School of Information Sciences
University of Pittsburgh

TELCOM2125: Network Science and Analysis

Konstantinos Pelechrinis

Spring 2015

Figures are taken from:
M.E.J. Newman, “Networks: An Introduction”

Part 3: The Large-Scale Structure of Networks

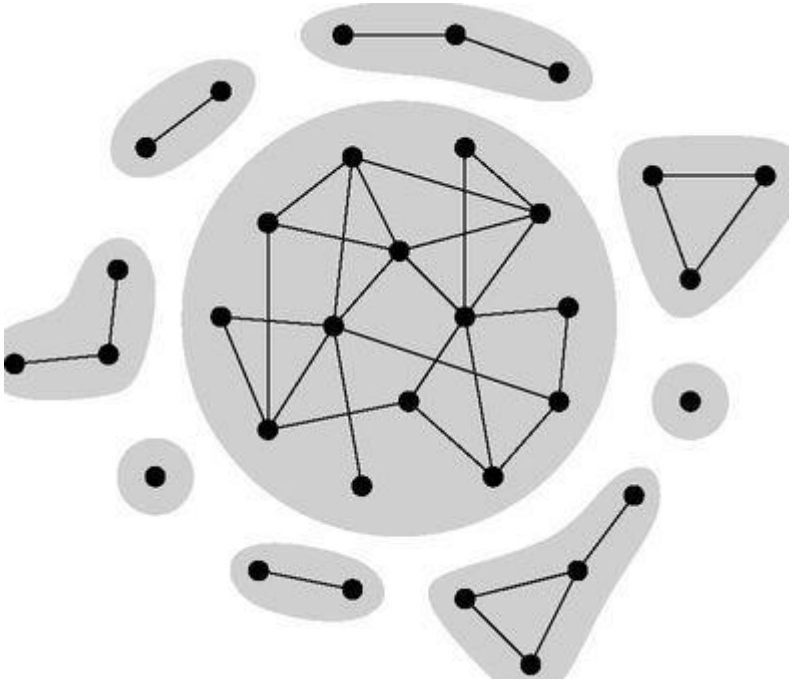
Statistics for real networks

	Network	Type	n	m	c	S	ℓ	α	C	C_{WS}	
Social	Film actors	Undirected	449 913	25 516 482	113.43	0.980	3.48	2.3	0.20	0.78	0.
	Company directors	Undirected	7 673	55 392	14.44	0.876	4.60	–	0.59	0.88	0.
	Math coauthorship	Undirected	253 339	496 489	3.92	0.822	7.57	–	0.15	0.34	0.
	Physics coauthorship	Undirected	52 909	245 300	9.27	0.838	6.19	–	0.45	0.56	0.
	Biology coauthorship	Undirected	1 520 251	11 803 064	15.53	0.918	4.92	–	0.088	0.60	0.
	Telephone call graph	Undirected	47 000 000	80 000 000	3.16			2.1			
	Email messages	Directed	59 812	86 300	1.44	0.952	4.95	1.5/2.0		0.16	
	Email address books	Directed	16 881	57 029	3.38	0.590	5.22	–	0.17	0.13	0.
	Student dating	Undirected	573	477	1.66	0.503	16.01	–	0.005	0.001	−0.
	Sexual contacts	Undirected	2 810					3.2			
Information	WWW nd.edu	Directed	269 504	1 497 135	5.55	1.000	11.27	2.1/2.4	0.11	0.29	−0.
	WWW AltaVista	Directed	203 549 046	1 466 000 000	7.20	0.914	16.18	2.1/2.7			
	Citation network	Directed	783 339	6 716 198	8.57			3.0/–			
	Roget's Thesaurus	Directed	1 022	5 103	4.99	0.977	4.87	–	0.13	0.15	0.
	Word co-occurrence	Undirected	460 902	16 100 000	66.96	1.000		2.7		0.44	
Technological	Internet	Undirected	10 697	31 992	5.98	1.000	3.31	2.5	0.035	0.39	−0.
	Power grid	Undirected	4 941	6 594	2.67	1.000	18.99	–	0.10	0.080	−0.
	Train routes	Undirected	587	19 603	66.79	1.000	2.16	–		0.69	−0.
	Software packages	Directed	1 439	1 723	1.20	0.998	2.42	1.6/1.4	0.070	0.082	−0.
	Software classes	Directed	1 376	2 213	1.61	1.000	5.40	–	0.033	0.012	−0.
	Electronic circuits	Undirected	24 097	53 248	4.34	1.000	11.05	3.0	0.010	0.030	−0.
	Peer-to-peer network	Undirected	880	1 296	1.47	0.805	4.28	2.1	0.012	0.011	−0.
Biological	Metabolic network	Undirected	765	3 686	9.64	0.996	2.56	2.2	0.090	0.67	−0.
	Protein interactions	Undirected	2 115	2 240	2.12	0.689	6.80	2.4	0.072	0.071	−0.
	Marine food web	Directed	134	598	4.46	1.000	2.05	–	0.16	0.23	−0.
	Freshwater food web	Directed	92	997	10.84	1.000	1.90	–	0.20	0.087	−0.
	Neural network	Directed	307	2 359	7.68	0.967	3.97	–	0.18	0.28	−0.

Components

- **What are the component sizes in a real-world network?**
  - Typically there is a large component that fills most of the network
    - ✓Even more than 90% of the nodes
  - Rest of the network is divided in many smaller components disconnected from each other

- **The large components can arise either due to the nature of the network (e.g., Internet), or due to the way the network was measured (e.g., Web)**



Components

- **Can a network have more than one large components that fill a sizable fraction of the entire graph?**
  - Usually the answer is no!
    - ✓ If a network of n nodes was divided into two components of size n/2 each, there would be $0.25n^2$ vertex pairs such that one vertex belongs to the first component and the other one to the second. Given that if *any* of these pairs gets at some point connected through an edge, it is highly unlikely that not one such pair would be connected → it is very unlikely to have two large components
- **Are there networks with no large component?**
  - There can be but not of interest
    - ✓ E.g., immediate family ties

Components in directed networks

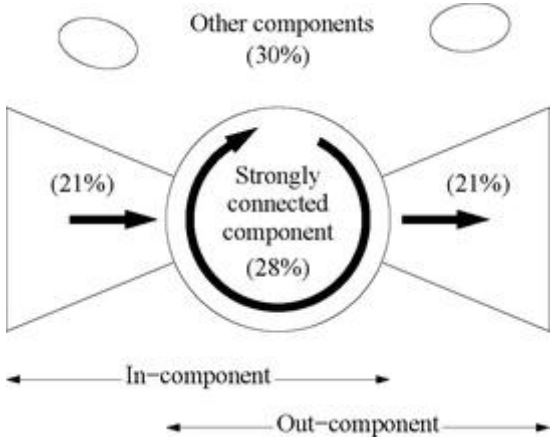
- **The weakly connected components of directed networks *behave* similar to the components of undirected networks**
  - There is a large weakly connected component and many smaller ones

- **The situation is similar with strongly connected components**
  - There is a large strongly connected component and a selection of smaller ones
  - However, the large connected component is typically not "as large" as in the case of undirected networks
    - ✓ E.g., For the WWW, the largest connected component fills about a quarter of the network*

* Broder et al., "Graph structure in the Web", in Computer Networks, 2000

Components in directed networks

- **Associated with a strongly connected component is an out- and in-component**
- **Out- and in- components are supersets of the strongly connected component and can contain many vertices that themselves lie outside the strongly connected component**
  - E.g., in the WWW the portion of in- and out-components that lie outside the largest strongly connected component each also occupy about a quarter of the network
    - ✓"Bow tie" structure
- **Not all directed networks have large strongly connected components**
  - E.g., acyclic networks (citation network)



The small world effect

- **In many networks the typical network distances between vertices are surprisingly small**
  - Small world effect

- **In math terms the small world effect is a hypothesis that the mean distance l is "small"**
  - Typically networks have been found to have mean distance less than 20 – or in many cases less than 10 – even though the networks themselves have millions of nodes
    - ✓ Implications such as rumor spread in a social networks, response time in the Internet etc.

The small world effect

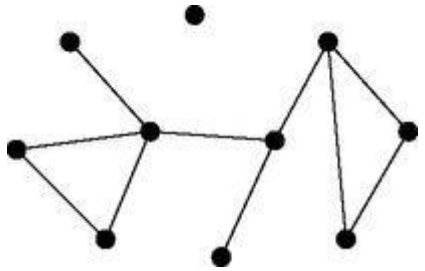
- **The small world effect is not surprising**
  - Mathematical models for networks suggest that the mean path length l in a network increases slowly with the number n of vertices in the network
    - ✓ l~log(n)

- **Similarly, the diameter of a network is relatively small as well**
  - Scales logarithmically as well with the number of vertices
  - However, it is not a very useful metric (extreme case)

- **Funneling**
  - Most of the shortest path of vertex i go through one or two of its neighbors

Degree distributions

- **We define p_k to be the fraction of vertices in a network that have degree k**
  - p_k is essentially the probability that a randomly selected node of the network will have degree k

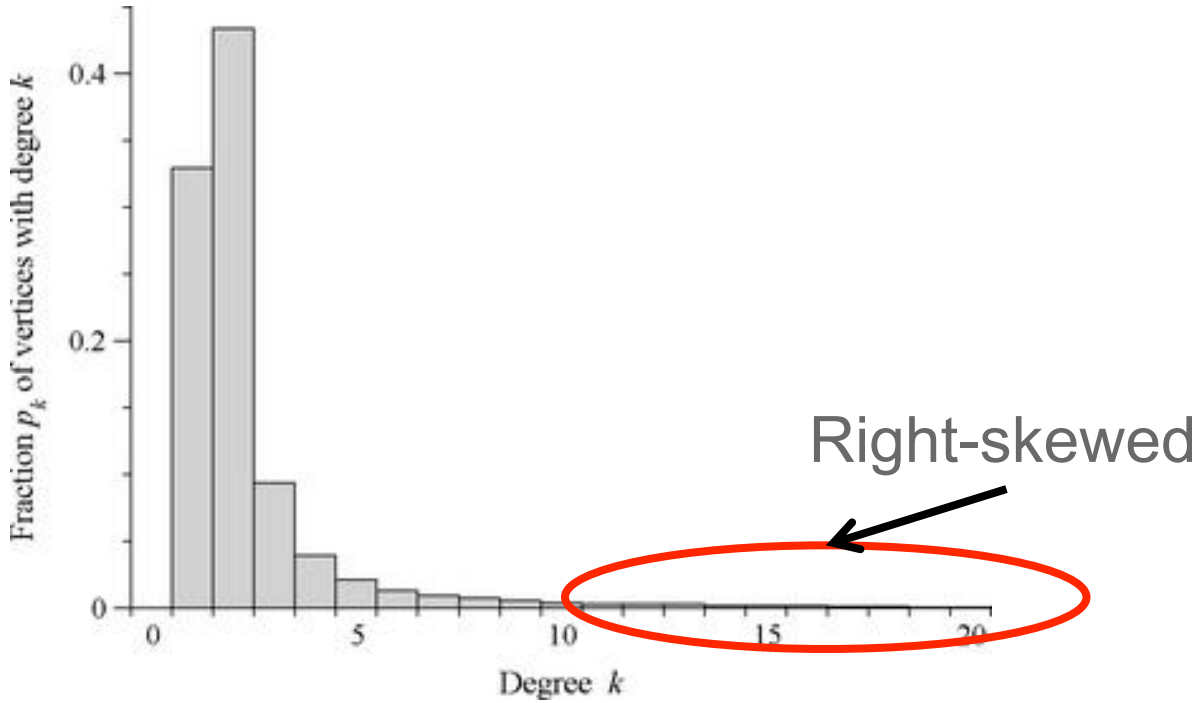


$$p_0 = \frac{1}{10}, p_1 = \frac{2}{10}, p_2 = \frac{4}{10}, p_3 = \frac{2}{10}, p_4 = \frac{1}{10}, p_k = 0 \forall k \geq 5$$

- **Degree distribution does not capture the whole structure of the network!**

Degree distribution

The Internet

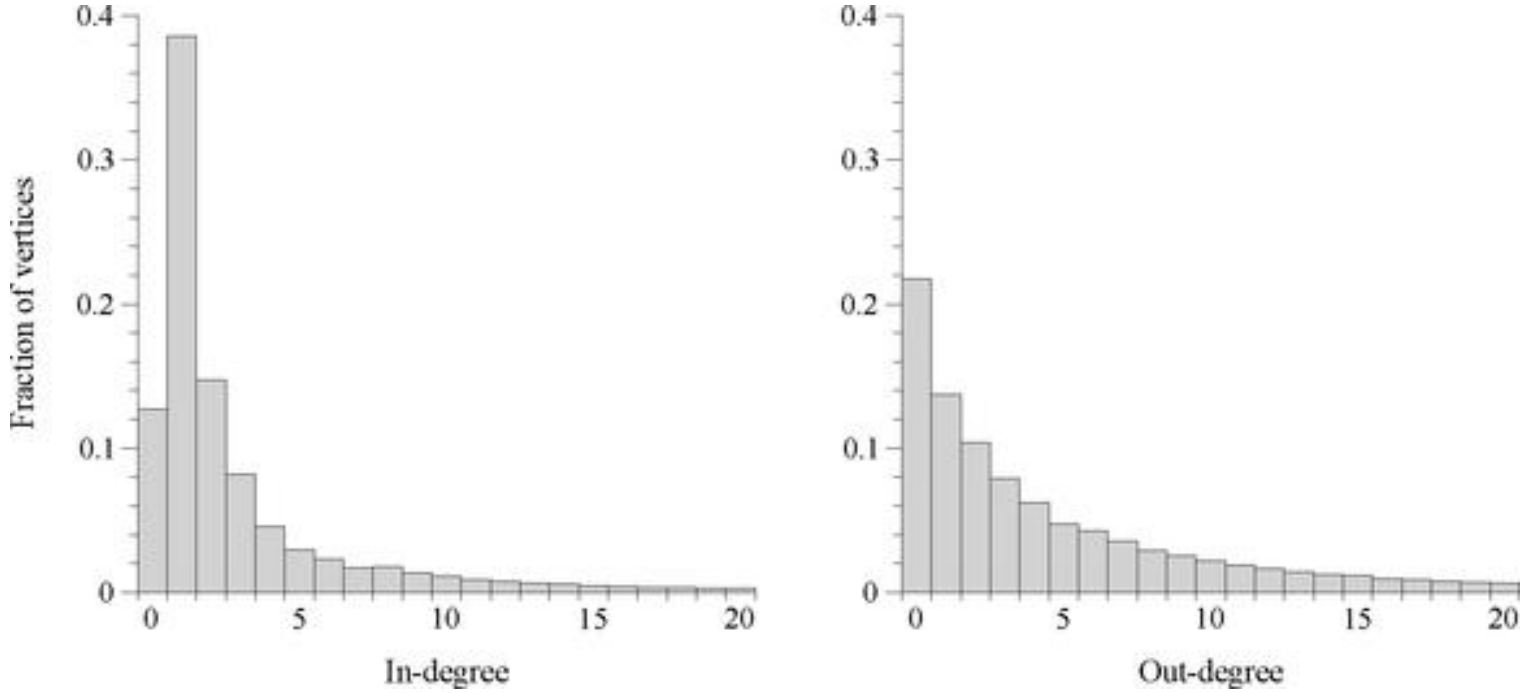


- **Many nodes with small degrees, few with extremely high**
  - Largest degree is 2407 → Since there are 19956 nodes in total this node is connected to 12% of the whole network in one hop
    - ✓Such nodes are called hubs

Degree distribution in directed networks

- **For directed networks we have both in- and out-degree distributions**

The Web



Degree distribution in directed networks

- **In directed networks we can also define a joint in- and out-degree distribution p_{jk}**
  - p_{jk} is the fraction of vertices that have simultaneously an in-degree j and an out-degree k

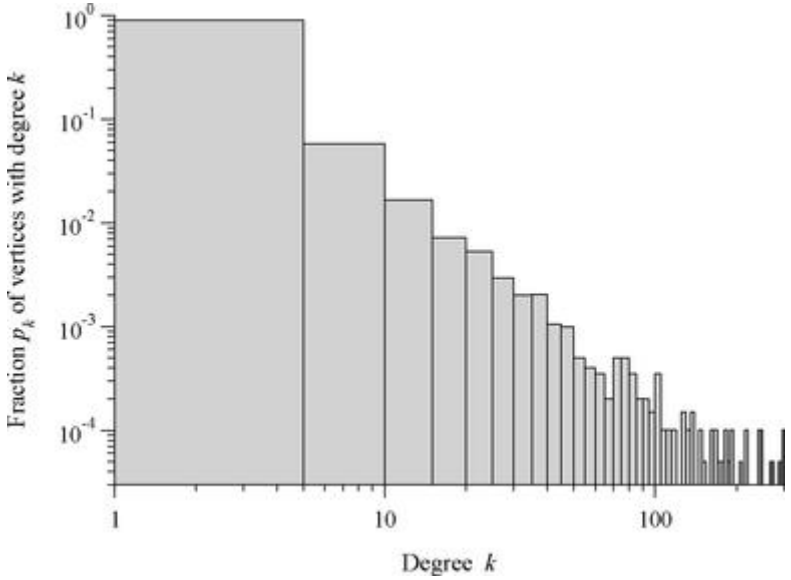
- **The joint distribution can allow to identify correlations between the in- and out-degrees**
  - This is not possible with the two separate, one-dimension, degree distributions

Power laws and scale free networks

- **If we plot the degree distribution for the Internet in log-log scale we get a straight line figure**

$$\ln p_k = -\alpha \ln k + c \Rightarrow p_k = Ck^{-\alpha}, \alpha > 0, C = e^c$$

- **Distributions of the above form are called power law**
- **α is called the exponent**
  - Typically 2≤α≤3
    - ✓Values slightly outside this range are also possible
- **The constant C is in general not interesting**
  - Used for normalization



Power laws and scale free networks

- **Real networks do not follow power law degree distribution over the whole range of k**
  - Usually when we say a degree distribution follows a power law we refer to its tail
    - ✓ Deviations from power law can appear for high values of k as well
      - E.g., cut-offs that limit the maximum degree of vertices in the tail

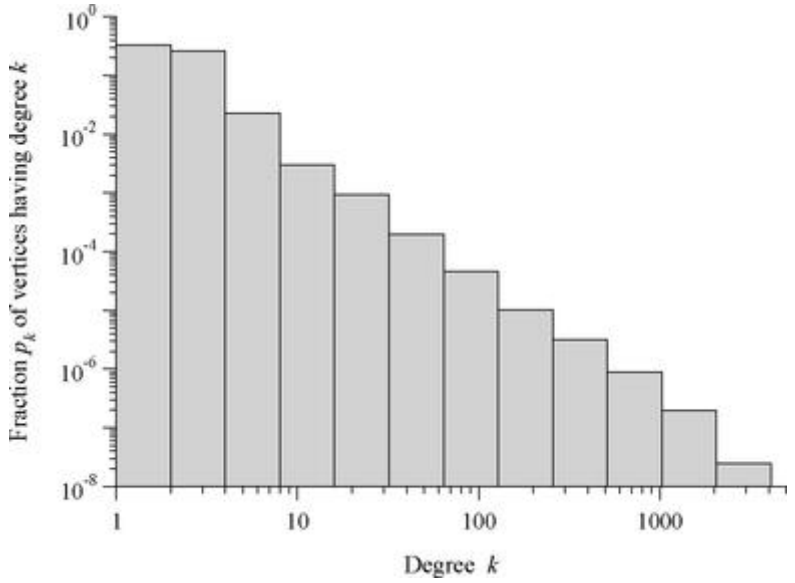
- **Networks that follow power law degree distribution are often referred to as scale-free networks**
  - Identifying scale-free from non scale-free networks is not trivial
    - ✓ Simplest – but not very accurate - strategy → log-log plot is a straight line

Detecting and visualizing power laws

- **Simply scaling logarithmically the axis of the histogram will give poor statistics at the tail of the distribution**
  - In every bin there will be only a few samples → large statistical fluctuations in the number of samples from bin to bin
- **We could use larger bins to reduce the noise at the tail**
  - However, this reduces the detail captured from the histogram (especially at the right side)
- **Try to get the best of both worlds → different bin sizes in different parts of the histogram**
  - Careful at normalizing the bins correctly!

Logarithmic binning

- **In this scheme each bin is made wider than its predecessor by a constant factor a**
  - The n-th bin will cover the range: $a^{n-1} \leq k < a^n$
  - The most common choice for a is 2
- **When plotted in log-log scale, the bins *appear* to have equal width**
- **We do not plot degree zero**



Cummulative distribution function

- **Another way to visualize power laws is through the cummulative distribution function (CDF) P_k**

$$P_k = \sum_{k'=k}^{\infty} p_{k'}$$

  - P_k is the fraction of vertices that have degree k or greater

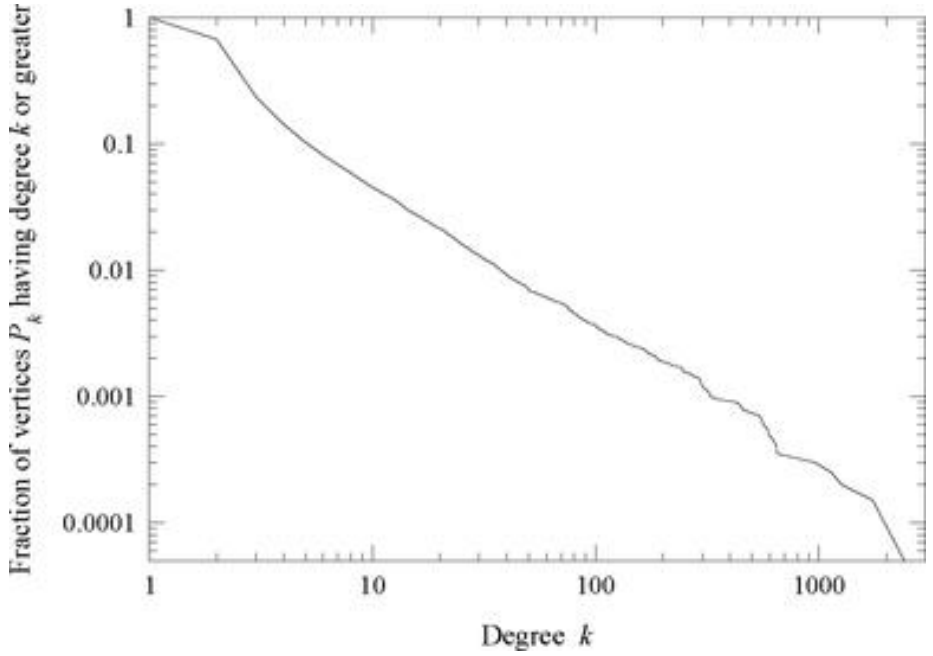
- **Let's assume that the degree distribution follows power law at the tail (i.e., for $k \geq k_{min}$). Then:**

$$P_k = C\sum_{k'=k}^{\infty} k'^{-\alpha} \cong C\int_{k}^{\infty} k'^{-\alpha} dk' = \frac{C}{\alpha - 1} k^{-(\alpha-1)}, \alpha > 1, k \geq k_{\min}$$

  - CDF follows a power law as well!

Cummulative distribution function

- **Hence, we can visualize the CDF in log-log scales**
  - No need for binning
    - ✓ Hence, we are not throwing any information as we do when we bin the normal histogram
  - Easy to compute from data



CDF disadvantages

- **Less easy to interpret as compared to normal histograms**

- **Successive points on a CDF plot are correlated**
  - Adjacent values are not independent
  - Not appropriate to extract the value of the exponent by standard techniques (e.g., least squares) that assume independence between data points
    - ✓ In general it is not a good practice to fit straight line to either CDF or normal histograms
      - Biased estimations for different reasons

Power law exponent

- **The exponent α can be estimated from:**

$$\alpha = 1 + N[\sum_i \ln \frac{k_i}{k_{\min} - \frac{1}{2}}]^{-1}$$

  - k_{min} is the minimum degree for which the power law holds
  - N is the number of vertices with degree greater than or equal to k_{min}

- **The statistical error on the estimation of α is:**

$$\sigma = \sqrt{N}[\sum_i \ln \frac{k_i}{k_{\min} - \frac{1}{2}}]^{-1} = \frac{\alpha - 1}{\sqrt{N}}$$

Properties of power-law distributions

- **Power laws appear in a wide variety of places**
  - Size of city population
  - Earthquakes
  - Use of words of a given language
  - Number of papers scientists write
  - Number of hits on web pages
  - Personal names
  - Sales of books
  - Income of people
  - …

Normalization

- Constant C is computed through the requirement that the sum of all probabilities for the different degrees must be 1:

$$\sum_{k=0}^{\infty} p_k = 1$$

- In a pure power-law distribution degree of zero is not allowed. Hence, the above sum should start from k=1. Then:

$$C = \frac{1}{\sum_{k=1}^{\infty} k^{-\alpha}} = \frac{1}{\zeta(\alpha)} \Rightarrow p_k = \frac{k^{-\alpha}}{\zeta(a)}, k > 0, p_0 = 0$$

- If the distribution deviates for small values of k, then the above constant is not correct!

Normalization

- **When we are interested in the tail of the distribution, we can discard the rest of the data**
  - We normalize over only the tail, starting from the minimum value k_{min} for which the power-law holds:

$$p_k = \frac{k^{-\alpha}}{\sum_{k=k_{\min}}^{\infty} k^{-\alpha}} = \frac{k^{-\alpha}}{\zeta(\alpha, k_{\min})}$$

→ Incomplete zeta function

- **If we approximate the sum over k at the tail of the distribution with an integral we have:**

$$C \cong \frac{1}{\int_{k_{\min}}^{\infty} k^{-\alpha}\, dk} = (\alpha - 1) k_{\min}^{\alpha - 1} \Rightarrow p_k \cong \frac{\alpha - 1}{k_{\min}} \left(\frac{k}{k_{\min}}\right)^{-\alpha}$$

Moments

- **The m-th moment of the degree distribution is given by:**

$$< k^m >= \sum_{k=0}^{\infty} k^m p_k$$

- **If a power-law is followed by the degree distribution at the tail we have:**

$$< k^m >= \sum_{k=0}^{k_{\min}-1} k^m p_k + C \sum_{k=k_{\min}}^{\infty} k^{m-\alpha}$$

  - Approximating the sum at the tail of the distribution with an integral we have:

$$< k^m > \cong \sum_{k=0}^{k_{\min}-1} k^m p_k + C \int_{k_{\min}}^{\infty} k^{m-\alpha} dk = \sum_{k=0}^{k_{\min}-1} k^m p_k + \frac{C}{m-\alpha+1} [k^{m-\alpha+1}]_{k_{\min}}^{\infty}$$

  - If ***m-α+1≥0*** then the m-th moment is not well defined (i.e., diverges)

Moments

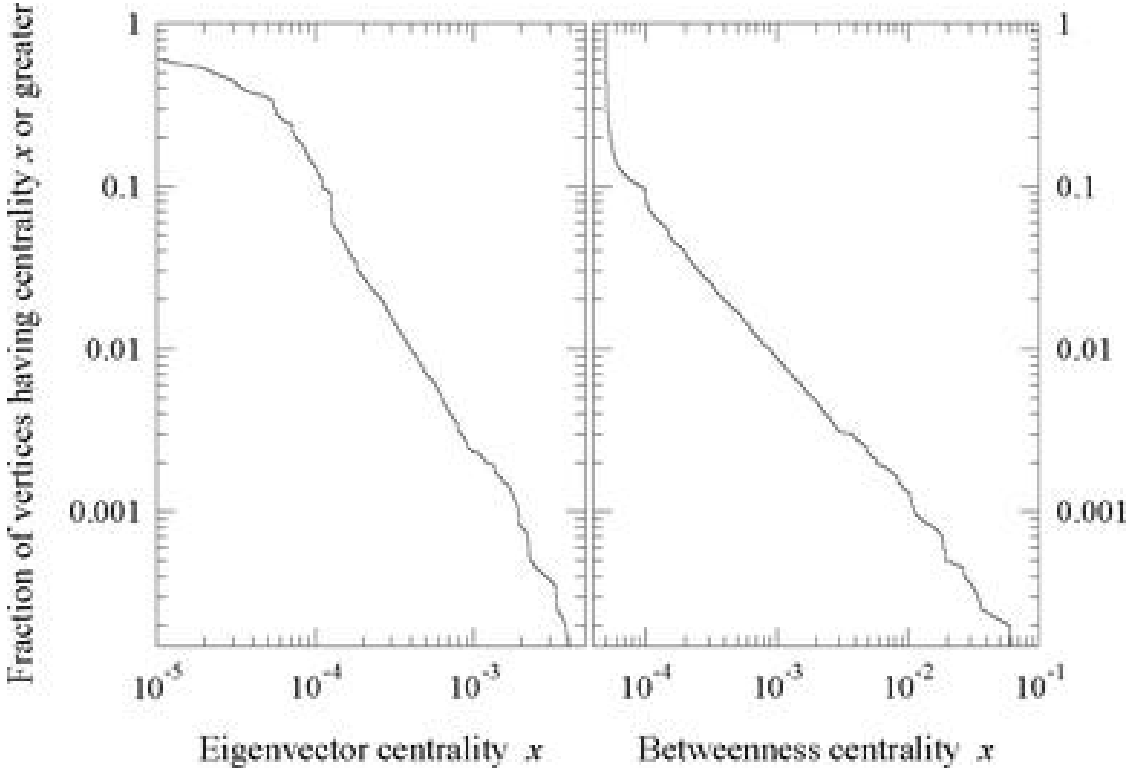
- **Given that real networks exhibit exponent 2≤α≤3, for these the second moment is not defined**
  - This is true even if the power-law holds only for the tail of the distribution

- **What does this mean for network datasets?**
  - In any real network all the moments of the degree distribution will actually be finite and calculated by: $< k^m >= \frac{1}{n} \sum_{i=1}^{n} k_i^m$
    - ✓ Since k_i is finite, the sum is finite too
    - ✓ Also since we have finite, simple networks, the maximum value that the degree can get is k=n → $< k^m > \sim [k^{m-\alpha+1}]_{k_{\min}}^{n} \sim n^{m-\alpha+1}$
  - What does the divergence mean then?

Distributions of other centrality measures

- **Eigenvector and betweeness centrality often have a highly right-skewed distribution**
  - Not necessarily power law though

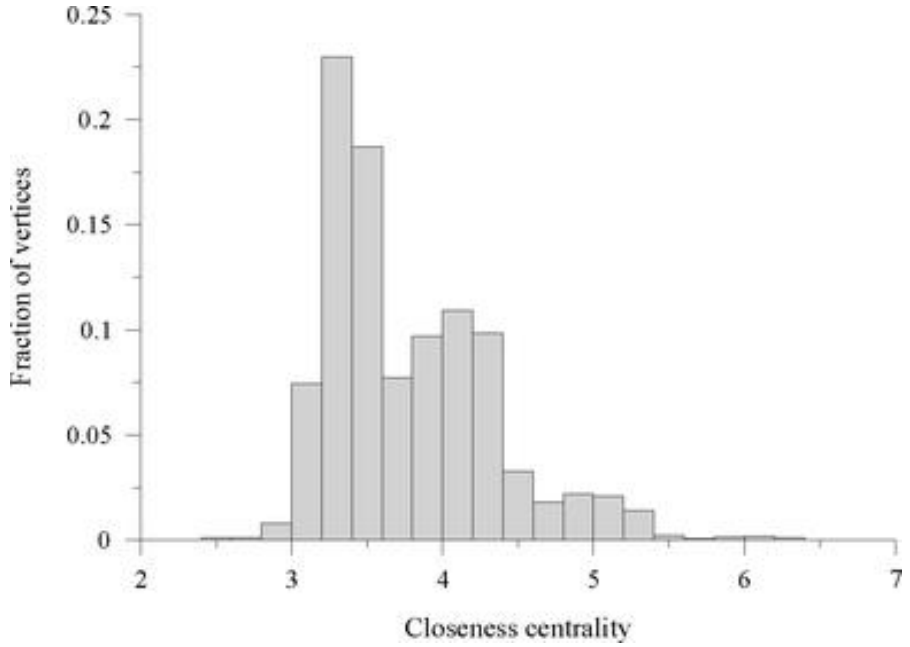
Internet



Distributions of other centrality measures

- **Closeness centrality does not exhibit skewed distribution**
  - Closeness centrality takes values in a small range (1 to logn) and hence there cannot be long tail

Internet



Clustering coefficients

- **Given a network with a specific degree distribution the expected clustering coefficient is given by:**

$$C = \frac{1}{n}\frac{[< k^2 > - < k >^2]^2}{< k >^3}$$

  - In general when the two first moments of the degree distribution are finite, then as we increase n, the clustering coefficient takes very small values

- **However, in many real networks the clustering coefficient takes much different values (lower or larger) from the expected one**

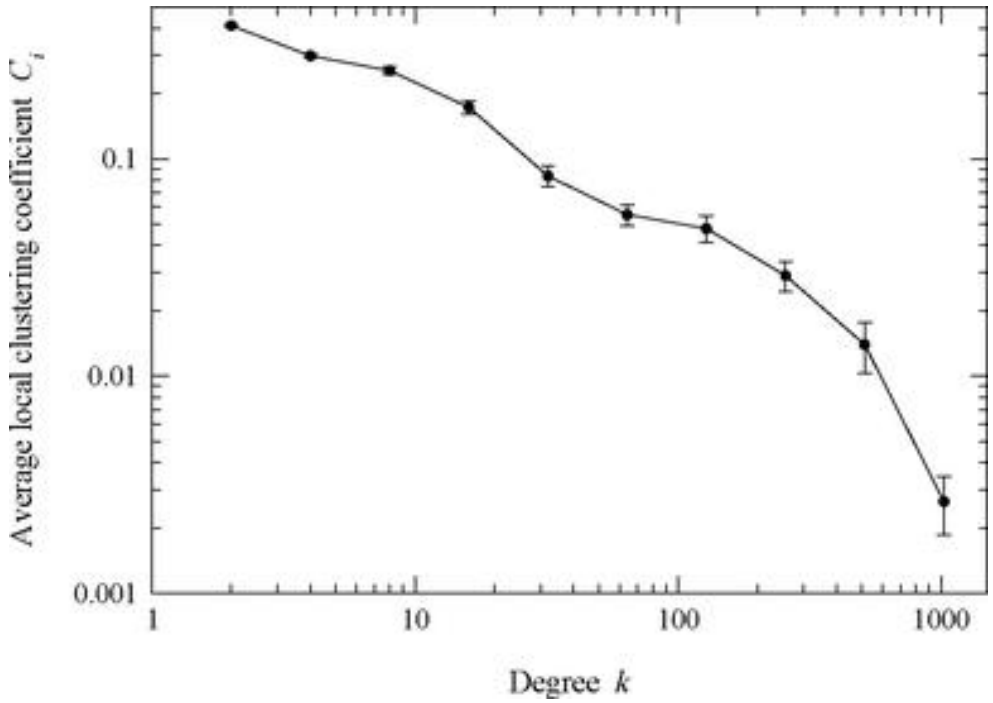
  - The exact reason for this phenomenon is not well understood, but it may be connected with the formation of groups or communities

    - ✓ E.g., in social networks → triadic closure

Local clustering coefficient

- **If we calculate the clustering coefficient of all vertices of a network an interesting pattern occurs**
  - On average vertices of higher degree exhibit lower local clustering

Internet



Assortative mixing by degree

- **In general the absolute values of the assortativity coefficient r are not large**
- **Technological networks tend to have negative r, while social networks tend to have positive r**
- **In general, for simple networks, due to the limited number of possible edges between high degree nodes, one should expect (in the absence of other biases) disassortative mixing**
  - Social networks?
    - ✓ There might be other biases in place that cause a slight assortative mixing