

Online Social Networks and Media

Network Measurements and Models

Measuring and Modeling Networks

- There are networks everywhere
- What do they look like?
 - How do you measure and describe a billion node network?
- What are the process that generate them?
 - Can we create models for real-life networks?
- These two questions are related: We need to measure the characteristics that we want to model

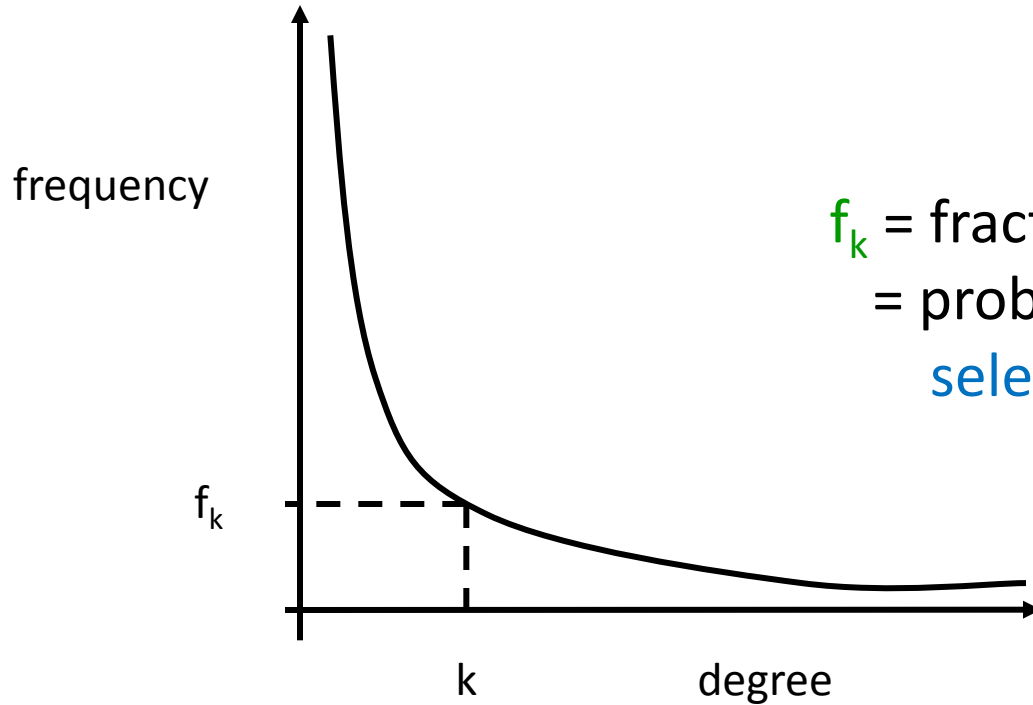
Before we start

- Wait, there is a model for generating graphs!
- The Erdős-Renyi $G_{n,p}$ random graph model:
 - n : the number of vertices
 - p : probability of generating an edge
 - for each pair (i,j) , generate the edge (i,j) independently with probability p
- A very well studied model in graph theory!
 - As we will see, not good enough in our case

Measuring Networks

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

Degree distributions



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

It all started with some Greeks

- Faloutsos, Faloutsos, Faloutsos, “On the power-law relationships of the internet topology”, SIGCOMM 1999.

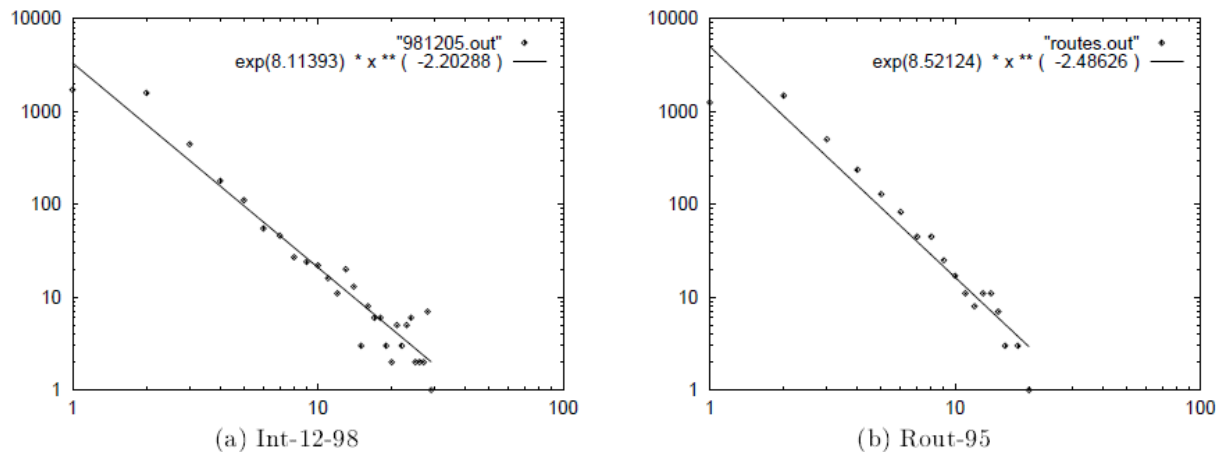


Figure 6: The outdegree plots: Log-log plot of frequency f_d versus the outdegree d .

- Degree distributions for the internet graph

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, $z=np$

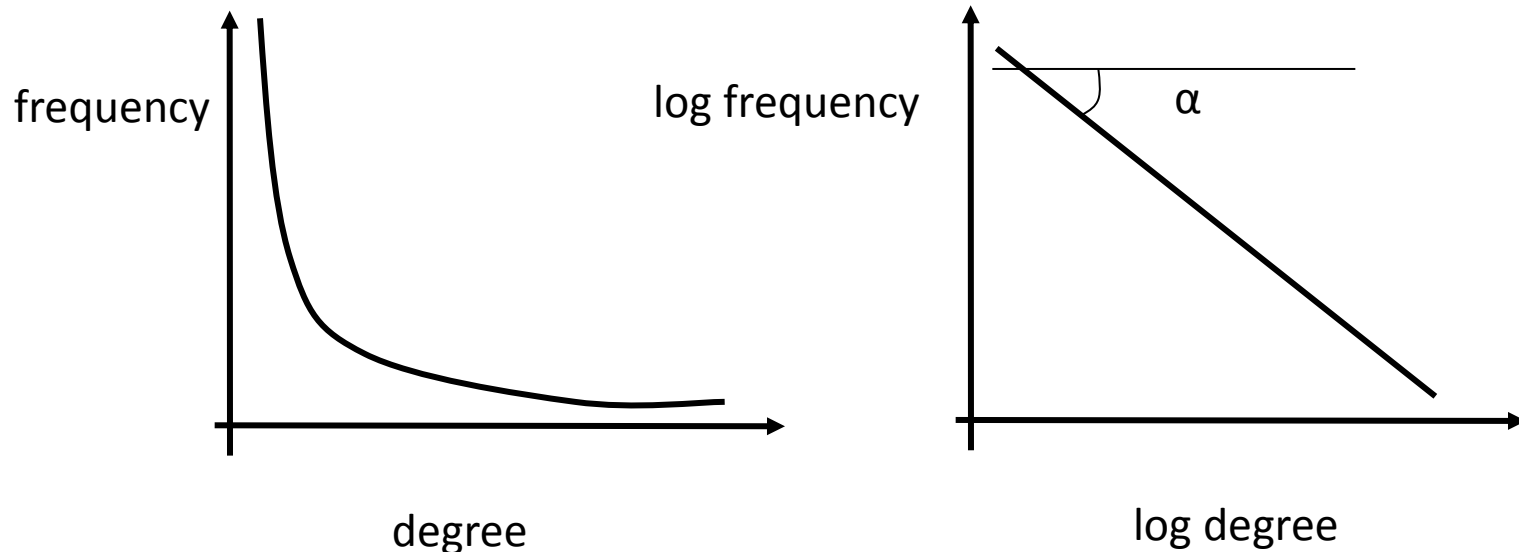
$$p(k) = \frac{z^k}{k!} e^{-z}$$

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

Power-law signature

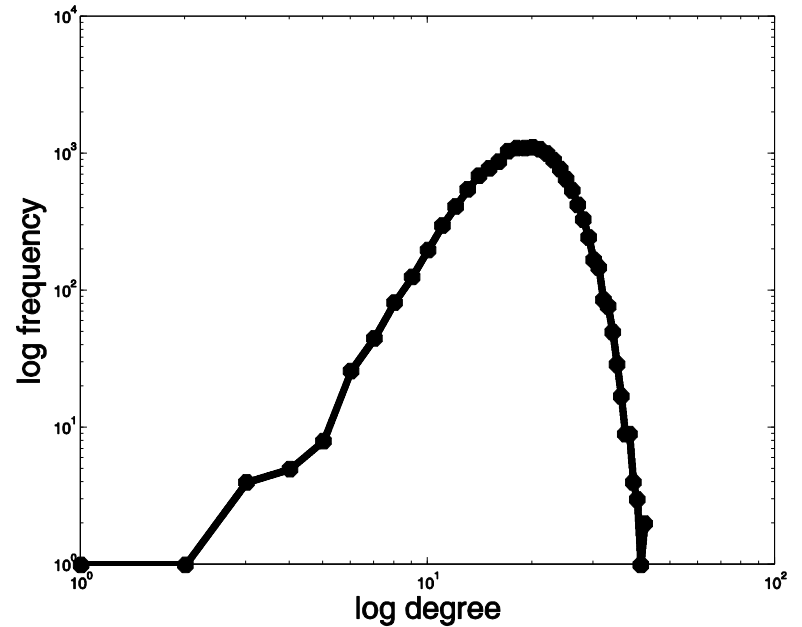
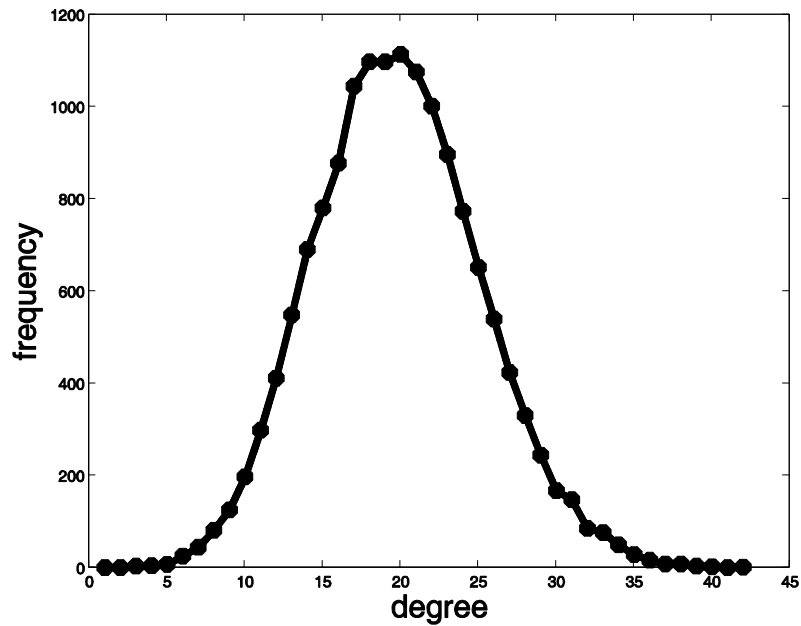
- Power-law distribution gives a line in the **log-log plot**

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

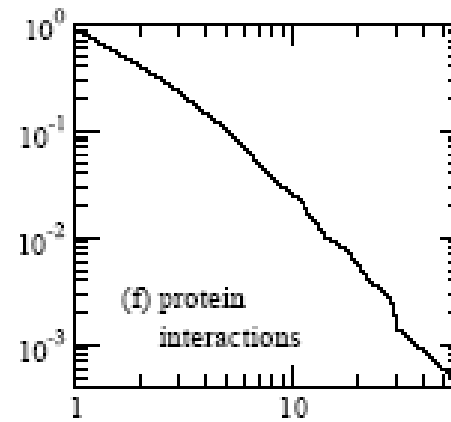
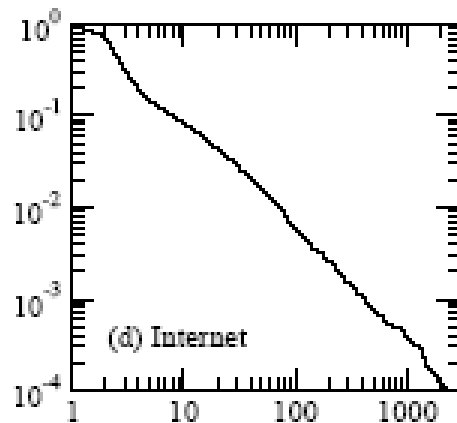
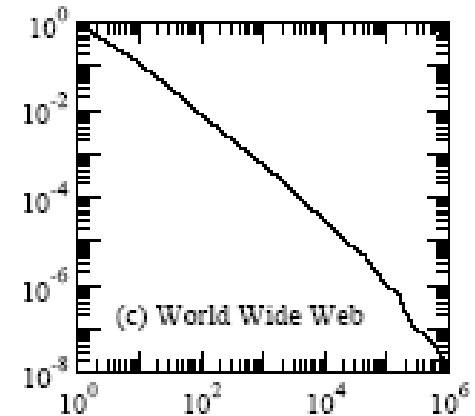
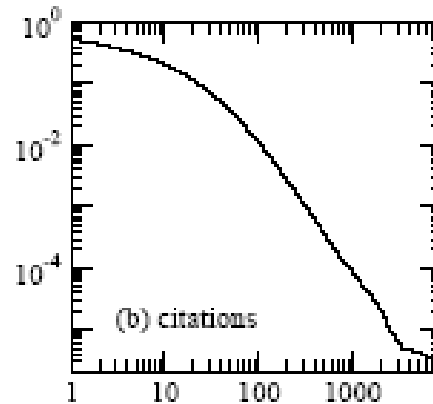
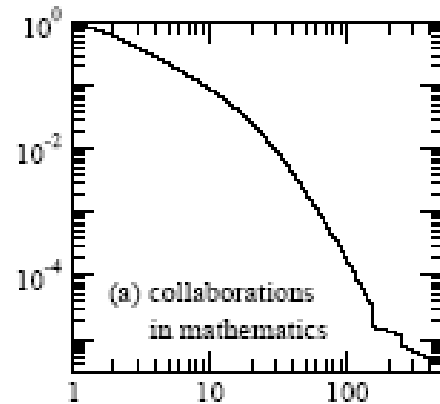


- α : power-law exponent (typically $2 \leq \alpha \leq 3$)

A random graph example



Power-laws appear in all networks!

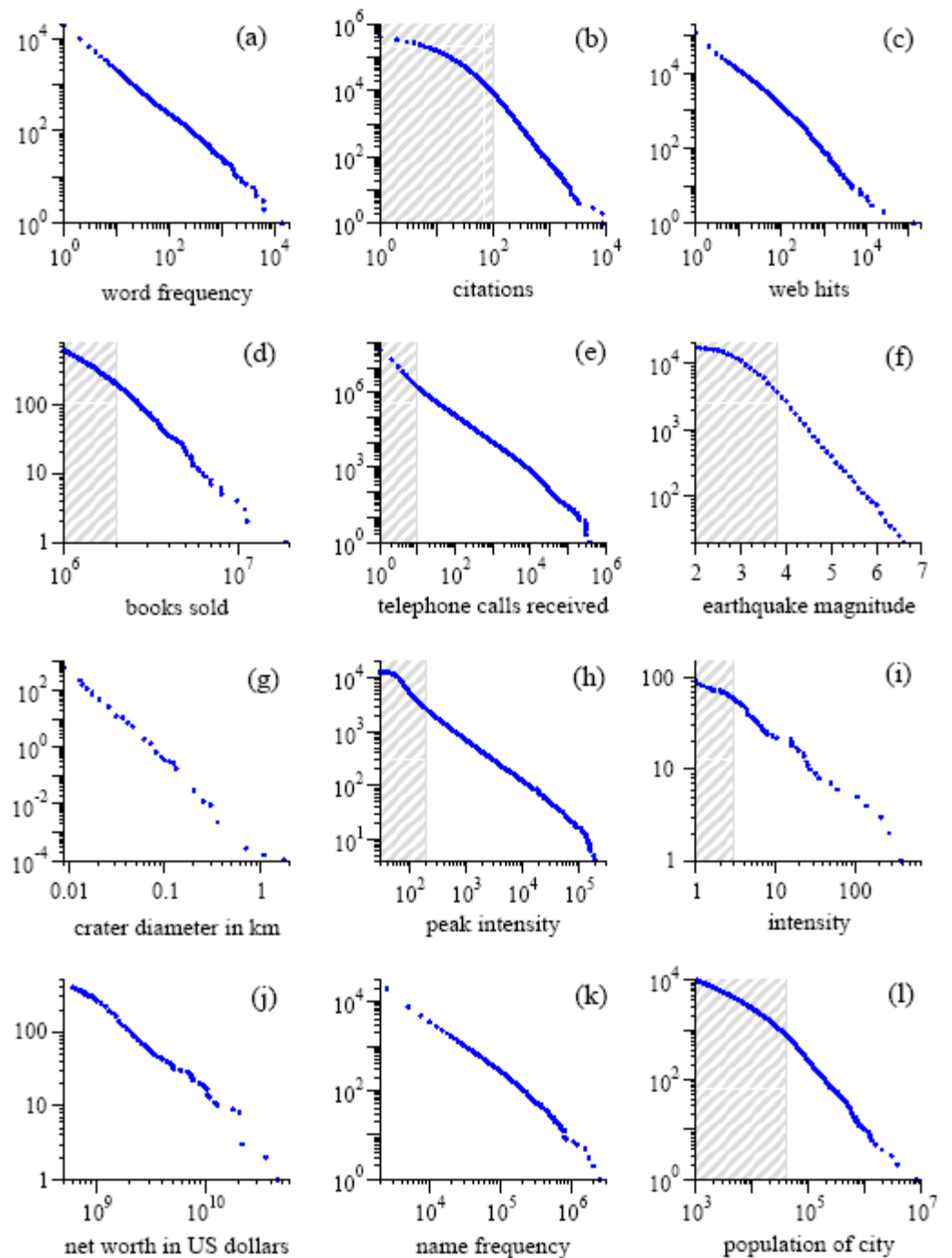


Taken from [Newman 2003]

And not only in networks!

| quantity | minimum x_{\min} | exponent α |
|------------------------------------|-----------------------|----------------------|
| (a) frequency of use of words | 1 | 2.20(1) |
| (b) number of citations to papers | 100 | 3.04(2) |
| (c) number of hits on web sites | 1 | 2.40(1) |
| (d) copies of books sold in the US | 2 000 000 | 3.51(16) |
| (e) telephone calls received | 10 | 2.22(1) |
| (f) magnitude of earthquakes | 3.8 | 3.04(4) |
| (g) diameter of moon craters | 0.01 | 3.14(5) |
| (h) intensity of solar flares | 200 | 1.83(2) |
| (i) intensity of wars | 3 | 1.80(9) |
| (j) net worth of Americans | \$600m | 2.09(4) |
| (k) frequency of family names | 10 000 | 1.94(1) |
| (l) population of US cities | 40 000 | 2.30(5) |

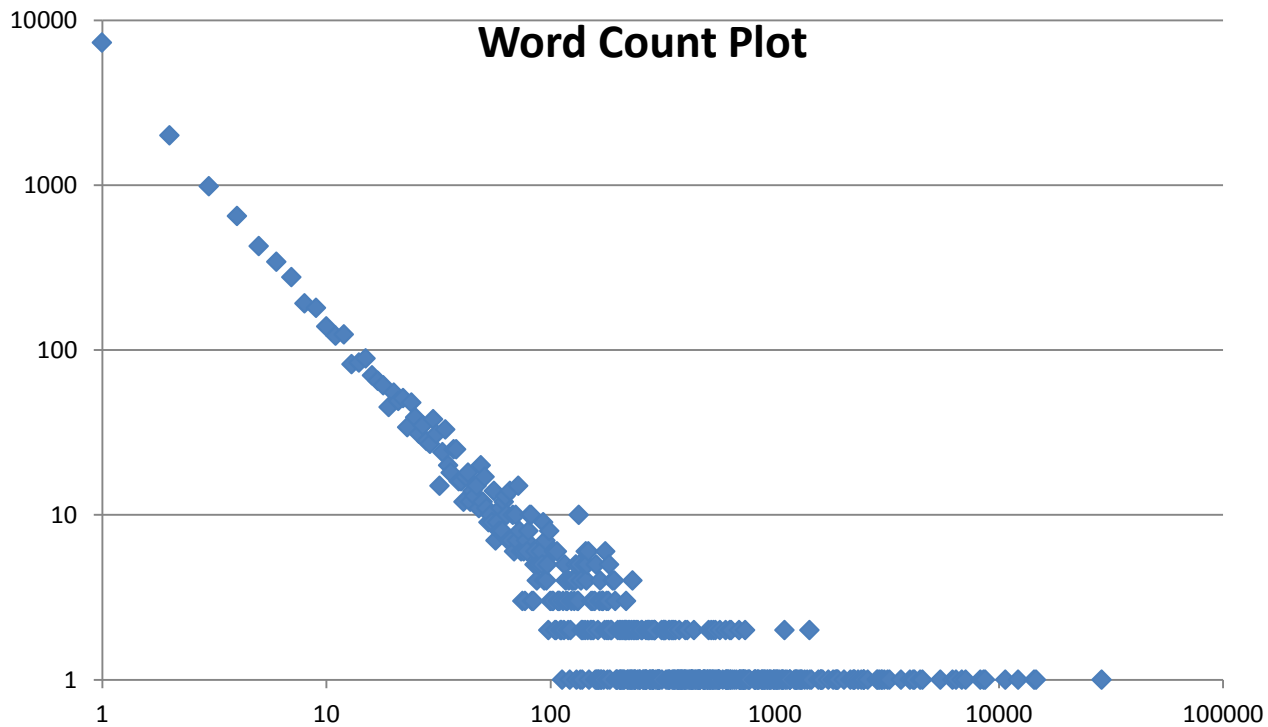
TABLE I Parameters for the distributions shown in Fig. 4. The labels on the left refer to the panels in the figure. Exponent values were calculated using the maximum likelihood method of Eq. (5) and Appendix B, except for the moon craters (g), for which only cumulative data were available. For this case the exponent quoted is from a simple least-squares fit and should be treated with caution. Numbers in parentheses give the standard error on the trailing figures.



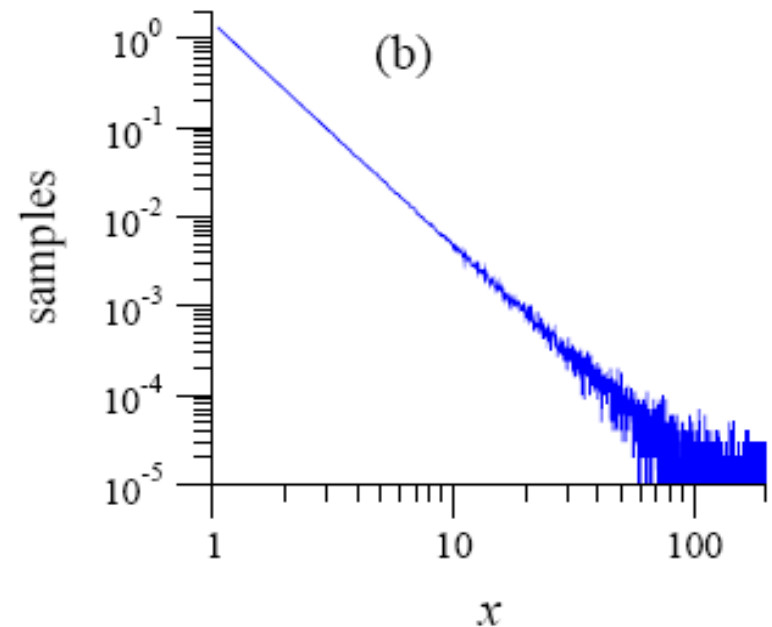
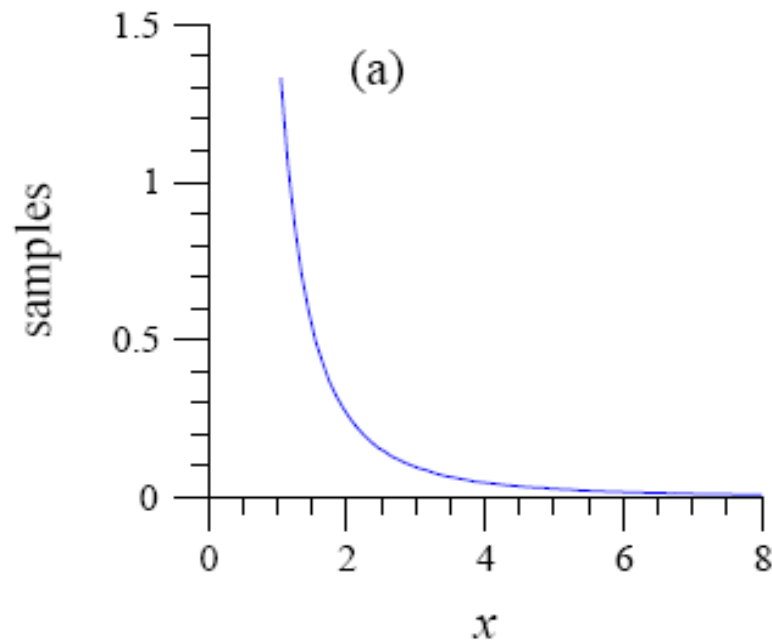
Measuring power-laws

- 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



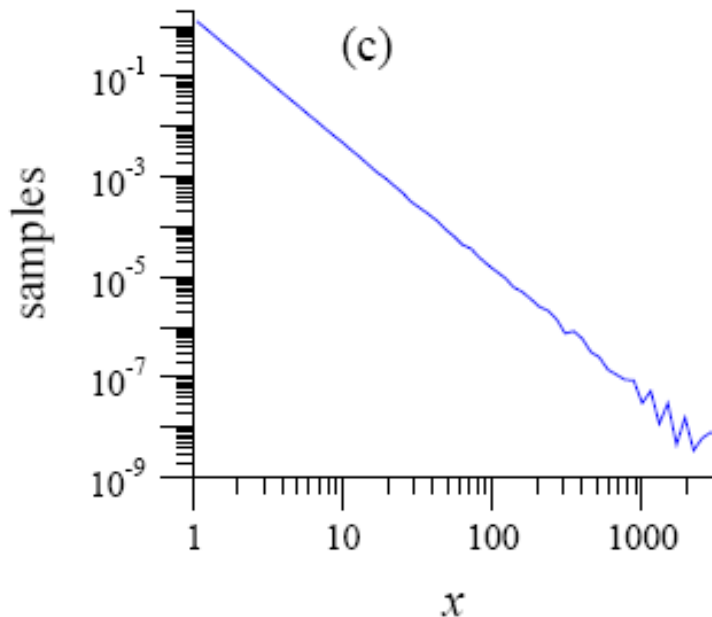
Measuring power laws



Simple binning produces a noisy plot

Logarithmic binning

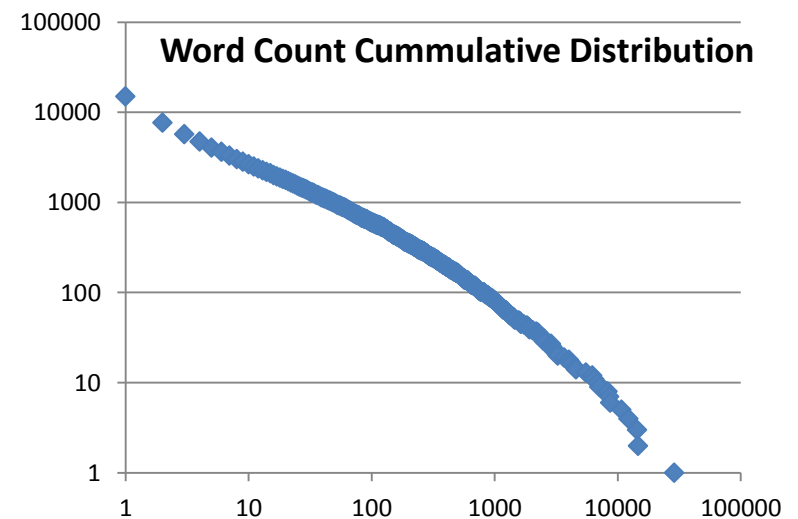
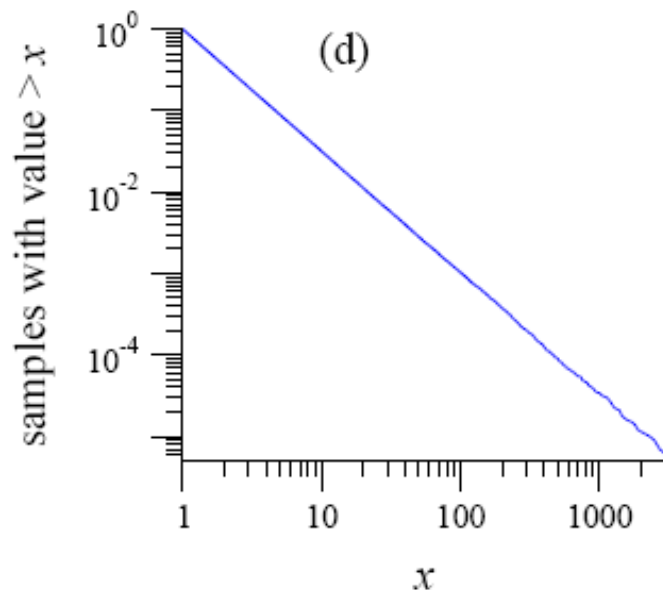
- Exponential binning
 - Create bins that grow **exponentially** in size
 - In each bin **divide** the **sum of counts** by the **bin length** (number of **observations per bin unit**)



Still some noise at the tail

Cumulative distribution

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



Pareto distribution

- A random variable follows a Pareto distribution if

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

- 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

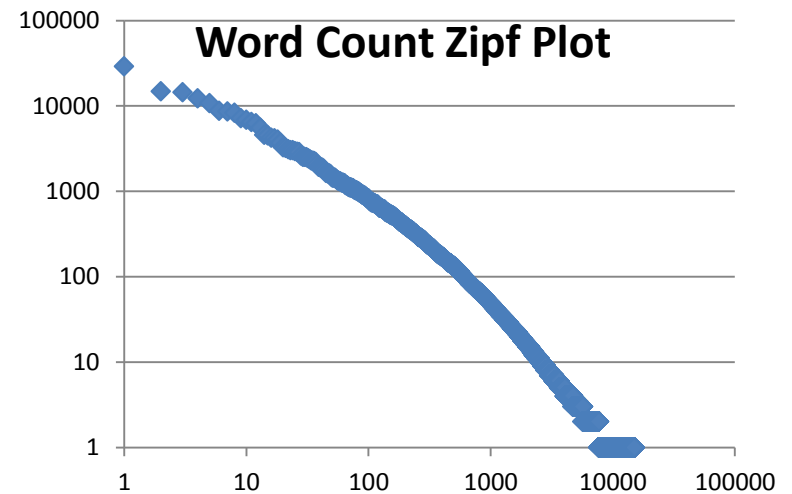
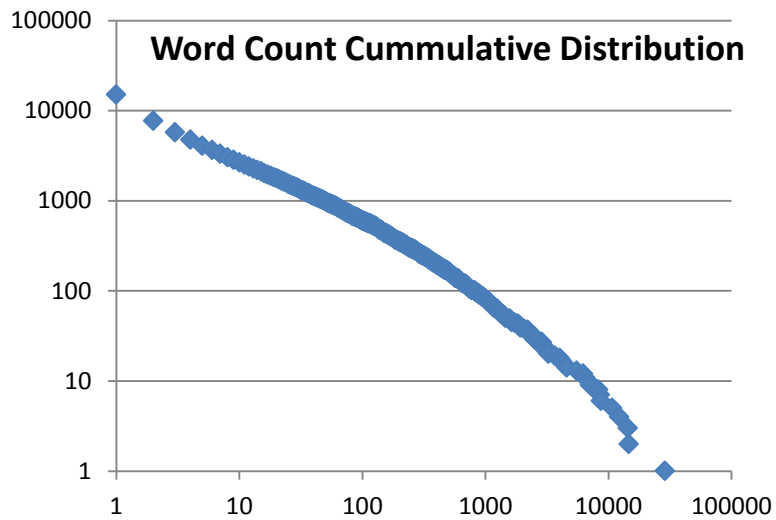
$$x_r \approx r^{-\gamma}$$

- 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 \mathbf{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 $P(\alpha | \mathbf{x})$

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

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

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

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

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

Average/Expected degree

- For power-law distributed degree

- if $\alpha \geq 2$, it is a constant

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

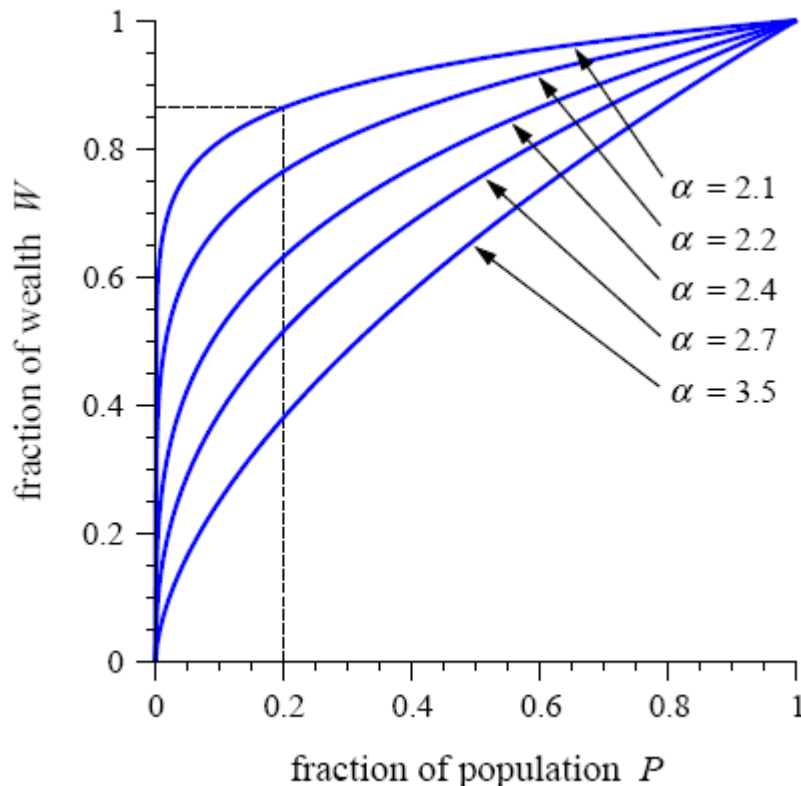
- if $\alpha < 2$, it diverges

- 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

The 80/20 rule

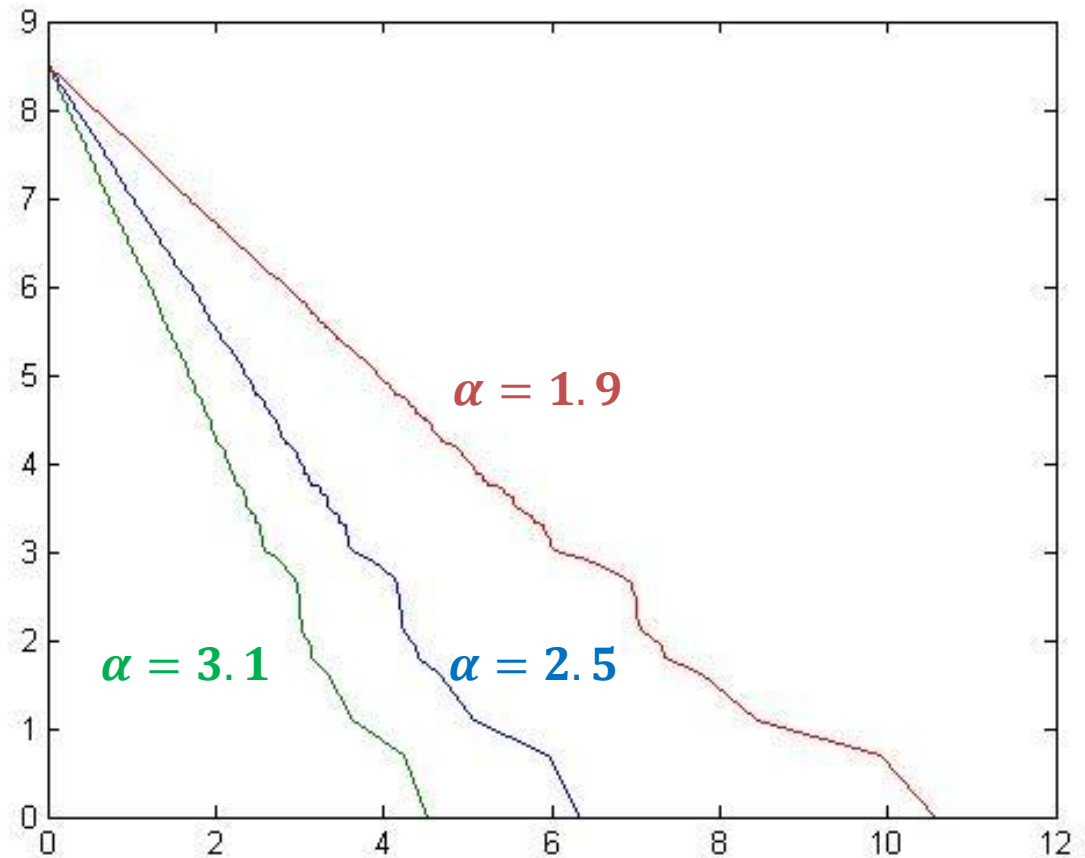
- **Top-heavy**: Small fraction of values collect most of distribution mass



- 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

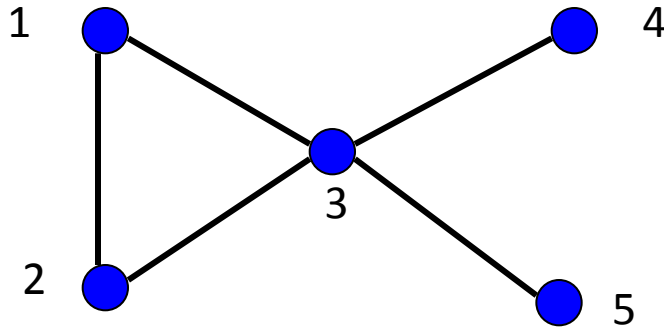
Clustering (Transitivity) coefficient

- 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

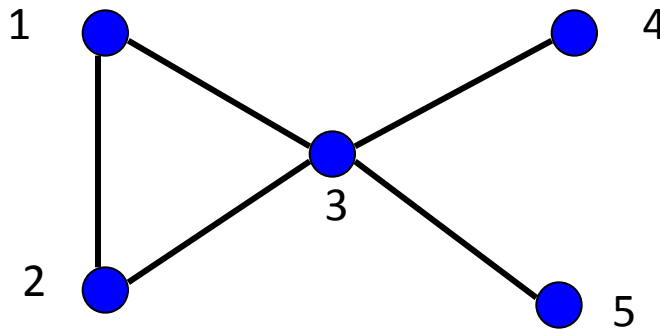
- Clustering coefficient for node i

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

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

- The mean of the ratios

Example



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

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

- 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 nodes, 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 Erdős number

Measuring the small world phenomenon

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

- **Diameter:** $d = \max_{i,j} d_{ij}$

- **Characteristic path length:**

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

Problem if no path between two nodes

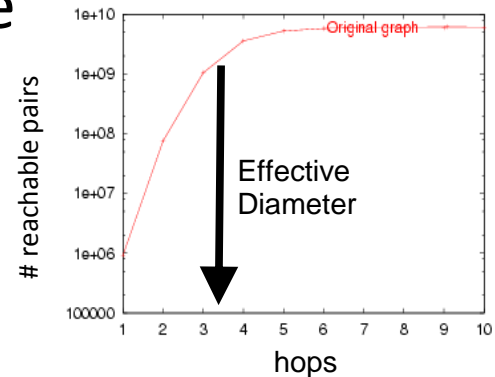
- **Harmonic mean**

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

- Also, distribution of all shortest paths

Effective Diameter

- Disconnected components or isolated long paths can throw off the computation of the diameter.
- **Effective diameter**: the **interpolated value** where 90% of node pairs are reachable



- Computation:
 - $f(d)$: for **integer** d , the fraction of pairs in the graph that have distance less or equal to D
 - $f(x)$: for **real** x : $d - 1 < x < d$, $f(x) = \frac{f(d) - f(d-1)}{x - d}$
 - **Effective Diameter**: the **real value** x such that $f(x) = 0.9$

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

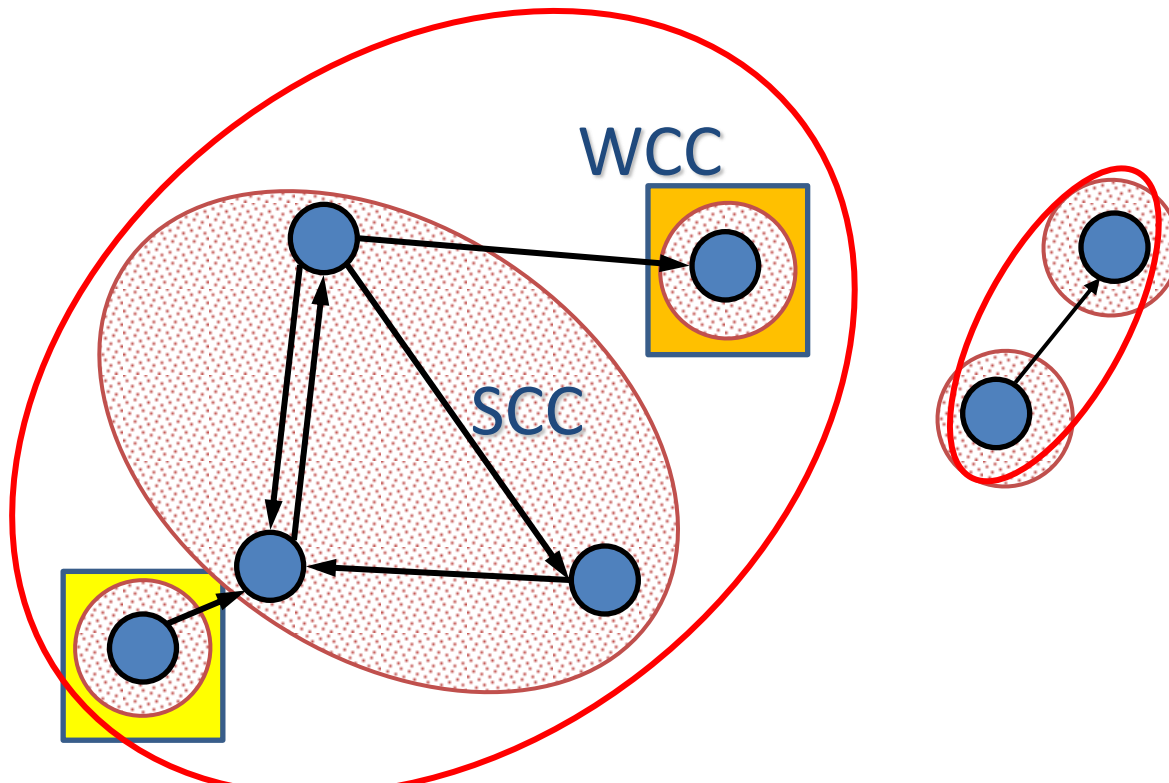
- For all real networks there are (on average) **short paths** between nodes of the network.
 - Largest path found in the IMDB actor network: 7
- Is this interesting?
 - **Random graphs** also have **small diameter**
($d = \log n / \log \log n$ when $z = \omega(\log n)$)
- **Short paths are not surprising** and should be combined with other properties
 - ease of navigation
 - high clustering coefficient

Connected components

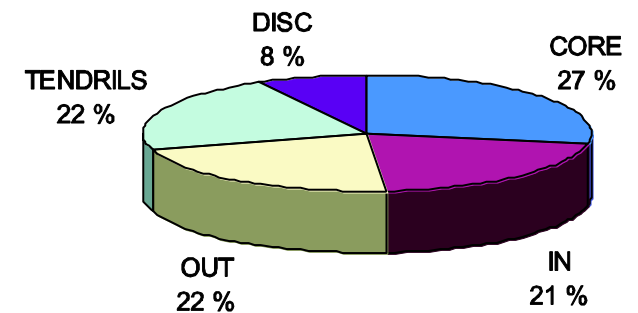
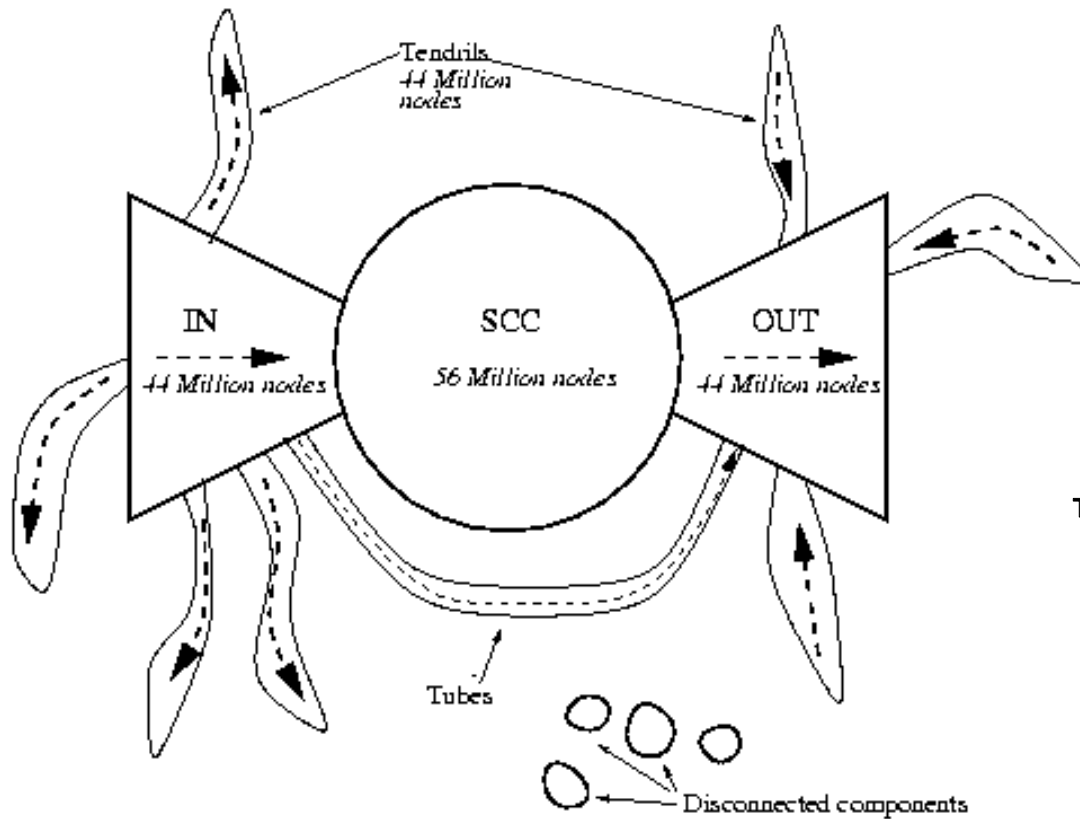
- For undirected graphs, the size and distribution of the **connected components**
 - is there a **giant component**?
 - Most known real undirected networks have a giant component
- For directed graphs, the size and distribution of **strongly** and **weakly connected components**

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)



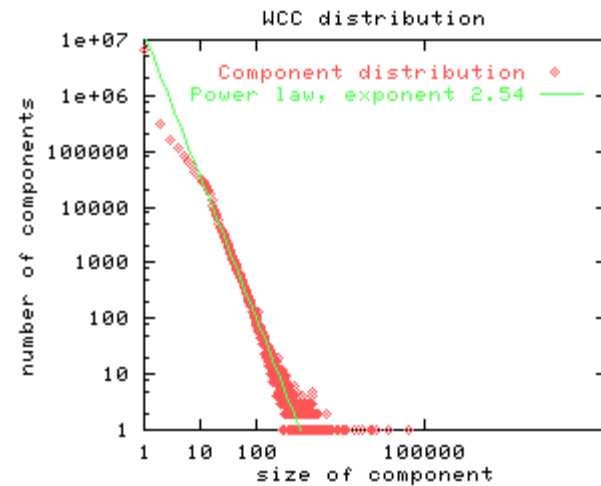
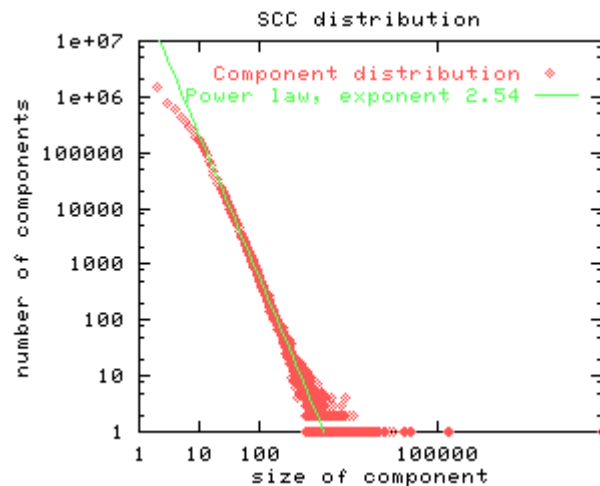
The bow-tie structure of the Web



The largest weakly connected component contains 90% of the nodes

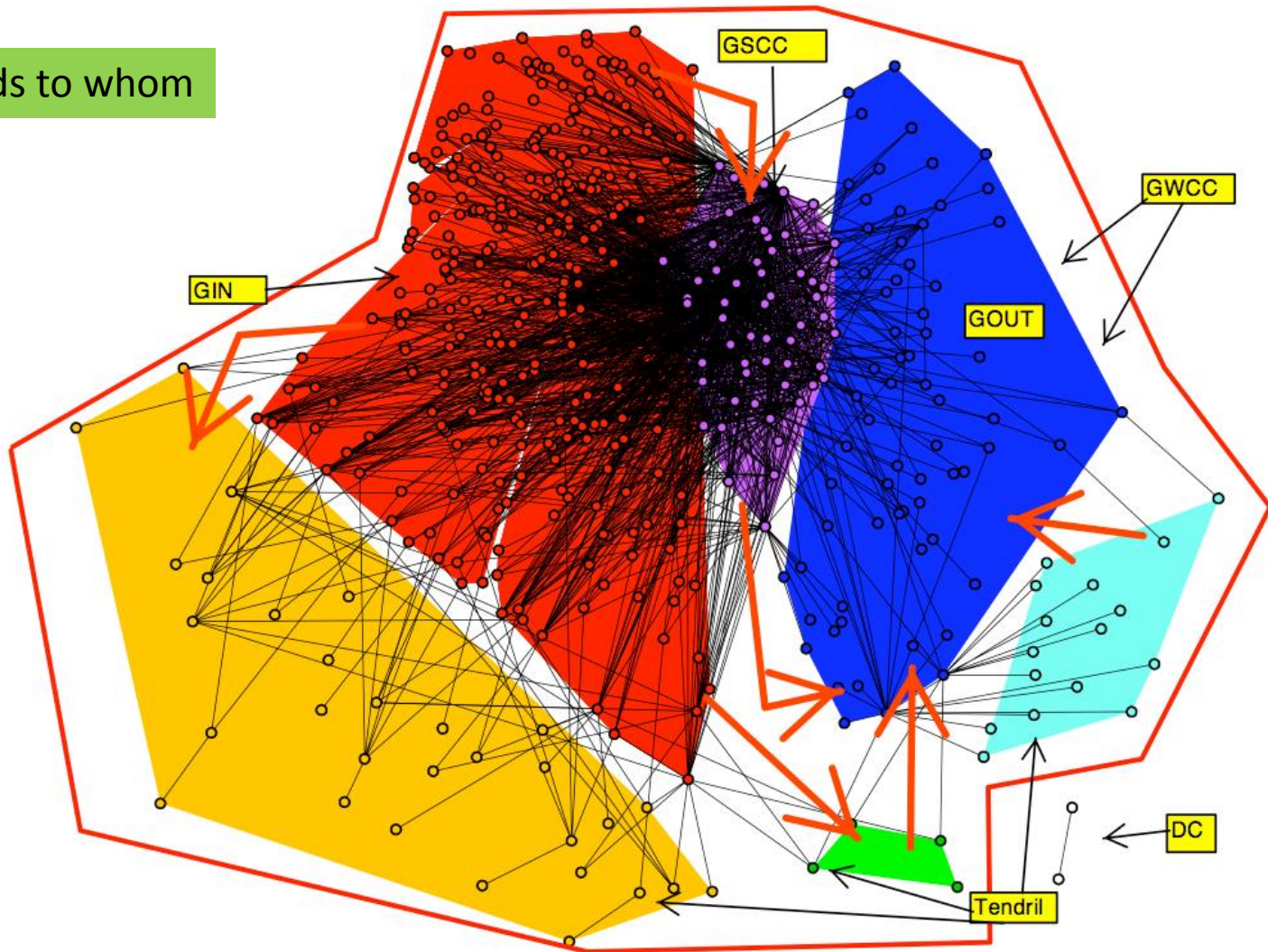
SCC and WCC distribution

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



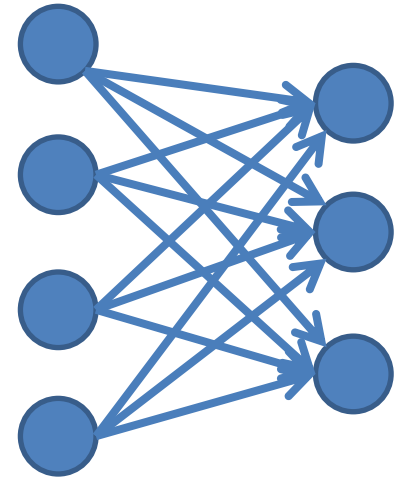
Another bow-tie

Who lends to whom



Web Cores

- **Cores:** Small complete bipartite graphs (of size 3×3 , 4×3 , 4×4)
 - 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)

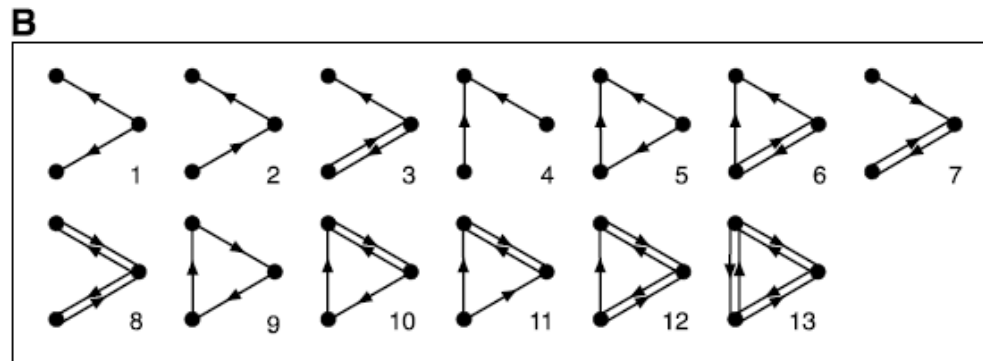


Motifs

- 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

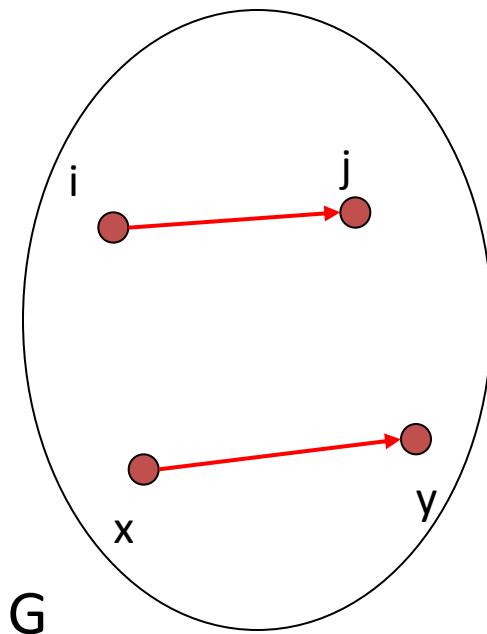


Finding interesting motifs

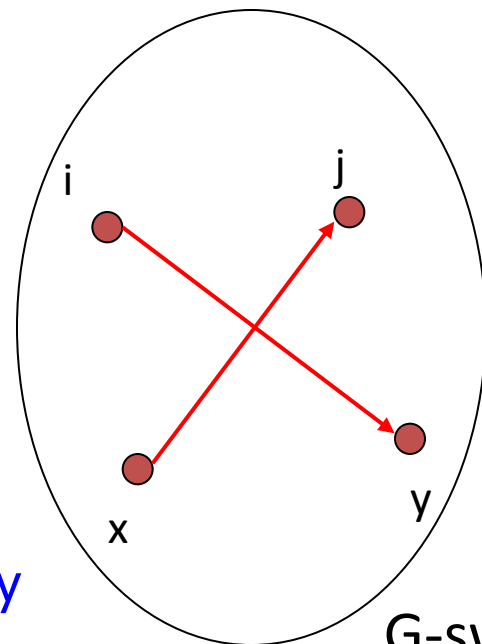
- Sample a part of the graph of size S
- Count the frequency of the motifs of interest
- Compare against the frequency of the motif in a random graph with the same number of nodes **and** the same degree distribution

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



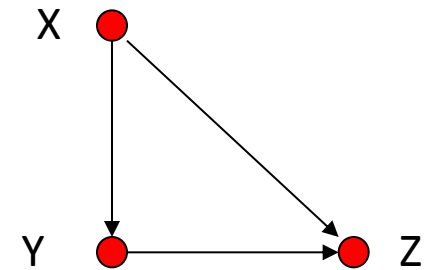
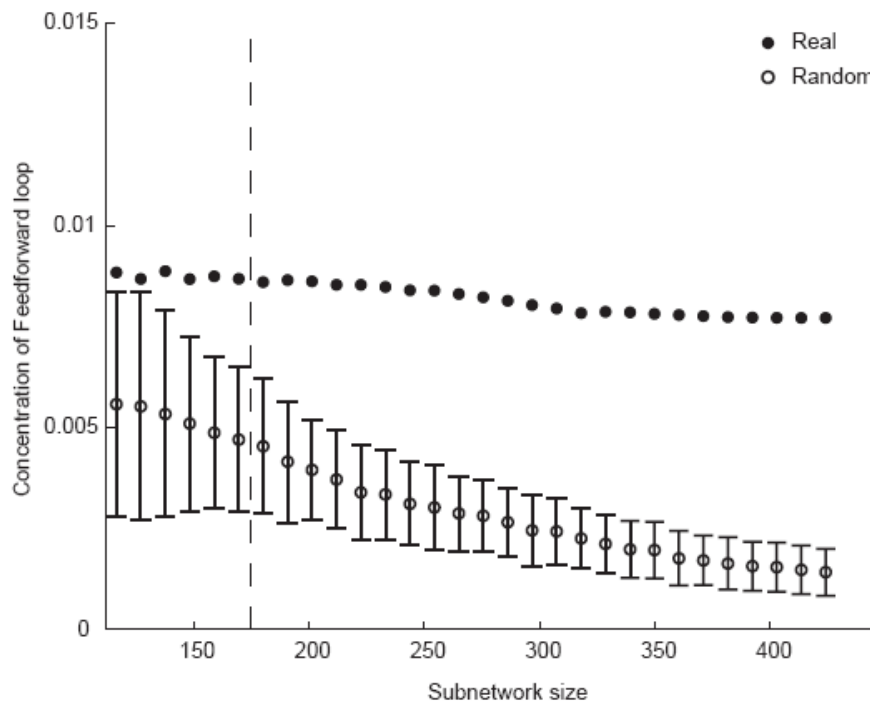
degrees of i,j,x,y
are preserved



G-swapped

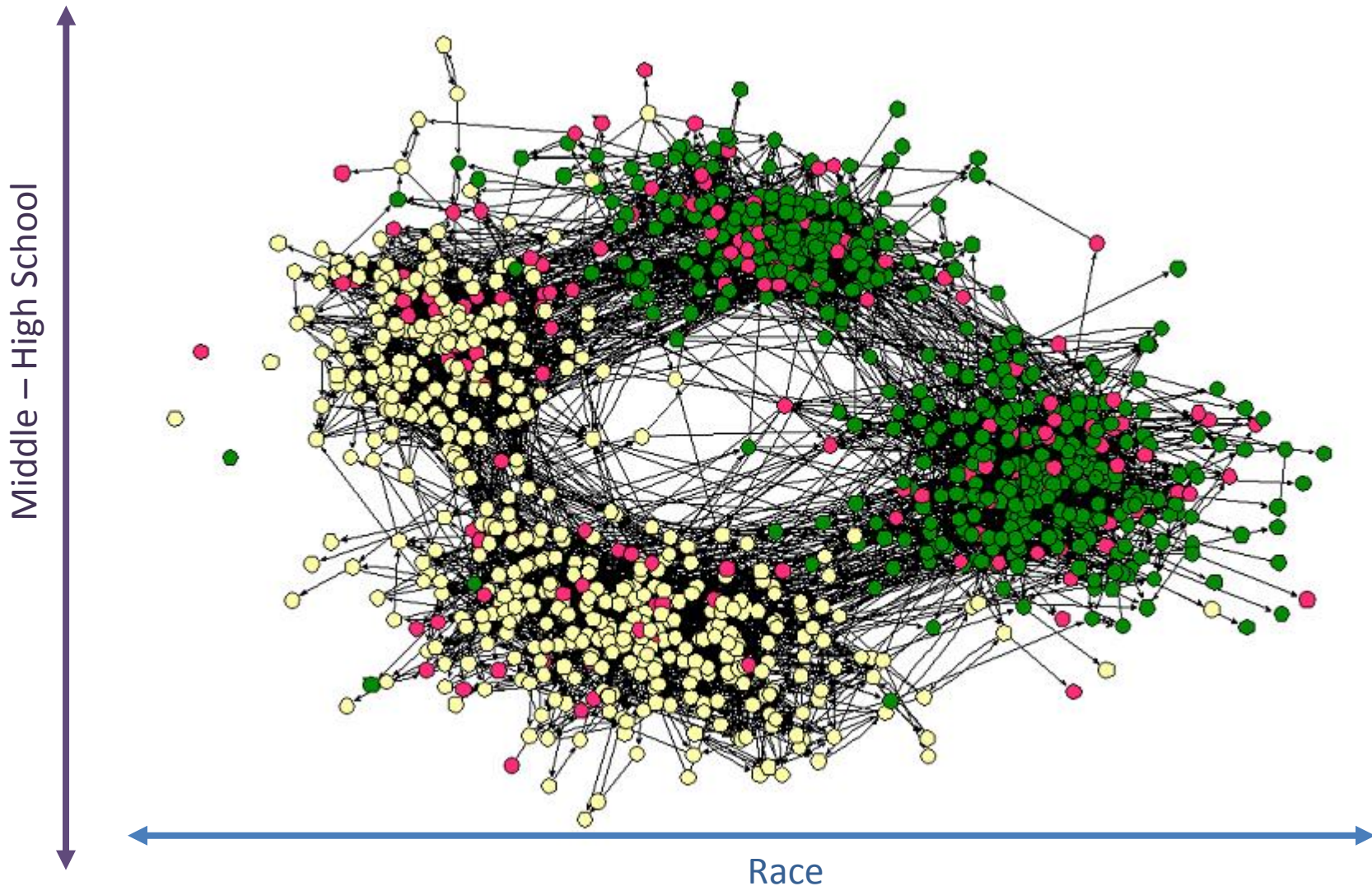
The feed-forward loop

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

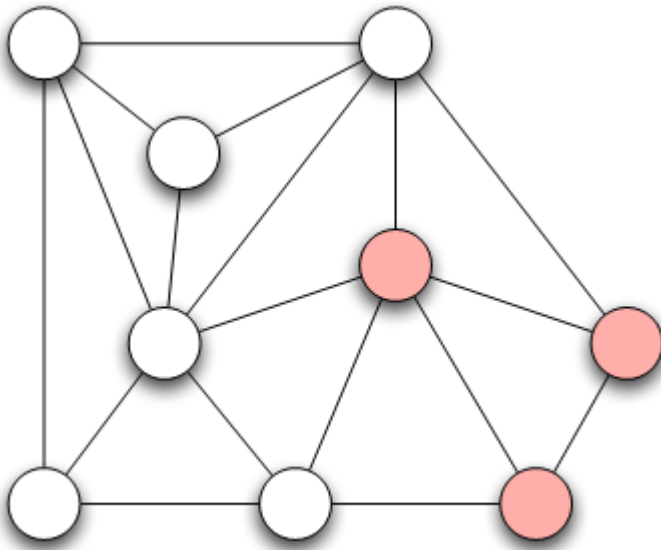


Homophily

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



Measuring Homophily



If the fraction of cross-gender edges is significantly less than expected, then there is evidence for homophily

gender male with probability p
gender female with probability q

Probability of cross-gender edge?

$$\frac{\#cross_gender_edges}{\#edges} \ll 2pq$$

Measuring Homophily

- “significantly” less than
- Inverse homophily
- Characteristics with more than two values:
 - Number of heterogeneous edges (edge between two nodes that are different)

Mechanisms Underlying Homophily: Selection and Social Influence

Selection: tendency of people to form friendships with others who are like them

Socialization or Social Influence: the existing social connections in a network are influencing the individual characteristics of the individuals

Social Influence as the inverse of Selection

Mutable & immutable characteristics

The Interplay of Selection and Social Influence

Longitudinal studies in which the social connections and the behaviors within a group are tracked over a period of time

Why?

- Study teenagers, scholastic achievements/drug use (peer pressure and selection)
- Relative impact?
- Effect of possible interventions (example, drug use)

The Interplay of Selection and Social Influence

Christakis and Fowler on obesity, 12,000 people over a period of 32-years

People more similar on obesity status to the network neighbors than if assigned randomly

Why?

- (i) Because of selection effects, choose friends of similar obesity status,
 - (ii) Because of confounding effects of homophily according to other characteristics that correlate with obesity
 - (iii) Because changes in the obesity status of person's friends was exerting an influence that affected her
- (iii) As well -> “contagion” in a social sense

Tracking Link Formation in Online Data: interplay between selection and social influence

- Underlying social network
- Measure for behavioral similarity

Wikipedia

Node: Wikipedia editor who maintains a user account and user talk page

Link: if they have communicated with one writing on the user talk page of the other

Editor's behavior: set of articles she has edited

Neighborhood overlap in the bipartite affiliation network of editors and articles consisting only of edges between editors and the articles they have edited

$$\frac{|N_A \cap N_B|}{|N_A \cup N_B|}$$

FACT: Wikipedia editors who have communicated are significantly more similar in their behavior than pairs of Wikipedia editors who have not (homomophily), **why?**

Selection (editors form connections with those have edited the same articles) vs Social Influence (editors are led to the articles of people they talk to)

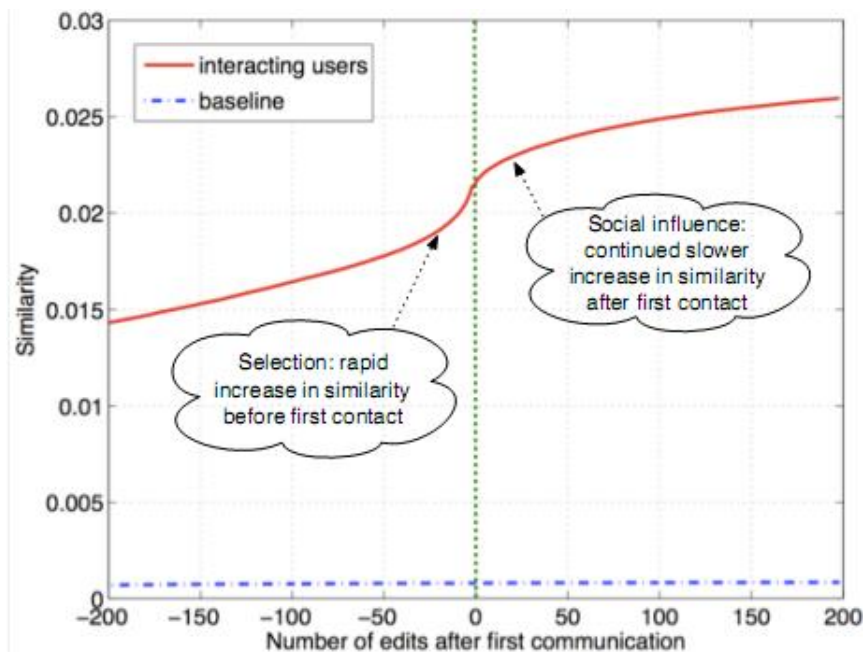
Tracking Link Formation in Online Data: interplay between selection and social influence

Actions in Wikipedia are time-stamped

For each pair of editors A and B who have ever communicated,

- Record their similarity over time
- Time 0 when they first communicated -- Time moves in discrete units, advancing by one “tick” whenever either A or B performs an action on Wikipedia
- Plot one curve for each pair of editors

Average, single plot: average level of similarity relative to the time of first interaction



Similarity is clearly increasing both before and after the moment of first interaction (both selection and social influence)

Not symmetric around time 0 (particular role on similarity): Significant increase before they meet

Blue line shows similarity of a random pair (non-interacting)

References

- M. E. J. Newman, *Power laws, Pareto distributions and Zipf's law*, *Contemporary Physics*.
- 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).

NETWORK MODELS

What is a network model?

- Informally, a network model is a **process** (radomized or deterministic) for generating a graph of arbitrary size.
- Models of **static** graphs
 - **input**: a set of parameters Π , and the size of the graph n
 - **output**: a graph $G(\Pi, n)$
- Models of **evolving** graphs
 - **input**: a set of parameters Π , and an initial graph G_0
 - **output**: a graph G_t for each time t

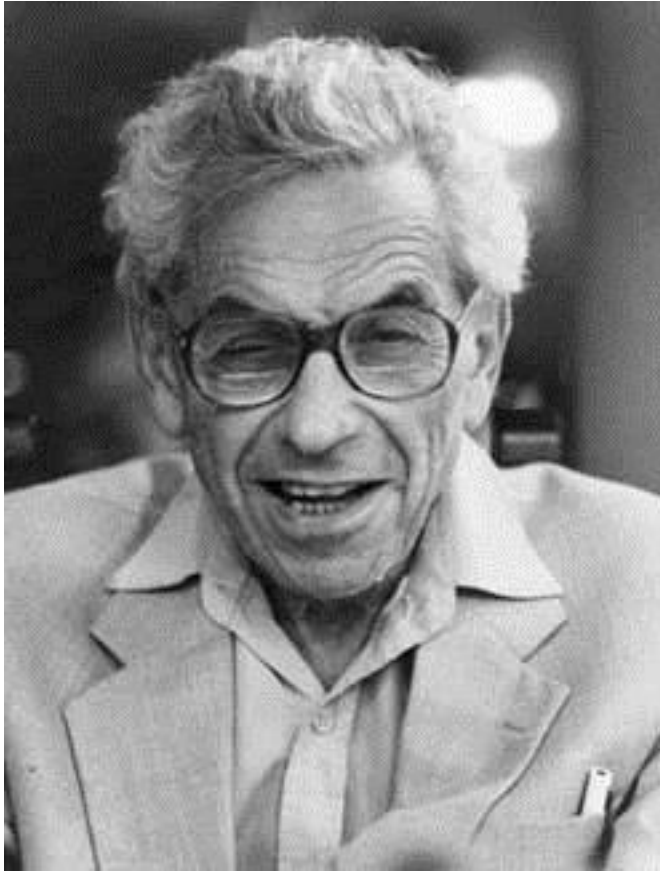
Families of random graphs

- A deterministic model \mathbf{D} defines a single graph for each value of n (or t)
- A randomized model \mathbf{R} defines a probability space $\langle G_n, P \rangle$ where G_n is the set of all graphs of size n , and P a probability distribution over the set G_n (similarly for t)
 - we call this a family of random graphs \mathbf{R} , or a random graph \mathbf{R}

Why do we care?

- Creating models for real-life graphs is important for several reasons
 - Create data for **simulations** of processes on networks
 - Identify the **underlying mechanisms** that govern the network generation
 - **Predict** the evolution of networks

Erdős-Renyi Random graphs



Paul Erdős (1913-1996)

Erdős-Renyi Random Graphs

- The $G_{n,p}$ model
 - **input**: the number of vertices n , and a parameter p , $0 \leq p \leq 1$
 - **process**: for each pair (i,j) , generate the edge (i,j) independently with probability p
- Related, but not identical: The $G_{n,m}$ model
 - **process**: select m edges uniformly at random

Graph properties

- A property **P** holds **almost surely (a.s.)** (or for **almost every** graph), if

$$\lim_{n \rightarrow \infty} P[G \text{ has } P] = 1$$

- Evolution of the graph: which properties hold as the parameters of the graph model change?
 - different from the evolving graphs over time that we saw before
- **Threshold phenomena**: Many properties appear suddenly. That is, there exist a parameter θ_c (e.g., the probability p_c) such that for $\theta < \theta_c$ the property does not hold a.s. and for $\theta > \theta_c$ the property holds a.s.

The giant component

- Let $z=np$ be the average degree
- If $z < 1$, then almost surely, the largest component has size at most $O(\ln n)$
- if $z > 1$, then almost surely, the largest component has size $\Theta(n)$. The second largest component has size $O(\ln n)$
- if $z = \omega(\ln n)$, then the graph is almost surely connected.

The phase transition

- When $z=1$, there is a phase transition
 - The largest component is $O(n^{2/3})$
 - The sizes of the components follow a power-law distribution.

Random graphs degree distributions

- The degree distribution follows a **binomial**

$$p(k) = B(n; k; p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Assuming $z=np$ is fixed, as $n \rightarrow \infty$, $B(n, k, p)$ is approximated by a **Poisson** distribution

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

- Highly concentrated around the mean, with a tail that drops **exponentially**

Other properties

- Clustering coefficient
 - $C = p$
- Diameter (maximum path)
 - $L = \log n / \log z$

Phase transitions

- **Phase transitions** (a.k.a. Threshold Phenomena, Critical phenomena) are observed in a variety of natural or human processes, and they have been studied extensively by Physicists and Mathematicians
 - Also, in popular science: “**The tipping point**”
- Examples
 - Water becoming ice
 - Percolation
 - Giant components in graphs
- In all of these examples, the transition from one state to another (e.g., from water to ice) happens almost instantaneously when a parameter crosses a **threshold**
- At the threshold value we have **critical phenomena**, and the appearance of **Power Laws**
 - There is no characteristic scale.

Random graphs and real life

- A beautiful and elegant theory studied exhaustively
- Random graphs had been used as idealized network models
- Unfortunately, they don't capture reality...

Departing from the ER model

- We need models that better capture the characteristics of real graphs
 - degree sequences
 - clustering coefficient
 - short paths

Graphs with given degree sequences

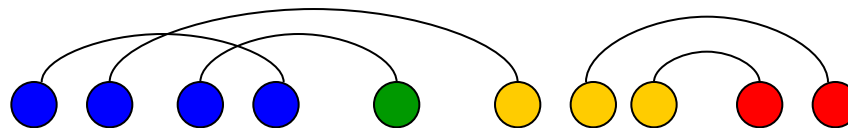
- The configuration model
 - input: the degree sequence $[d_1, d_2, \dots, d_n]$
 - process:
 - Create d_i copies of node i
 - Take a **random matching** (pairing) of the copies
 - self-loops and multiple edges are allowed
- Uniform distribution over the graphs with the given degree sequence

Example

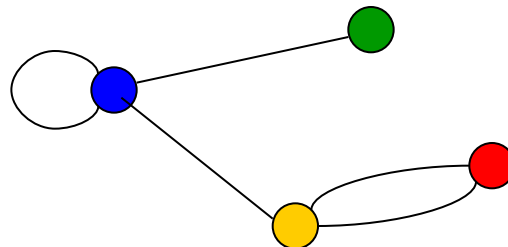
- Suppose that the degree sequence is



- Create multiple copies of the nodes



- Pair the nodes uniformly at random
- Generate the resulting network



Power-law graphs

- The critical value for the exponent α is

$$\alpha = 3.4788\dots$$

- The clustering coefficient is

$$C \propto n^{-\beta} \quad \beta = \frac{3\alpha - 7}{\alpha - 1}$$

- When $\alpha < 7/3$ the clustering coefficient **increases** with n

Graphs with given **expected** degree sequences

- Input: the degree sequence $[d_1, d_2, \dots, d_n]$
- m = total number of edges
- Process: generate edge (i,j) with probability $d_i d_j / m$
 - preserves the **expected** degrees
 - easier to analyze

However...

- The problem is that these models are too contrived
- It would be more interesting if the network structure **emerged** as a side product of a stochastic process rather than fixing its properties in advance.

Preferential Attachment in Networks

- First considered by [Price 65] as a model for citation networks (directed)
 - each new paper is generated with m citations (mean)
 - new papers cite previous papers with probability proportional to their in-degree (citations)
 - what about papers without any citations?
 - each paper is considered to have a “default” a citations
 - probability of citing a paper with degree k , proportional to $k+a$
- Power law with exponent $\alpha = 2+a/m$

Practical Issues

- The model is equivalent to the following:
 - With probability $m/(m+a)$ link to a node with probability proportional to the degree.
 - With probability $a/(m+a)$ link to a node selected uniformly at random.
- How do we select a node with probability proportional to the degree in practice:
 - Maintain a list with the endpoints of all the edges seen so far, and select a node from this list uniformly at random
 - Append the list each time new edges are created.

Barabasi-Albert model

- The BA model (undirected graph)
 - **input**: some initial subgraph G_0 , and m the number of edges per new node
 - **the process**:
 - nodes arrive one at the time
 - each node connects to m other nodes selecting them with probability proportional to their degree
 - if $[d_1, \dots, d_t]$ is the degree sequence at time t , the node $t+1$ links to node i with probability

$$\frac{d_i}{\sum_i d_i} = \frac{d_i}{2mt}$$

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

The mathematicians point of view

[Bollobas-Riordan]

- Self loops and multiple edges are allowed
- For the **single edge** problem:
 - At time t , a new vertex v , connects to an existing vertex u with probability
$$\frac{d_u}{2t-1}$$
 - it creates a self-loop with probability
$$\frac{1}{2t-1}$$
- If m edges, then they are inserted **sequentially**, as if inserting m nodes
 - the problem reduces to studying the single edge problem.

Preferential attachment graphs

- Expected diameter
 - if $m = 1$, the diameter is $\Theta(\log n)$
 - if $m > 1$, the diameter is $\Theta(\log n / \log \log n)$
- Expected clustering coefficient is small

$$\mathbb{E}[C^{(2)}] = \frac{m-1}{8} \frac{\log^2 n}{n}$$

Weaknesses of the BA model

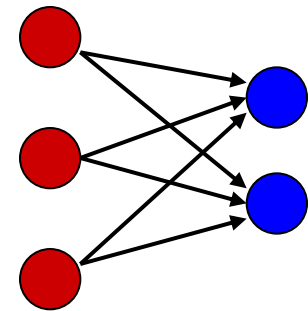
- Technical issues:
 - It is not directed (not good as a model for the Web) and when directed it gives acyclic graphs
 - It focuses mainly on the (in-) degree and does not take into account other parameters (out-degree distribution, components, clustering coefficient)
 - It correlates age with degree which is not always the case
- Academic issues
 - the model rediscovers the wheel
 - preferential attachment is not the answer to every power-law
 - what does “scale-free” mean exactly?
- Yet, it was a breakthrough in the network research, that popularized the area

Variations of the BA model

- Many **variations** have been considered some in order to address the problems with the vanilla BA model
 - edge rewiring, **appearance** and **disappearance**
 - **fitness** parameters
 - **variable** mean degree
 - **non-linear** preferential attachment
 - surprisingly, only linear preferential attachment yields power-law graphs

Empirical observations for the Web graph

- In a large scale experimental study by Kumar et al, they observed that the Web contains a large number of small **bipartite cliques** (cores)
 - the **topical** structure of the Web



a $K_{3,2}$ clique

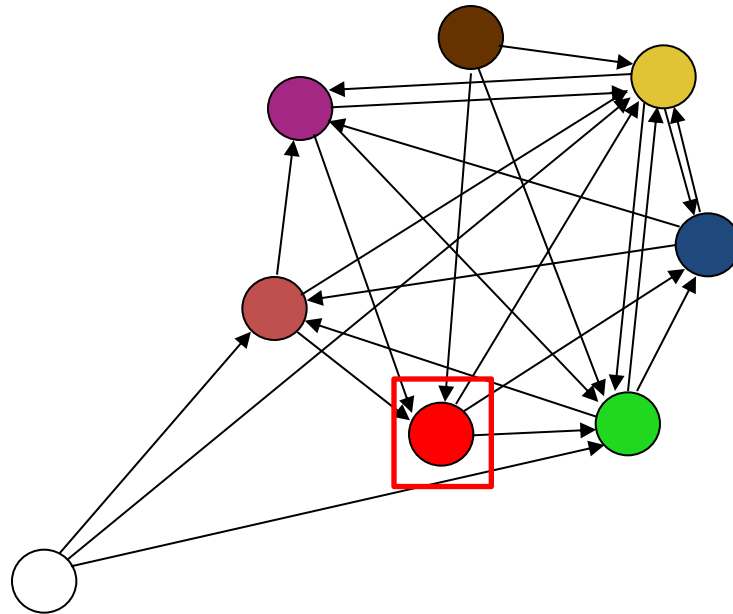
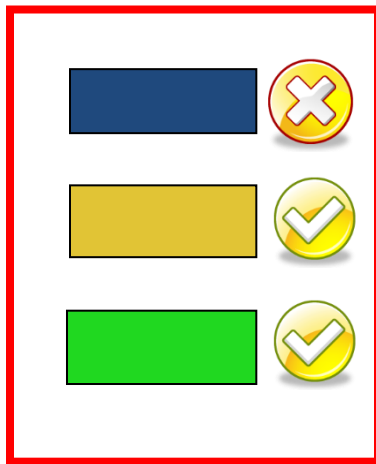
- Such subgraphs are highly unlikely in random graphs
- They are also unlikely in the BA model
- Can we create a model that will have high concentration of small cliques?

Copying model

- Input:
 - the out-degree d (constant) of each node
 - a parameter α
- The process:
 - Nodes arrive one at the time
 - A new node selects uniformly one of the existing nodes as a **prototype**
 - The new node creates d outgoing links. For the i^{th} link
 - with probability α it copies the i -th link of the prototype node
 - with probability $1 - \alpha$ it selects the target of the link uniformly at random

An example

- $d = 3$



Copying model properties

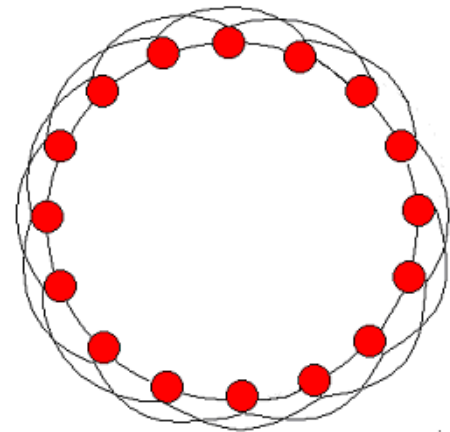
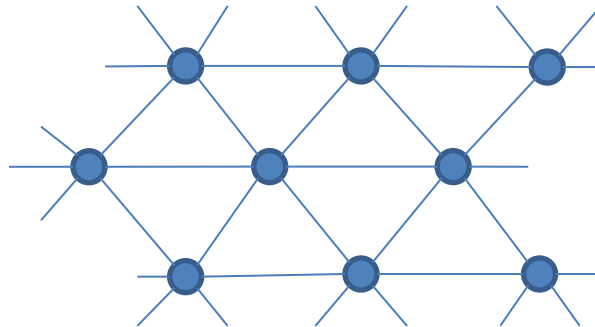
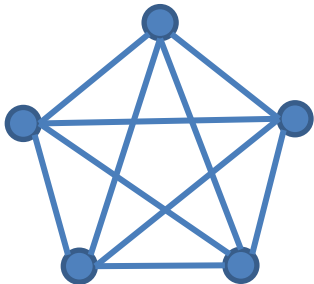
- Power law degree distribution with exponent $\beta = (2-\alpha)/(1-\alpha)$
- Number of bipartite cliques of size $i \times d$ is ne^{-i}
- The model has also found applications in biological networks
 - copying mechanism in gene mutations

Small world Phenomena

- So far we focused on obtaining graphs with power-law distributions on the degrees. What about other properties?
 - **Clustering coefficient**: real-life networks tend to have high clustering coefficient
 - **Short paths**: real-life networks are “small worlds”
 - this property is easy to generate
 - Can we combine these two properties?

Clustering Coefficient

- How can you create a graph with high clustering coefficient?



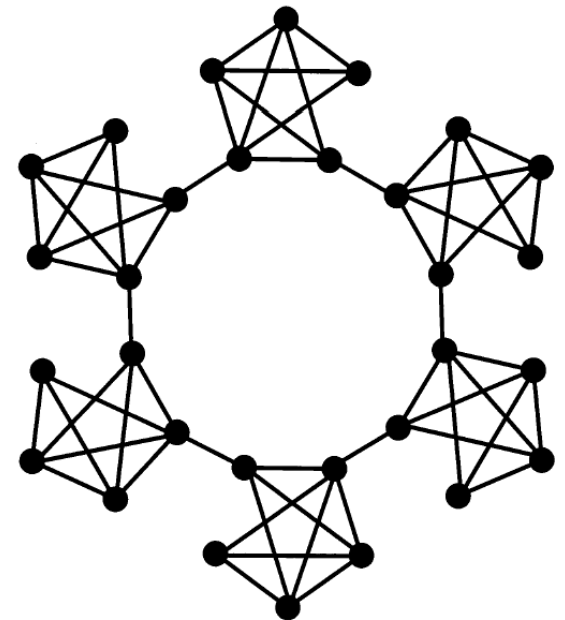
- High clustering coefficient but long paths

Small-world Graphs

- According to Watts [W99]
 - Large networks ($n \gg 1$)
 - Sparse connectivity (avg degree $z \ll n$)
 - No central node ($k_{\max} \ll n$)
 - Large clustering coefficient (larger than in random graphs of same size)
 - Short average paths ($\sim \log n$, close to those of random graphs of the same size)

The Caveman Model [W99]

- The random graph
 - edges are generated completely at random
 - low avg. path length $L \leq \log n / \log z$
 - low clustering coefficient $C \sim z/n$
- The Caveman model
 - edges follow a structure
 - high avg. path length $L \sim n/z$
 - high clustering coefficient $C \sim 1 - O(1/z)$
- Can we interpolate between the two?



Mixing order with randomness

- Inspired by the work of Solomonoff and Rapoport
 - nodes that share neighbors should have higher probability to be connected
- Generate an edge between i and j with probability proportional to R_{ij}

$$R_{ij} = \begin{cases} 1 & \text{if } m_{ij} \geq z \\ \left(\frac{m_{ij}}{z}\right)^\alpha (1-p) + p & \text{if } 0 < m_{ij} < z \\ p & \text{if } m_{ij} = 0 \end{cases}$$

m_{ij} = number of common neighbors of i and j

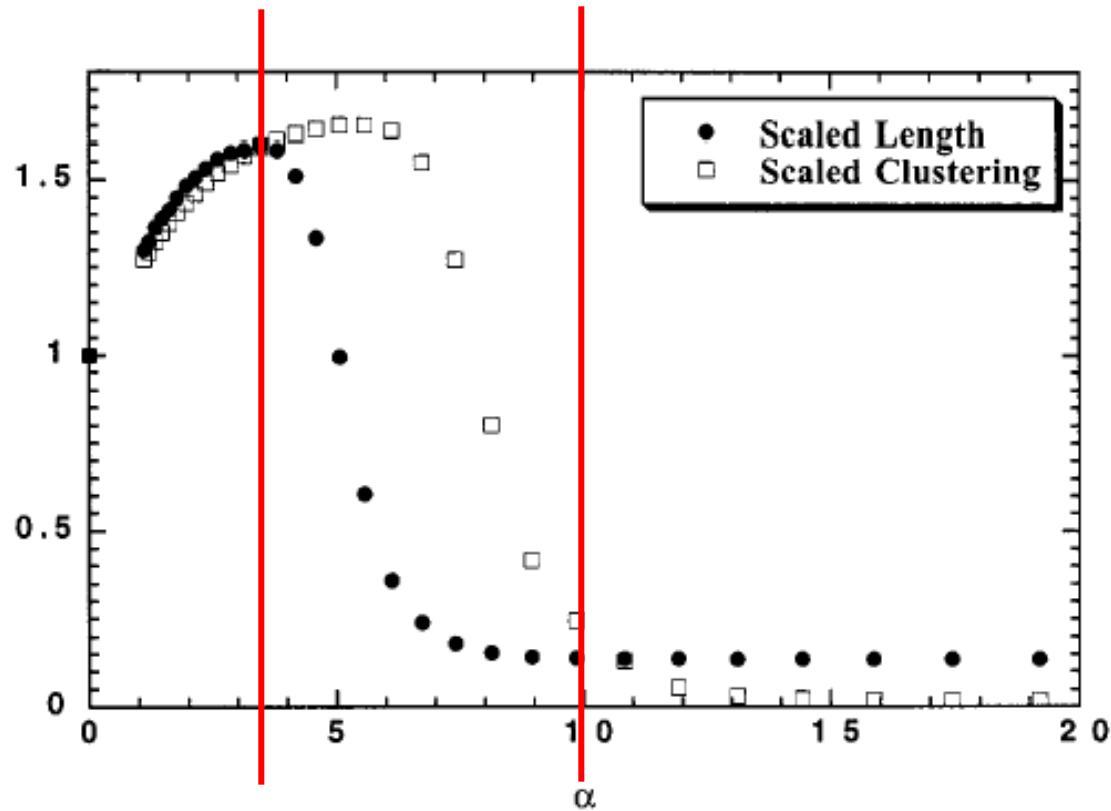
p = very small probability

- When $\alpha \rightarrow \infty$, edges are determined by common neighbors
- When $\alpha = 0$, edges are independent of common neighbors
- For intermediate values we obtain a combination of order and randomness

Algorithm

- Start with a ring
- For $i = 1 \dots n$
 - Select a vertex j with probability proportional to R_{ij} and generate an edge (i,j)
- Repeat until z edges are added to each vertex

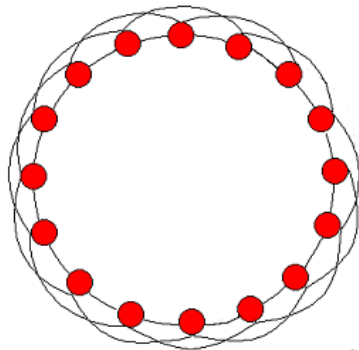
Clustering coefficient – Avg path length



small world graphs

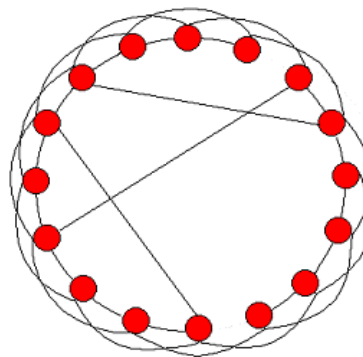
Watts and Strogatz model [WS98]

- Start with a ring, where every node is connected to the next z nodes
- With probability p , **rewire** every edge (or, **add a shortcut**) to a uniformly chosen destination.
 - Granovetter, “The strength of weak ties”

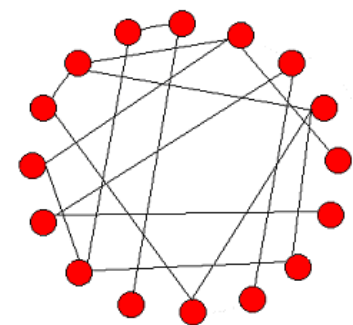


order

$p = 0$



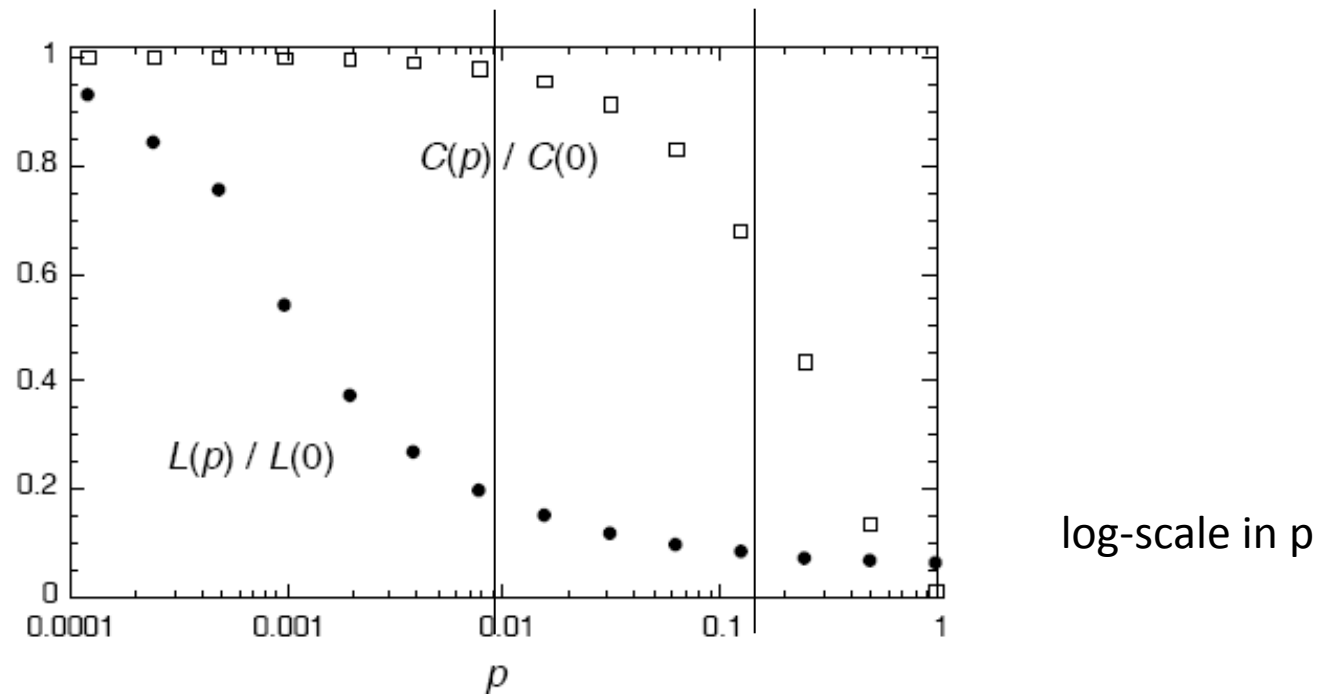
$0 < p < 1$



randomness

$p = 1$

Clustering Coefficient – Characteristic Path Length



When $p = 0$, $C = 3(k-2)/4(k-1) \sim 3/4$
 $L = n/k$

For small p , $C \sim 3/4$
 $L \sim \log n$

Graph Theory Results

- Graph theorist failed to be impressed. Most of these results were known.

Network models and temporal evolution

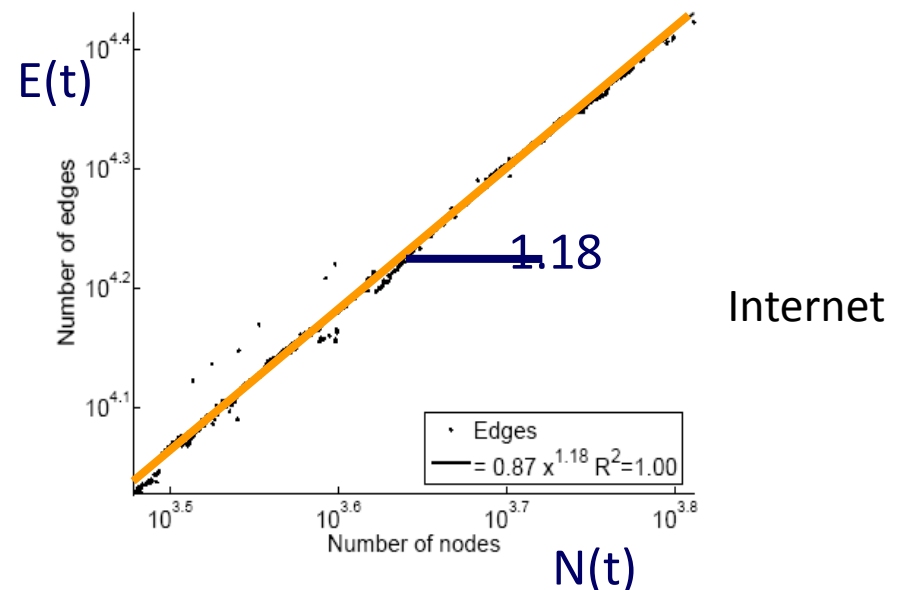
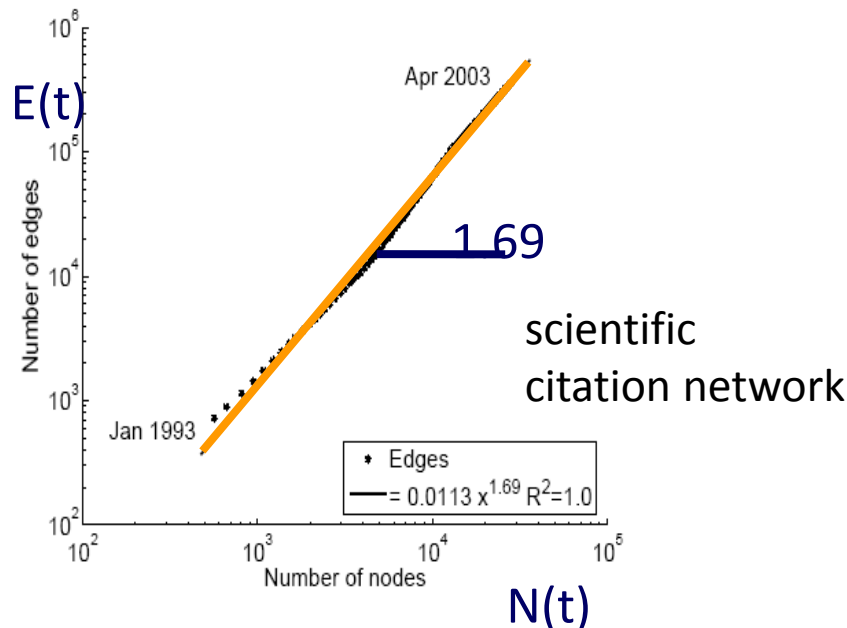
- For most of the existing models it is assumed that
 - number of edges grows linearly with the number of nodes
 - the diameter grows at rate $\log n$, or $\log \log n$
- What about real graphs?
 - Leskovec, Kleinberg, Faloutsos 2005

Densification laws

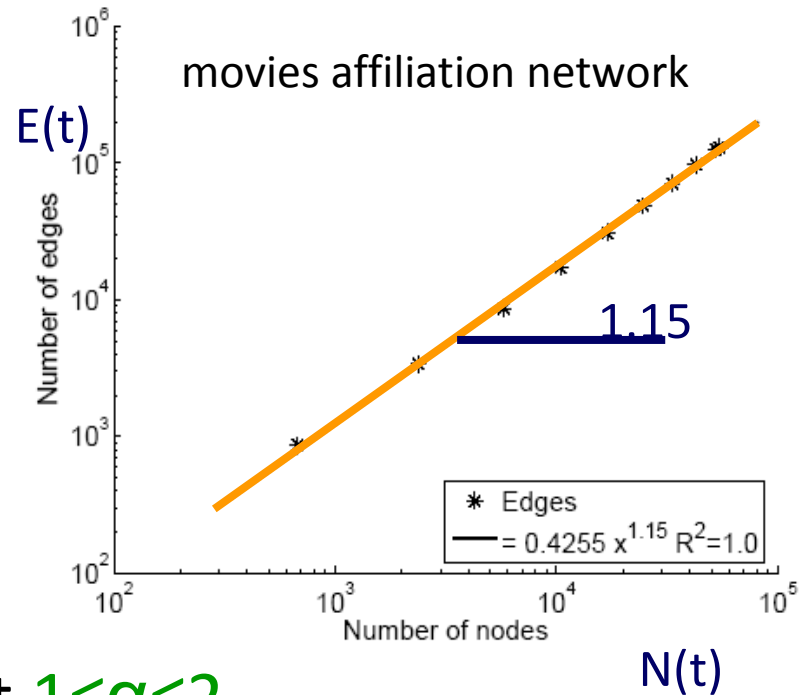
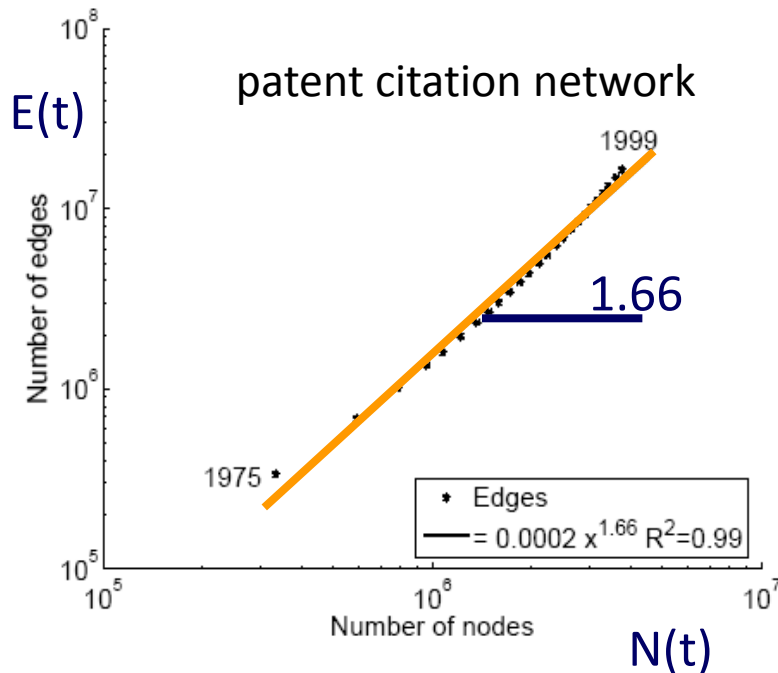
- In real-life networks the average degree increases! – networks become **denser**!

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

α = densification exponent



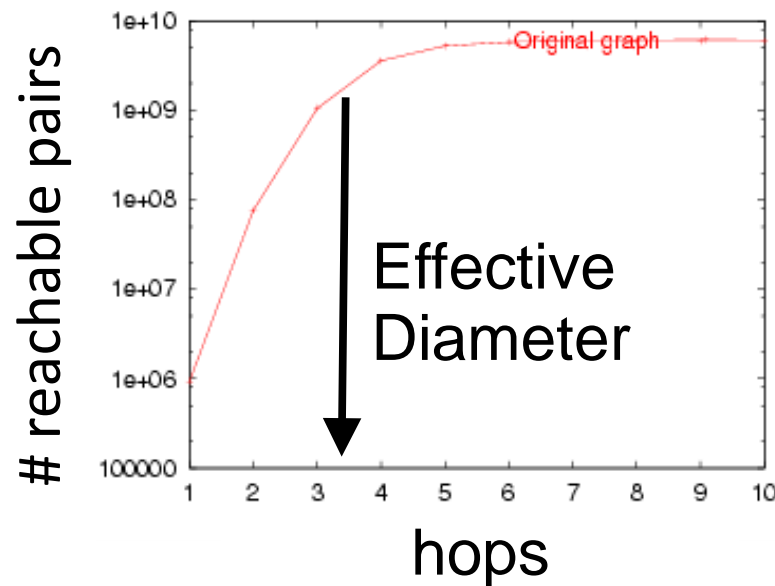
More examples



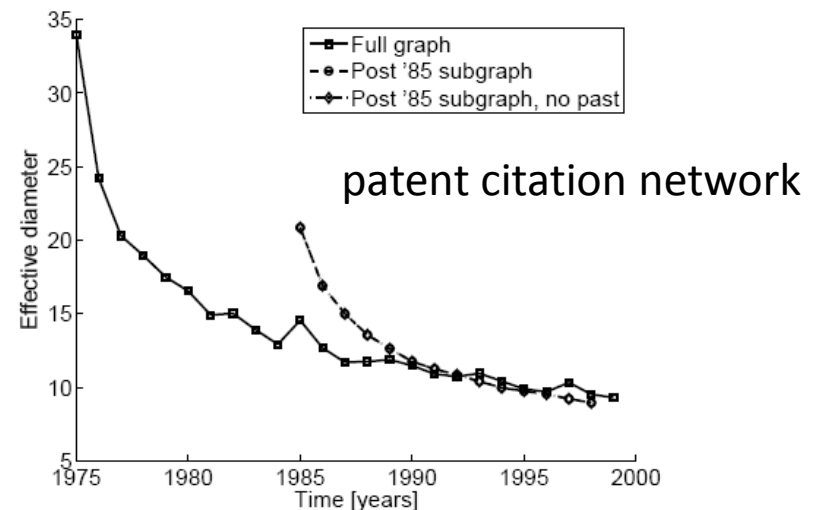
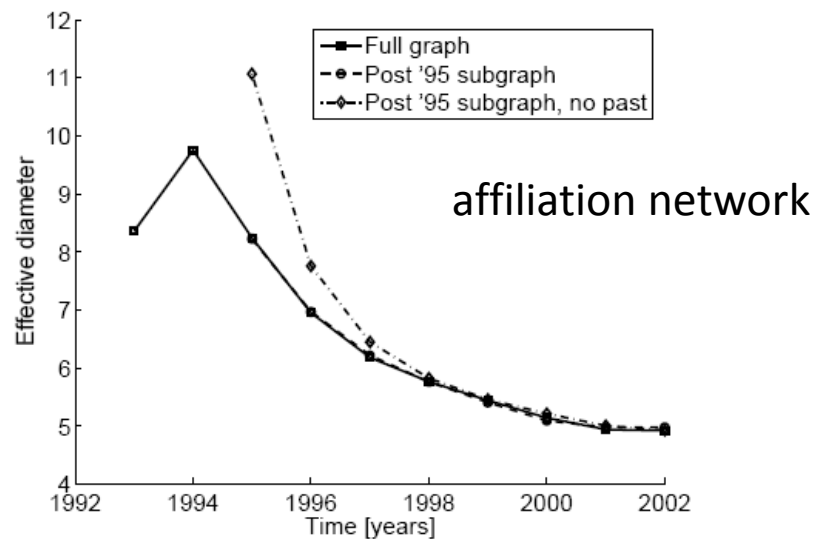
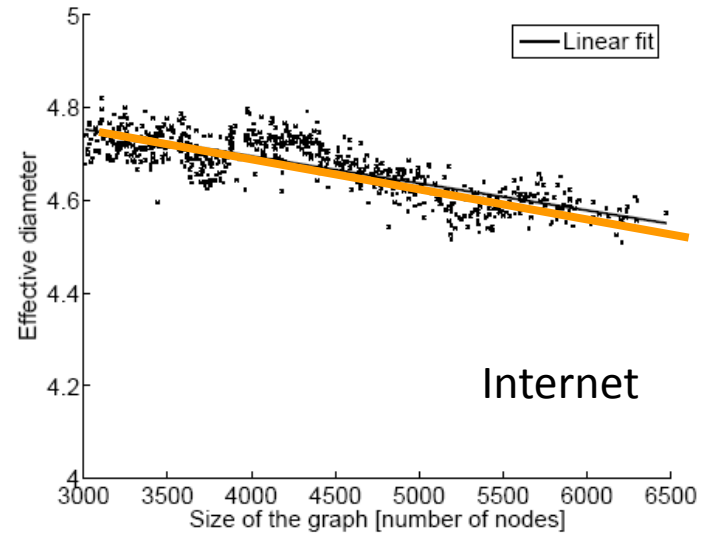
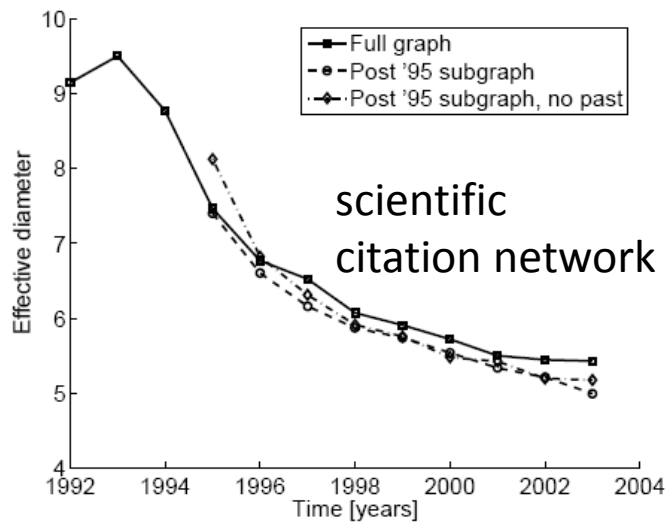
- The densification exponent $1 \leq \alpha \leq 2$
 - $\alpha = 1$: linear growth – constant out degree
 - $\alpha = 2$: quadratic growth - clique

What about diameter?

- **Effective diameter**: the interpolated value where 90% of node pairs are reachable



Diameter shrinks

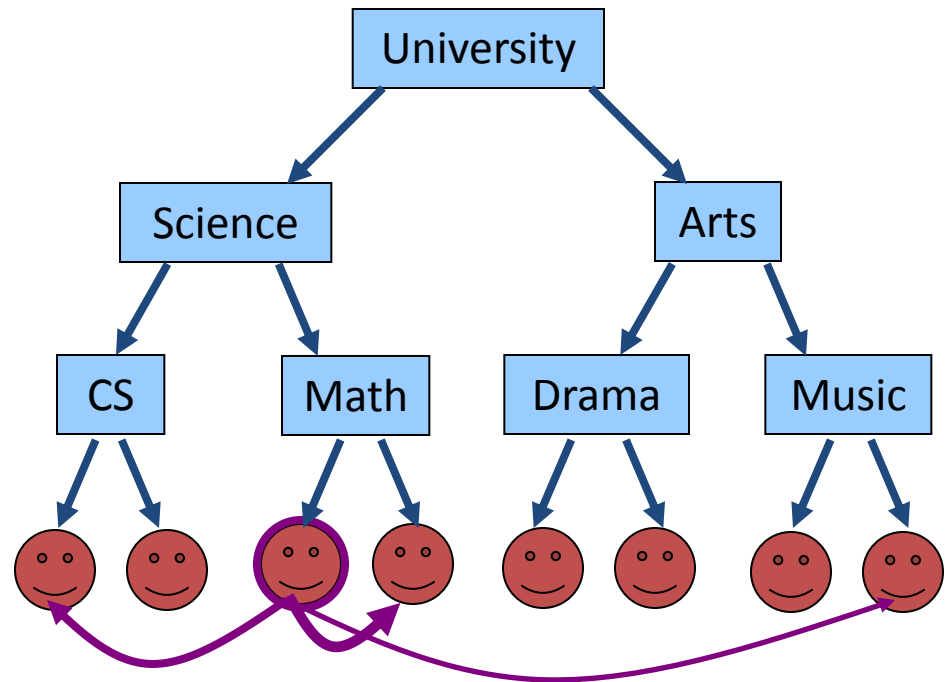


Densification – Possible Explanation

- Existing graph generation models do not capture the **Densification Power Law** and **Shrinking diameters**
- Can we find a simple model of **local** behavior, which naturally leads to observed phenomena?
- Two proposed models
 - **Community Guided Attachment** – obeys Densification
 - **Forest Fire model** – obeys Densification, Shrinking diameter (and Power Law degree distribution)

Community structure

- Let's assume the **community structure**
- One expects many within-group friendships and fewer cross-group ones
- How hard is it to **cross communities?**



Self-similar university
community structure

Fundamental Assumption

- The cross-community linking probability of nodes at tree-distance h (the height of the least common ancestor) is **scale-free**
- We propose cross-community linking probability:

$$f(h) = c^{-h}$$

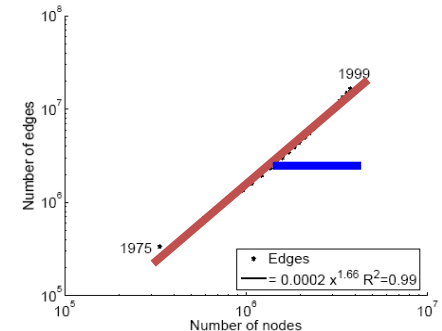
where: $c \geq 1$... the **Difficulty constant**

h ... tree-distance

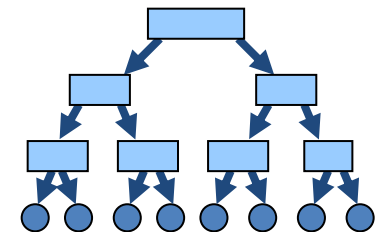
Densification Power Law

- Theorem: The Community Guided Attachment leads to Densification Power Law with exponent

$$a = 2 - \log_b(c)$$



- α ... densification exponent $E(t) \propto N(t)^\alpha$
- b ... community structure branching factor
- c ... difficulty constant



Difficulty Constant

- Theorem:

$$a = 2 - \log_b(c)$$

- Gives any non-integer Densification exponent
- If $c = 1$: easy to cross communities
 - Then: $\alpha = 2$, quadratic growth of edges – near clique
- If $c = b$: hard to cross communities
 - Then: $\alpha = 1$, linear growth of edges – constant out-degree

Room for Improvement

- Community Guided Attachment explains **Densification Power Law**
- Issues:
 - Requires explicit **Community structure**
 - Does not obey **Shrinking Diameters**
- The "Forrest Fire" model

“Forest Fire” model – Wish List

- We want:
 - no explicit Community structure
 - Shrinking diameters
 - and:
 - “Rich get richer” attachment process, to get heavy-tailed in-degrees
 - “Copying” model, to lead to communities
 - Community Guided Attachment, to produce
Densification Power Law

“Forest Fire” model – Intuition

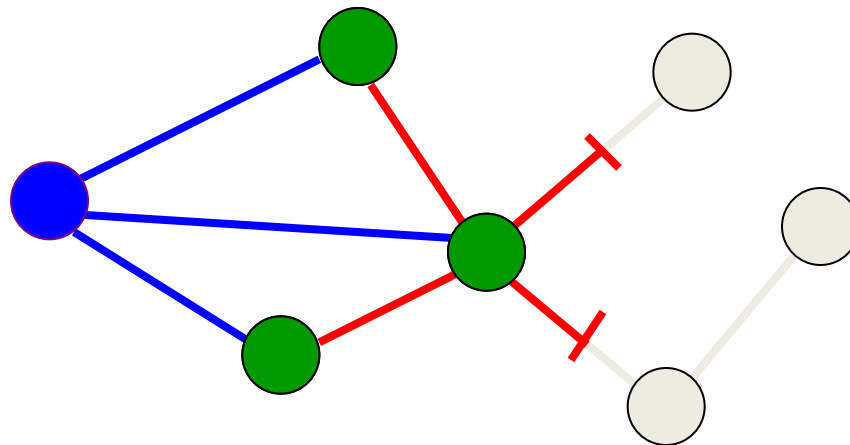
- How do authors identify references?
 1. Find first paper and cite it
 2. Follow a few citations, make citations
 3. Continue recursively
 4. From time to time use bibliographic tools (e.g. Google Scholar) and chase back-links

“Forest Fire” model – Intuition

- How do people make friends in a new environment?
 1. Find first a person and make friends
 2. From time to time get introduced to his friends
 3. Continue recursively
- Forest Fire model imitates exactly this process

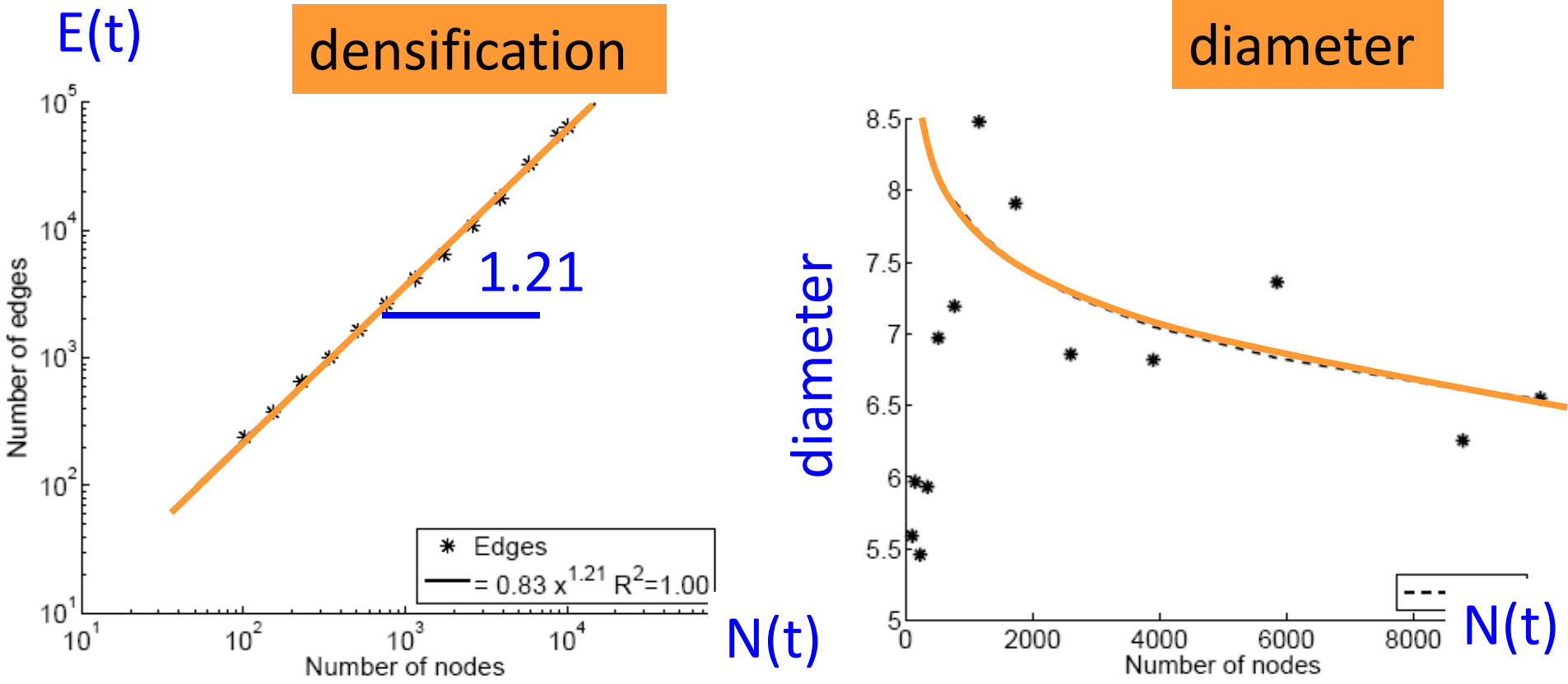
“Forest Fire” – the Model

- A node arrives
- Randomly chooses an “ambassador”
- Starts burning nodes (with probability p) and adds links to burned nodes
- “Fire” spreads recursively



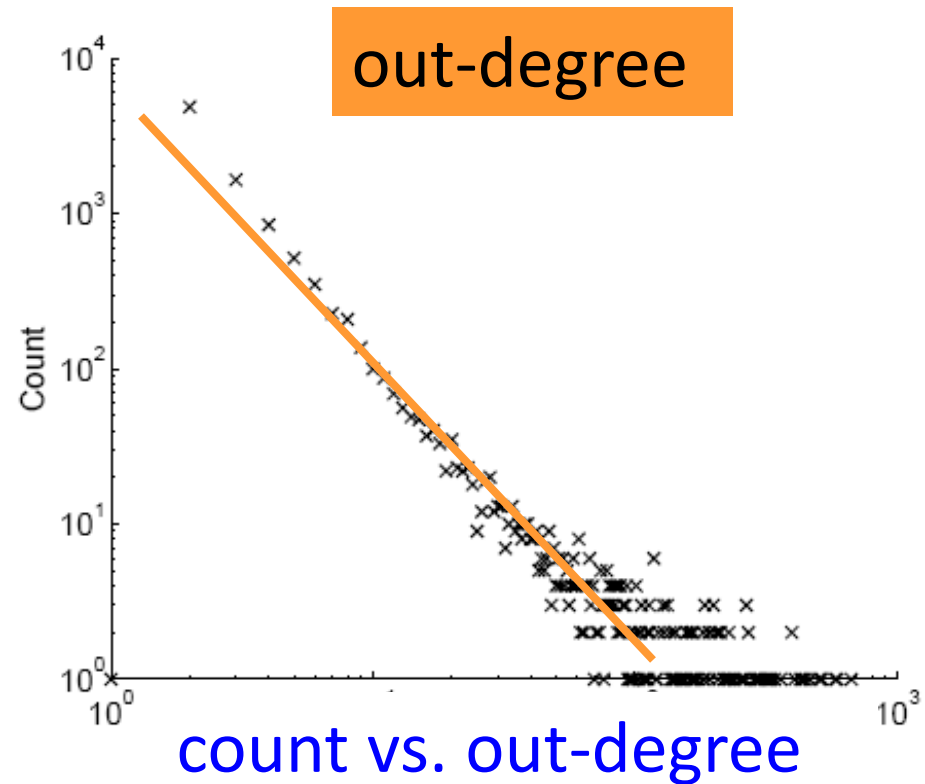
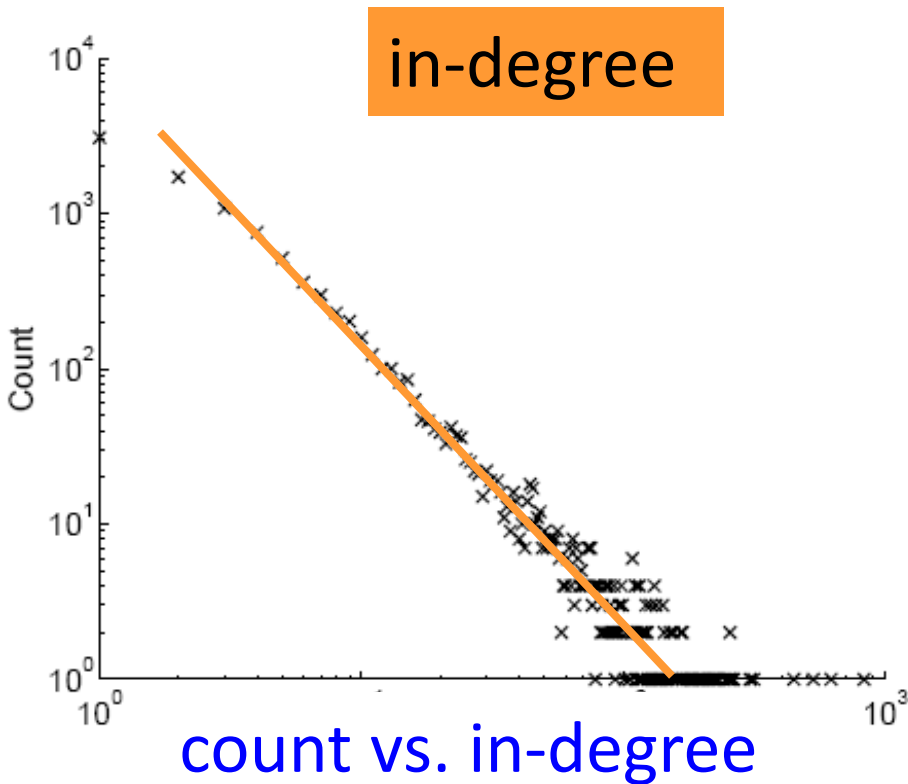
Forest Fire in Action (1)

- Forest Fire generates graphs that **Densify** and have **Shrinking Diameter**



Forest Fire in Action (2)

- Forest Fire also generates graphs with **heavy-tailed degree distribution**



Forest Fire model – Justification

- **Densification Power Law:**
 - Similar to Community Guided Attachment
 - The probability of linking decays exponentially with the distance – Densification Power Law
- **Power law out-degrees:**
 - From time to time we get large fires
- **Power law in-degrees:**
 - The fire is more likely to reach hubs

Forest Fire model – Justification

- Communities:
 - Newcomer copies neighbors' links
- Shrinking diameter

Acknowledgements

- Many thanks to Jure Leskovec for his slides from the KDD 2005 paper

References

- M. E. J. Newman, [The structure and function of complex networks](#), SIAM Reviews, 45(2): 167-256, 2003
- R. Albert and L.A. Barabasi, [Statistical Mechanics of Complex Networks](#), Rev. Mod. Phys. 74, 47-97 (2002).
- B. Bollobas, [Mathematical Results in Scale-Free random Graphs](#)
- D.J. Watts. Networks, [Dynamics and Small-World Phenomenon](#), American Journal of Sociology, Vol. 105, Number 2, 493-527, 1999
- Watts, D. J. and S. H. Strogatz. [Collective dynamics of 'small-world' networks](#). Nature 393:440-42, 1998
- Michael T. Gastner and M. E. J. Newman, [Optimal design of spatial distribution networks](#), *Phys. Rev. E* **74**, 016117 (2006).J.
- Leskovec, J. Kleinberg, C. Faloutsos. [Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations](#). Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2005.