Online Social Networks and Media

Network Measurements

Measuring Networks

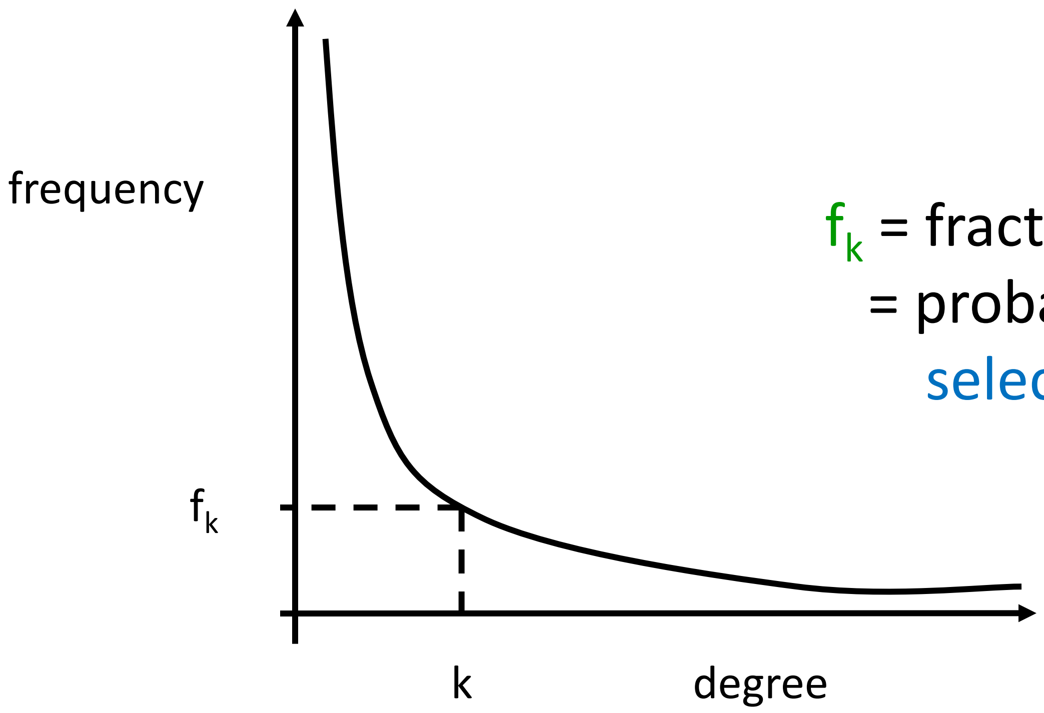
- Degree distributions and power-laws
- Clustering Coefficient
- Small world phenomena
- Components
- Motifs
- Homophily

The basic random graph model

- The measurements on real networks are usually compared against those on "random networks"

- The basic $G_{n,p}$ (Erdös-Renyi) random graph model:
  - n : the number of vertices
  - $0 \leq p \leq 1$
  - for each pair (i,j), generate the edge (i,j) independently with probability p
  - Expected degree of a node: $z = np$

Degree distributions



f_k = fraction of nodes with degree k
= probability of a randomly selected node to have degree k

- Problem: find the probability distribution that best fits the observed data

Power-law distributions

- The degree distributions of most real-life networks follow a power law

$$p(k) = Ck^{-\alpha}$$

- Right-skewed/Heavy-tail distribution
  - there is a non-negligible fraction of nodes that has very high degree (hubs)
  - scale-free: no characteristic scale, average is not informative

- In stark contrast with the random graph model!
  - Poisson degree distribution, z=np

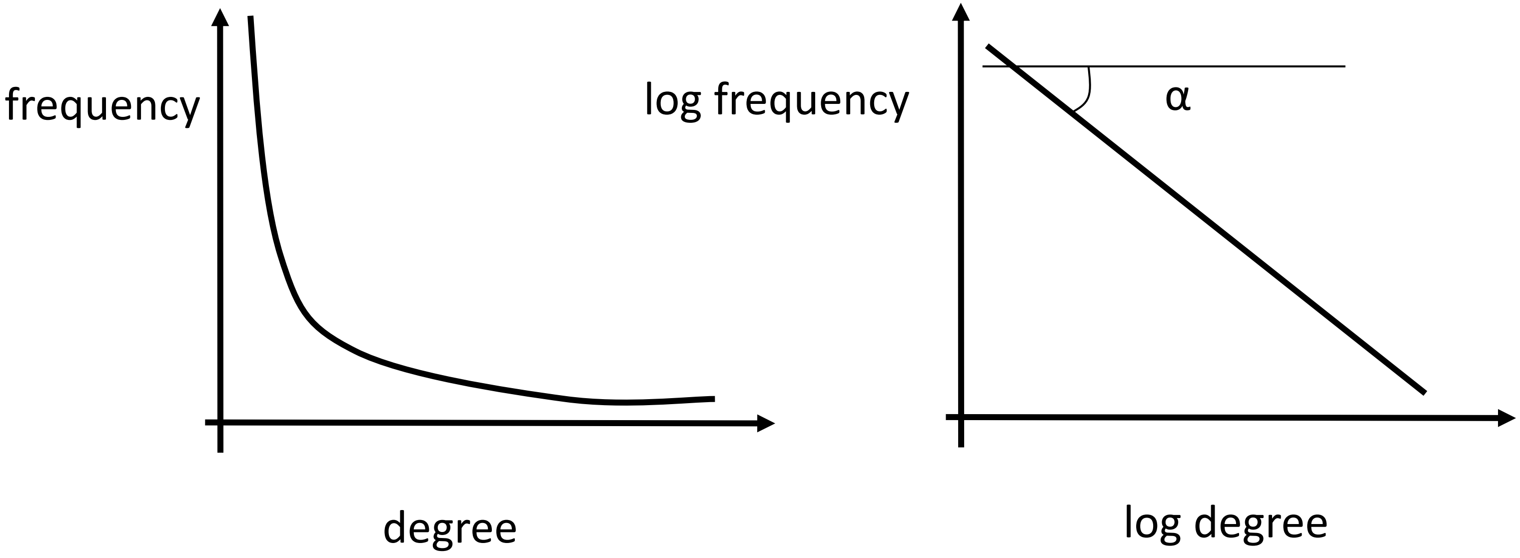
$$p(k) = P(k;z) = \frac{z^k}{k!} e^{-z}$$

  - highly concentrated around the mean
  - the probability of very high degree nodes is exponentially small

Power-law signature

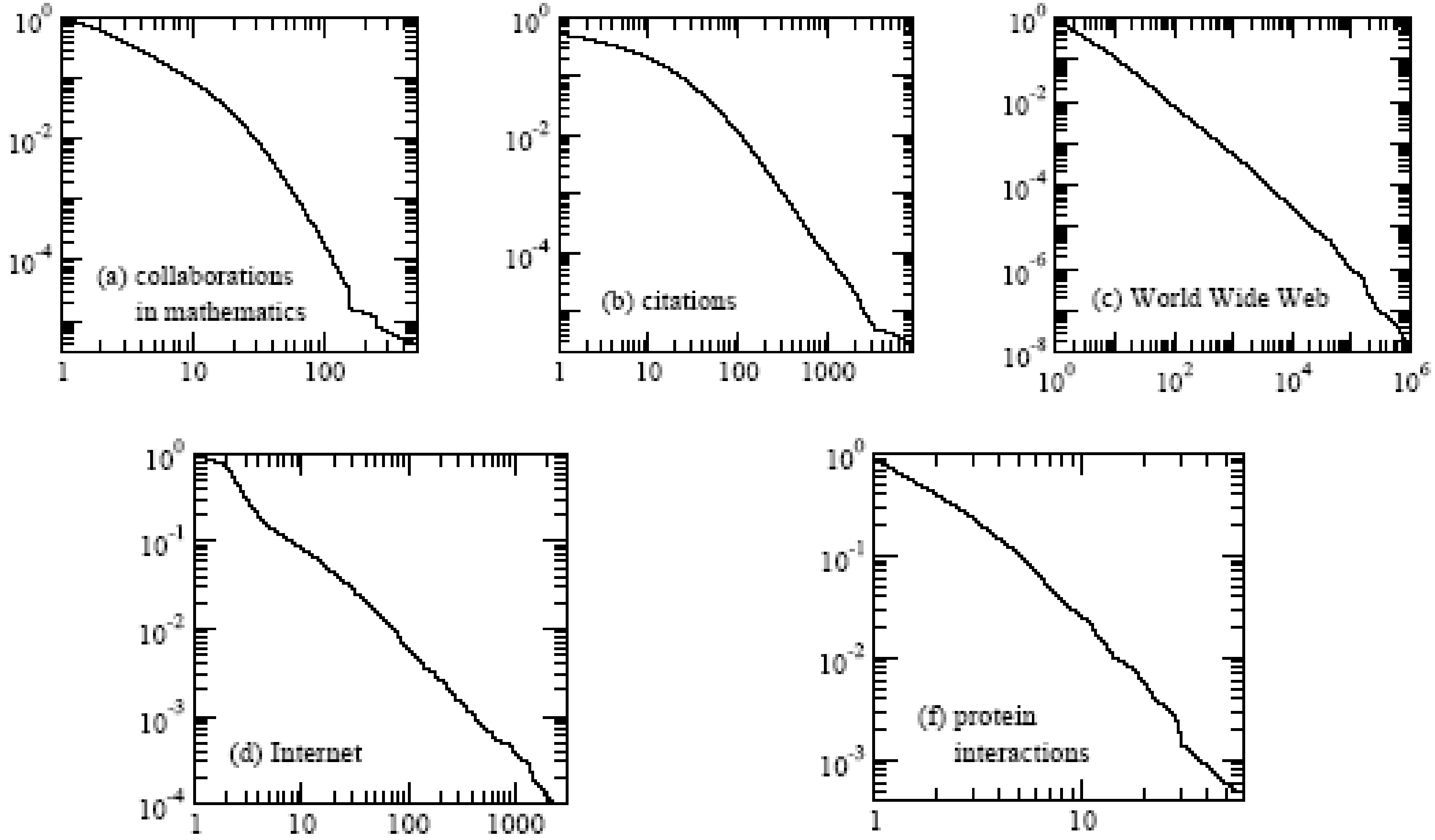
- Power-law distribution gives a line in the log-log plot

$$\log p(k) = -\alpha \log k + \log C$$



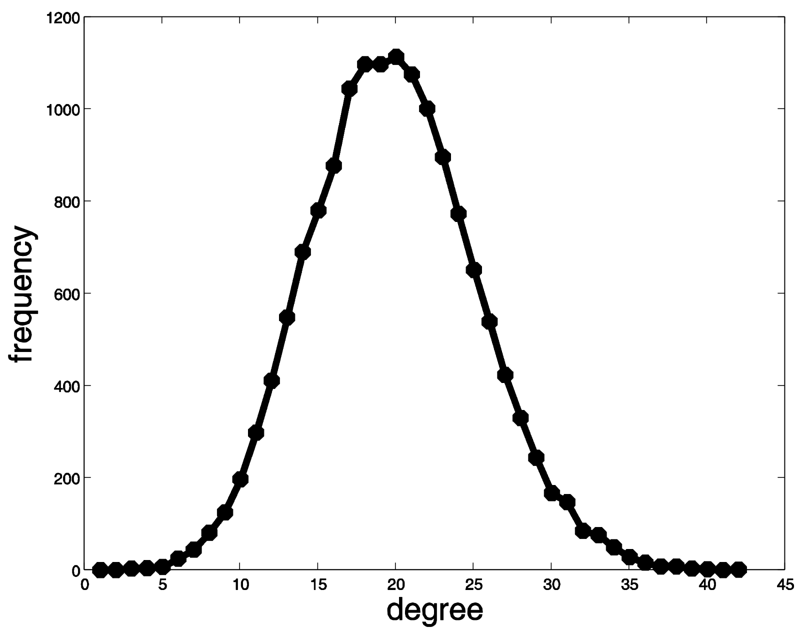
- α : power-law exponent (typically $2 \leq \alpha \leq 3$)

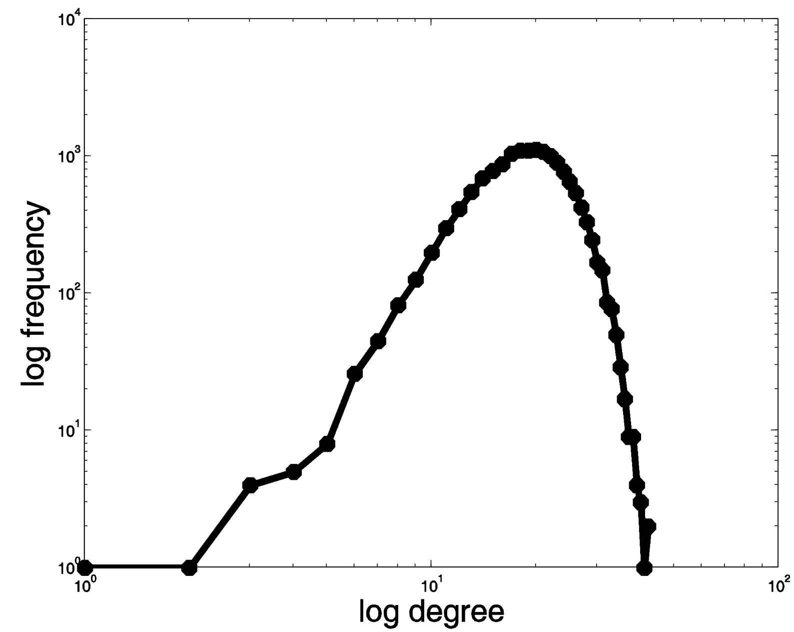
Examples



Taken from [Newman 2003]

A random graph example





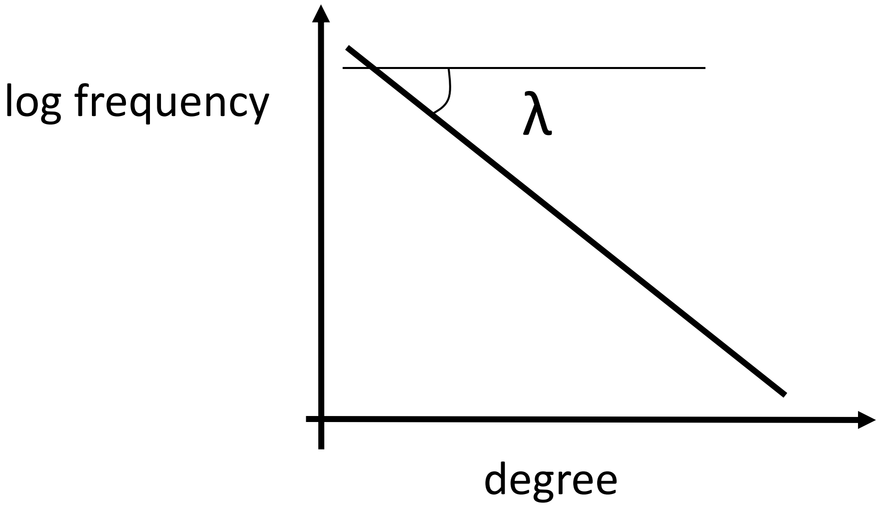
Exponential distribution

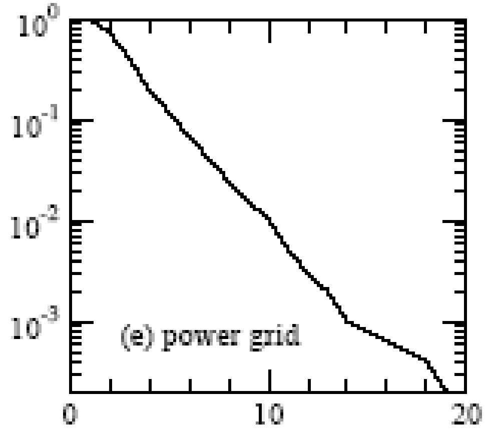
- Observed in some technological or collaboration networks

$$p(k) = \lambda e^{-\lambda k}$$

- Identified by a line in the log-linear plot

$$\log p(k) = -\lambda k + \log \lambda$$





Measuring power-laws

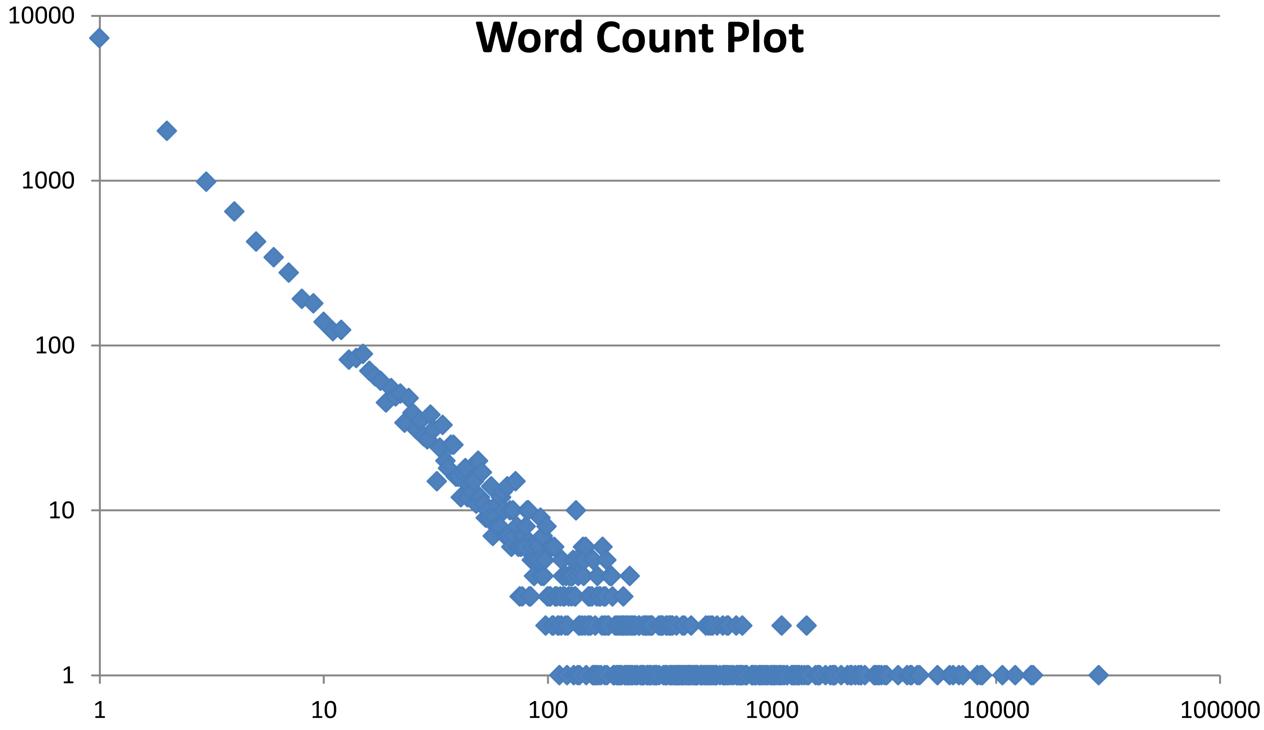
- How do we create these plots? How do we measure the power-law exponent?

- Collect a set of measurements:
  - E.g., the degree of each page, the number of appearances of each word in a document, the size of solar flares(continuous)

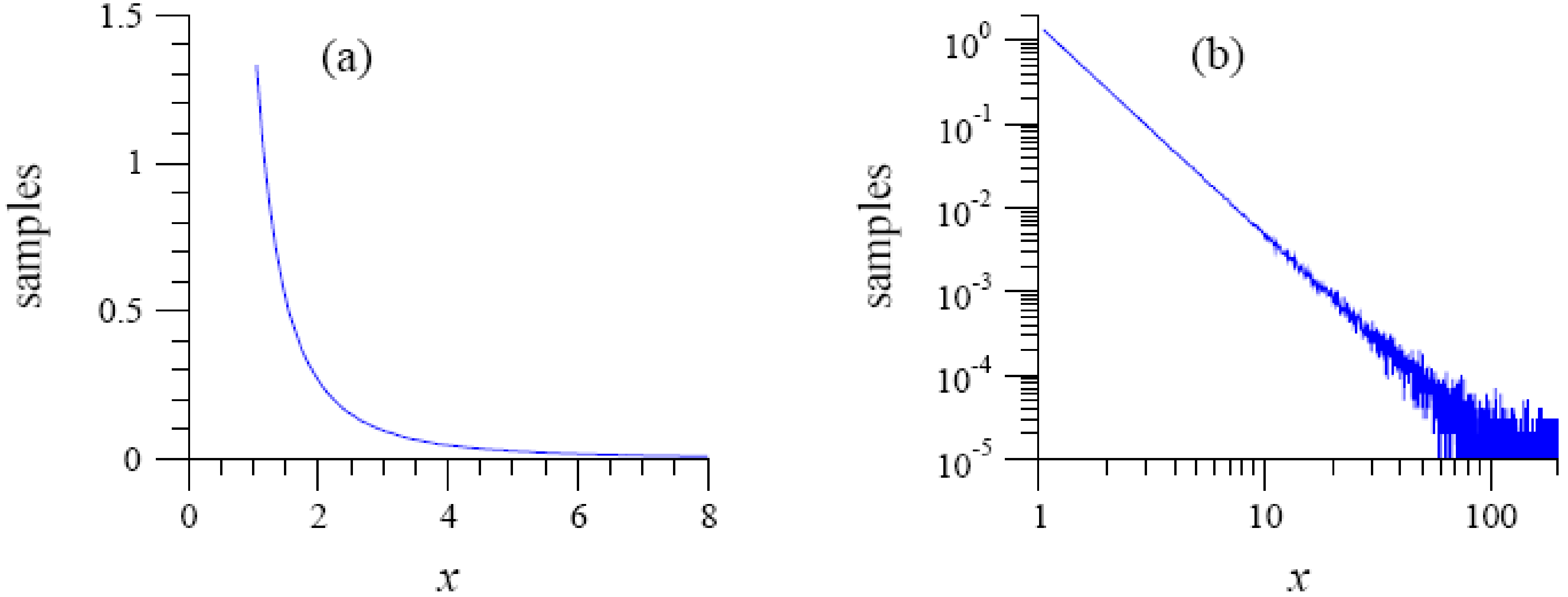
- Create a value histogram
  - For discrete values, number of times each value appears
  - For continuous values (but also for discrete):
    - Break the range of values into bins of equal width
    - Sum the count of values in the bin
    - Represent the bin by the mean (median) value

- Plot the histogram in log-log scale
  - Bin representatives vs Value in the bin

Discrete Counts



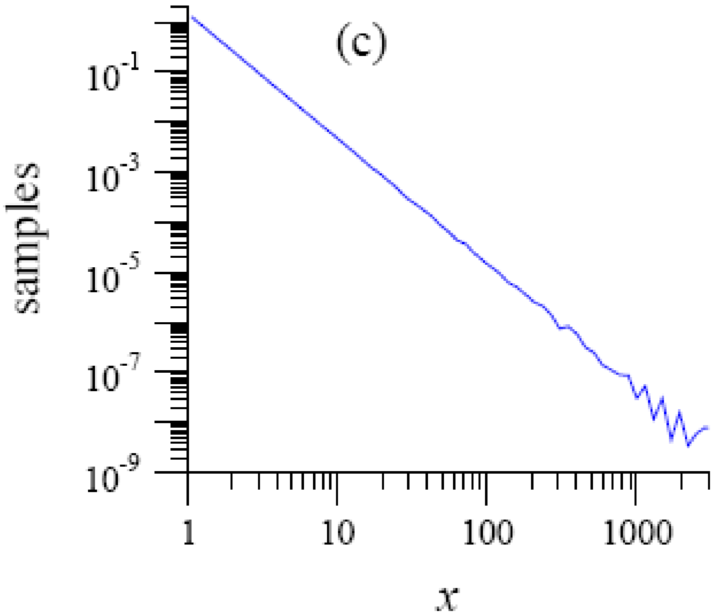
Measuring power laws



Simple binning produces a noisy plot

Logarithmic binning

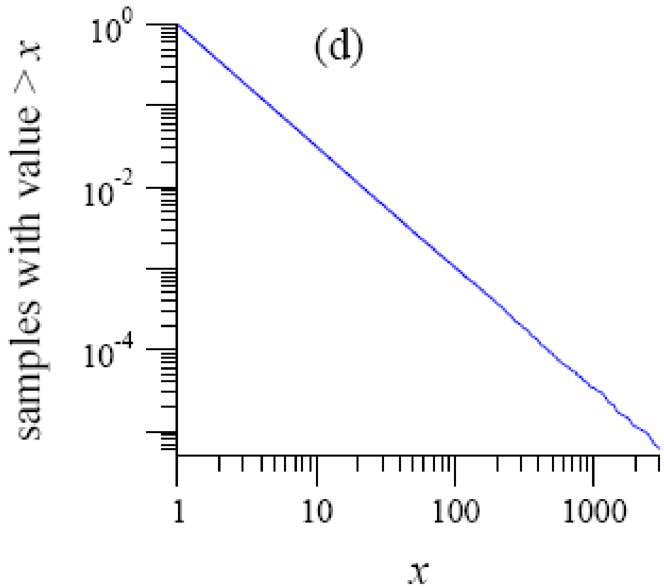
- Exponential binning
  - Create bins that grow exponentially in size
  - In each bin divide the sum of counts by the bin length (number of observations per bin unit)

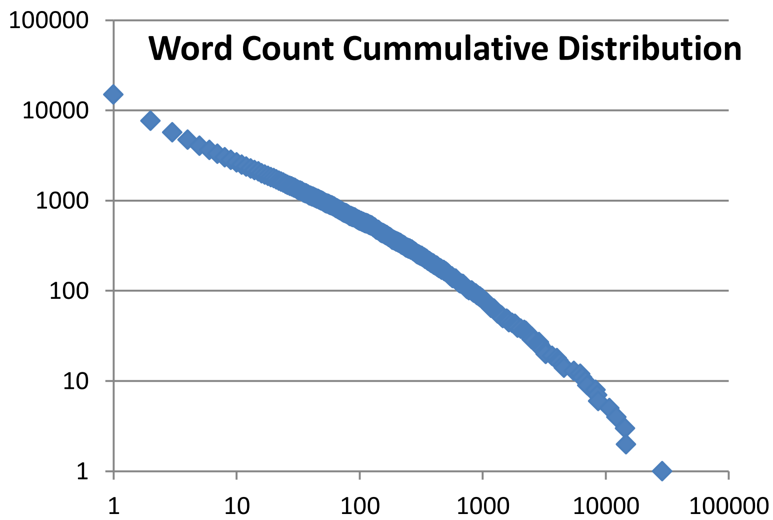


Still some noise at the tail

Cumulative distribution

- Compute the cumulative distribution
  - P[X≥x]: fraction (or number) of observations that have value at least x
  - It also follows a power-law with exponent α-1





Pareto distribution

- A random variable follows a Pareto distribution if

$$P[X \geq x] = C' x^{-\beta} \qquad x \geq x_{min}$$

- Power law distribution with exponent $\alpha = 1 + \beta$

Zipf plot

- There is another easy way to see the power-law, by doing the Zipf plot
  - Order the values in decreasing order
  - Plot the values against their rank in log-log scale
    - i.e., for the r-th value x_r, plot the point $(\log(r),\log(x_r))$
  - If there is a power-law you should see something like a straight line

Zipf's Law

- A random variable X follows Zipf's law if the r-th largest value x_r satisfies

$$x_r \approx r^{-\gamma}$$

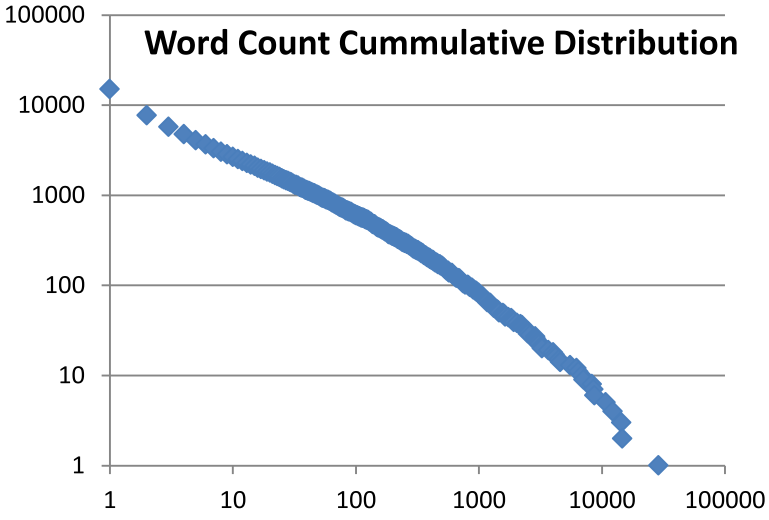
- Same as Pareto distribution

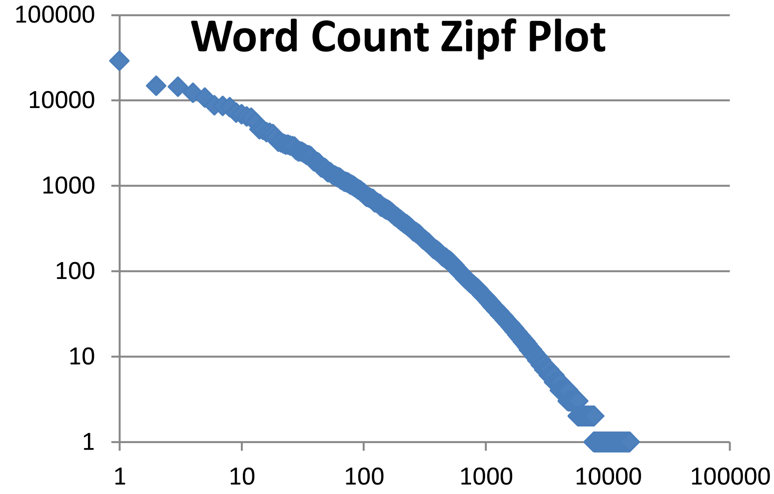
$$P[X \geq x] \approx x^{-1/\gamma}$$

- X follows a power-law distribution with $\alpha = 1 + 1/\gamma$

- Named after Zipf, who studied the distribution of words in English language and found Zipf law with exponent 1

Zipf vs Pareto





Computing the exponent

- Maximum likelihood estimation
  - Assume that the set of data observations **x** are produced by a power-law distribution with some exponent α
    - Exact law: $p(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha}$
  - Find the exponent that maximizes the probability P(α|**x**)

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

Collective Statistics (M. Newman 2003)

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW nd.edu	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law (or "–" if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient r, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.

Power Laws - Recap

- A (continuous) random variable X follows a power-law distribution if it has density function

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

- A (continuous) random variable X follows a Pareto distribution if it has cumulative function

$$P[X \geq x] = Cx^{-\beta}$$ power-law with $\alpha=1+\beta$

- A (discrete) random variable X follows Zipf's law if the frequency of the r-th largest value satisfies

$$p_r = Cr^{-\gamma}$$ power-law with $\alpha=1+1/\gamma$

Average/Expected degree

- For power-law distributed degree
  - if $\alpha \geq 2$, it is a constant

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

  - if $\alpha < 2$, it diverges
    - The expected value goes to infinity as the size of the network grows
- The fact that $\alpha \geq 2$ for most real networks guarantees a constant average degree as the graph grows

Maximum degree

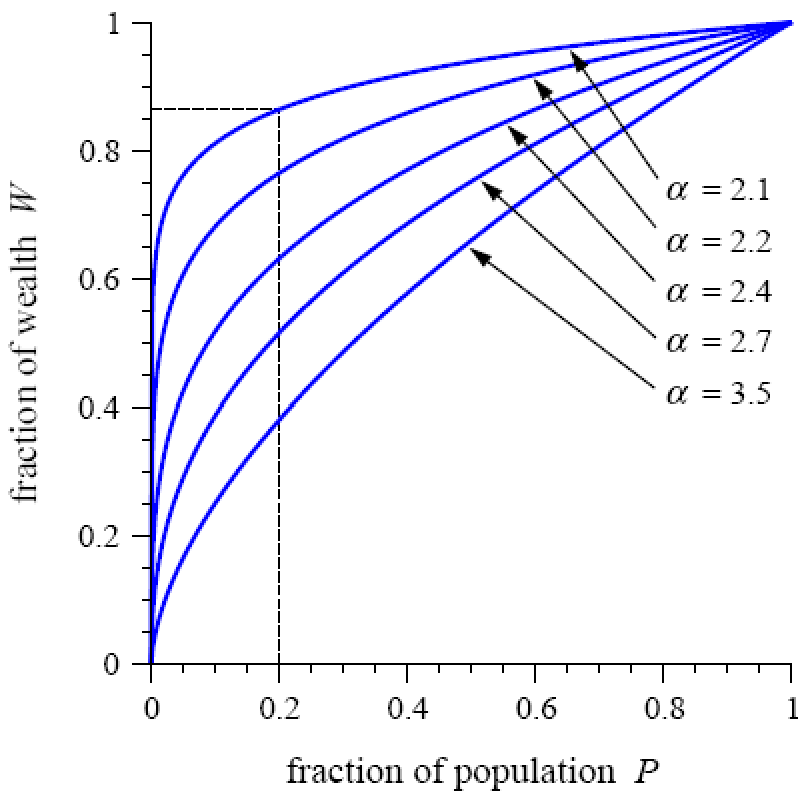
- For random graphs, the maximum degree is highly concentrated around the average degree z
- For power law graphs

$$k_{max} \approx n^{1/(\alpha-1)}$$

- Rough argument: solve $nP[X \geq k]=1$

The 80/20 rule

- Top-heavy: Small fraction of values collect most of distribution mass

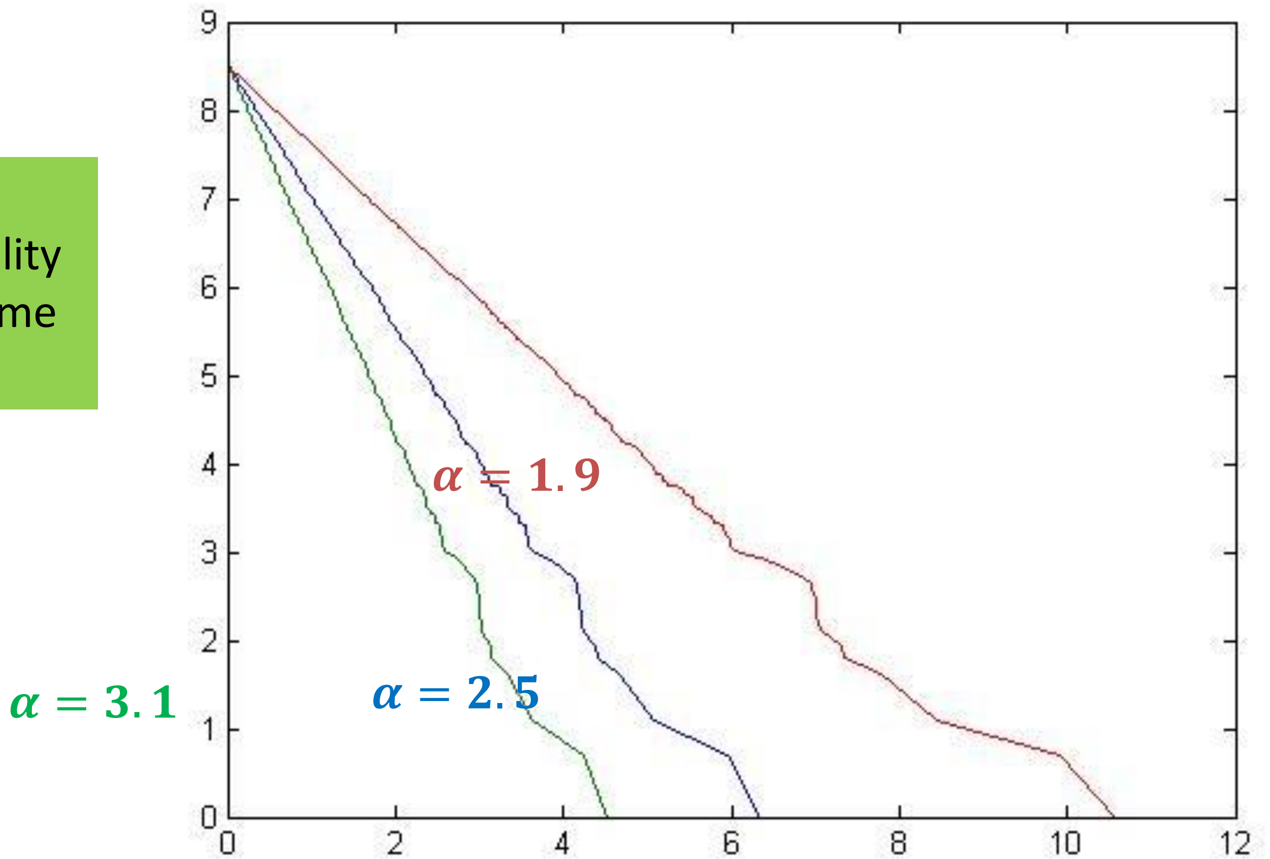


- This phenomenon becomes more extreme when $\alpha < 2$
- 1% of values has 99% of mass

- E.g. name distribution

The effect of exponent

As the exponent increases the probability of observing an extreme value decreases



Generating power-law values

- A simple trick to generate values that follow a power-law distribution:
  - Generate values r uniformly at random within the interval [0,1]
  - Transform the values using the equation
  $$x = x_{min}(1-r)^{-1/(\alpha-1)}$$
  - Generates values distributed according to power-law with exponent α

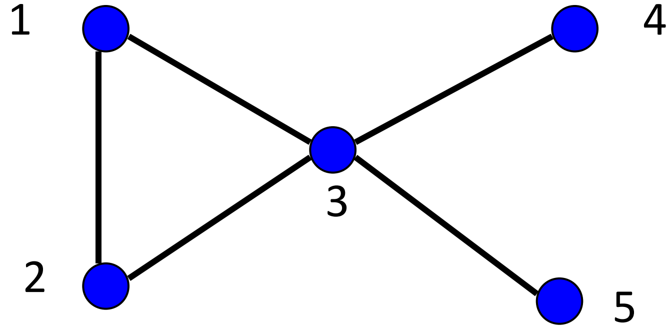
Clustering (Transitivity) coefficient

- Measures the density of triangles (local clusters) in the graph
- Two different ways to measure it:

$$C^{(1)} = \frac{\sum_i \text{triangles centered at node } i}{\sum_i \text{triples centered at node } i}$$

- The ratio of the means

Example



$$C^{(1)} = \frac{3}{1+1+6} = \frac{3}{8}$$

Clustering (Transitivity) coefficient

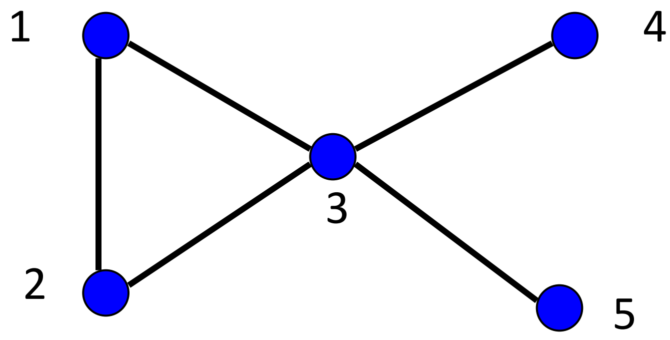
- Clustering coefficient for node i

$$C_i = \frac{\text{triangles centered at node } i}{\text{triples centered at node } i}$$

$$C^{(2)} = \frac{1}{n} C_i$$

- The mean of the ratios

Example



$$C^{(2)} = \frac{1}{5}\left(1+1+1/6\right) = \frac{13}{30}$$

$$C^{(1)} = \frac{3}{8}$$

- The two clustering coefficients give different measures
- $C^{(2)}$ increases with nodes with low degree

Collective Statistics (M. Newman 2003)

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW nd.edu	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law (or "–" if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient r, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.

Clustering coefficient for random graphs

- The probability of two of your neighbors also being neighbors is p, independent of local structure
  - clustering coefficient C = p
  - when the average degree z=np is constant C =O(1/n)

Table 1: Clustering coefficients, C, for a number of different networks; n is the number of node, z is the mean degree. Taken from [146].

Network	n	z	C measured	C for random graph
Internet [153]	6,374	3.8	0.24	0.00060
World Wide Web (sites) [2]	153,127	35.2	0.11	0.00023
power grid [192]	4,941	2.7	0.080	0.00054
biology collaborations [140]	1,520,251	15.5	0.081	0.000010
mathematics collaborations [141]	253,339	3.9	0.15	0.000015
film actor collaborations [149]	449,913	113.4	0.20	0.00025
company directors [149]	7,673	14.4	0.59	0.0019
word co-occurrence [90]	460,902	70.1	0.44	0.00015
neural network [192]	282	14.0	0.28	0.049
metabolic network [69]	315	28.3	0.59	0.090
food web [138]	134	8.7	0.22	0.065

Small worlds

- Millgram's experiment: Letters were handed out to people in Nebraska to be sent to a target in Boston
- People were instructed to pass on the letters to someone they knew on first-name basis
- The letters that reached the destination followed paths of length around 6
- Six degrees of separation: (play of John Guare)

- Also:
  - The Kevin Bacon game
  - The Erdös number

- Small world project: http://smallworld.columbia.edu/index.html

Measuring the small world phenomenon

- d_{ij} = shortest path between i and j
- Diameter:

$$d = \max_{i,j} d_{ij}$$

- Characteristic path length:

$$\ell = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}$$

- Harmonic mean

Problem if no path between two nodes

$$\ell^{-1} = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}^{-1}$$

- Also, distribution of all shortest paths

Collective Statistics (M. Newman 2003)

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW nd.edu	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law (or "–" if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient r, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.

Small worlds in real networks

- For all real networks there are (on average) short paths between nodes of the network.
  - Largest path found in the IMDB actor network: 7

- Is this interesting?
  - Random graphs also have small diameter ($d=\log n/\log\log n$ when $z=\omega(\log n)$)

- Short paths are not surprising and should be combined with other properties
  - ease of navigation
  - high clustering coefficient

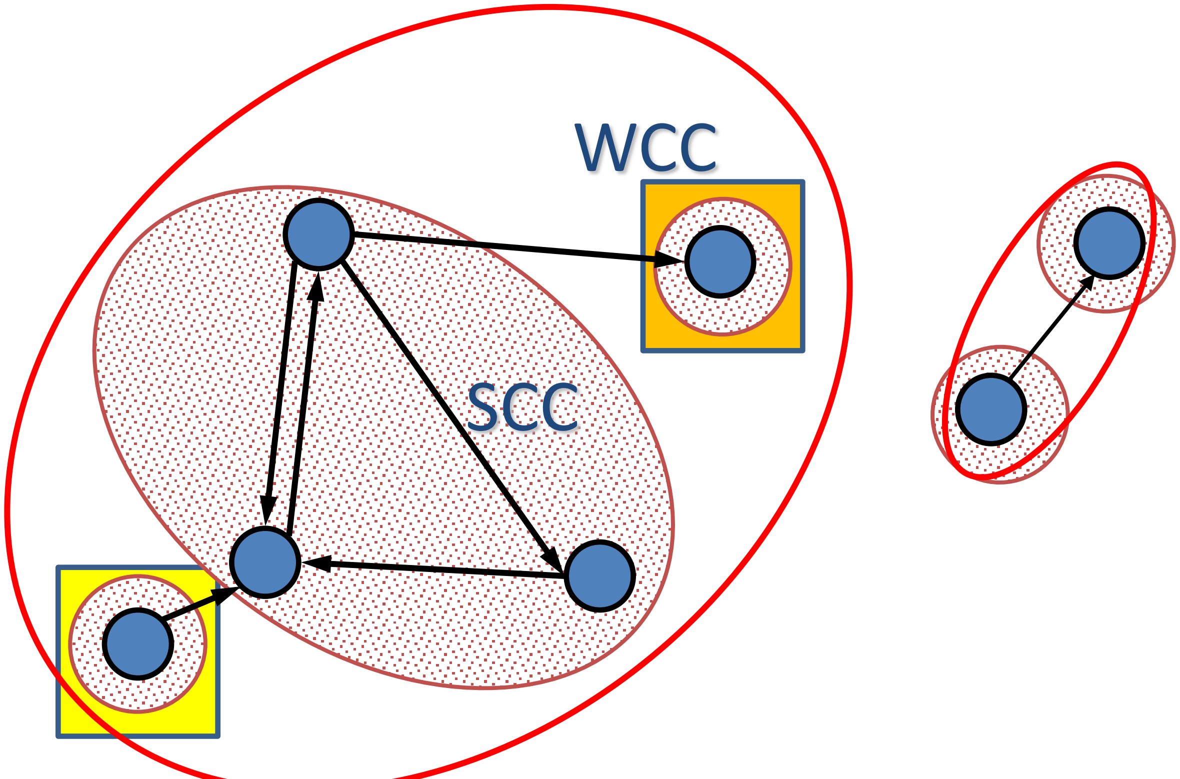
Connected components

- For undirected graphs, the size and distribution of the connected components
  - is there a giant component?
  - Most known real undirected networks have a giant component

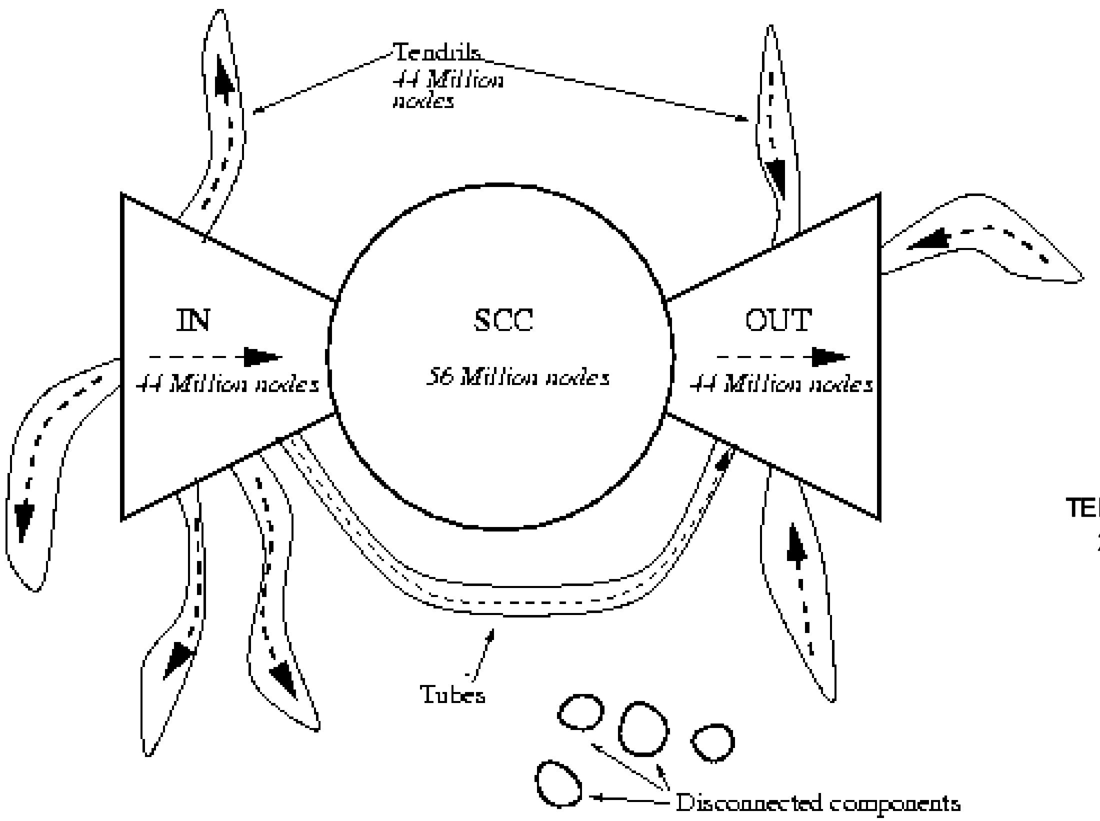
- For directed graphs, the size and distribution of strongly and weakly connected components

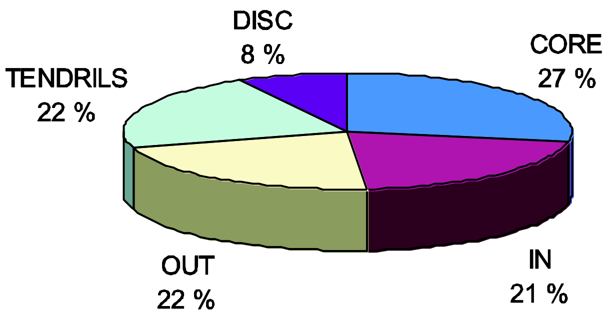
Connected components – definitions

- Weakly connected components (WCC)
  - Set of nodes such that from any node can go to any node via an undirected path
- Strongly connected components (SCC)
  - Set of nodes such that from any node can go to any node via a directed path.
  - IN: Nodes that can reach the SCC (but not in the SCC)
  - OUT: Nodes reachable by the SCC (but not in the SCC)



The bow-tie structure of the Web

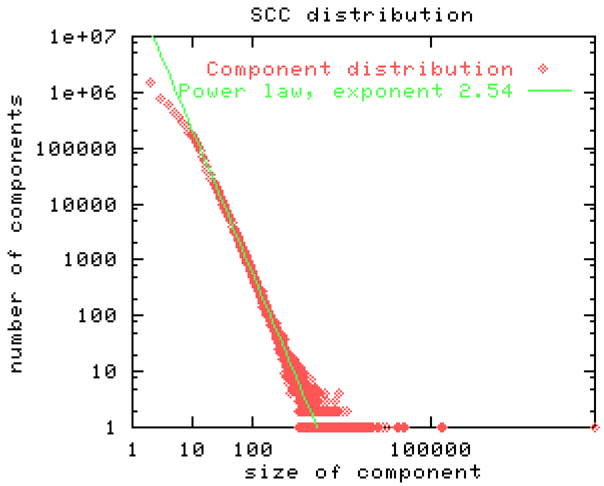


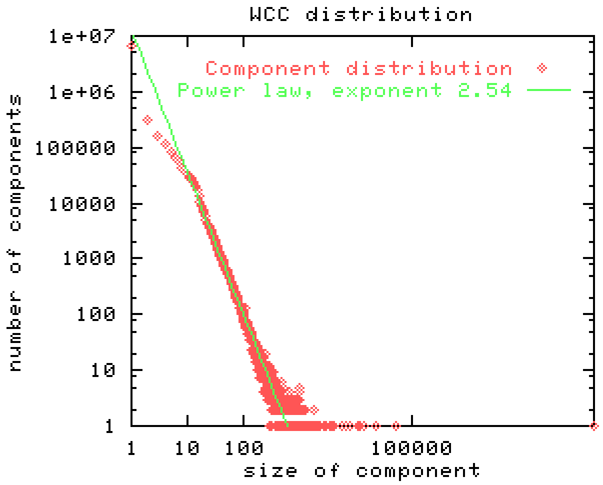


The largest weakly connected component contains 90% of the nodes

SCC and WCC distribution

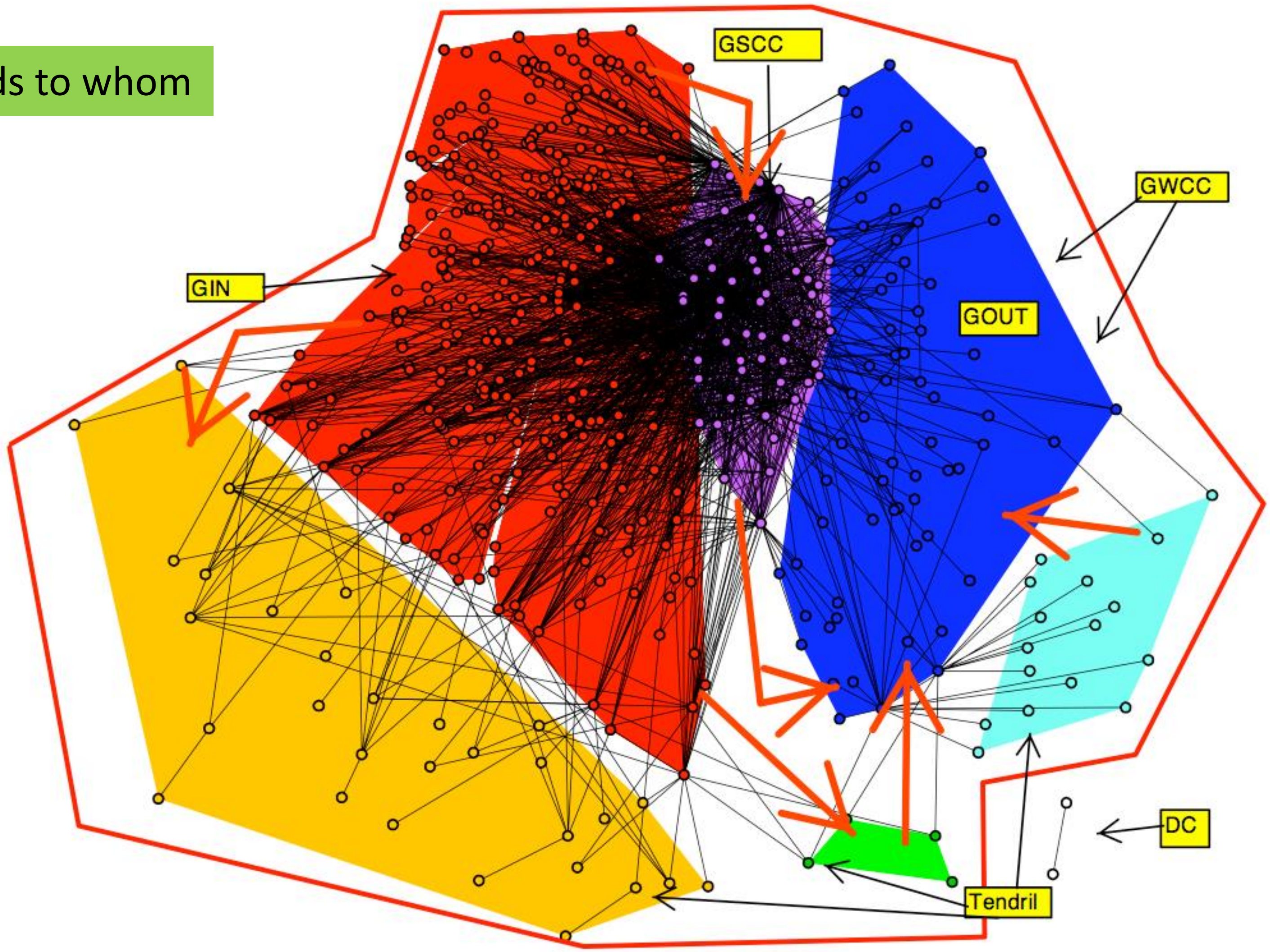
- The SCC and WCC sizes follows a power law distribution
  - the second largest SCC is significantly smaller





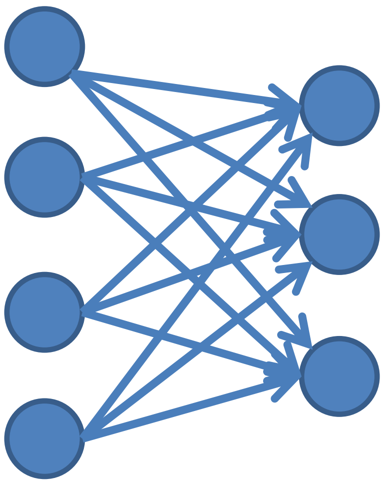
Another bow-tie

Who lends to whom



Web Cores

- Cores: Small complete bipartite graphs (of size 3x3, 4x3, 4x4)
  - Similar to the triangles for undirected graphs
- Found more frequently than expected on the Web graph
- Correspond to communities of enthusiasts (e.g., fans of japanese rock bands)



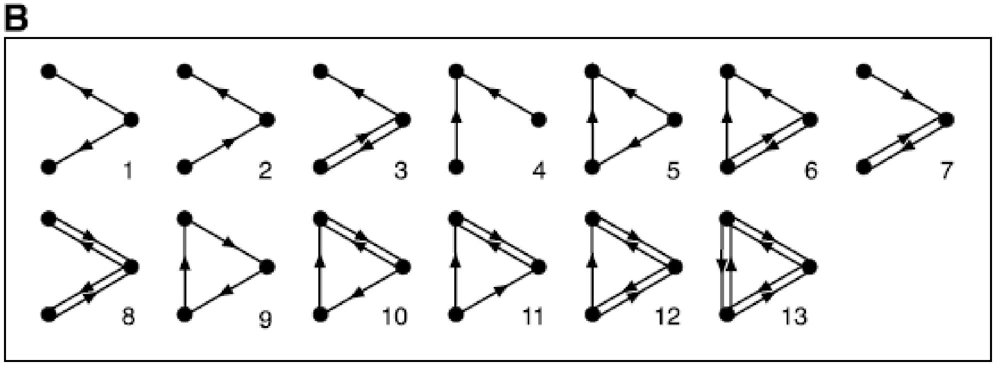
Motifs

- Most networks have the same characteristics with respect to global measurements
  - can we say something about the local structure of the networks?

- Motifs: Find small subgraphs that over-represented in the network

Example

- Motifs of size 3 in a directed graph

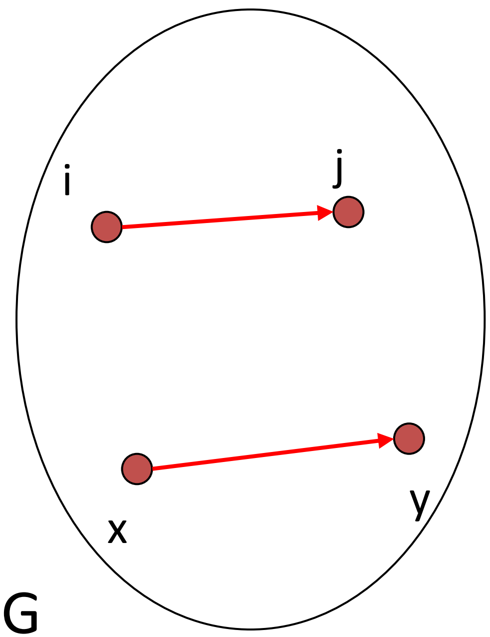


Finding interesting motifs

- Sample a part of the graph of size S
- Count the frequency of the motifs of interest
- Compare against the frequency of the motif in a random graph with the same number of nodes and the same degree distribution

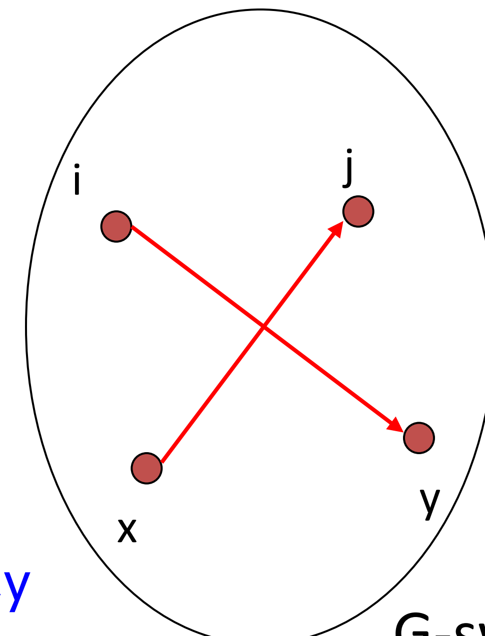
Generating a random graph

- Find edges (i,j) and (x,y) such that edges (i,y) and (x,j) do not exist, and swap them
  - repeat for a large enough number of times



G

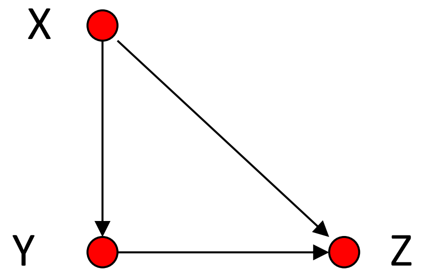
degrees of i,j,x,y are preserved

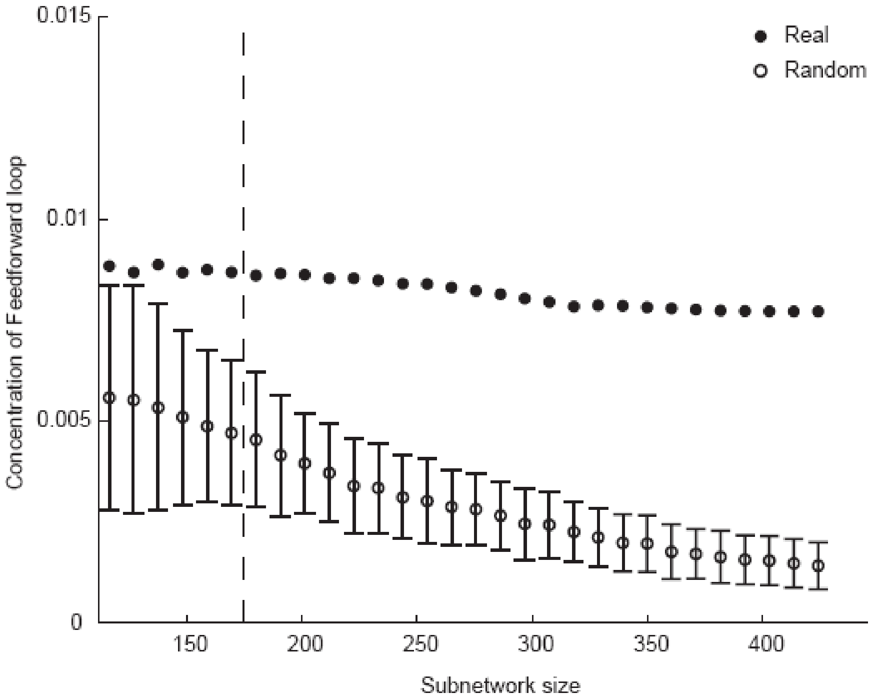


G-swapped

The feed-forward loop

- Over-represented in gene-regulation networks
  - a signal delay mechanism

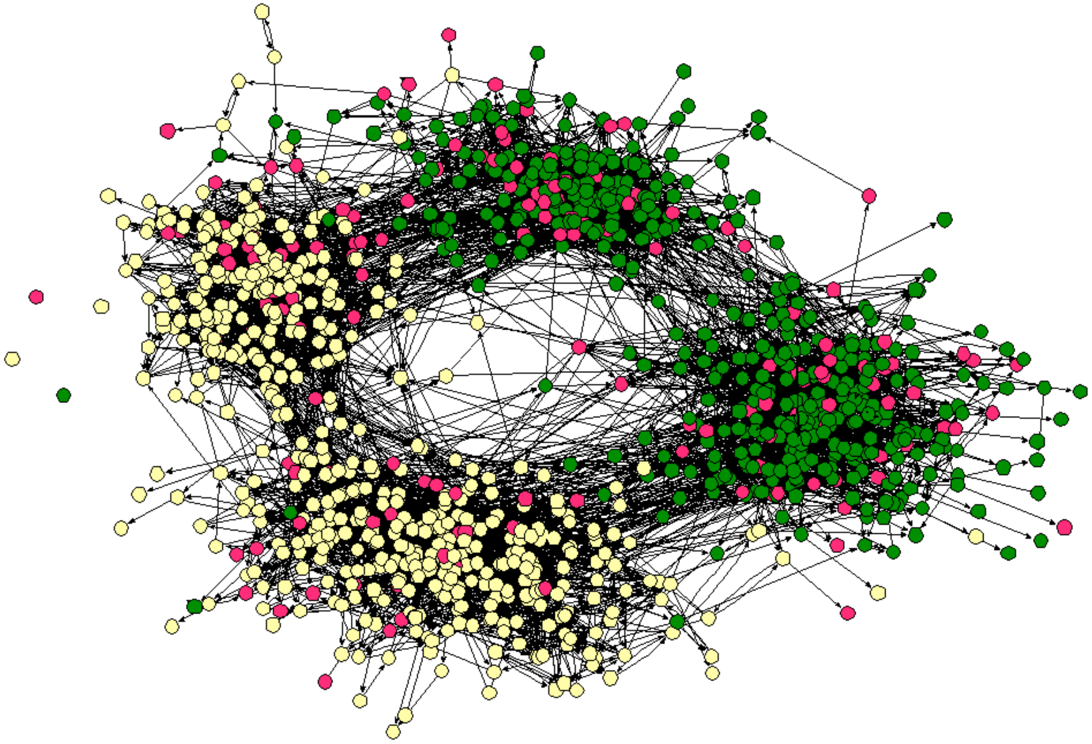




Milo et al. 2002

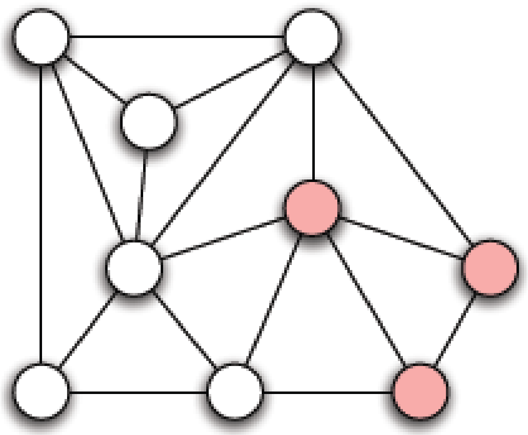
Homophily

- Love of the same: People tend to have friends with common interests
  - Students separated by race and age



Measuring homophily

- Friendships in elementary school



- The connections of people with the same interests should be higher than on a random experiment

References

- M. E. J. Newman, The structure and function of complex networks, SIAM Reviews, 45(2): 167-256, 2003
- R. Albert and A.-L. Barabási, Statistical mechanics of complex networks, *Reviews of Modern Physics* **74**, 47-97 (2002).
- S. N. Dorogovstev and J. F. F. Mendez, Evolution of Networks: From Biological Nets to the Internet and WWW.
- Michalis Faloutsos, Petros Faloutsos and Christos Faloutsos. On Power-Law Relationships of the Internet Topology. ACM SIGCOMM 1999.
- E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, Hierarchical organization of modularity in metabolic networks, *Science* **297**, 1551-1555 (2002).
- R Milo, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii & U Alon, Network Motifs: Simple Building Blocks of Complex Networks. Science, 298:824-827 (2002).
- R Milo, S Itzkovitz, N Kashtan, R Levitt, S Shen-Orr, I Ayzenshtat, M Sheffer & U Alon, Superfamilies of designed and evolved networks. Science, 303:1538-42 (2004).