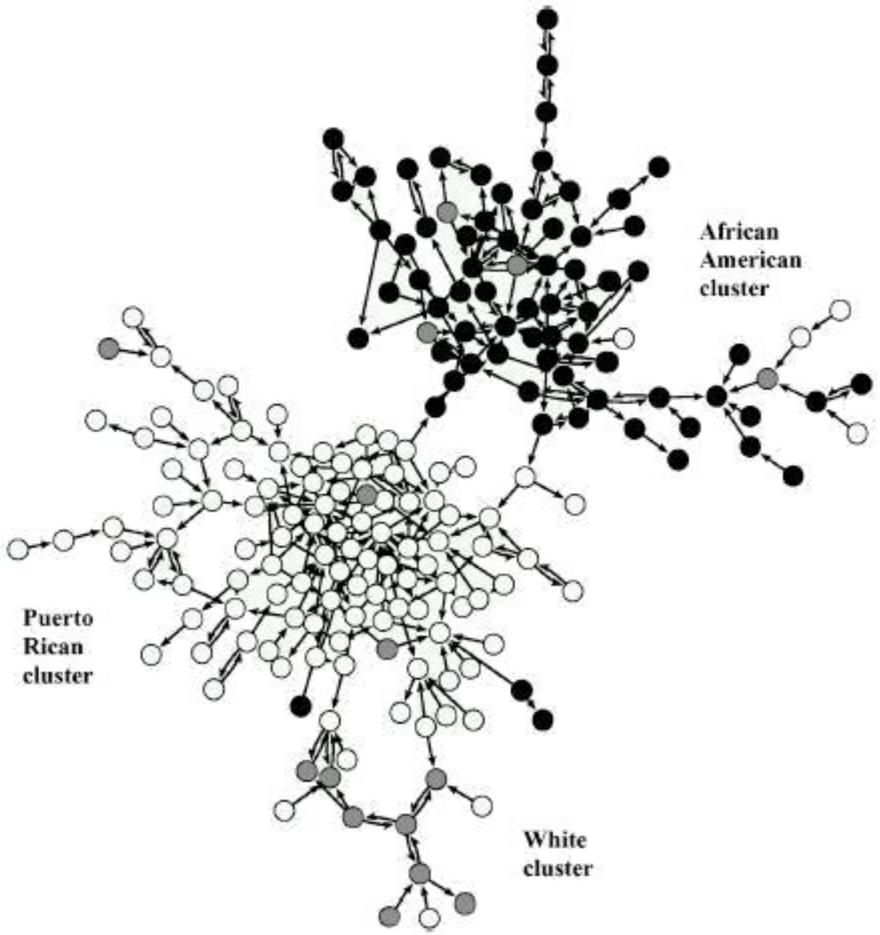
Lecture 8: Network Models III

A few things about Evolving Networks

- What can we do with graphs?
  - What patterns or “laws” hold for most real-world graphs?
  - How do the graphs evolve over time?
  - Can we generate synthetic but “realistic” graphs?



“Needle exchange” networks of drug users

Evolution of Graphs

- How do graphs evolve over time?

- Conventional Wisdom:
  - Constant average degree: the number of edges grows linearly with the number of nodes
  - Slowly growing diameter: as the network grows the distances between nodes grow

- Findings:
  - Densification Power Law: networks are becoming denser over time
  - Shrinking Diameter: diameter is decreasing as the network grows

Evolution of aggregate network metrics

- As individual nodes and edges come and go, how do aggregate features change?
  - degree distribution?
  - clustering coefficient?
  - average shortest path?

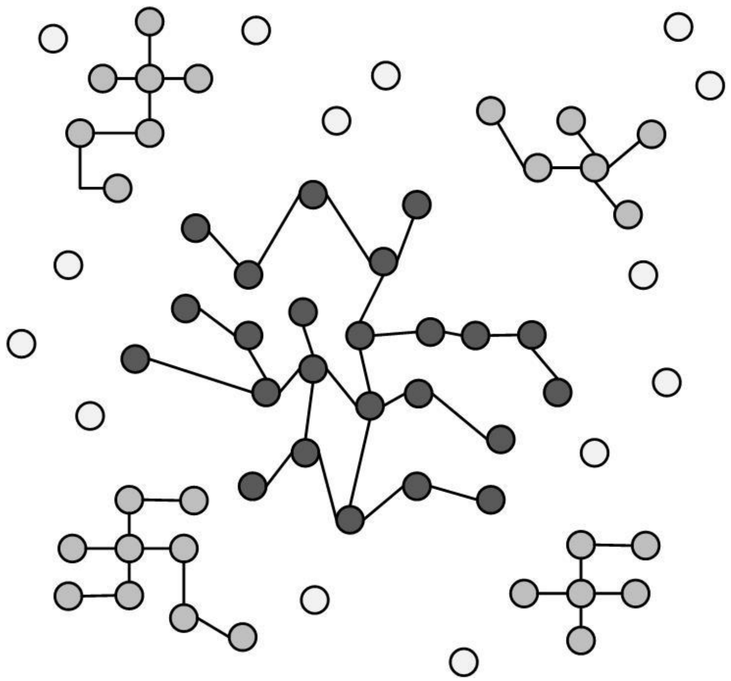
Network Growth Patterns

- Network Segmentation
- Graph Densification
- Diameter Shrinkage

Network Growth Patterns: Network Segmentation

In evolving networks, segmentation takes place, where the large network is decomposed over time into three parts:

- **Giant Component**: As network connections stabilize, a giant component of nodes is formed, with a large proportion of network nodes and edges falling into this component.
- **Small Component**: These are isolated parts of the network that form star, cycle, tree structures.
- **Singletons**: These are isolated nodes disconnected from all nodes in the network.



Network Growth Patterns: Graph Densification

- The density of the graph increases as the network grows
  - The number of edges increases faster than the number of vertices does

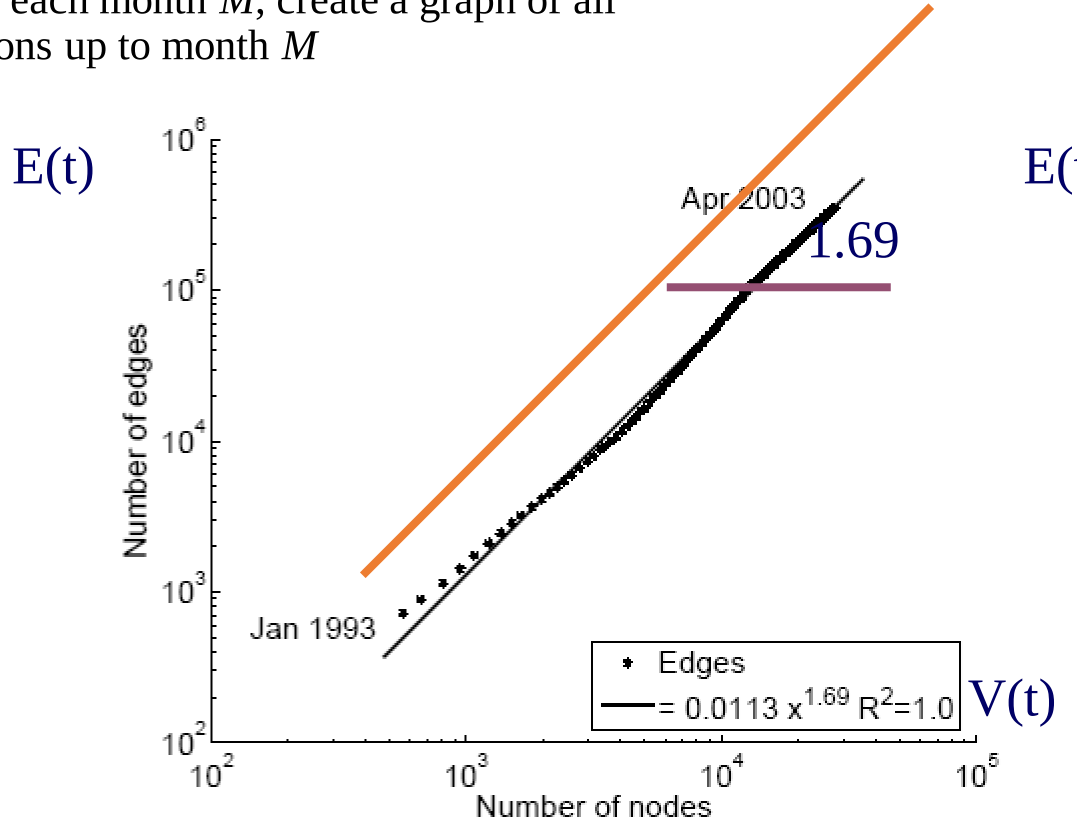
$$E(t) \propto V(t)^{\alpha}$$

- Densification exponent $1 \leq \alpha \leq 2$
  - $\alpha = 1$: linear growth – constant out-degree
  - $\alpha = 2$: quadratic growth – clique

E(t) and *V(t)* are numbers of edges and vertices respectively at time *t*

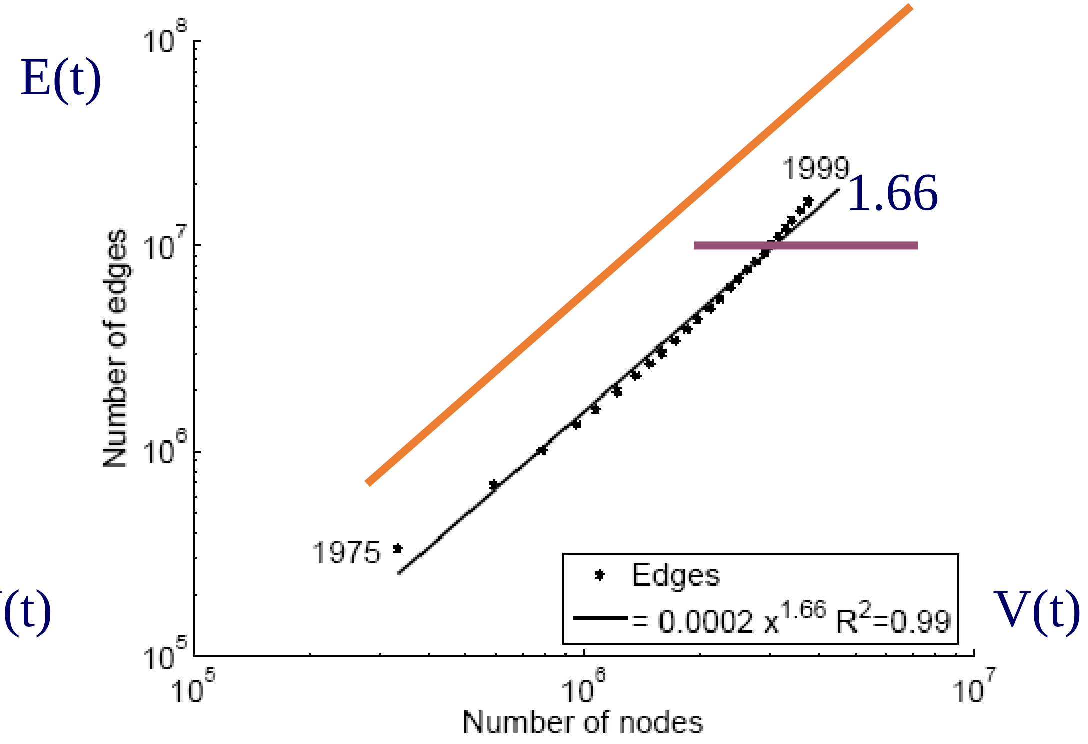
Network Growth Patterns: Graph Densification

- Citations among physics papers
- 1992: 1,293 papers, 2,717 citations
- 2003: 29,555 papers, 352,807 citations
- For each month *M*, create a graph of all citations up to month *M*



Physics Citations

- Citations among patents granted
- 1975: 334,000 nodes, 676,000 edges
- 1999: 2.9 million nodes, 16.5 million edges
- Each year is a data point

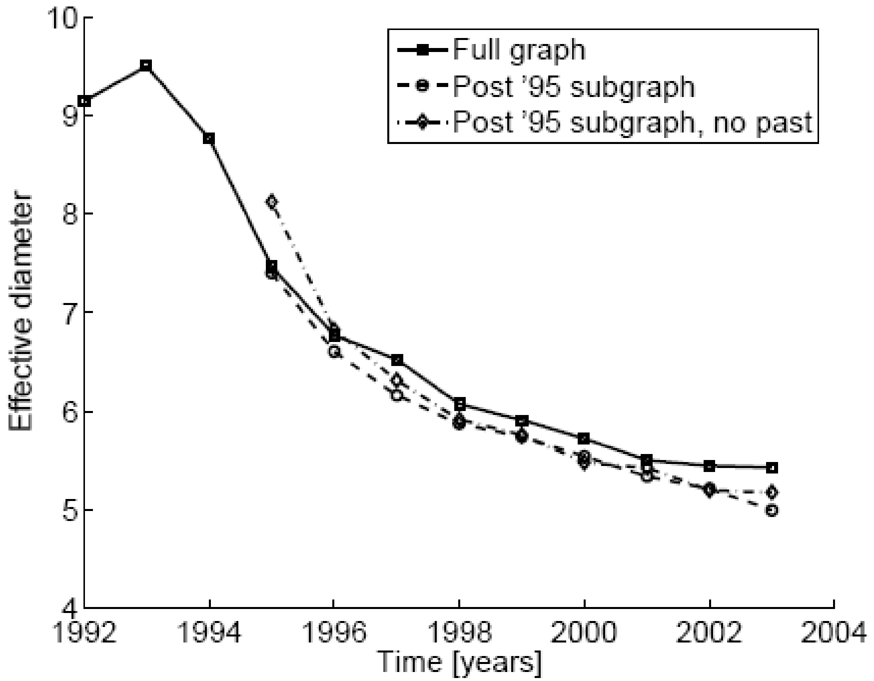


Patent Citations

Network Growth Patterns: Diameter Shrinking

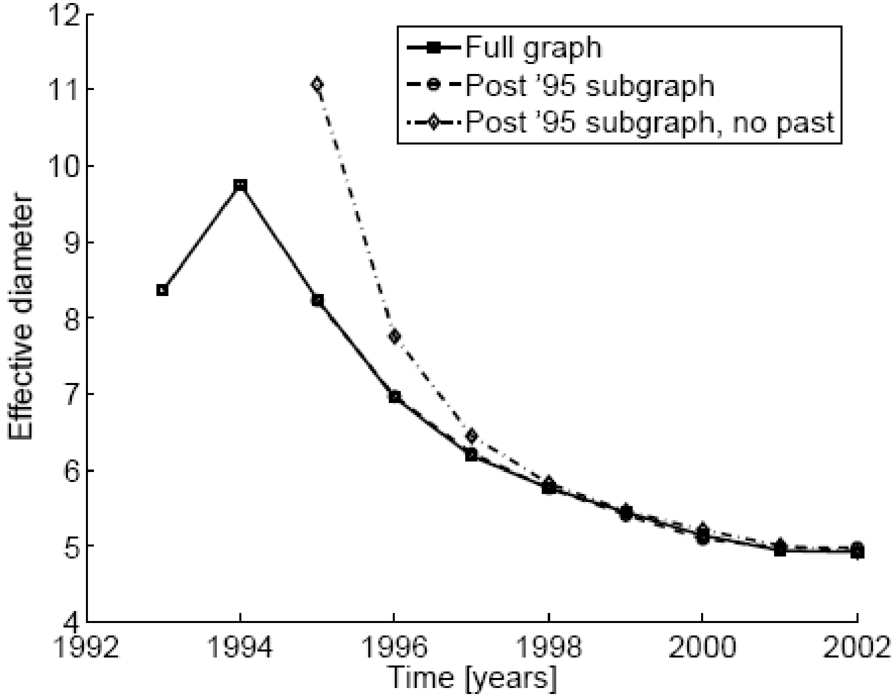
In networks diameter shrinks over time:

- Citations among physics papers
- 1992 –2003
- One graph per year



ArXiv citation graph

- Graph of collaborations in physics
  - authors linked to papers
- 10 years of data



Affiliation Network

Evolution of the Diameter

- Prior work on Power Law graphs hints at Slowly growing diameter:
  - diameter ~ O(log N)
  - diameter ~ O(log log N)
- However, Diameters shrinks over time
  - As the network grows the distances between nodes slowly decrease
- There are several factors that could influence the *Shrinking diameter*
  - Effective Diameter:
    - Distance at which 90% of pairs of nodes is reachable
  - Problem of "Missing past"
    - How do we handle the citations outside the dataset?
  - Disconnected components
  - ….

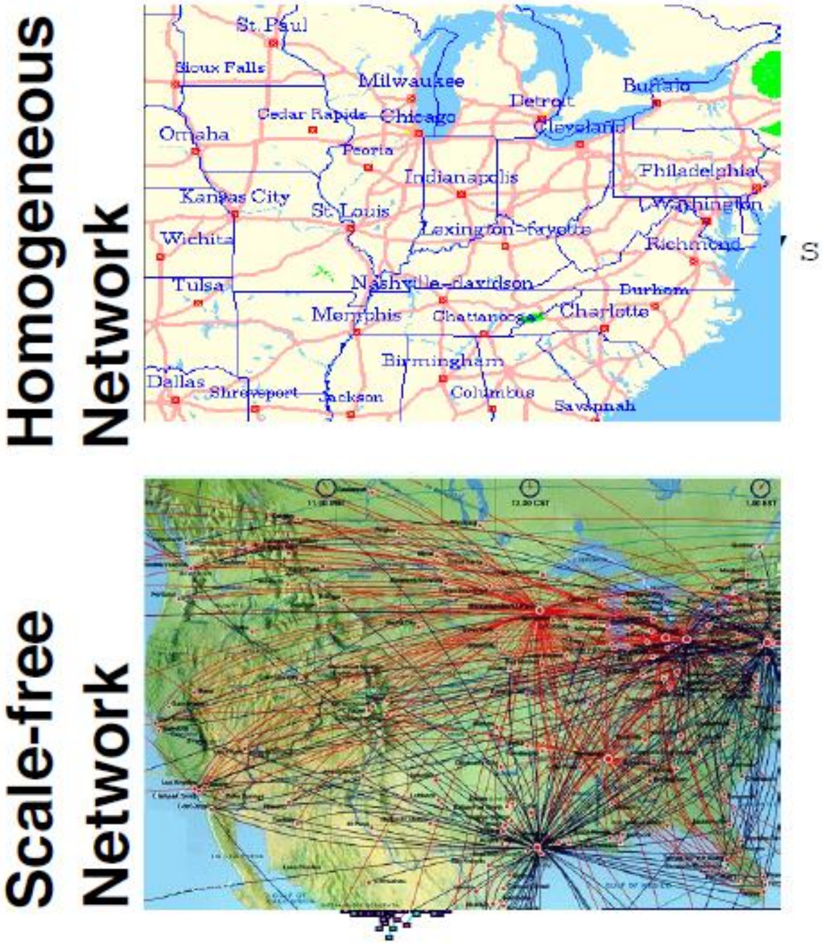
Why is all this important?

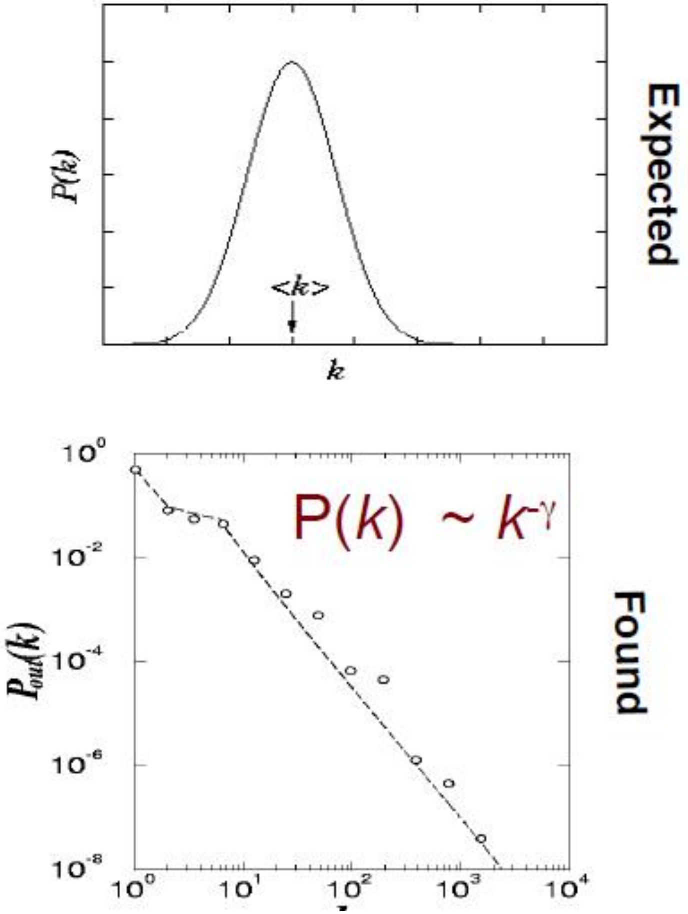
- Gives insight into the graph formation process:
  - Anomaly detection – abnormal behavior, evolution
  - Predictions – predicting future from the past
  - Simulations of new algorithms
  - Graph sampling – many real world graphs are too large to deal with

Scale-Free Networks

Scale-Free Networks

What does the difference mean?

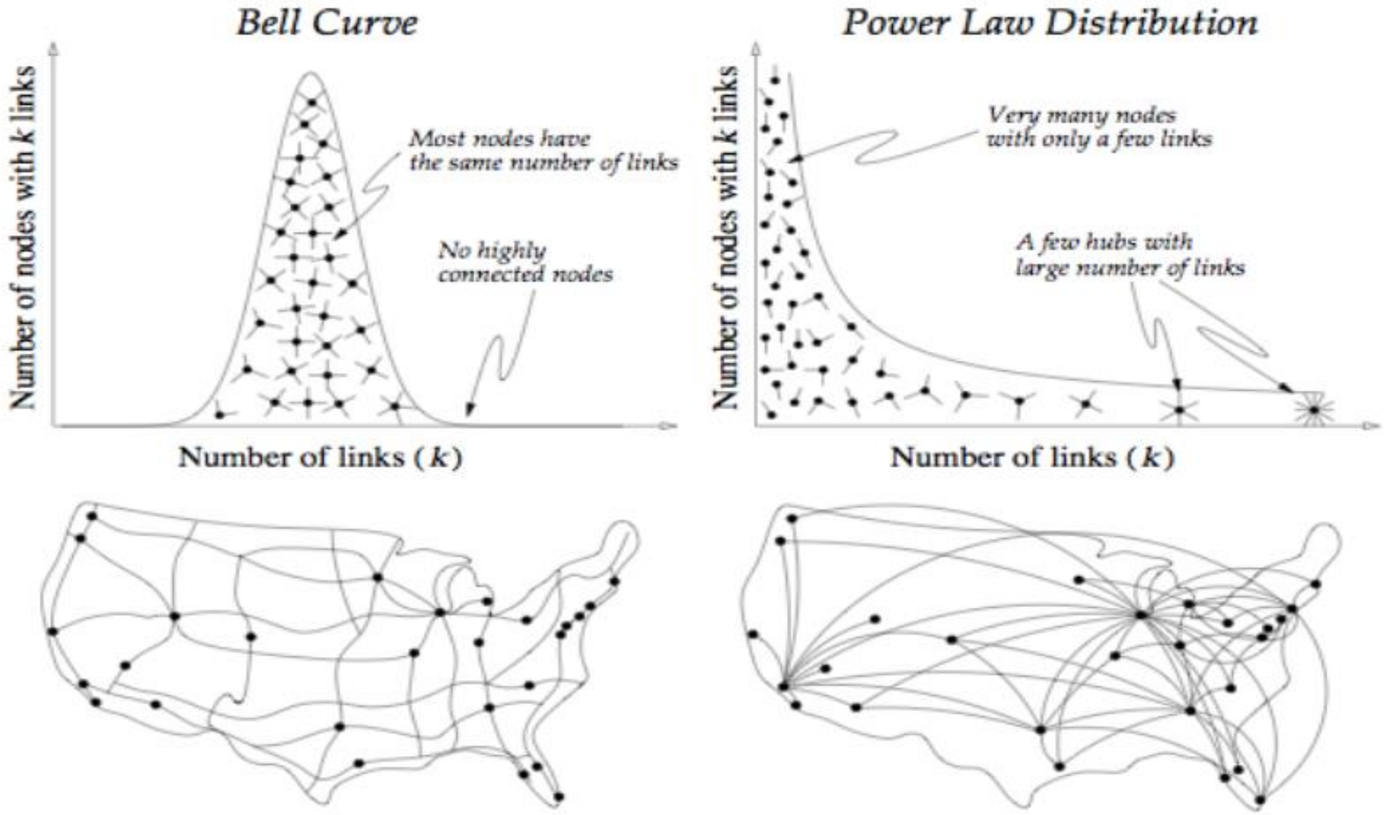




R. Albert, H. Jeong, A-L Barabasi, *Nature*, 401 130 (1999).

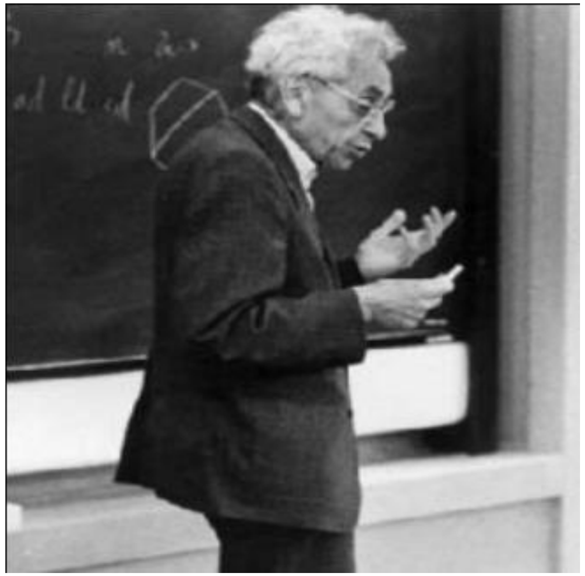
Scale-Free Networks

HOMOGENEOUS AND POWER-LAW DEGREE DISTRIBUTIONS



Scale-Free Networks

Another Example: ERDOS NUMBER



Paul Erdős (1913-1996)

Erdős has better centrality in his network than Bacon has in his.

Number of links required to connect scholars to Erdős, via co-authorship of papers!

Erdős wrote 1500+ papers with 507 coauthors.

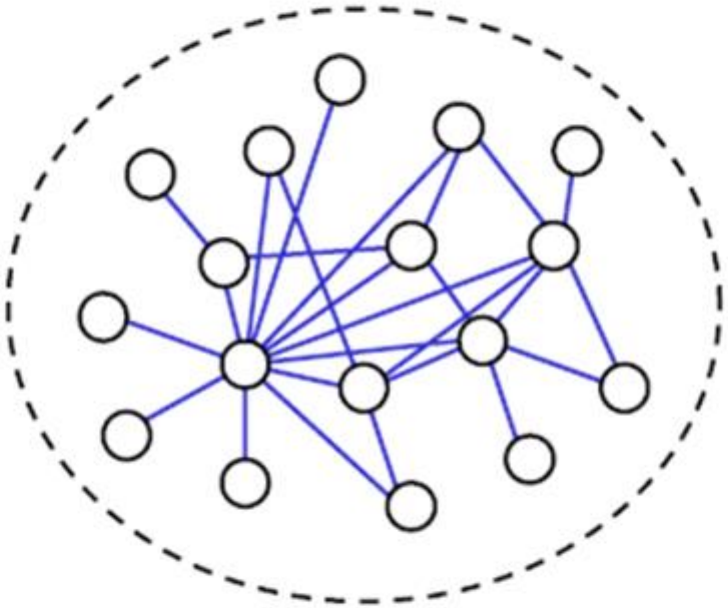
Jerry Grossman's (Oakland Univ.) website allows mathematicians to compute their Erdős numbers:

http://www.oakland.edu/enp/

Connecting path lengths, among mathematicians only:

– average is ***4.65***
– maximum is ***13***

Scale-Free Networks



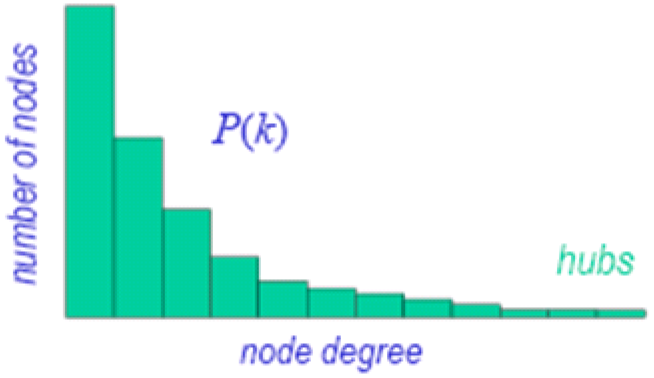
A schematic scale-free network

- in a *scale-free network* the degree distribution follows a POWER-LAW:

$$P(k) \sim k^{-\gamma}$$

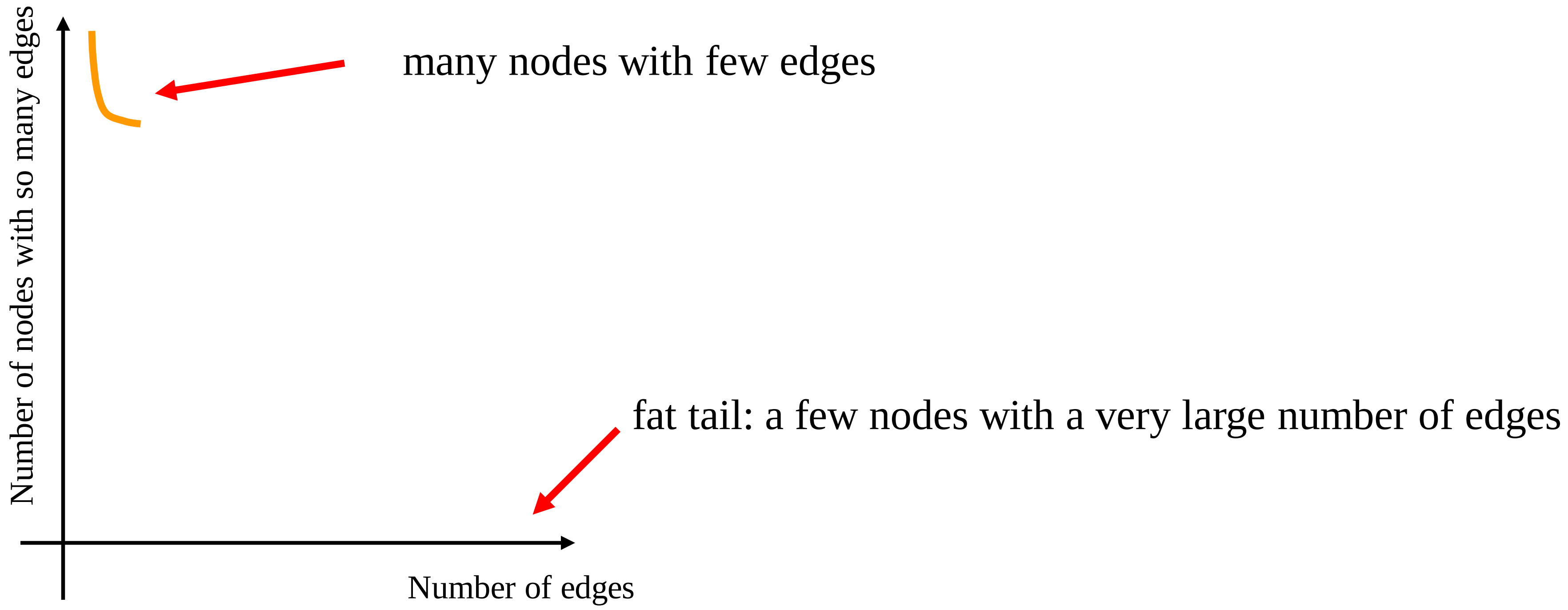
- there exists a small number of highly connected nodes, called *hubs* (tail of the distribution)

- the great majority of nodes have few connections (head of the distribution)



Scale-Free Networks

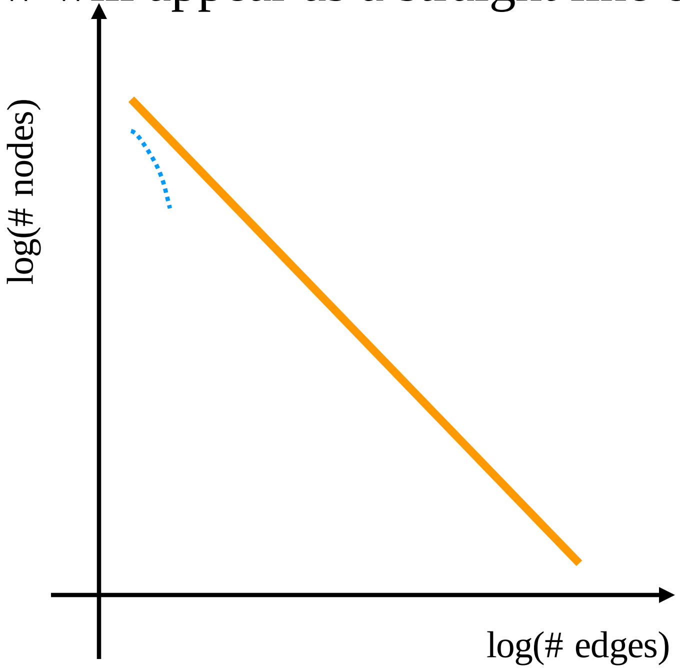
- Many real world networks contain hubs: highly connected nodes
- Usually the distribution of edges is extremely skewed



→ No “typical” number of edges

Is it really a power-law?

- A power-law will appear as a straight line on a log-log plot:

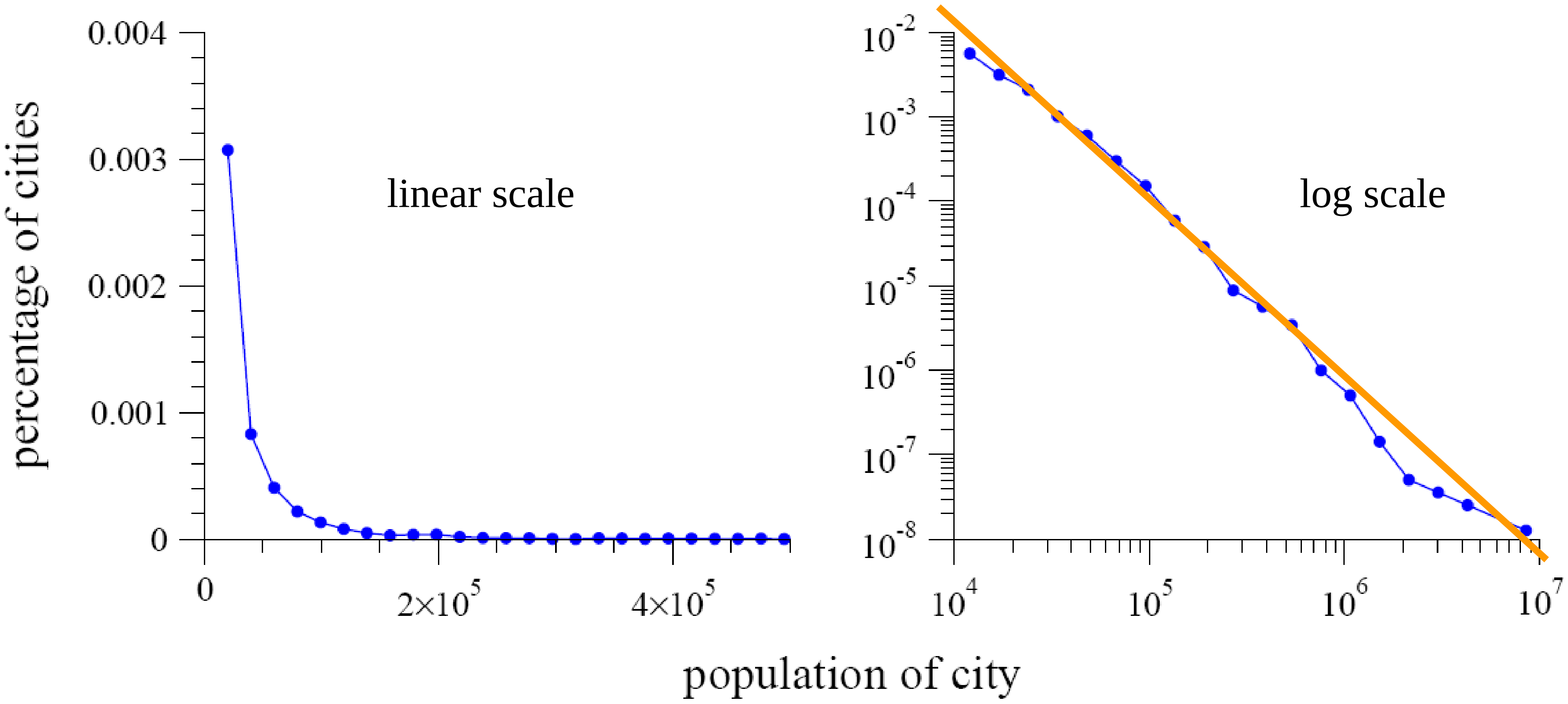


- A deviation from a straight line could indicate a different distribution:
  - exponential
  - lognormal

What is a heavy tailed-distribution?

- Right skew
  - normal distribution (not heavy tailed)
    - e.g. heights of human males: centered around 180cm (5’11’’)
  - Zipf’s or power-law distribution (heavy tailed)
    - e.g. city population sizes: NYC 8 million, but many, many small towns
- High ratio of max to min
  - human heights
    - tallest man: 272cm (8’11”), shortest man: (1’10”) *ratio: 4.8*
      from the Guinness Book of world records
  - city sizes
    - NYC: pop. 8 million, Duffield, Virginia pop. 52, *ratio: 150,000*

Review: Power-Law Distribution



- high skew (asymmetry)
- straight line on a log-log plot

Review: Power-Law Distribution

Pareto Distribution (1896)

If X is a random variable with a Pareto distribution, then the probability that X is greater than some number x is given by

$$\Pr[X \geq x] = \left(\frac{x^{-\alpha}}{k}\right)$$

where α and k are positive parameters.

Vilfredo Pareto. *Cours d'économie politique professé a l'université de Lausanne. Vol. I' 1896; Vol. II, 1897.*

Review: Power-Law Distribution

Zipf's Law

Zipf's law is an empirical law formulated using mathematical statistics, refers to the fact that many types of data studied in the physical and social sciences can be approximated with a Zipfian distribution. It states that, in a corpus of natural language utterances, the frequency of any word is roughly inversely proportional to its rank in the frequency table. So, the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc.

$$p(r) \sim r^{-\alpha} \; ; \quad \alpha \sim 1$$

K.G. Zipf, *The Psycho-Biology of Language*, Cambridge (Mass), 1935; *Human Behavior and the Principle of least Effort*, 1949.

Review: Power-Law Distribution

Zipf and Pareto

- The phrase "The ***r***-th largest city has ***n*** inhabitants" is equivalent to saying "***r*** cities have ***n*** or more inhabitants".

- This is exactly the definition of the Pareto distribution, except the x and y axes are flipped.
  - for Zipf, ***r*** is on the x-axis and ***n*** is on the y-axis,
  - for Pareto, ***r*** is on the y-axis and ***n*** is on the x-axis.

- Simply inverting the axes,
  - if the rank exponent is ***b***, i.e.
  - $n \sim r^{-b}$ for Zipf, (n = income, r = rank of person with income n)
  - then the Pareto exponent is ***1/b*** so that
  - $r \sim n^{-1/b}$ (n = income, r = number of people whose income is n or higher)

Review: Power-Law Distribution

Zipf's Law and city sizes (~1930)

Rank(k)	City	Population (1990)	Zips's Law $10{,}000{,}000/k$	Modified Zipf's law: (Mandelbrot) $5{,}000{,}000\Big/\left(k-\frac{2}{5}\right)^{3/4}$
1	Now York	7,322,564	10,000,000	7,334,265
7	Detroit	1,027,974	1,428,571	1,214,261
13	Baltimore	736,014	769,231	747,693
19	Washington DC	606,900	526,316	558,258
25	New Orleans	496,938	400,000	452,656
31	Kansas City	434,829	322,581	384,308
37	Virgina Beach	393,089	270,270	336,015
49	Toledo	332,943	204,082	271,639
61	Arlington	261,721	163,932	230,205
73	Baton Rouge	219,531	136,986	201,033
85	Hialeah	188,008	117,647	179,243
97	Bakersfield	174,820	103,270	162,270

source: Luciano Pietronero

Review: Power-Law Distribution

- Straight line on a log-log plot

$$\ln\big(p(x)\big) = C - \alpha \ln(x)$$

- Exponentiate both sides to get that *p(x)*, the probability of observing an item of size "*x*" is given by

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

Normalization constant (probabilities over all *x* must sum to 1)

power law exponent α

Review: Power-Law Distribution

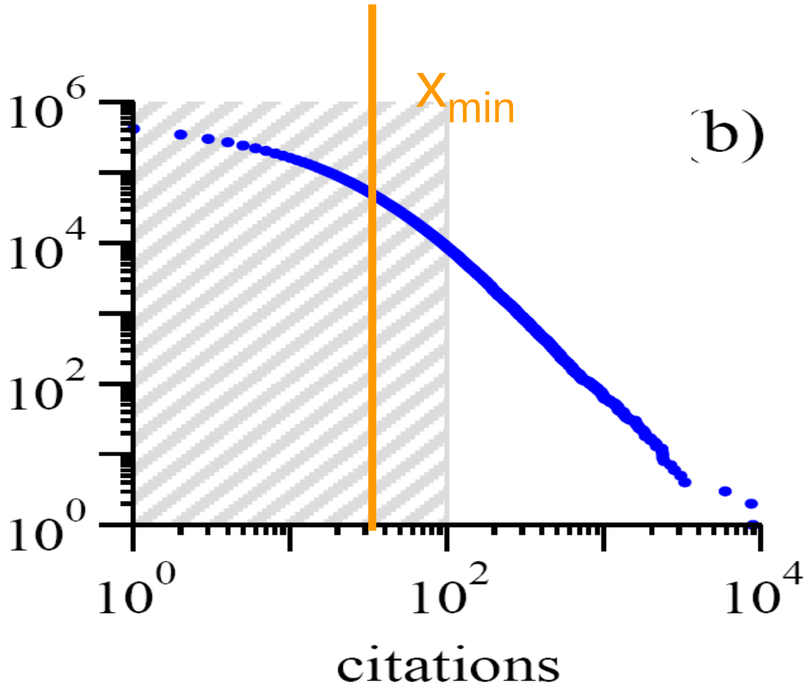
Where to start fitting?

- some data exhibit a power law only in the tail
- after binning or taking the cumulative distribution you can fit to the tail
- so need to select an x_{min} the value of x where you think the power-law starts
- certainly x_{min} needs to be greater than 0, because $x^{-\alpha}$ is infinite at $x = 0$

Review: Power-Law Distribution

Example:

- Distribution of citations to papers
- power law is evident only in the tail
  - $x_{min} > 100$ citations



Source: MEJ Newman, 'Power laws, Pareto distributions and Zipf's law

Review: Power-Law Distribution

Maximum likelihood fitting

You have to be sure you have a power-law distribution

*** This will just give you an exponent but not a goodness of fit*

$$\alpha = 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}\right]^{-1}$$

where x_i are your data points, there are n of them

"Finite Sample Corrections for Parameters Estimation and Significance Testing", B.K. Teh et al., Frontiers in Applied Mathematics and Statistics, 2018. (Front. Appl. Math. Stat. | doi: 10.3389/fams.2018.00002)

Review: Power-Law Distribution

Some exponents for real world data

	x_{min}	exponent α
frequency of use of words	1	2.20
number of citations to papers	100	3.04
number of hits on web sites	1	2.40
copies of books sold in the US	2 000 000	3.51
telephone calls received	10	2.22
magnitude of earthquakes	3.8	3.04
diameter of moon craters	0.01	3.14
intensity of solar flares	200	1.83
intensity of wars	3	1.80
net worth of Americans	\$600m	2.09
frequency of family names	10 000	1.94
population of US cities	40 000	2.30

Review: Power-Law Distribution

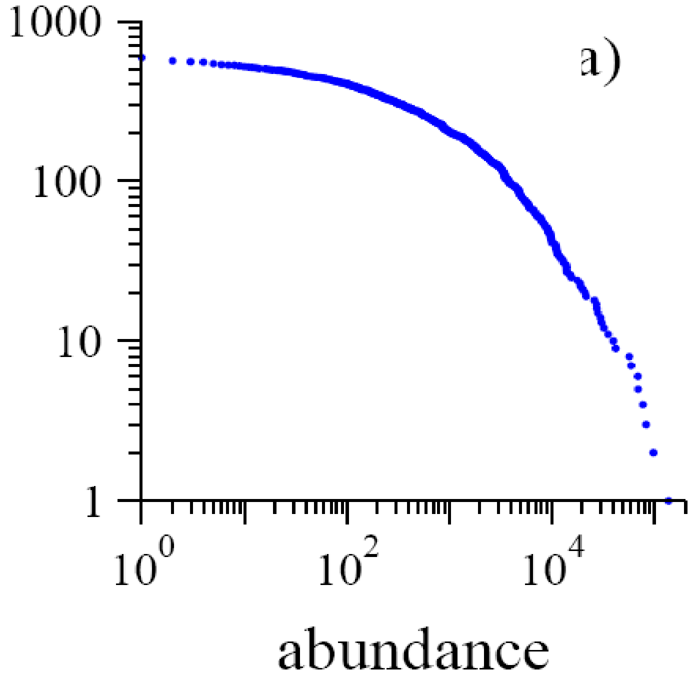
Many real world networks are power law

	exponent α (in/out degree)
film actors	2.3
telephone call graph	2.1
email networks	1.5/2.0
sexual contacts	3.2
WWW	2.3/2.7
AS-level Internet	2.5
Router-level Internet	2.0
Link-level Internet	2.7
peer-to-peer	2.1
metabolic network	2.2
protein interactions	2.4

Review: Power-Law Distribution

Not everything is a power law!

Example: Number of sightings of 591 bird species in the North American Bird survey in 2003.



Another example: size of wildfires (in acres)

Source: MEJ Newman, 'Power laws, Pareto distributions and Zipf's law

Review: Power-Law Distribution

Not every network is power law distributed !

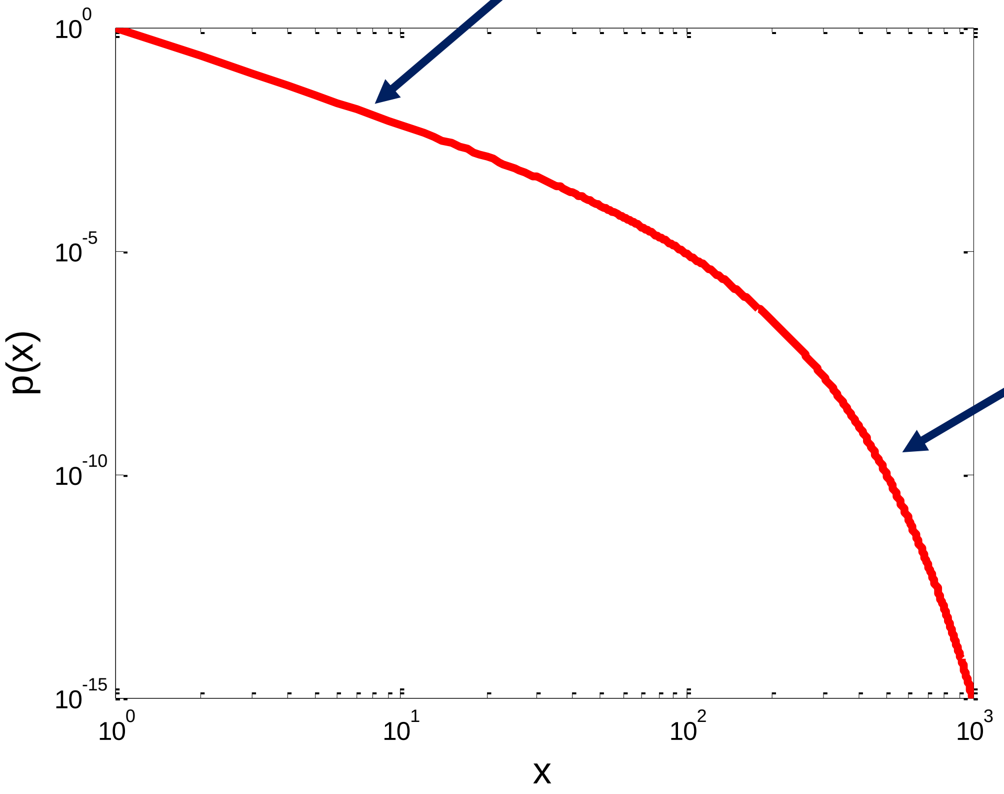
- reciprocal, frequent email communication
- power grid
- Roget’s thesaurus
- company directors…

Another common distribution: Power-law with an exponential cutoff

$$p(x) \sim x^{-\alpha} e^{-x/k_0}$$

starts out as a power law

ends up as an exponential



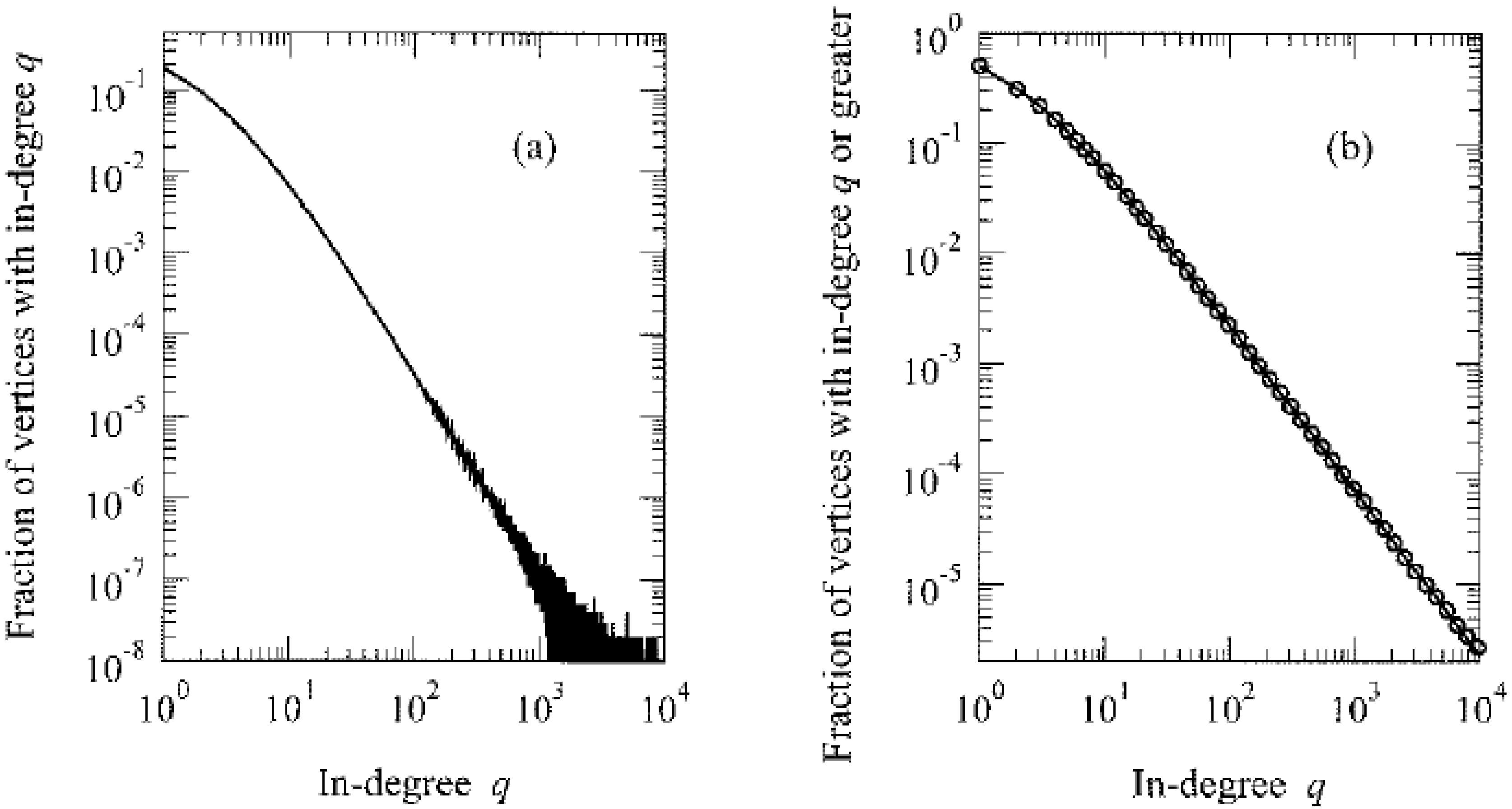
but could also be a lognormal or double exponential…

Preferential Attachment in Networks

- First considered by [Price, 1965] as a model for citation networks
  - each new paper is generated with *m* citations (mean)
  - new papers cite previous papers with probability proportional to their indegree (citations)
  - what about papers without any citations?
    - each paper is considered to have a “default” citation, here denoted by *a*
    - probability of citing a paper with degree *k*, proportional to *k*+*a*

- Power law with exponent α = 2+*a*/*m*, *i.e.* $P(k) \sim (k+a)^{-\alpha}$ when k is large, or simply, $P(k) \sim k^{-\alpha}$

Preferential Attachment in Networks



Degree distribution in Price's model of a growing network. (a) A histogram of the in-degree distribution for a computer-generated network with $m = 3$ and $a = 1.5$ which was grown until it had $n = 10^8$ vertices. (b) The cumulative distribution function for the same network.

The Barabasi-Albert model

- Undirected model

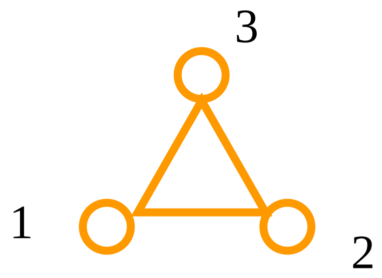
- Each node connects to other nodes with probability proportional to their degree
  - the process starts with some initial subgraph
  - each node comes with m edges

- Results in power-law with exponent $\alpha = 3$

The Barabasi-Albert model

Generating BA Graphs:

- Very simple algorithm to implement
  - start with an initial set of n_0 fully connected nodes
    - e.g. $n_0 = 3$



1 1 2 2 2 3 3 4 5 6 6 7 8 ….

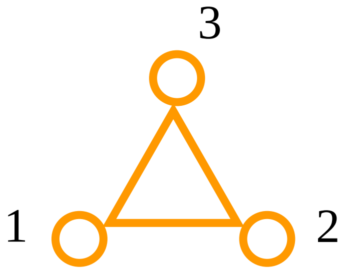
  - now add new vertices one by one, each one with exactly m edges
  - each new edge connects to an existing vertex in proportion to the number of edges that vertex already has → ***preferential attachment***
  - easiest if you keep track of edge endpoints in one large array and select an element from this array at random
    - the probability of selecting any one vertex will be proportional to the number of times it appears in the array – which corresponds to its degree

The Barabasi-Albert model

Generating BA Graphs:

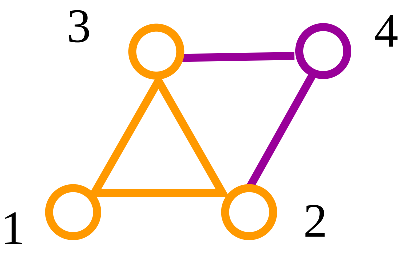
- To start, each vertex has an equal number of edges (2)
  - the probability of choosing any vertex is 1/3

1 1 2 2 3 3

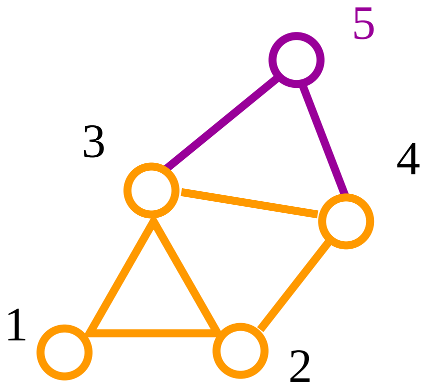


- We add a new vertex, and it will have *m* edges, here take *m*=2
  - draw 2 random elements from the array – suppose they are 2 and 3

1 1 2 2 2 3 3 3 4 4



- Now the probabilities of selecting 1,2,3,or 4 are 1/5, 3/10, 3/10, 1/5



- Add a new vertex, draw a vertex for it to connect from the array
  - etc.

1 1 2 2 2 3 3 3 3 4 4 4 5 5

The Barabasi-Albert model

Properties of the BA graph:

- The distribution is scale free with exponent $\alpha = 3$ $P(k) = 2\, m^2/k^3$

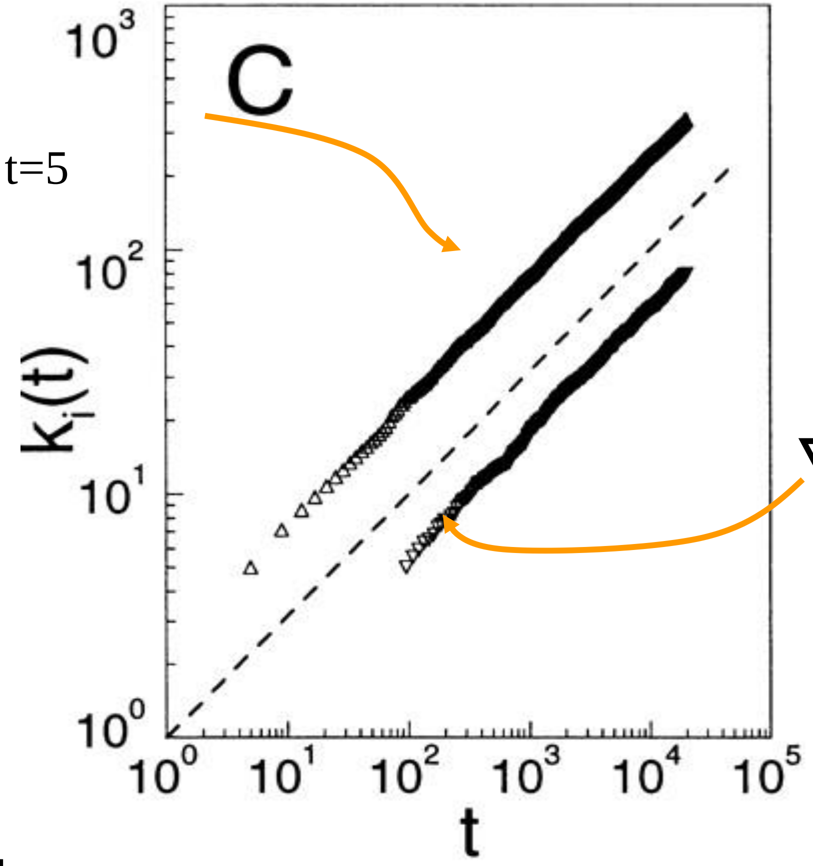
- The graph is connected
  - Every vertex is born with a link ($m = 1$) or several links ($m > 1$)
  - It connects to older vertices, which are part of the giant component

- The older the richer
  - Nodes accumulate links as time goes on
  - preferential attachment will prefer wealthier nodes, who tend to be older and had a head start

The Barabasi-Albert model

Time evolution of the connectivity of a vertex in the BA model:

△vertex introduced at time t=5



▽ vertex introduced at time t=95

- Younger vertex does not stand a chance:
- at t=95 older vertex has ~ 20 edges, and younger vertex is starting out with 5
- at t ~ 10,000 older vertex has 200 edges and younger vertex has 50

Source: Barabasi and Albert, 'Emergence of scaling in random networks’, Science 1999.

The Barabasi-Albert model

Properties of the BA graph:

- ***Degree Distribution:*** $$P(k) = \frac{2m^2}{k^3}$$

- ***Clustering Coefficient:*** $$C \sim \frac{(\ln N)^2}{N}$$

- ***Average Path Length:*** $$l \sim \frac{\ln N}{\ln(\ln N)}$$

The Barabasi-Albert model

Derivation of Degree Distribution:

The rate of change of the degree of node *i*,

$$\frac{dk_i(t)}{dt} = m\frac{k_i(t)}{\sum_j k_j(t)}$$

where *m* is the degree of each new added node. Assume one node is added at each time step, therefore, the summation in the denominator is given by $\sum_j k_j(t) = 2mt + k_0 n_0$, where n_0 and k_0 are the initial number of nodes and their degree respectively. The differential equation has the solution with the boundary condition $k_i(t_i) = m$. At large *t* and t_i >> *m*,

$$k_i(t) \approx m\left(\frac{t}{t_i}\right)^{1/2}$$

The Barabasi-Albert model

Derivation of Degree Distribution:

Taking the continuum limit, the degree distribution is obtained as

$$P(k,t) = \frac{1}{t+n_0}\int_0^t \delta[k-k_i(t)]dt_i = -\frac{1}{t+n_0}\left(\frac{dk_i(t)}{dt_i}\right)^{-1} |_{t_i=t_i(k,t)}$$

where $\delta[k-k_i(t)]$ is the Dirac delta function and $t_i(k,t)$ is the solution of the implicit equation $k = k_i(t)$. Substitute the solution of $k_i(t)$ into the above expression, one obtains

$$P(k,t) = 2m^2\left(\frac{t+n_0k_0/2m}{t+n_0}\right)\frac{1}{k^3} ,$$

In the limit $t \to \infty$,

$$P(k,t) \approx \frac{2m^2}{k^3}.$$

The Barabasi-Albert model

Derivation of Clustering Coefficient:

To calculate the clustering coefficient, one would calculate the number of triangles expected in the model. Denote the probability to have a link between nodes i and j by $P(i,j)$. Therefore, the probability that three nodes i, j, l form a triangle is $P(i,j)P(i,l)P(j,l)$. The expected number of triangles in which node l with degree k_l participates is given by the sum of the probabilities that node l participates in triangles with arbitrary chosen nodes i and j in the network. Using a continuous degree approximation, we have

$$N\Delta_l = \int_{i=1}^{N} di \int_{j=1}^{N} dj\, P(i,j)P(i,l)P(j,l)$$

To proceed, one needs to calculate $P(i,j)$. Denote the time when node j arrives with $t_j = j$, and the probability that node j links to node i with degree k_i is given by preferential attachment

$$P(i,j) = m\frac{k_i(j)}{\sum_{i=1}^{j} k_i(j)} = m\frac{k_i(j)}{2mj}$$

Assume here the initial number of nodes to be zero, or at large time t.

The Barabasi-Albert model

Derivation of Clustering Coefficient:

Recall that at large time *t*

$$k_i(t) \approx m\left(\frac{t}{t_i}\right)^{\frac{1}{2}} = m\left(\frac{j}{i}\right)^{\frac{1}{2}} \longrightarrow P(i,j) = \frac{m}{2}(ij)^{-\frac{1}{2}}$$

Substituting into the above equation,

$$N\Delta_l = \int_{i=1}^{N} di \int_{j=1}^{N} dj\, P(i,j)P(i,l)P(j,l) = \frac{m^3}{8}\int_{i=1}^{N} di \int_{j=1}^{N} dj\, (ij)^{-\frac{1}{2}}(il)^{-\frac{1}{2}}(jl)^{-\frac{1}{2}} = \frac{m^3}{8l}\int_{i=1}^{N}\frac{di}{i}\int_{j=1}^{N}\frac{dj}{j}$$

$$= \frac{m^3}{8l}(\ln N)^2$$

The clustering coefficient is given by

$$C_l = \frac{2N\Delta_l}{k_l(k_l-1)} = \frac{\frac{m^3}{4l}(\ln N)^2}{k_l(N)(k_l(N)-1)}$$

The Barabasi-Albert model

Derivation of Clustering Coefficient:

Since

$$k_l(N) = m\left(\frac{N}{l}\right)^{\frac{1}{2}}$$

which is the degree of node *l* at time $t = N$. Therefore, for large k_l, we have

$$k_l(N)(k_l(N) - 1) \approx k_l^2\,(N) = m^2\frac{N}{l}$$

which gives the clustering coefficient of the Barabasi-Albert model

$$C_l = \frac{m}{4}\frac{(\ln N)^2}{N}$$

independent of *l*.

What implications does this have?

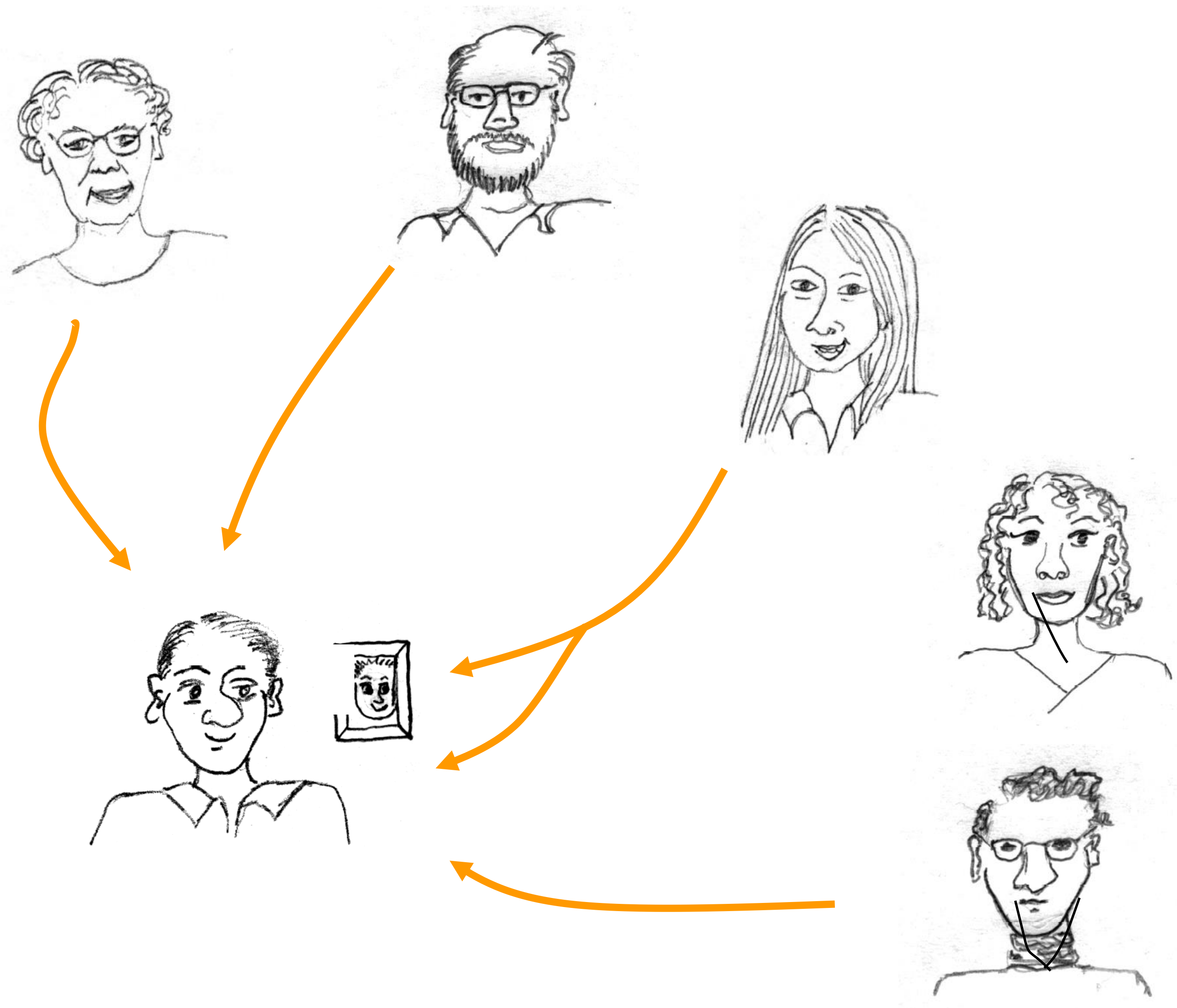
- Robustness
- Search
- Spread of disease
- Opinion formation
- Spread of computer viruses
- Gossip

The role of hubs in epidemics

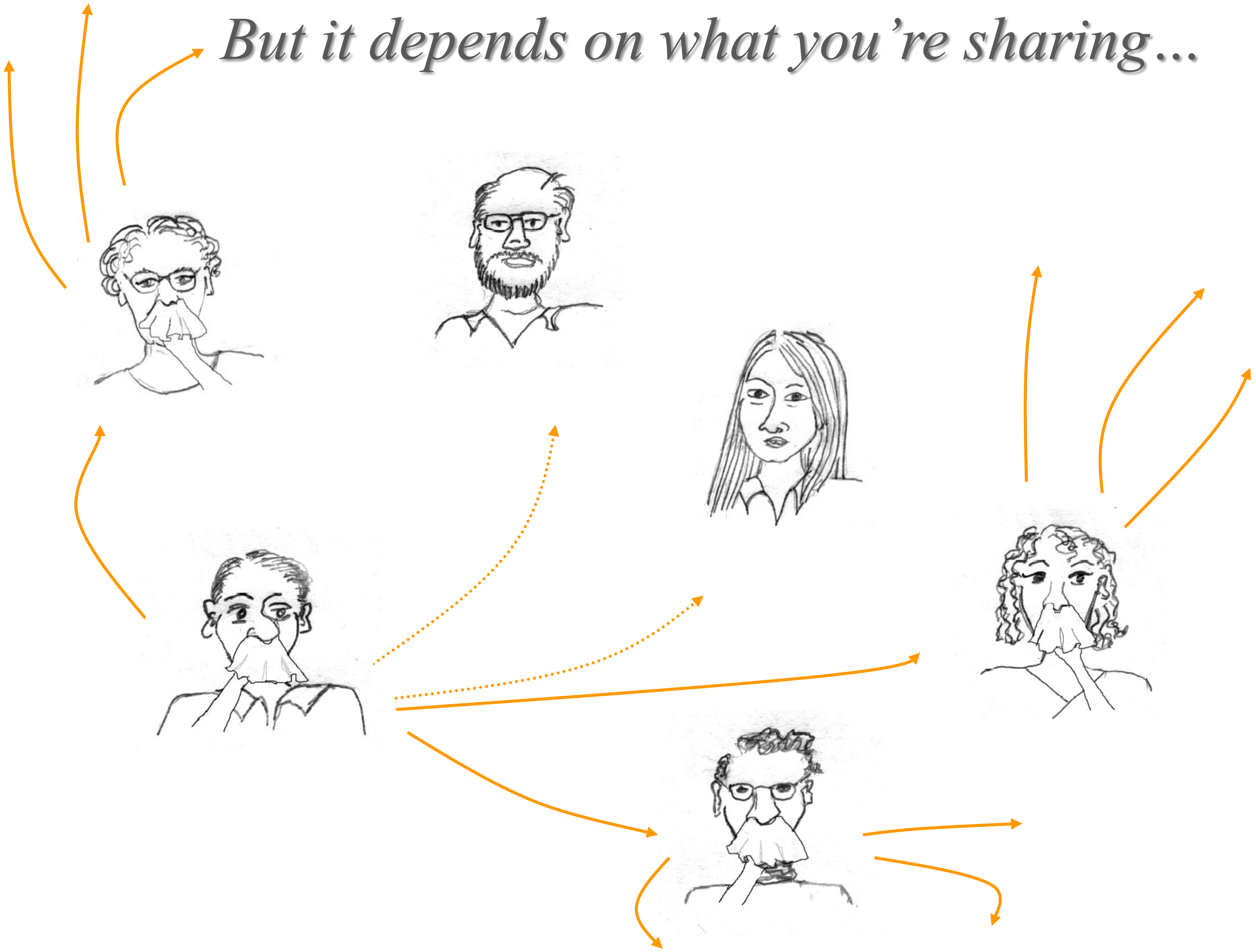
- In a power-law network, a virus can persist no matter how low its infectiousness

- Many real world networks do exhibit power-laws:
  - WWW
  - Transportation networks
  - email networks

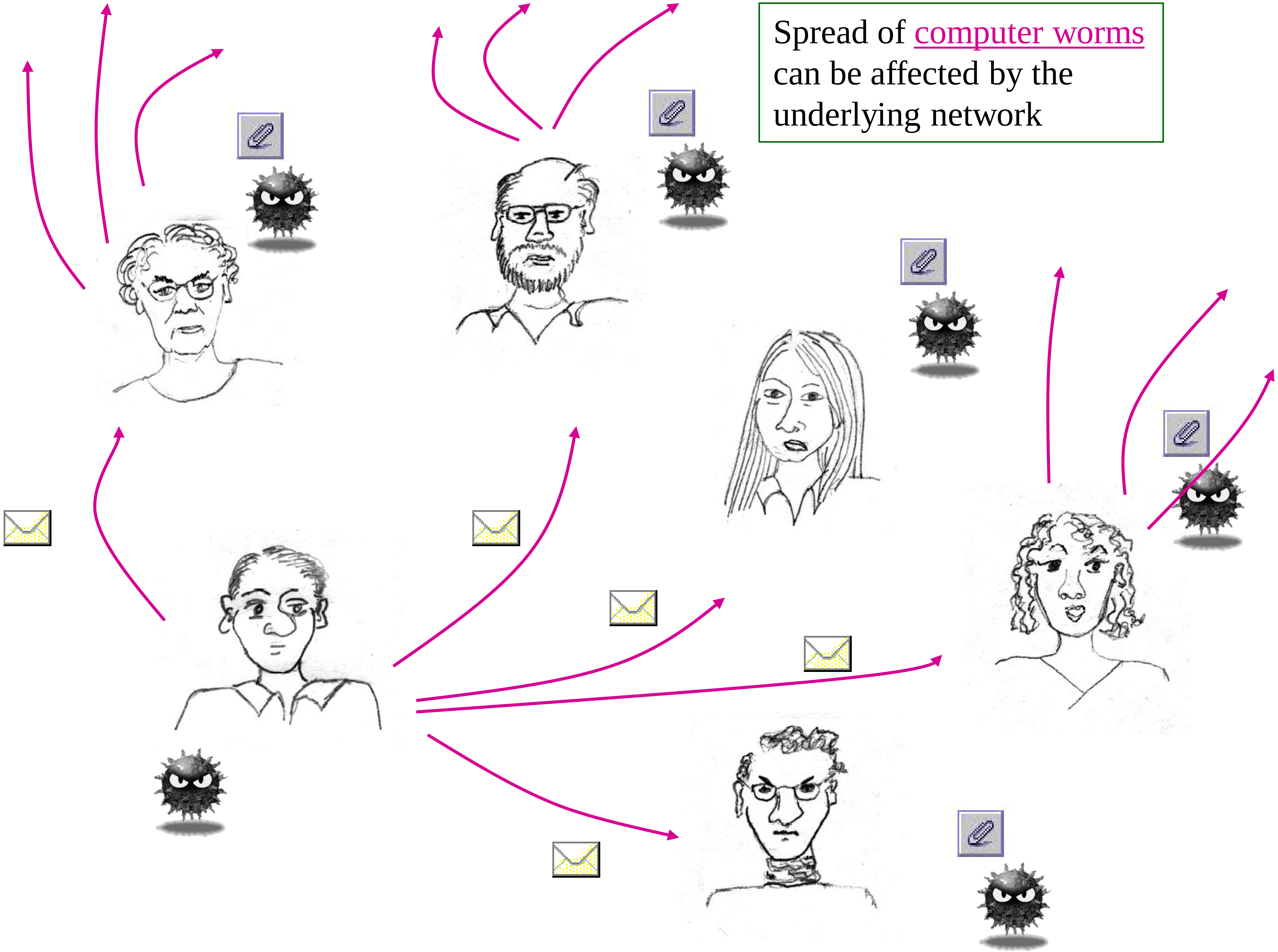
In social networks, it's nice to be a hub



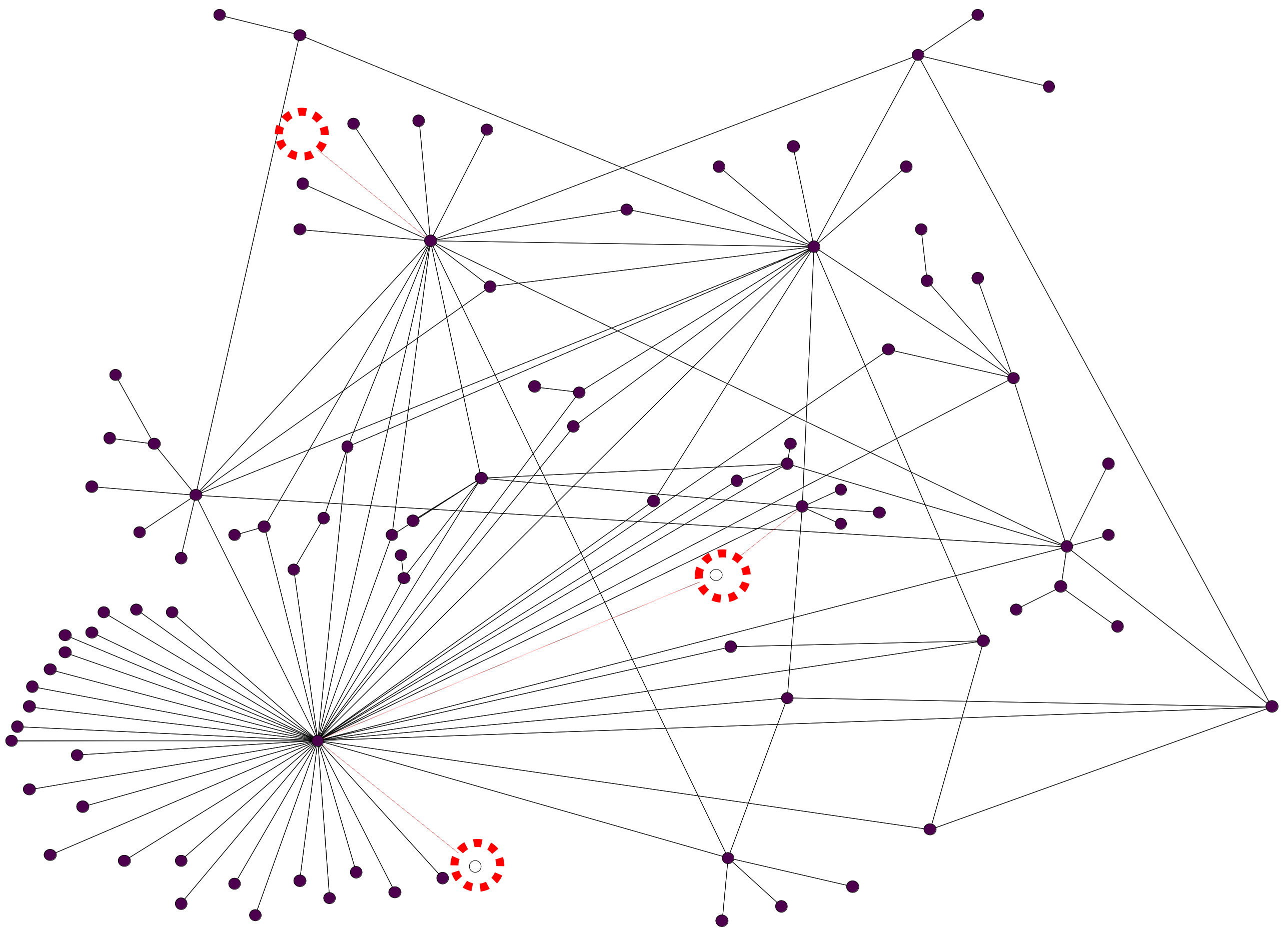
But it depends on what you're sharing…



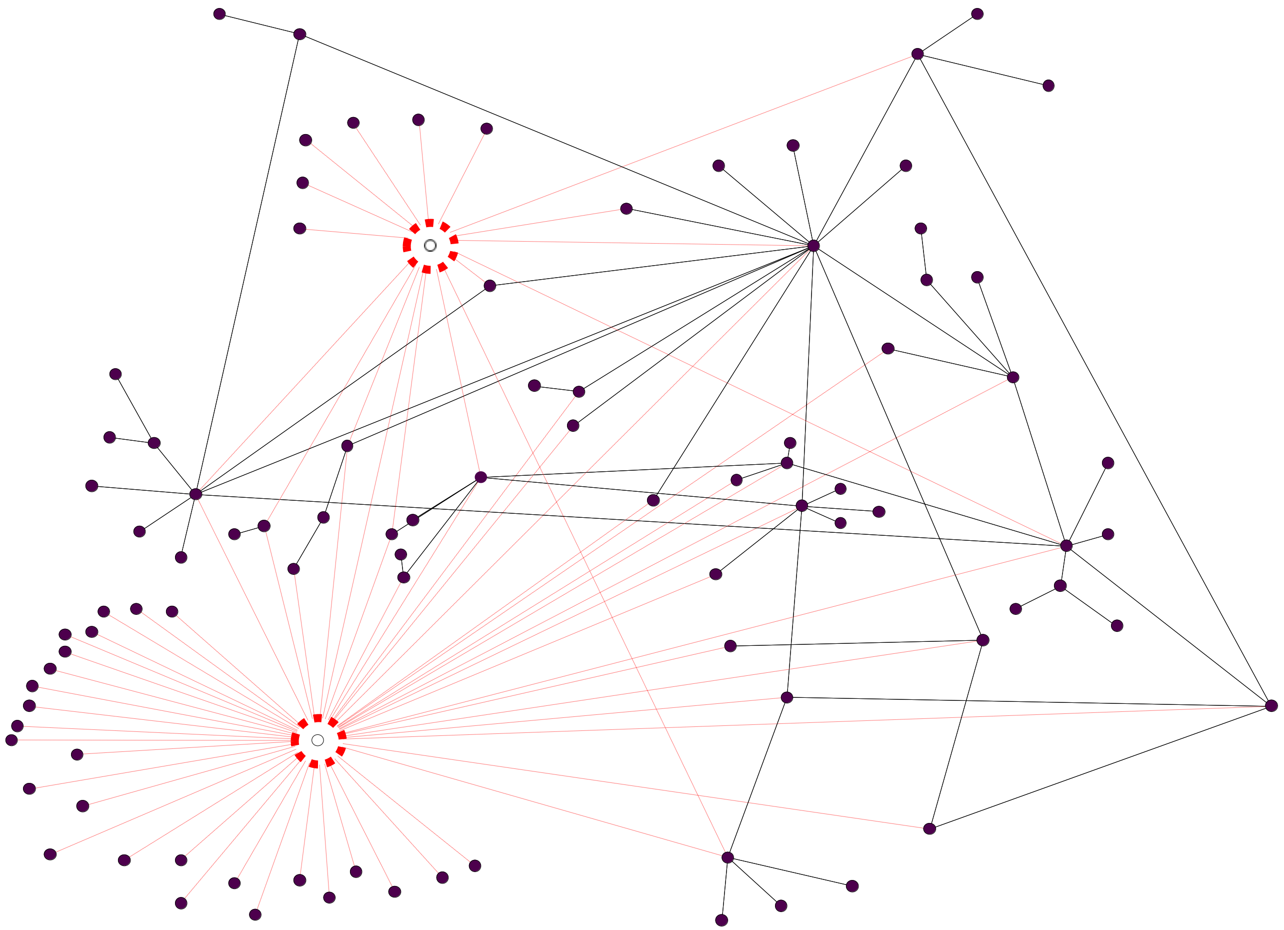
Spread of computer worms can be affected by the underlying network



Power-law networks are robust to random breakdown



But are vulnerable to targeted attack



Targeting and removing hubs can quickly break up the network

Summary

- power law distributions appear everywhere
- there are good and bad ways of fitting them
- some distributions are not power-law
- preferential attachment leads to power law networks…
- … but it's not the whole story, and not the only way of generating them

Questions to think about:

- BA networks are *not* clustered
  - *Can you think of a growth model of having preferential attachment and clustering at the same time?*

- What would the network look like if nodes are added over time, but not attached preferentially?

- Generalization: addition or removal of extra links at each time step; nonlinear preferential attachment; vertices of varying quality or attractiveness …

- What other processes might give rise to power law networks?

Network Models: Summary

Erdos-Renyi model

- short path lengths
- Poisson distribution (no hubs)
- no clustering

Barabasi-Albert scale-free model

- short path lengths
- power-law distribution for degrees
- robustness
- no clustering (may be fixed)

Watts-Strogatz Small World model

- short path lengths
- high clustering (N independent)
- almost constant degrees

Real-world networks

- short path lengths
- high clustering
- broad degree distributions, often power laws