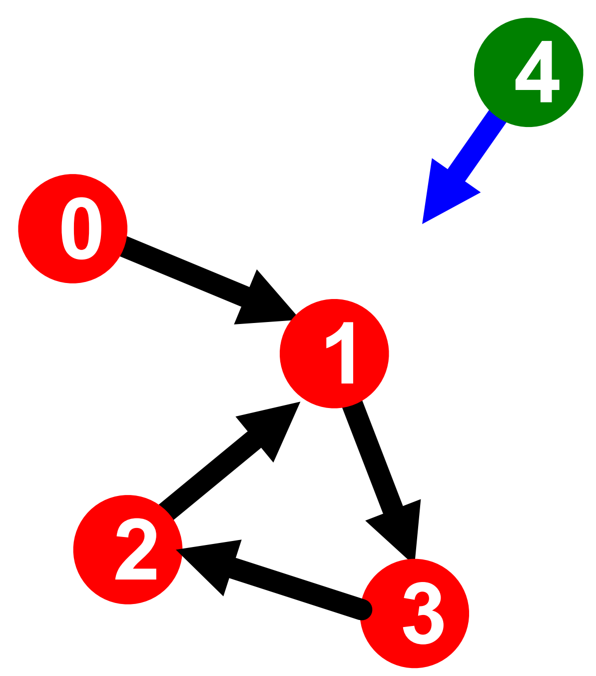
Time is now continuous, and degrees $d_i(t)$ evolve deterministically

Initial condition: $d_i(i) = 0$, as before

Growth equation:

Remember that before, prob that d_i increases is

$$\frac{p}{t} + \frac{(1-p)d_i(t)}{t}$$



Now:

$$\frac{dd_i}{dt} = \frac{p}{t} + \frac{(1-p)d_i}{t}$$

What is the rate of growth of d_i?

$$\frac{dd_i}{dt} = \frac{p + qd_i}{t}$$

$q = (1 - p)$

$$\frac{1}{p + qd_i}\frac{dd_i}{dt} = \frac{1}{t}$$

Divide by
$p + q\ d_i(t)$

$$\int \frac{1}{p + qd_i}\frac{dd_i}{dt}dt = \int \frac{1}{t}dt$$

integrate

$$\ln(p + qd_i) = q\ln t + c$$

$$p + qd_i = At^q$$

Exponentiate
and let $A = e^c$

$$\Rightarrow d_i(t) = \frac{1}{q}(At^q - p)$$

A=?

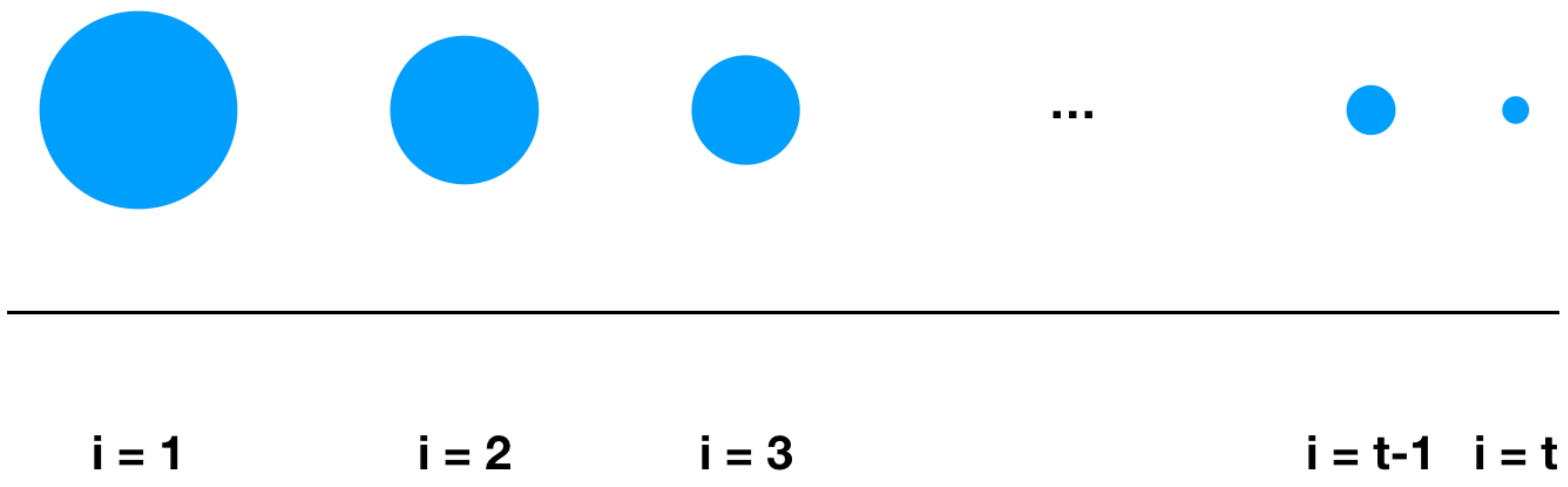
What is the constant A?

What is the value of constant A?

$$d_i(t) = \frac{1}{q}(At^q - p)$$

- **We know:** $d_i(i) = 0$
- **So:** $d_i(i) = \frac{1}{q}(Ai^q - p) = 0$
- $\Rightarrow A = \frac{p}{i^q}$
- **And so** $\Rightarrow d_i(t) = \frac{p}{q}\left(\left(\frac{t}{i}\right)^q - 1\right)$

Observation: Old nodes (small i values) have higher in-degrees $d_i(t)$



What is fraction of nodes with degree at least k?

Given k and time t, what fraction of all functions $d_i(t)$ satisfy $d_i(t) \geq k$?

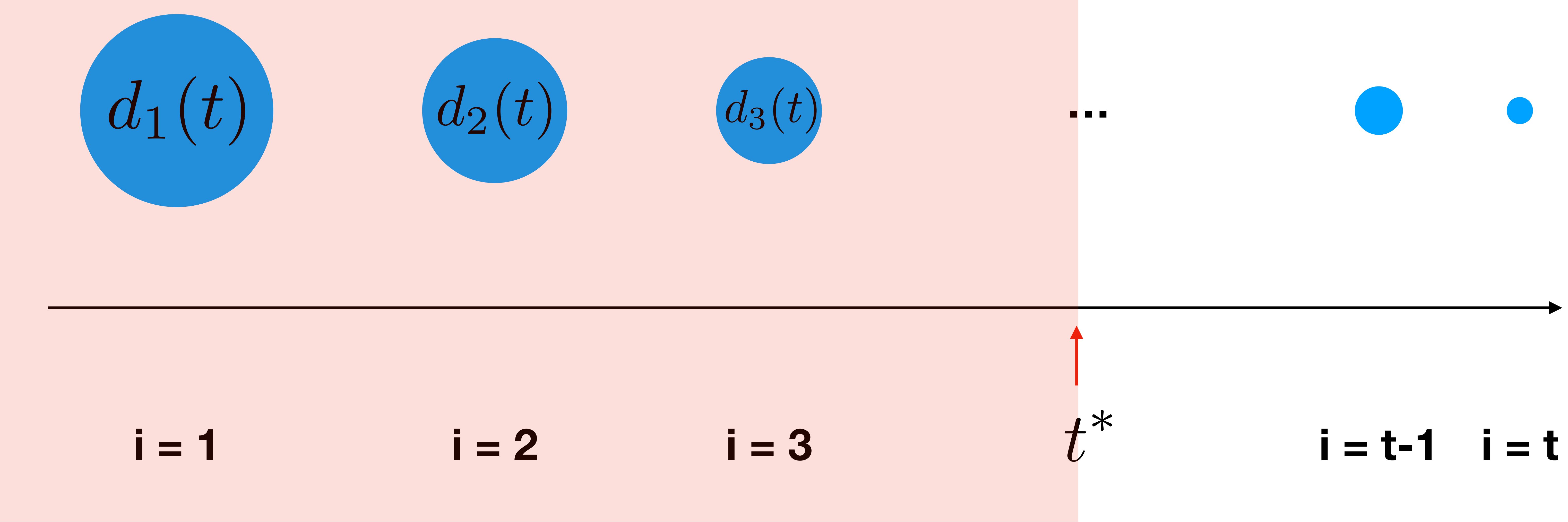
$$d_i(t) = \frac{p}{q}\left[\left(\frac{t}{i}\right)^q - 1\right] \geq k$$

Degree as a function of time

$$i \leq t\left[\frac{q}{p}k + 1\right]^{-1/q}$$

Rewrite in terms of i

t^*

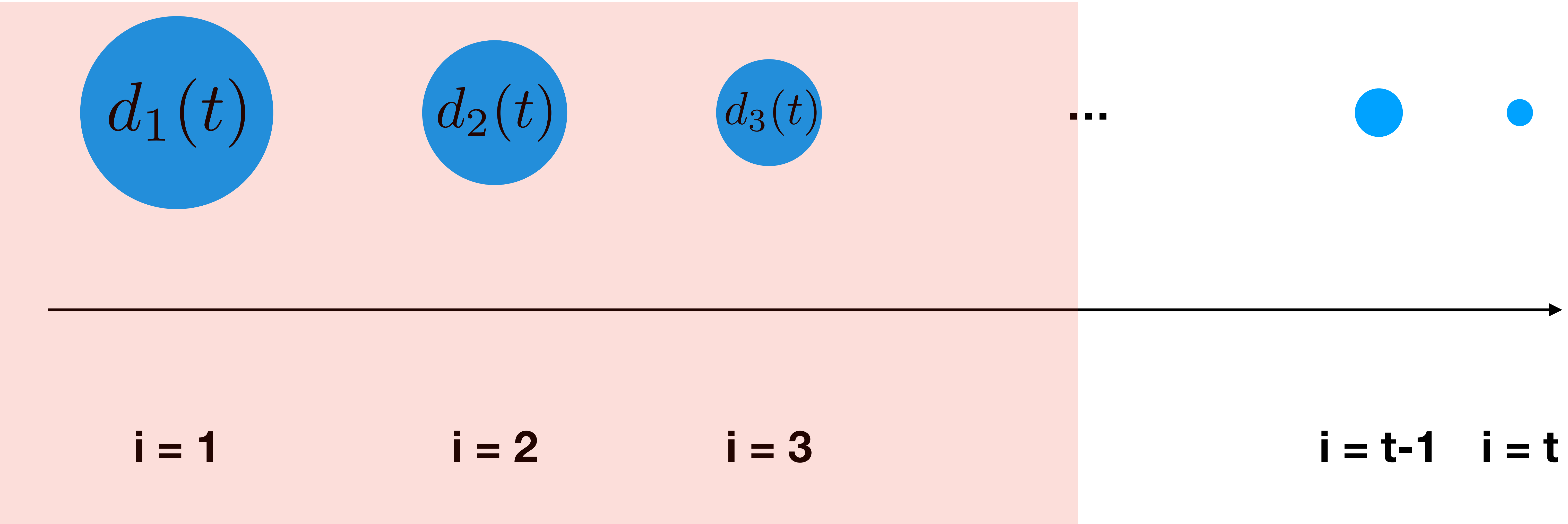


What is fraction of nodes with degree at least k?

Fraction that satisfy is: $i \leq \frac{t^*}{t}$

Recall that are t nodes at time t

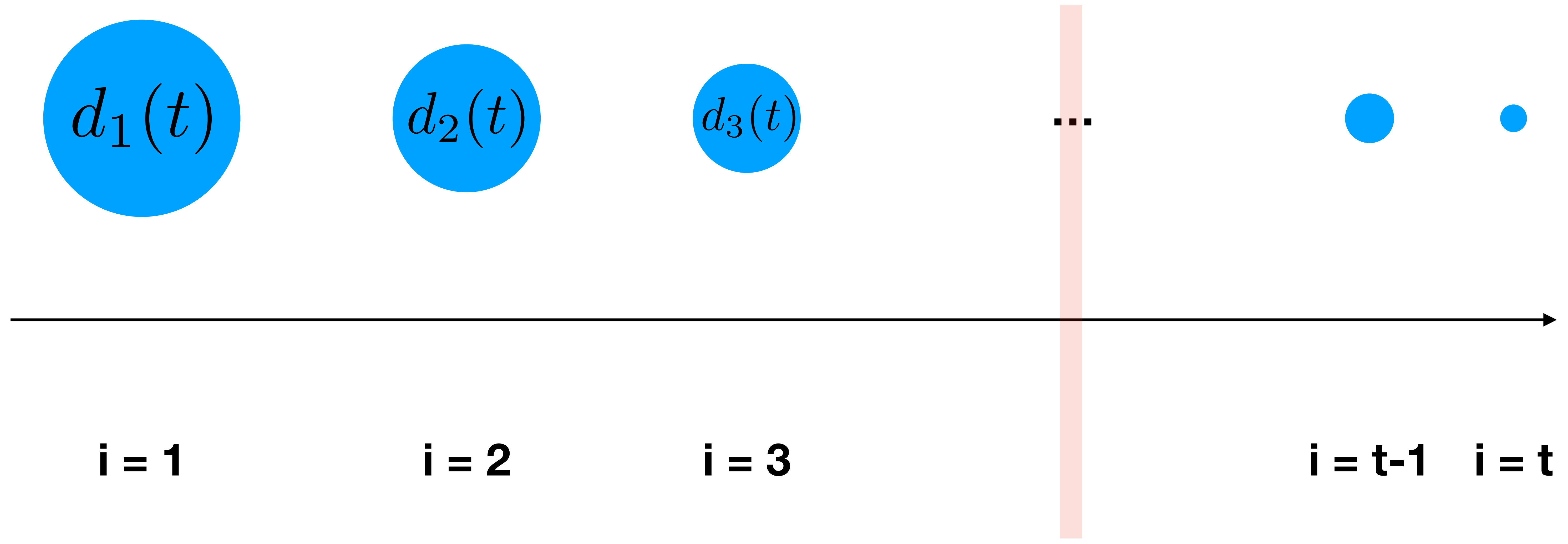
$$i \leq \frac{1}{t} t \left[\frac{q}{p} k + 1 \right]^{-1/q} = \left[\frac{q}{p} k + 1 \right]^{-1/q}$$



What is the fraction of nodes with degree exactly k?

$$F(k) = \left[\frac{q}{p}k + 1\right]^{-1/q} \quad \text{and} \quad f(k) = -dF/dk$$

$$\Rightarrow f(k) = \frac{1}{p}\left[\frac{q}{p}k + 1\right]^{-1-1/q}$$



We're done!!

$$\Rightarrow f(k) = \frac{1}{p}\left[\frac{q}{p}k + 1\right]^{-1-1/q}$$

Degree

Fraction of nodes with *k* in-links is proportional to $k^{-(1+1/q)}$

As we vary q (= 1-p):

- when q is close to 0, link formation is random choices, exponent goes to infinity (huge values rare)
- when q is close to 1, link formation is rich-get-richer, exponent goes to 2 (typical power law, huge values happen)

Preferential attachment: Good news

Preferential attachment gives
power-law degrees!

Intuitively reasonable process

Can tune ***p*** to get the observed exponent

On the web, ***P[node has degree d] ~ $d^{-2.1}$***

$2.1 = 1+1/(1-p)$ <u>$p \sim 0.1$</u>

Many models lead to Power-Laws

Copying mechanism (directed network)

Select a node and an edge of this node

Attach to the endpoint of this edge

Walking on a network (directed network)

The new node connects to a node, then to every first, second, … neighbor of this node

Attaching to edges

Select an edge and attach to both endpoints of this edge

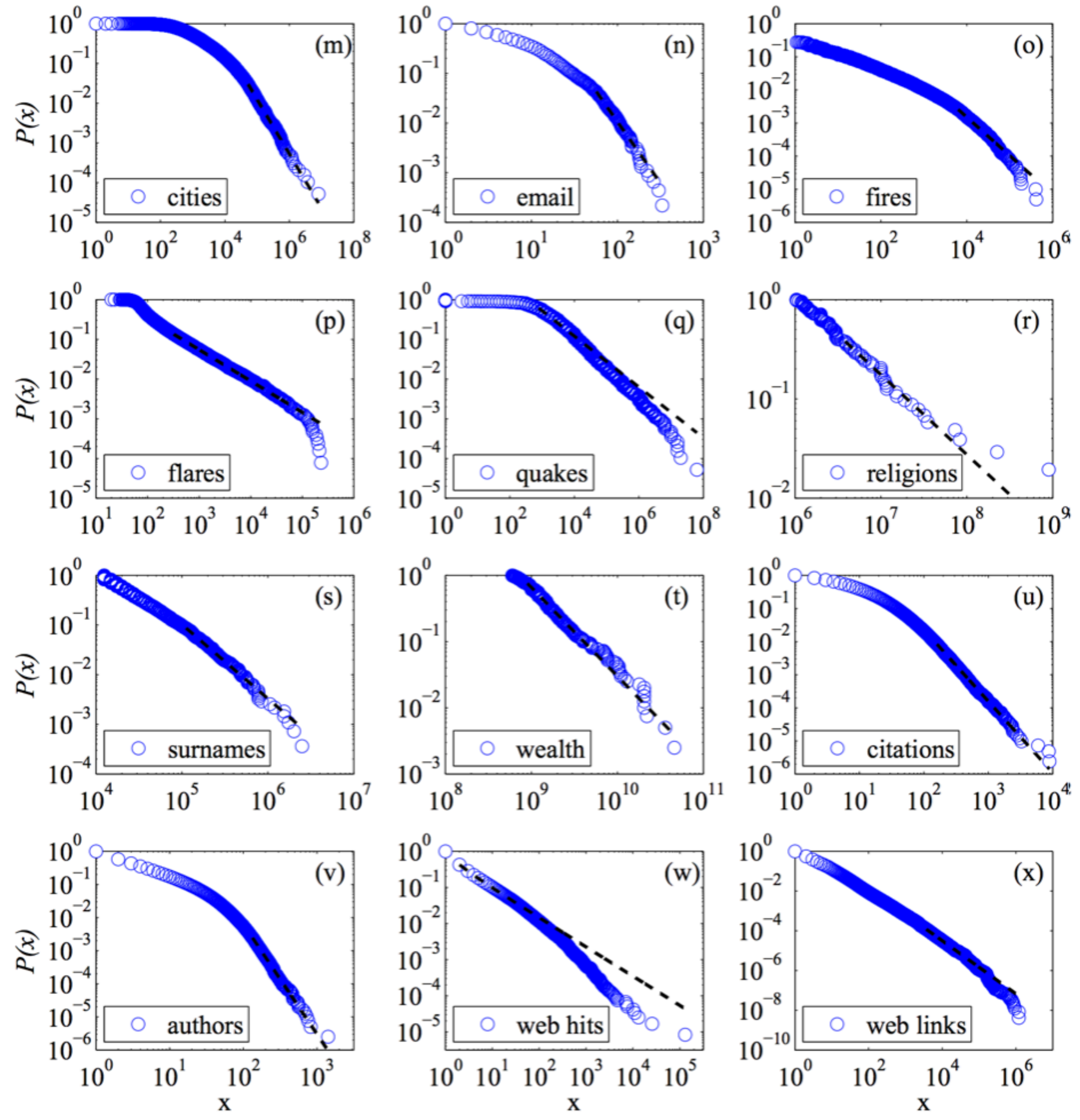
Node duplication

Duplicate a node with all its edges

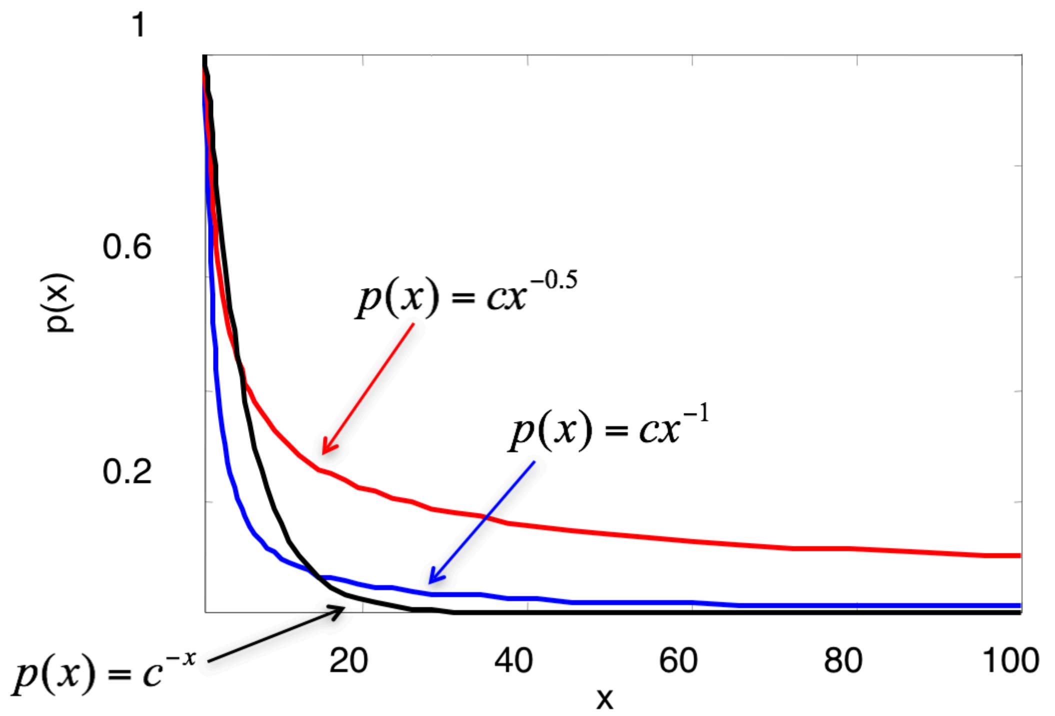
Randomly prune edges of new node

Power Laws

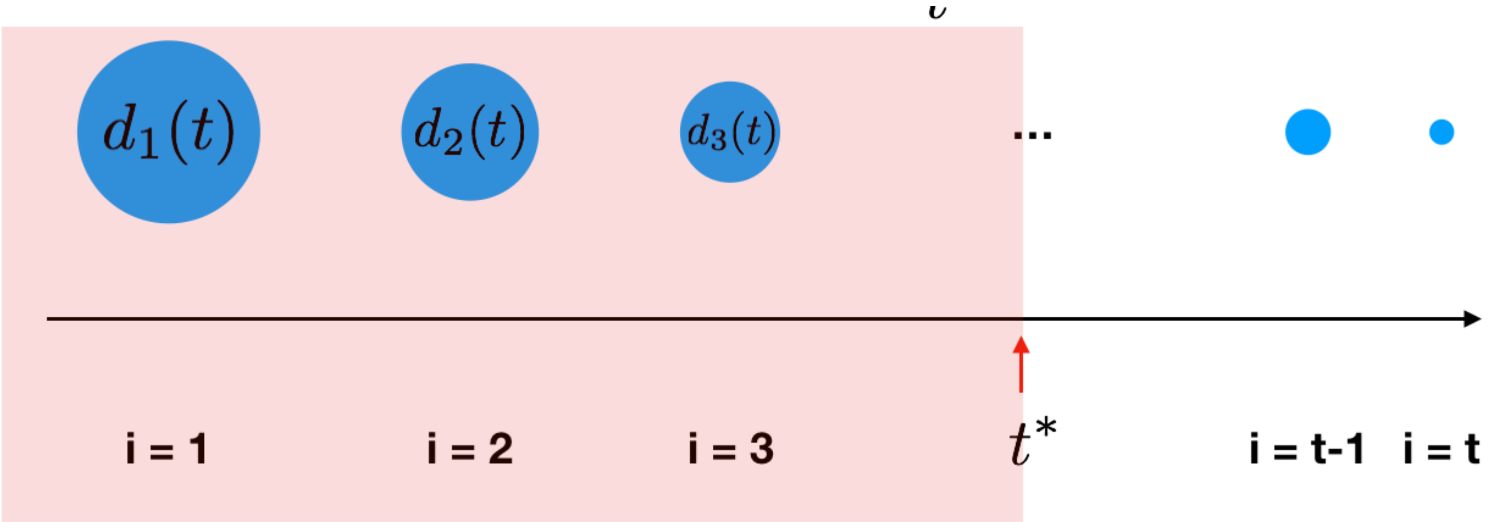
They’re everywhere

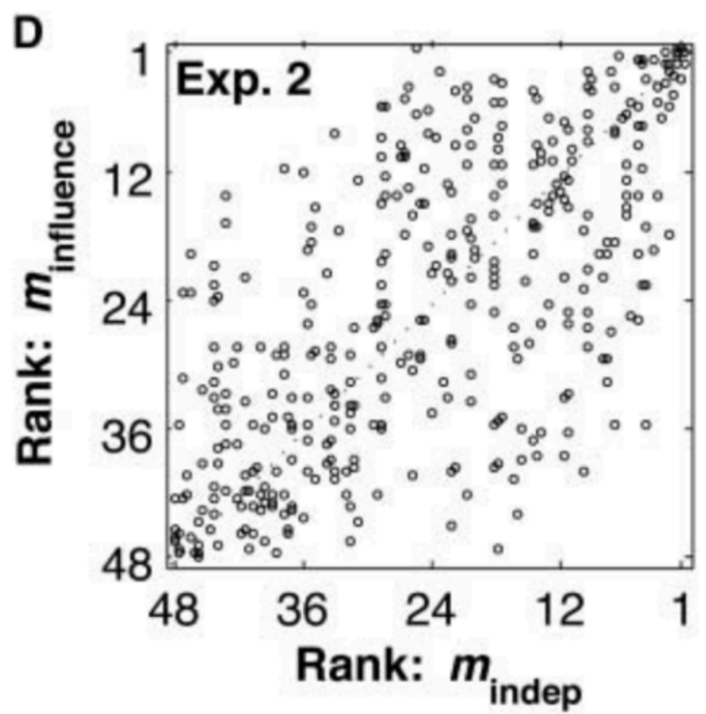


They’re “heavy-tailed”



They can arise from rich-get-richer dynamics





They mean the world is more unpredictable, and less meritocratic, than you might think