Freie Universität Berlin

Class 4 -
Clustering Coefficient and Network Models

Course: Computational Network Analysis

Prof. Dr. Claudia Müller-Birn

Institute of Computer Science, «Human-Centered Computing»

Feb 25, 2015

Recap

- You were introduced to many measures such as degree, mean degree, degree distribution, density, path, distance, eccentricity, components and some more.
- You discussed the power law degree distribution and its implications.
- You learnt about Milgram and the six degrees of separation.
- You now know about the bow-tie structure of the web.

Today's outline

- Global clustering coefficient (transitivity)
- Local clustering
- Network models
- Implications of these models

Dimensions of Analyzing Networks

Semantic Dimension of Network Analysis

Presentation of qualitative data (e.g. tabular with frequency, bar chart)	Presentation of quantitative data (e.g. histogram)	Measures of central tendency and variability (e.g. mean, range)

Syntactical Dimension of Network Analysis

Local structure	Global structure	*Partitions*
• Degree • Degree Centrality • Closeness Centrality • Betweenness Centrality • Local Clustering Coefficient	• Mean degree • Degree distribution • Density • Network Centralization • Global Clustering Coefficient • Components	• Local definition, such as clique, k-core, k-plex • Global definition with null model (modularity with modularity optimization and edge betweenness)

Dimensions of Analyzing Networks

Semantic Dimension of Network Analysis

Presentation of qualitative data (e.g. tabular with frequency, bar chart)	Presentation of quantitative data (e.g. histogram)	Measures of central tendency and variability (e.g. mean, range)

Syntactical Dimension of Network Analysis

Local structure	Global structure	*Partitions*
• Degree • Degree Centrality • Closeness Centrality • Betweenness Centrality • **Local Clustering Coefficient**	• Mean degree • Degree distribution • Density • Network Centralization • **Global Clustering Coefficient** • Components	• Local definition, such as clique, k-core, k-plex • Global definition with null model (modularity with modularity optimization and edge betweenness)

To add

The **degree sequence** need not necessarily be given in ascending order of degrees as said in the last lecture. For instance, in many cases the vertices are given numeric labels and their degrees are then listed in the order of the labels.

Newman, Mark: Networks: An Introduction. Oxford University Press, 2010.

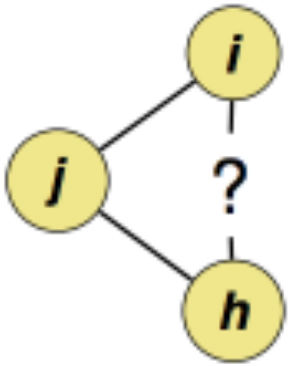
Global clustering coefficient (transitivity)

This chapter is mainly based on:
Mark E.J. Newman. Networks: An Introduction. Oxford University Press. 2010.

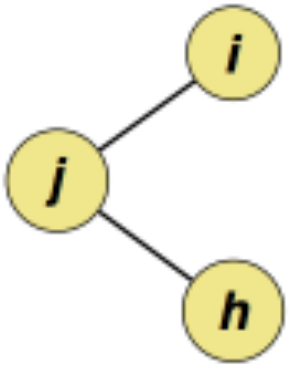
Transitivity

- Very important in networks derived from social media

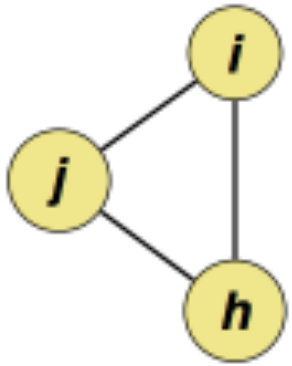
- In mathematics, a relation "*" is said to be transitive if a * b and b* c together imply a * c
  - Example: equality => a=b and b=c then it follows that a=c



Potentially transitive



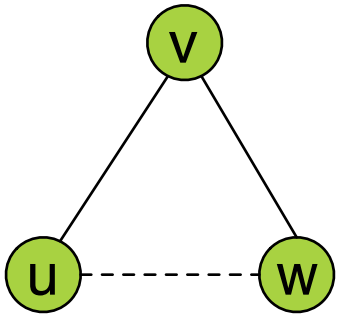
Intransitive



Transitive

Partial transitivity

- Again, a social network example
  - u knows v
  - v knows w
  - the latter makes it much more likely that u knows w also

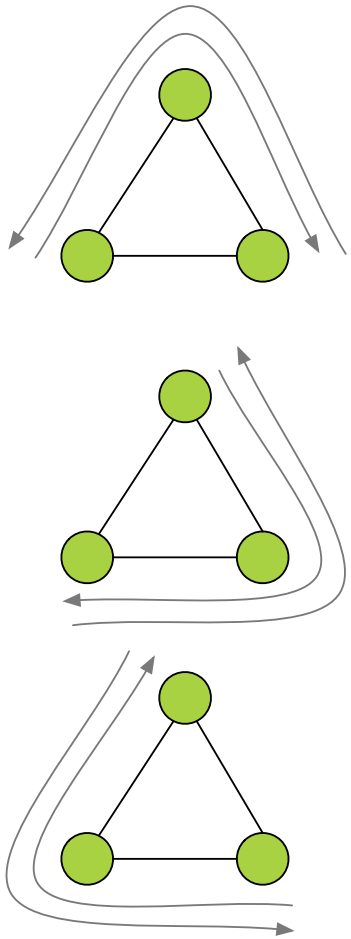


- Let's quantify this idea
  - u knows v and v knows w, then we have a path uvw of two edges in the network
  - If u also knows w then the path is closed

Calculating transitivity

- Note:
  - The path uvw is different to the path wvu and it is therefore counted separately
  - Closed paths are counted separately in each direction

$$C = \frac{(\text{number of triangles}) \times 6}{(\text{number of paths of length two})}$$



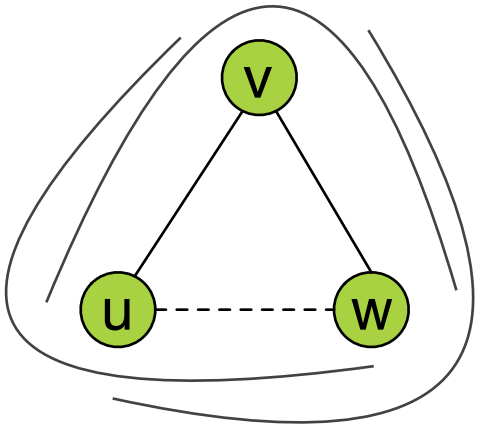
- Why 6?
  - Given is a triangle uvw
  - Then we have six paths of length 2: uvw, wvu, vwu, uwv, wuv, vuw
  - Each path is closed, therefore the number of closed paths is six times the number of a triangle

Another more common approach

- Fraction of pairs of people with a common friend who are themselves friends or equivalently as the mean probability that two people with a common friend are themselves friends

- Most often used definition

$$C = \frac{(\text{number of triangles}) \times 3}{(\text{number of connected triplets})}$$



- A “connected triples”?
  - Three vertices uvw with edges (u,v) and (v,w); the edge (u,w) can be present or not
  - The factor of three in the numerator arises because each triangle gets counted three times when we count the connected triples in the network
  - The triangle uvw for instance contains the triples uvw, vwu, and wuv

Measure for transitivity

- If C=1, then perfect transitivity
- If C=0, then no closed triads which happens in topologies such as trees

- For example, typical values for social networks are:
  - the network of film actor collaborations discussed earlier has been found to have C = 0.20;
  - a network of collaborations between biologists has been found to have C = 0.09;
  - a network of who sends email to whom in a large university has C = 0.16

- Some denser networks have even higher values, as high as 0.5 or 0.6.

- Technological and biological networks by contrast tend to have somewhat lower values
- For example, the Internet has a clustering coefficient of only about 0.01

Local clustering

This chapter is mainly based on:
Mark E.J. Newman. Networks: An Introduction. Oxford University Press. 2010.

Local clustering

- Represents the average probability that a pair of i's friends are friends of one another

- Clustering coefficient for a single vertex i

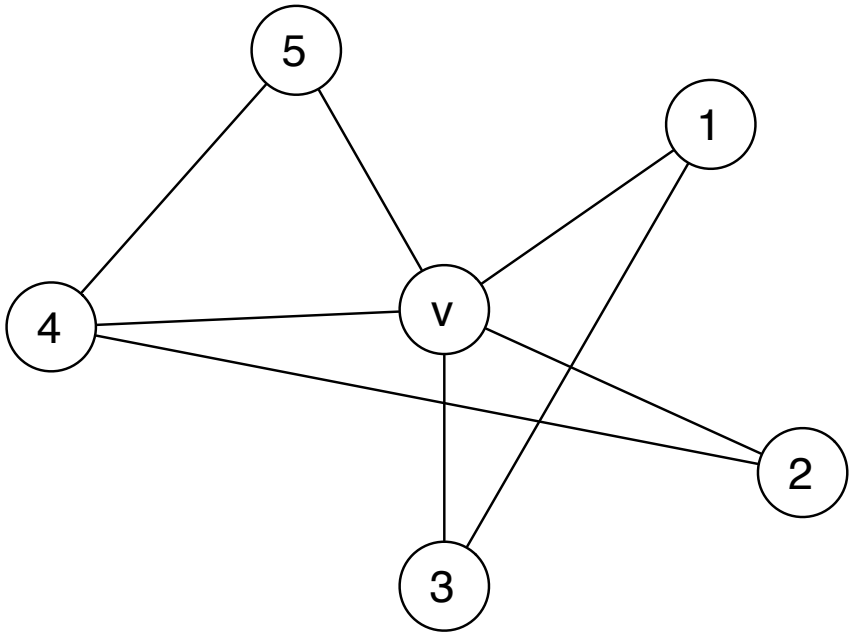
$$C_i = \frac{(\text{number of pairs of neighbors of i that are connected})}{(\text{number of pairs of neighbors of i})}$$

- Calculation
  - Go through all distinct pairs of vertices that are neighbors of i in the network
  - Count the number of such pairs that are connected to each other
  - Divide by total numbers of pairs, which is $\frac{1}{2}k_i(k_i - 1)$

    where k_i is the degree of a vertex i

Example

- Given a graph G = (V, E) and a vertex v ∈ V
- In the sample graph three of v's neighbors are connected to each other, namely (1-3) (2-4) and (4-5)
- There are a total of 10 pairs of neighbors, namely (1-2),(1-3),(1-4),(1-5),(2-3), (2-4),(2-5),(3-4),(3-5),(4-5)
- **Local CC: 3/10 = 0.3**



Local clustering

- Represents the average probability that a pair of i's friends are friends of one another

- Clustering coefficient for a single vertex i

$$C_i = \frac{(\text{number of pairs of neighbors of i that are connected})}{(\text{number of pairs of neighbors of i})}$$

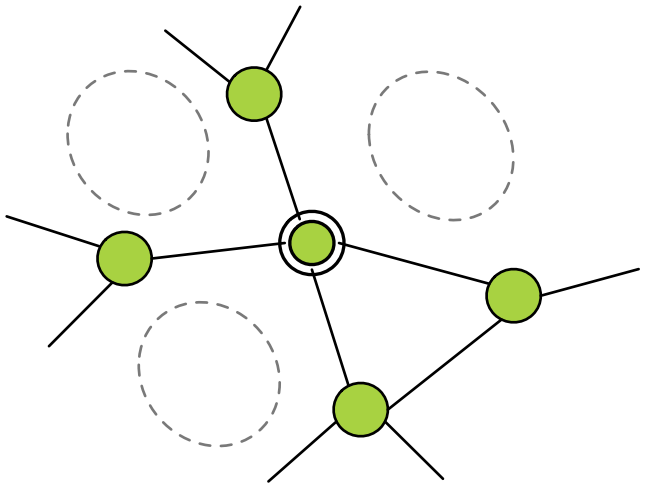
- Calculation
  - Go through all distinct pairs of vertices that are neighbors of i in the network
  - Count the number of such pairs that are connected to each other
  - Divide by total numbers of pairs, which is $\frac{1}{2}k_i(k_i - 1)$

    where k_i is the degree of a vertex i

- **Local clustering can be used as a probe for the existence of so-called structural holes**

What is the effect of missing connections?

- When the neighbors of a vertex are not connected the missing links are called **structural holes**



- First studied by Burt 1992

- Example
  - Reduce the number of alternative routes information or other traffic can take through a network
  - But, it can be a good thing for the central vertex i, because i has power over information between unconnected vertices, i.e. friends

- Measuring the number of structural hole allows to determine how influential a vertex is

Ronald S. Burt: Structural holes: The social structure of competition. Harvard University Press. Cambridge, MA. 1992.

Network average clustering coefficient

- Proposed by Watts and Strogatz (1998)

- Calculate a clustering coefficient for the entire network as mean of the local clustering coefficient of each vertex:

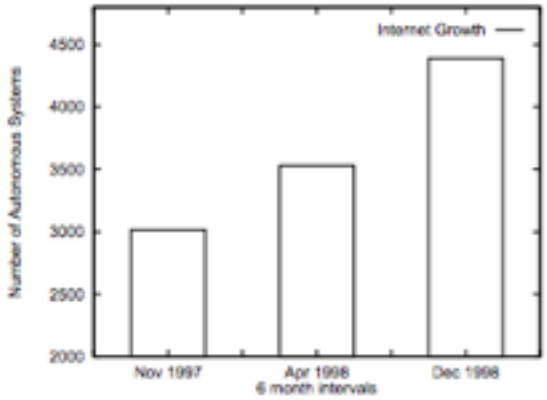
$$C_{WS} = \frac{1}{n} \sum_{i=1}^{n} C_i$$

- Tends to be dominated by vertices with low degree, since they have a small denominator

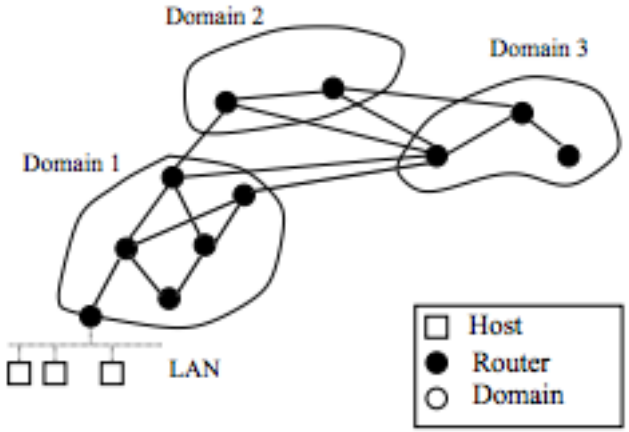
Duncan J. Watts, Steven H. Strogatz: Collective dynamics of 'small-world' networks. Nature. Vol 393: 4. 1998

Network models

Famous example of a degree distribution is power law degree distribution



Growth of the Internet



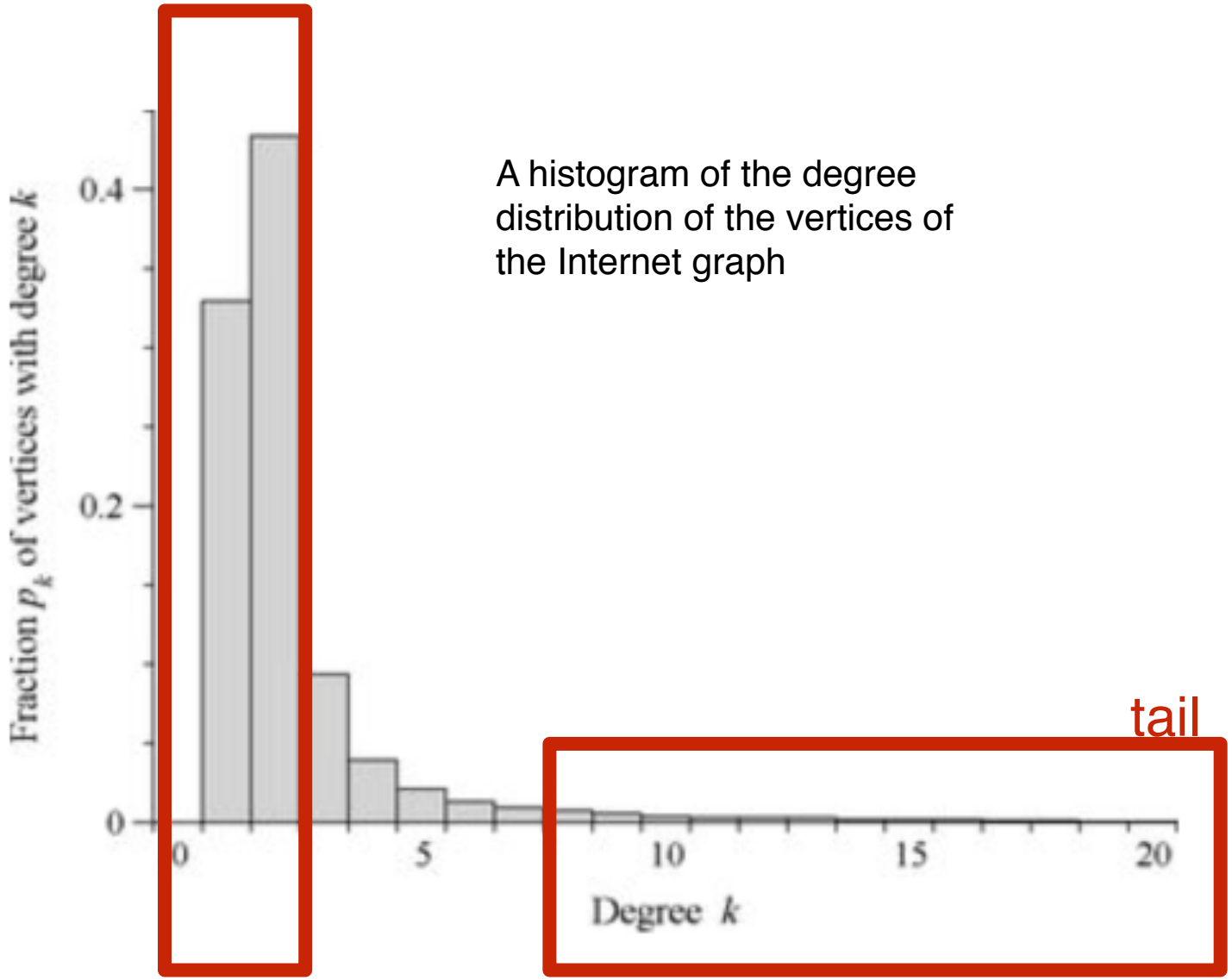
Structure of the Internet at the router level

	Int-11-97	Int-04-98	Int-12-98
nodes	3015	3530	4389
edges	5156	6432	8256
avg. outdegree	3.42	3.65	3.76
max. outdegree	590	745	979
diameter	9	11	10
avg. distance	3.76	3.77	3.75

The evolution of the Internet at the inter-domain level.

Faloutsos, Michalis, Petros Faloutsos, and Christos Faloutsos. "On power-law relationships of the internet topology." ACM SIGCOMM computer communication review. Vol. 29. No. 4. ACM, 1999.

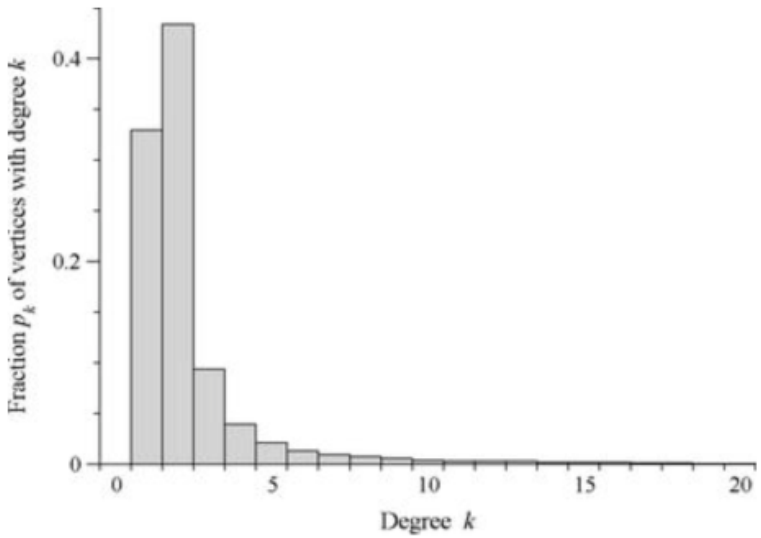
Degree distribution of the inter-domain topology of the Internet

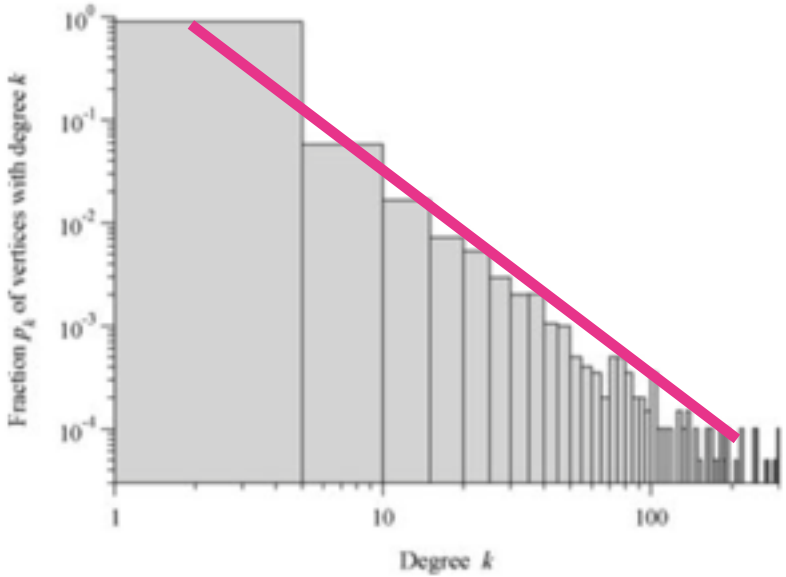


A histogram of the degree distribution of the vertices of the Internet graph

Newman, Mark: Networks: An Introduction. Oxford University Press, 2010.

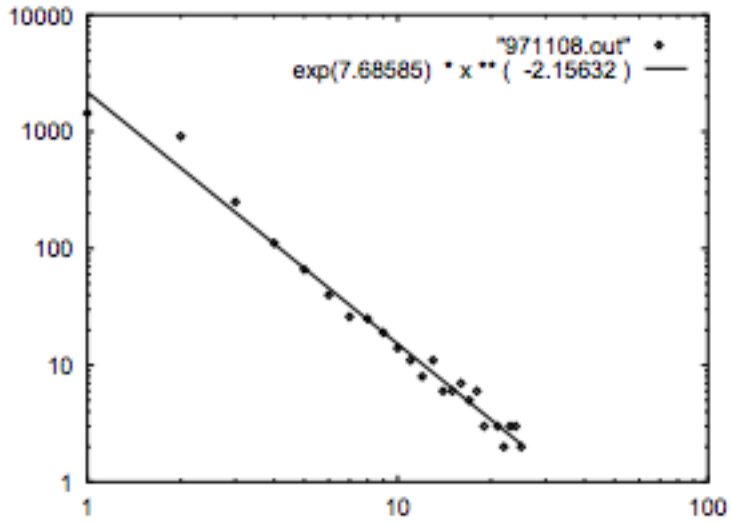
Power law distribution



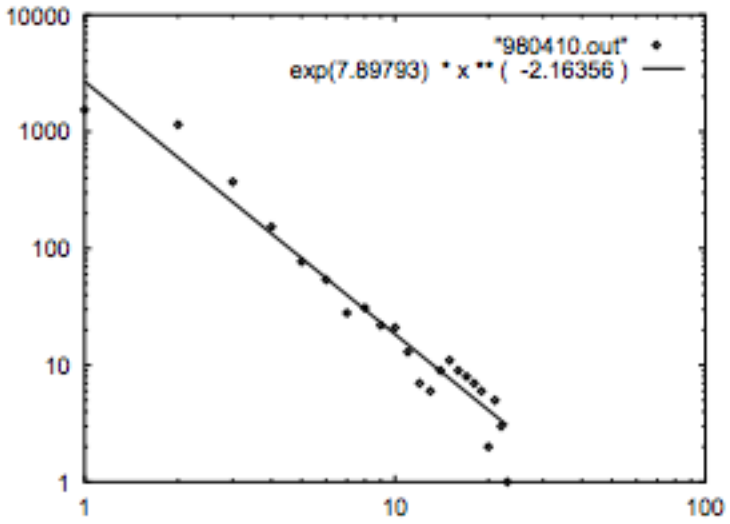


Newman, Mark: Networks: An Introduction. Oxford University Press, 2010.

Degree distribution of the inter-domain topology of the Internet



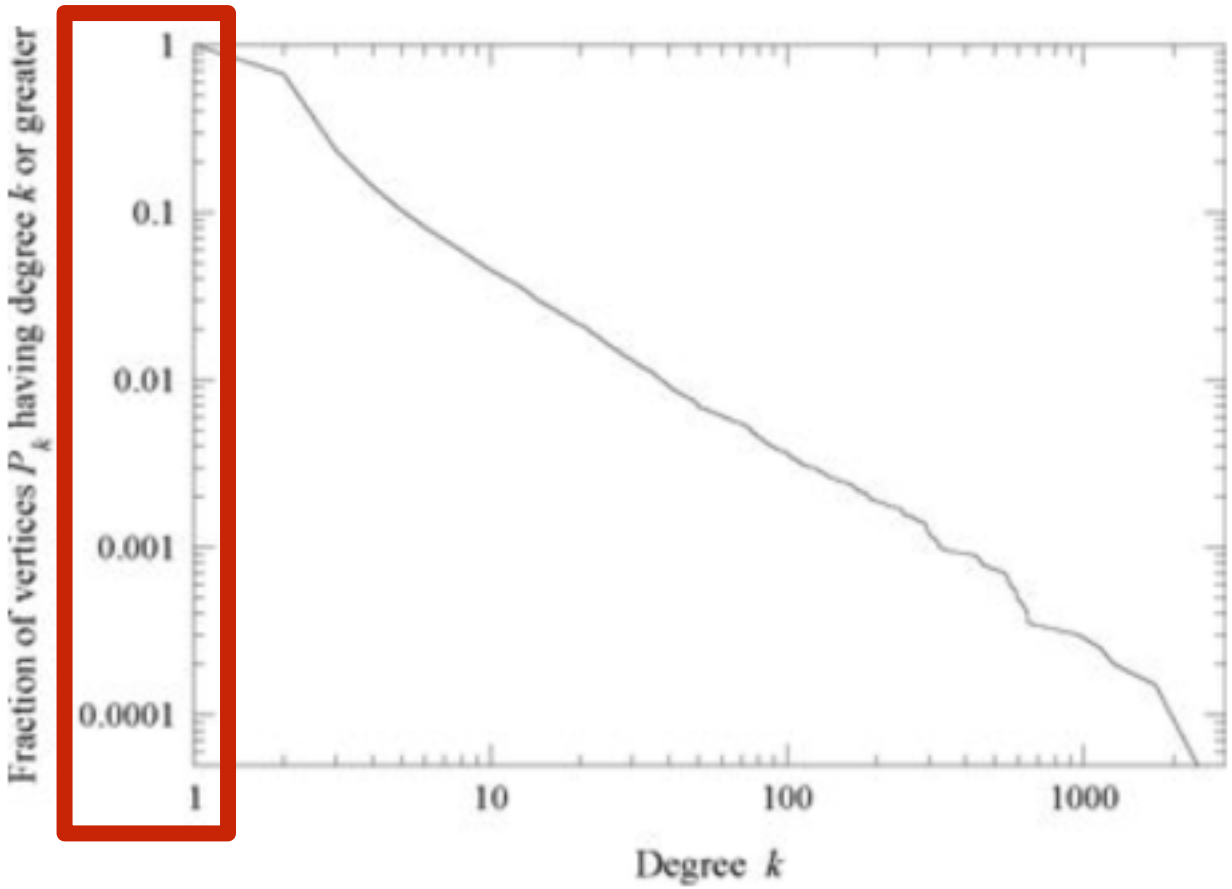
(a) Int-11-97



(b) Int-04-98

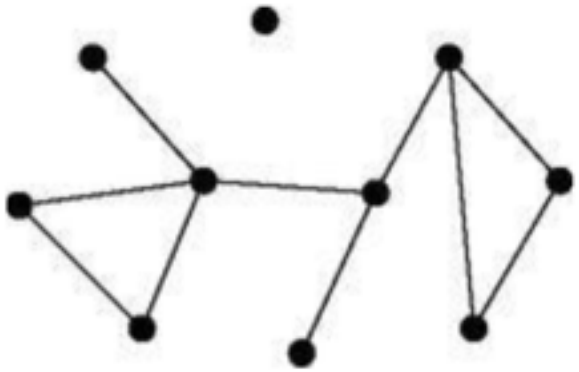
The out-degree plots: Log-log plot of frequency f_k versus the out-degree k

Cumulative degree distribution function



Newman, Mark: Networks: An Introduction. Oxford University Press, 2010.

Cumulative degree distribution function



n = 10 vertices, of which
1 has degree 0,
2 have degree
2 have degree 3, and
1 has degree 4

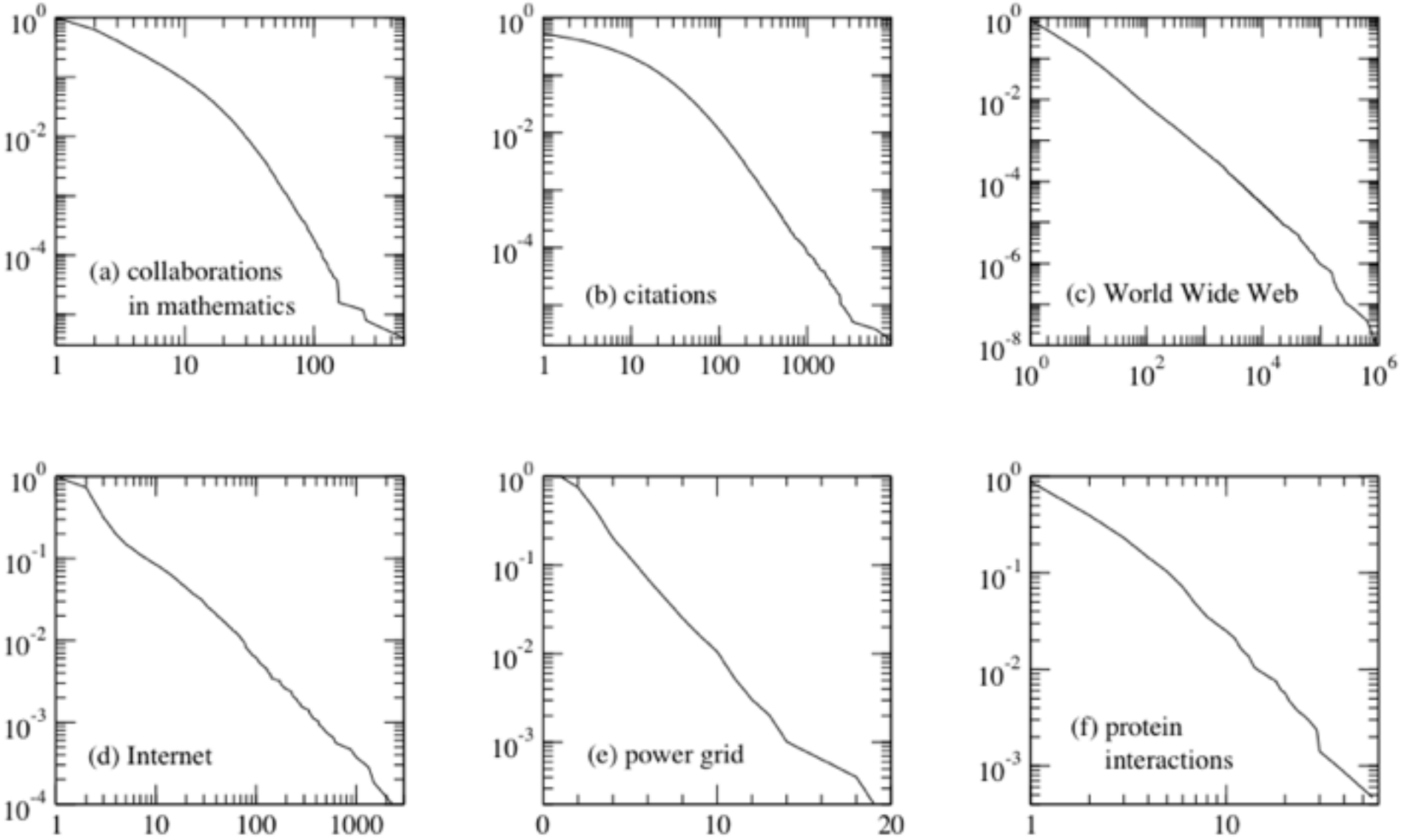
Thus the values of p_k for $k = 0,..., 4$ are $p_0 = 1/10$, $p_1 = 2/10$, $p_2 = 4/10$, $p_3 = 2/10$, $p_4 = 1/10$, and $p_k = 0$ for all $k > 4$.

Degrees sequence {4, 3, 3, 2, 2, 2, 2, 1, 1, 0}

Degree k	Rank r	$P_k = r/n$
4	1	0.1
3	2	0.2
3	3	0.3
2	4	0.4
2	5	0.5
2	6	0.6
2	7	0.7
1	8	0.8
1	9	0.9
0	10	1.0

Newman, Mark: Networks: An Introduction. Oxford University Press, 2010.

Cumulative degree distributions for different networks



Newman, M.E.J.: Structure and function of complex networks. SIAM Review. 45. 2003.

"How might a network come to have such a distribution?"

Newman, Mark: Networks: An Introduction. Oxford University Press, 2010.

Preferential attachment

- Price, who was interested in, among other things, the citation networks of scientific papers, was inspired by the work of economist Herbert Simon
- Simon proposed an explanation for the wealth distribution based on the idea that people who have money already gain more at a rate proportional to how much they already have.
- Price adapted Simon's methods, with relatively little change, to the network and coined it cumulative advantage
- In 1999, Barabási and Albert used the name "preferential attachment"
- What does it mean for a web page?
  - A page's popularity grows at a rate proportional to its current value, and hence exponentially with time. A page that gets a small lead over others will therefore tend to extend this lead.

Newman, Mark: Networks: An Introduction. Oxford University Press, 2010.

"What might be one reason for preferential attachment?"

Newman, Mark: Networks: An Introduction. Oxford University Press, 2010.

Experimental Study of Inequality and Unpredictability

- Set-up
  - Creation of an artificial music market comprising 14,341 participants, recruited mostly from a teen-interest World Wide Web site, who were shown a list of previously unknown songs from unknown bands
  - In real time, arriving participants were randomly assigned to one of two experimental conditions—independent and social influence—

- Independent condition
  - Participants made decisions about which songs to listen to, given only the names of the bands and their songs.
  - While listening to a song, they were asked to assign a rating from one star to five stars, after which they were given the opportunity (but not required) to download the song

- Social Influence condition
  - Additionally participants could also see how many times each song had been downloaded by previous participants

Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts. "Experimental study of inequality and unpredictability in an artificial cultural market." science 311.5762 (2006): 854-856.

Experimental Study of Inequality and Unpredictability (*cont.*)

Experiment 1

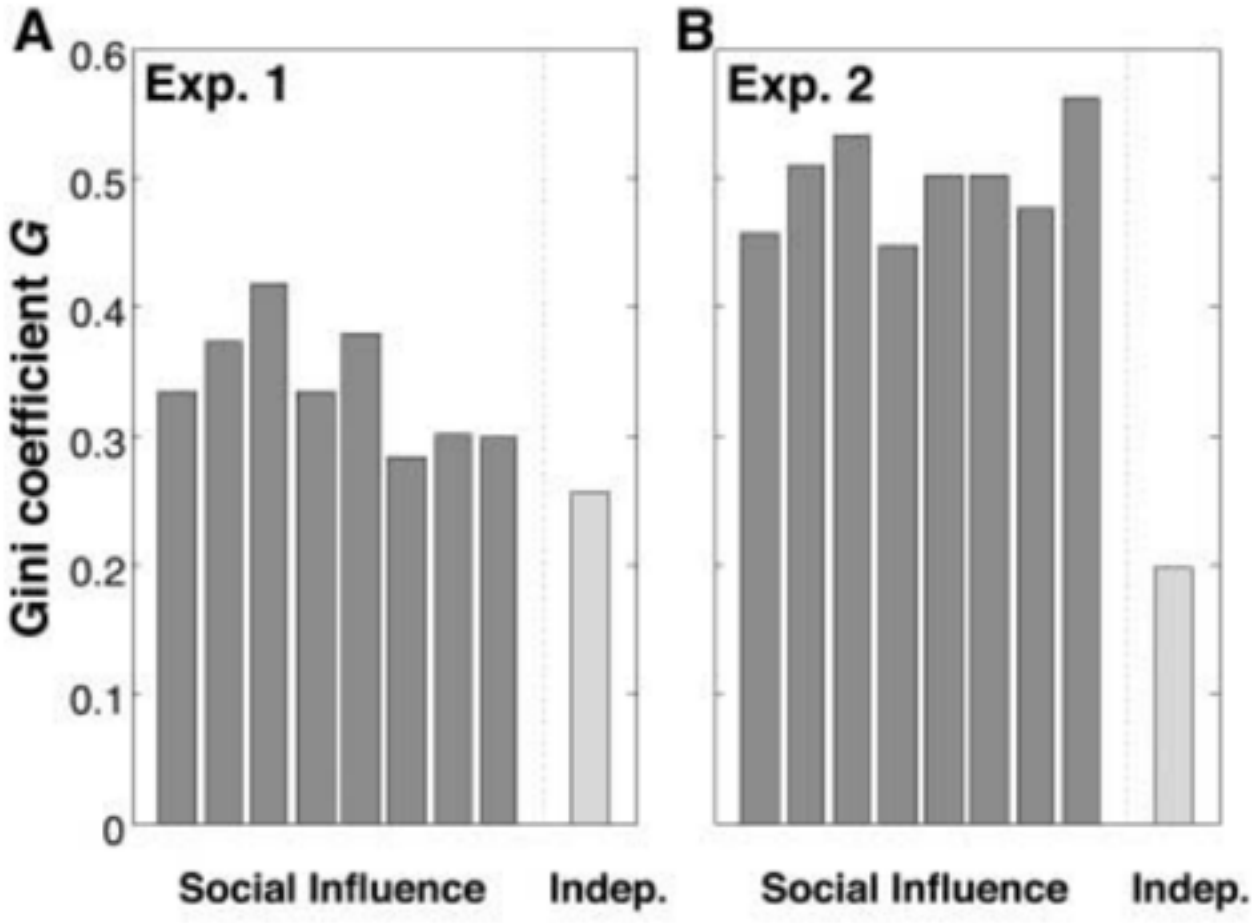
- Social influence condition
  - The songs, along with the number of previous downloads, were presented to the participants arranged in a 16 x 3 rectangular grid, where the positions of the songs were randomly assigned for each participant (i.e., songs were not ordered by download counts).
- Independent condition
  - Participants had the same presentation of songs, but without any information about previous downloads

Experiment 2

- Social influence condition
  - Songs were, with download counts, presented in one column in descending order of current popularity.
- Independent condition
  - Songs were also presented with the single column format, but without download counts and in an order that was randomly assigned for each participant

Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts. "Experimental study of inequality and unpredictability in an artificial cultural market." science 311.5762 (2006): 854-856.

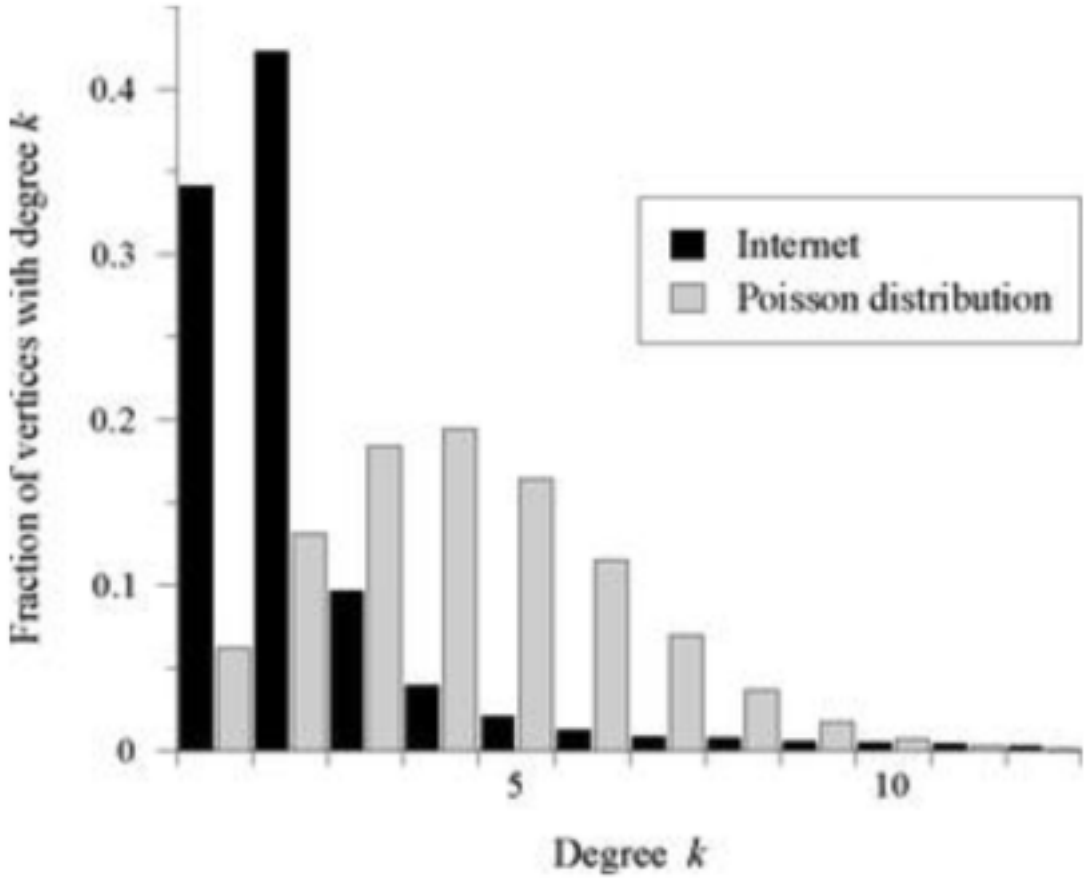
Experimental Study of Inequality and Unpredictability (*cont.*)



Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts. "Experimental study of inequality and unpredictability in an artificial cultural market." science 311.5762 (2006): 854-856.

Random graphs

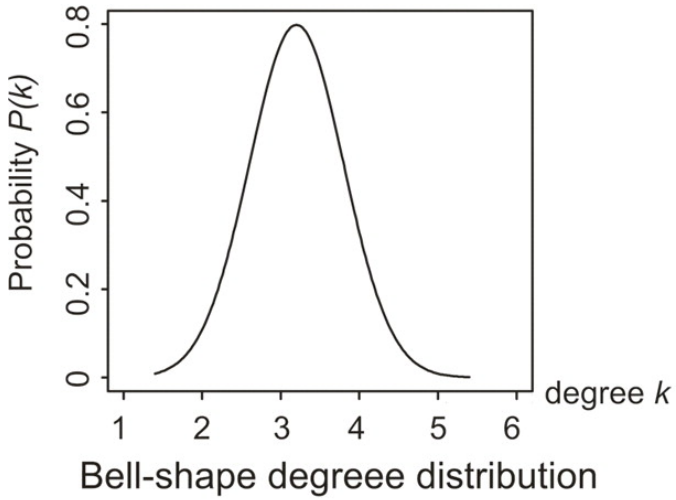
Degree distribution of the Internet and a Poisson random graph



Newman, Mark: Networks: An Introduction. Oxford University Press, 2010.

Barabási–Albert model

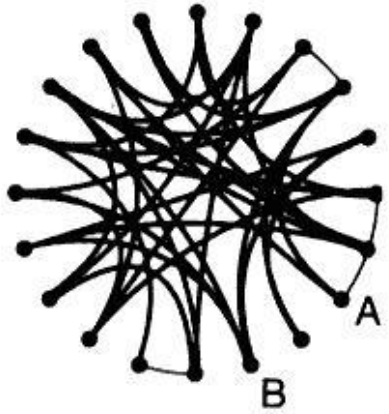
- Assumption is that the presence or absence of an edge between two vertices is independent of the presence or absence of any other edge, so that each edge may be considered to be present with independent probability *p*



- Each edge is present or absent with equal probability, therefore degree distribution is Poisson

Zhu X et al. Getting connected: analysis and principles of biological networks. Genes Dev. 2007;21:1010-1024

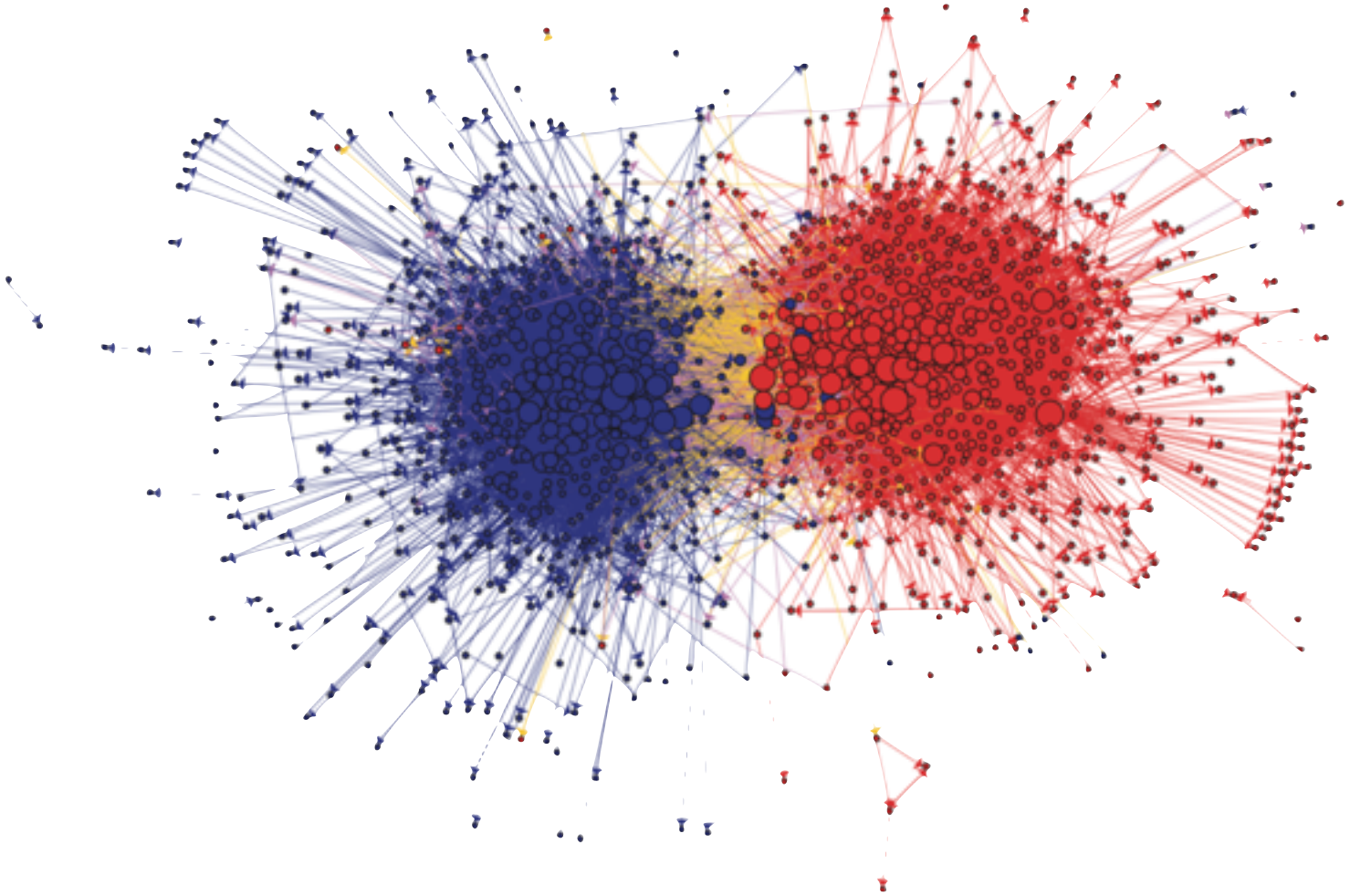
M. E. J. Newman, S. H. Strogatz, and D. J. Watts: Random graphs with arbitrary degree distributions and their applications

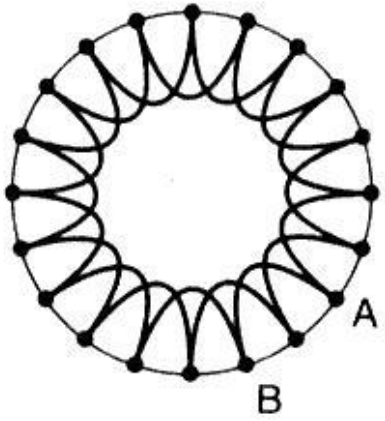


random graph:
all connections
random

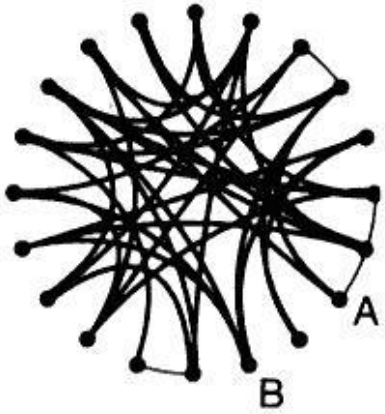
Watts, D.J., Strogatz, S.H.(1998) Collective dynamics of 'small-world' networks. Nature 393:440-442.

Politics revisited





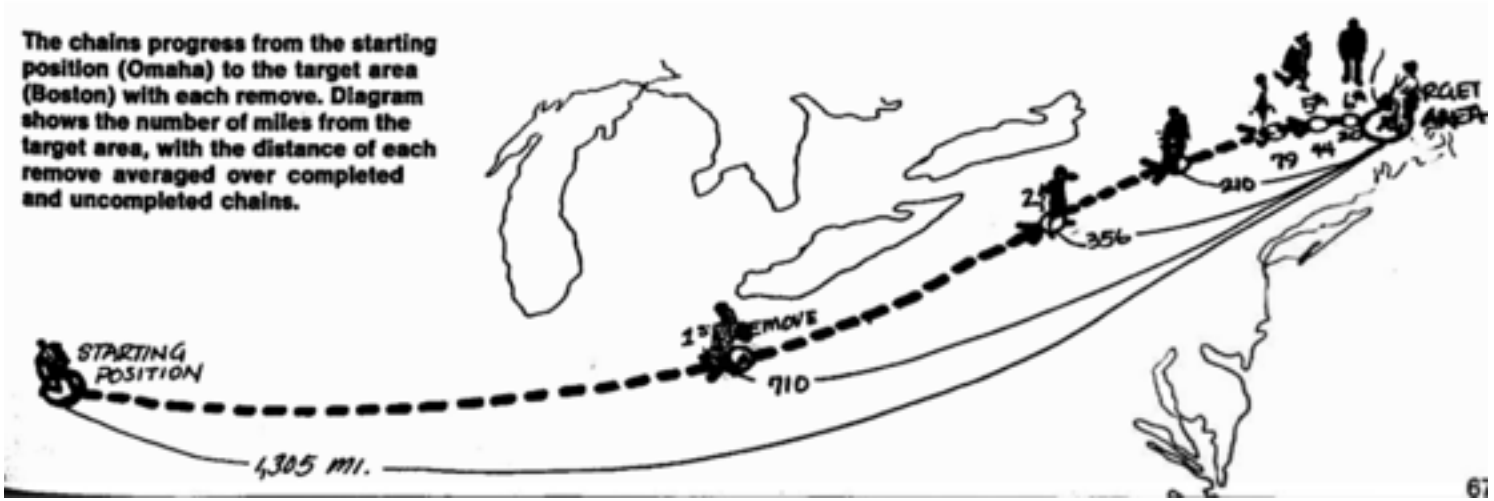
regular lattice:
my friend's friend is always my friend



random graph:
all connections random

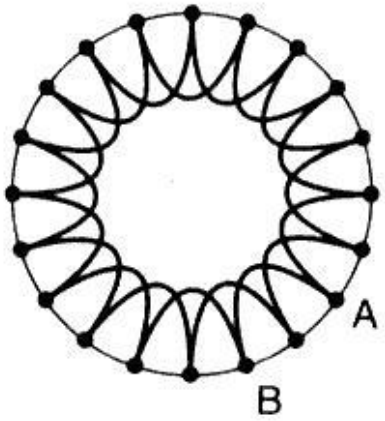
Watts, D.J., Strogatz, S.H.(1998) Collective dynamics of 'small-world' networks. Nature 393:440-442.

Milgram revisited

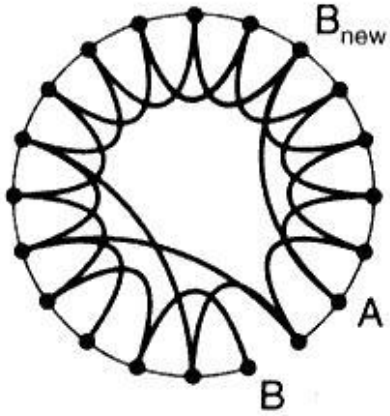


Watts-Strogatz model

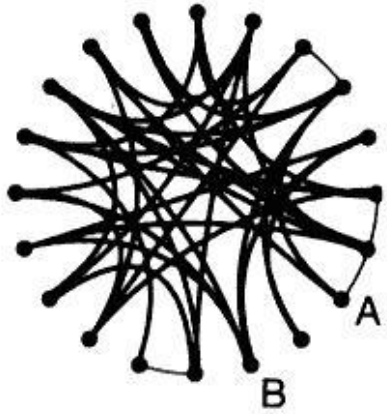
- A few random links in an otherwise structured graph make the network a small world: the average shortest path is short



regular lattice:
my friend's friend is always my friend



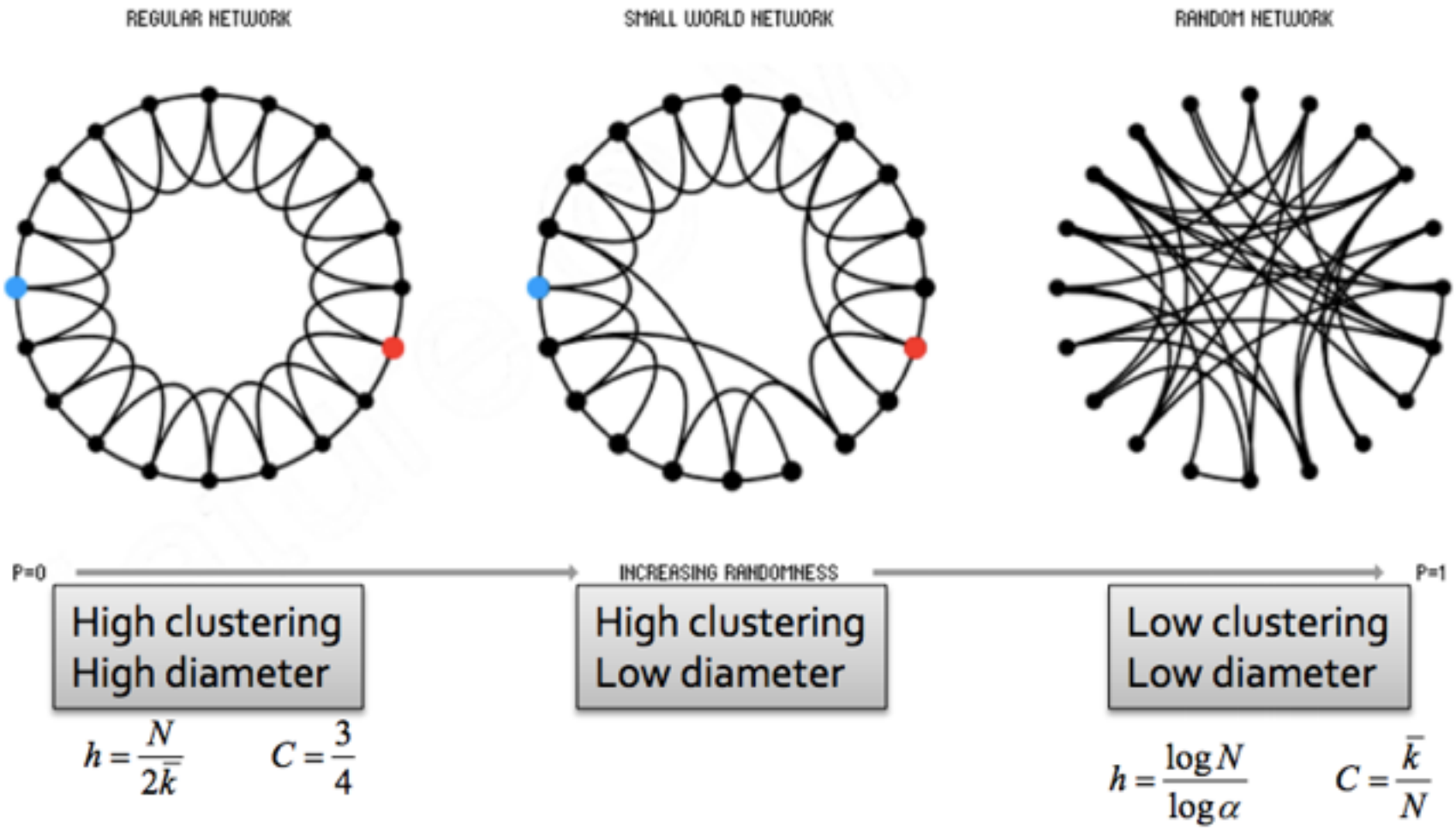
small world:
mostly structured with a few random connections



random graph:
all connections random

Watts, D.J., Strogatz, S.H.(1998) Collective dynamics of 'small-world' networks. Nature 393:440-442.

Watts-Strogatz model



C: Clustering Coefficient
h: Average shortest path length

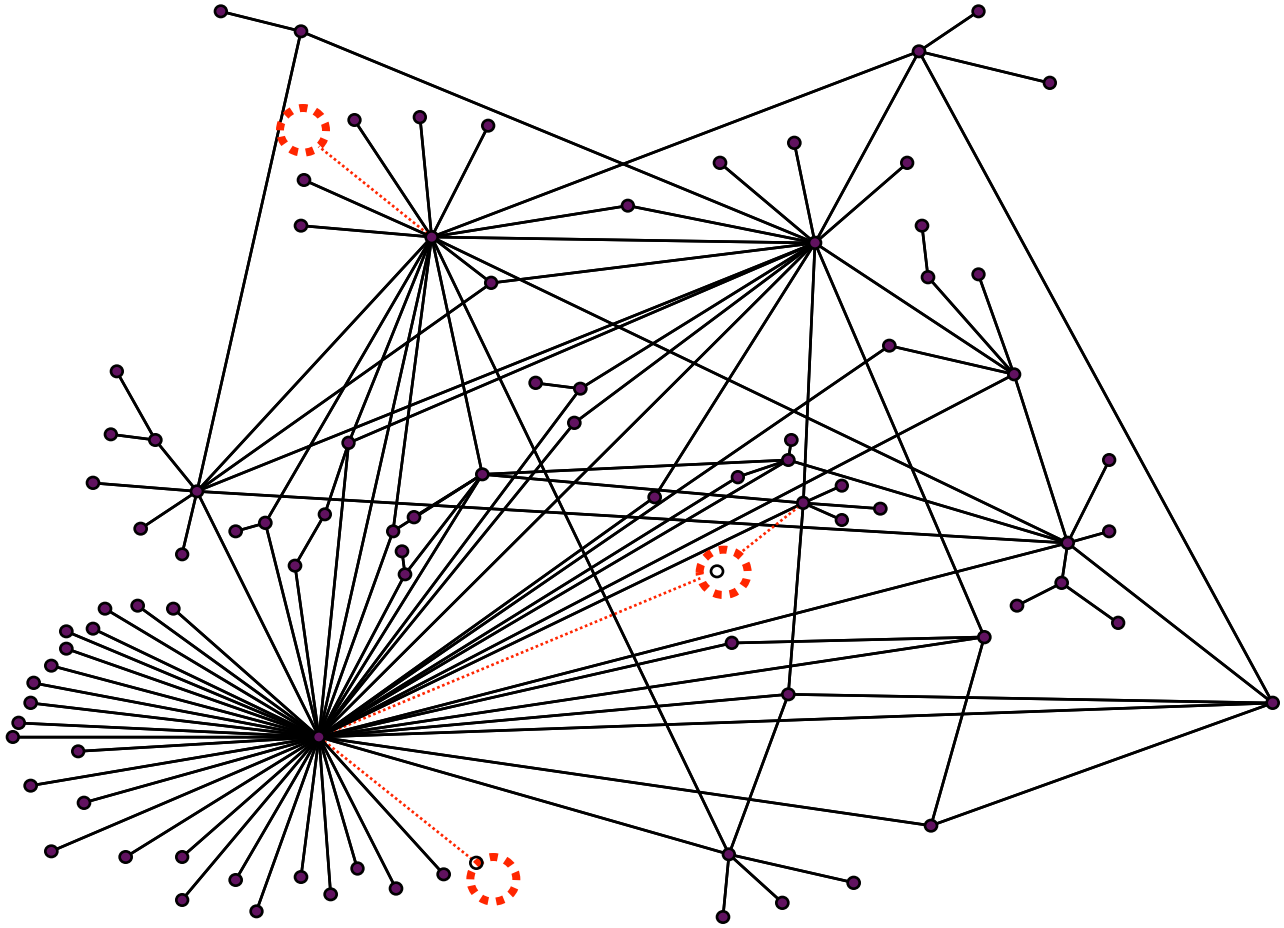
Implications

What can we learn from these models?

- Implications on
  - Robustness
  - Search
  - Spread of diseases
  - Opinion formation

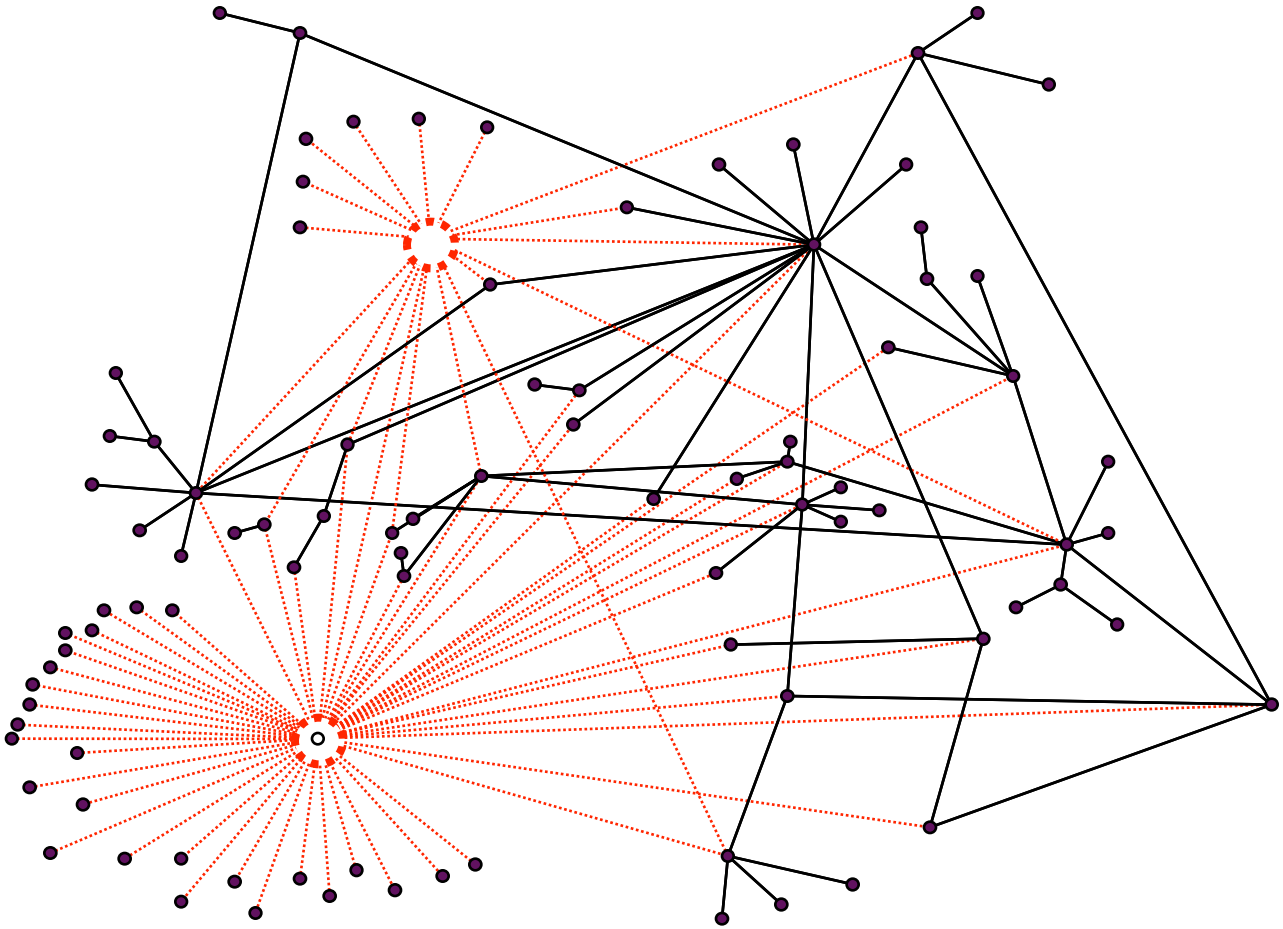
Slide taken from Lada Adamic, 2008

What implications does this have on robustness?



Slide taken from Lada Adamic, 2008

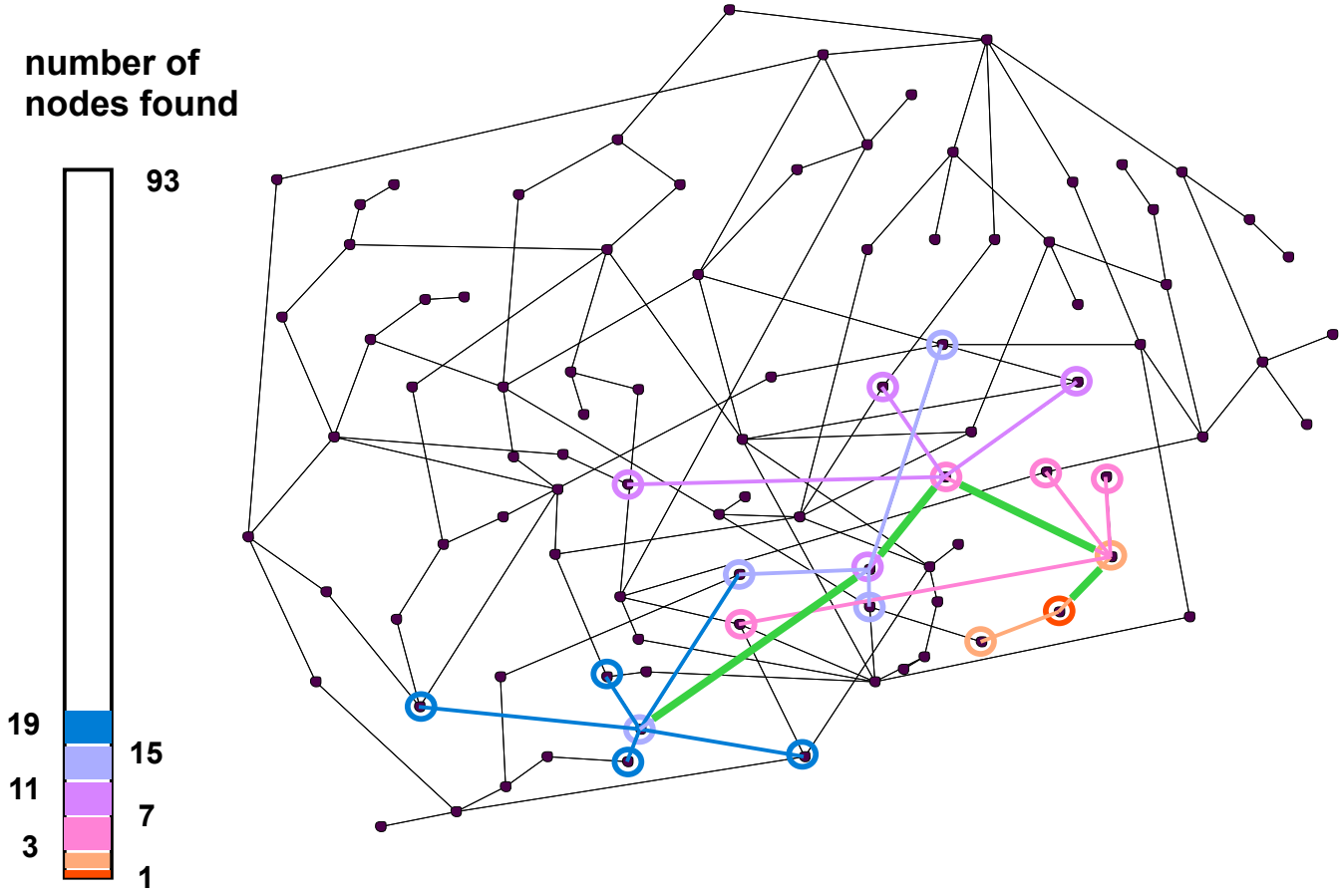
What implications does this have on robustness?



Slide taken from Lada Adamic, 2008

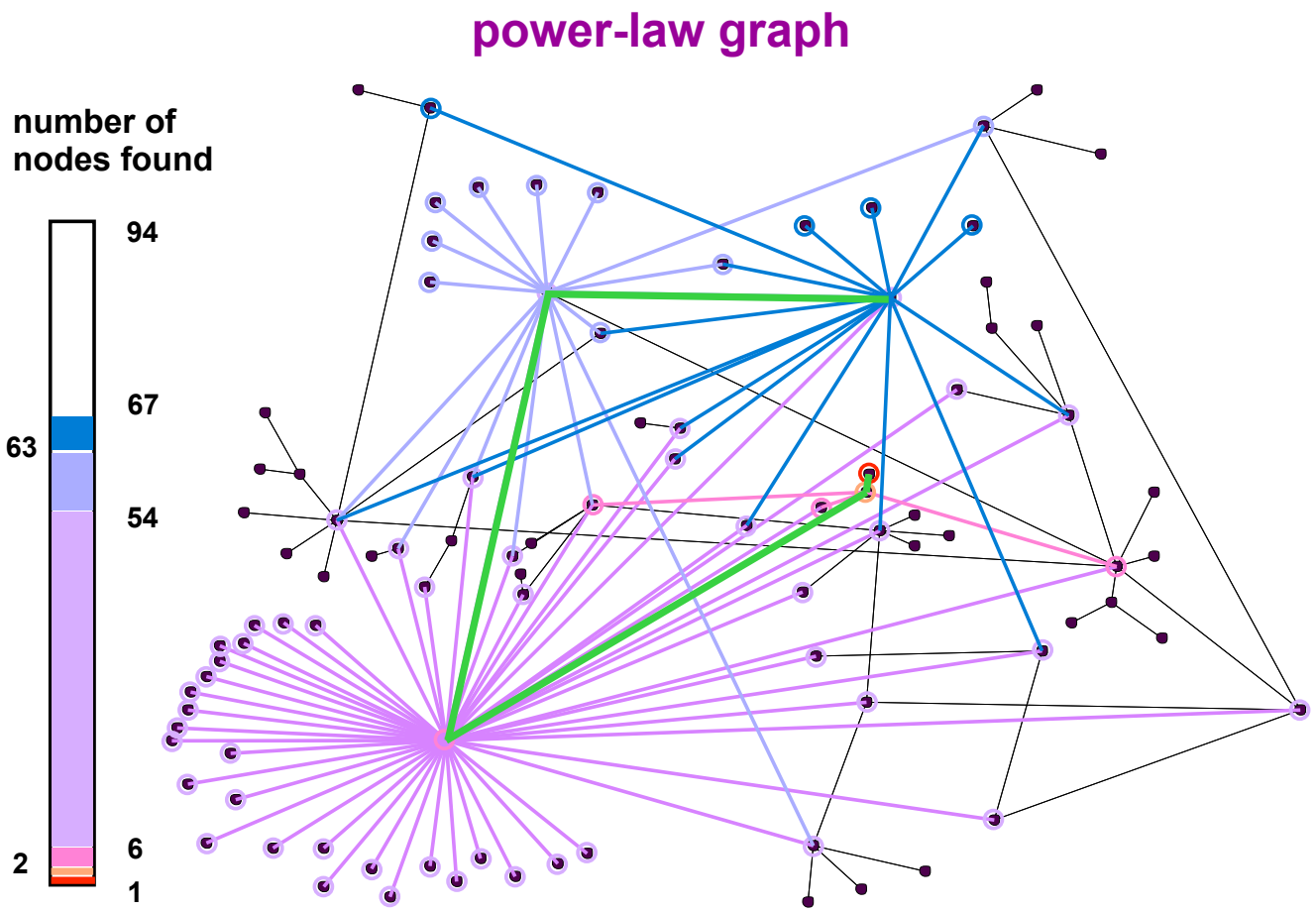
What implications does this have on search?

Poisson graph



Slide taken from Lada Adamic, 2008

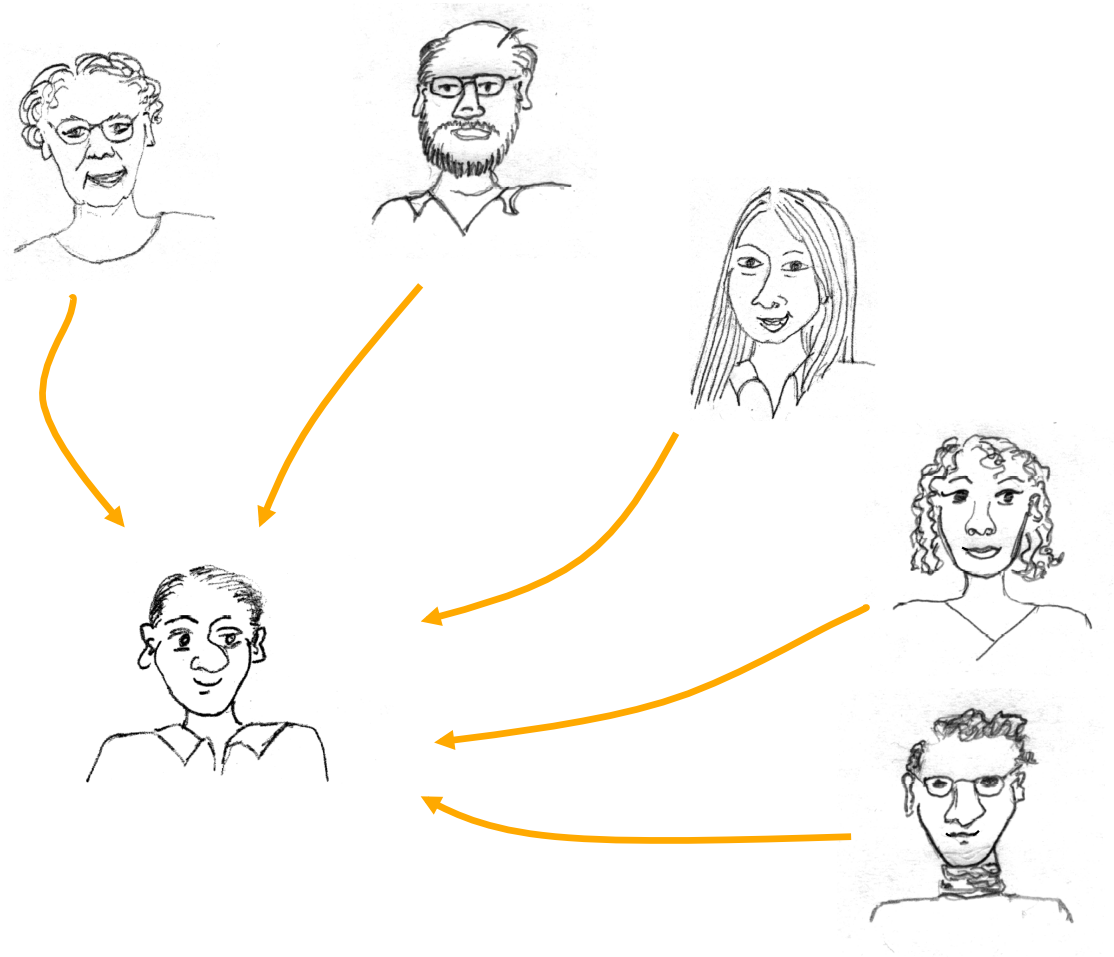
What implications does this have on search?



taken from Lada Adamic, 2008

What implications does this have on spread of diseases?

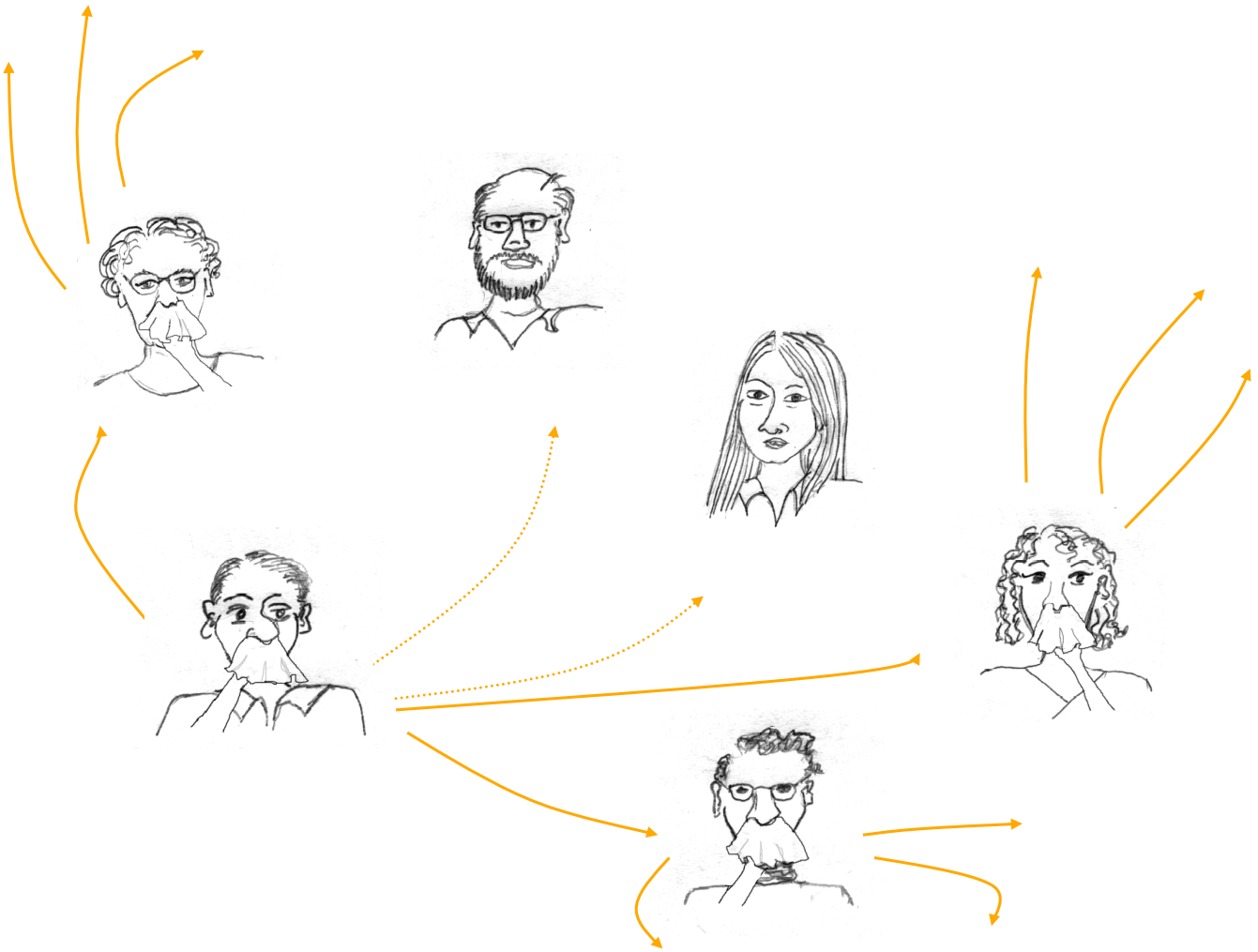
- In social networks, it's nice to be a hub...



Slide taken from Lada Adamic, 2008

What implications does this have on spread of diseases?

- But it depends on what you're sharing…



Questions?