Network Science: Real Networks and Universal Properties

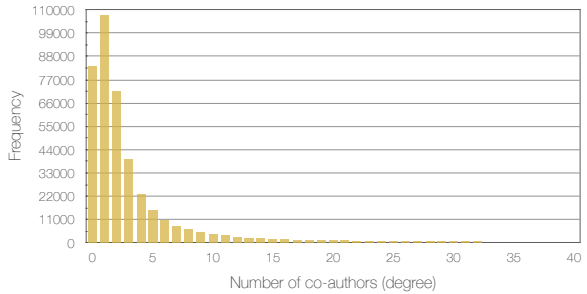
Ozalp Babaoglu
Dipartimento di Informatica — Scienza e Ingegneria
Università di Bologna
www.cs.unibo.it/babaoglu/

Universal structural properties

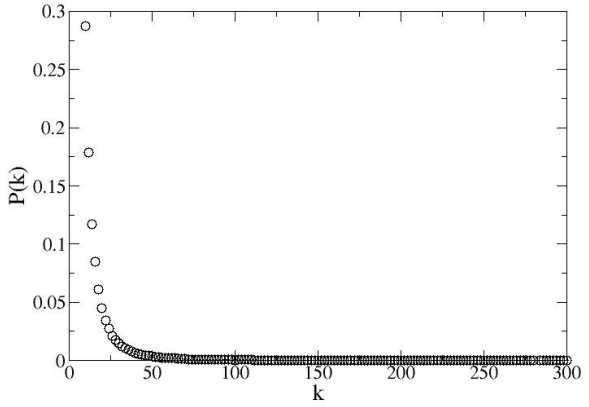
- Networks are typically very different at the microscopic level
- Are there macroscopic structural properties that are common to a large variety of real networks?
- Universal structural properties:
  - Heavy-tailed degree distributions — "hubs", "connectors"
  - Small diameter — "six degrees of separation"
  - Highly clustered — "friends of a friend are friends"
  - Well connected — only one giant component
- Need to make precise the notions "heavy", "small", "highly" and "well"
- Examine real networks to support or refute the claims

Degree distributions

- Recall the *Math Reviews co-authorship network*
  - 401,000 different authors (nodes)
  - 676,000 edges
  - Average number of co-authors per author is 3.36



Heavy-tailed distributions



Heavy-tailed distributions

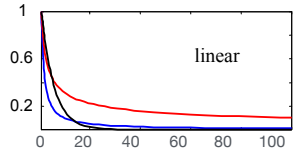
- Nodes with small degree are most frequent
- Fraction of high-degree nodes decreases but much more slowly than what is predicted by the random models with Poisson or Normal degree distributions which decay exponentially
- Typical of networks that have a few *hub* or *connector* nodes with very high degree and many nodes with small degree
- What are the *signatures* of heavy-tailed distributions?

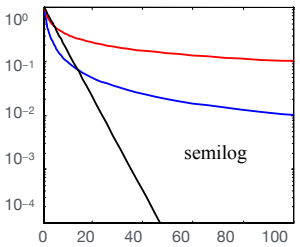
Plotting degree distributions

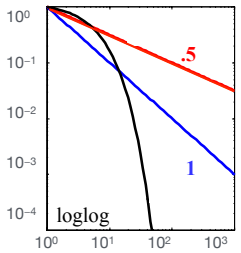
- Examine closely two different forms for the distribution function:
- Exponential: $f(x)=c^{-x}$
- Power-law: $f(x)=cx^{-\alpha}$
- Plot the two forms on different choices for the scales

Plotting degree distributions

Exponential
Power-law $\alpha=0.5$
Power-law $\alpha=1$







Plotting degree distributions

- A straight line on a log-log scale becomes the *signature* of power-law distributions

$$f(x) = cx^{-\alpha}$$

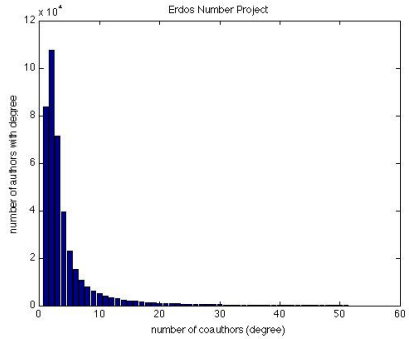
$$\log(f(x)) = \log(cx^{-\alpha})$$

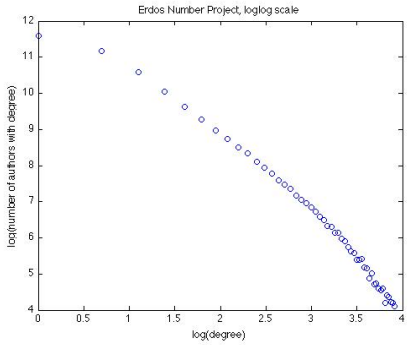
$$\log(f(x)) = \log(c) + \log(x^{-\alpha})$$

$$\log(f(x)) = \log(c) - \alpha\log(x)$$

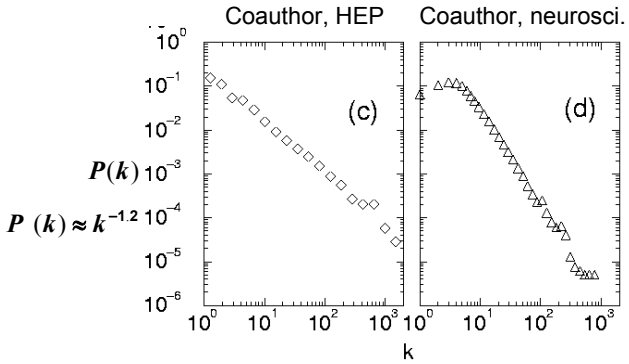
- If we plot $\log(f(x))$ as a function of $\log(x)$, we obtain a straight line with slope $-\alpha$

Power-law distributions in the wild
Math Reviews co-authorship





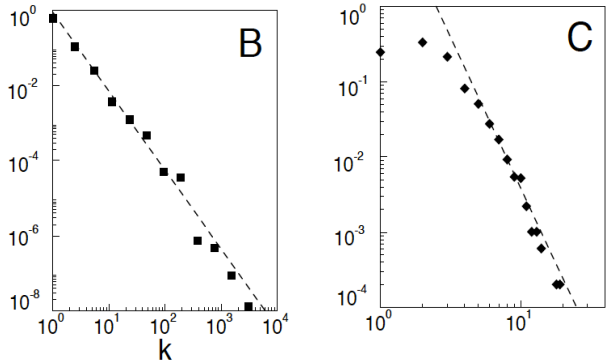
Power-law distributions in the wild
More co-authorships



$P(k) \approx k^{-1.2}$

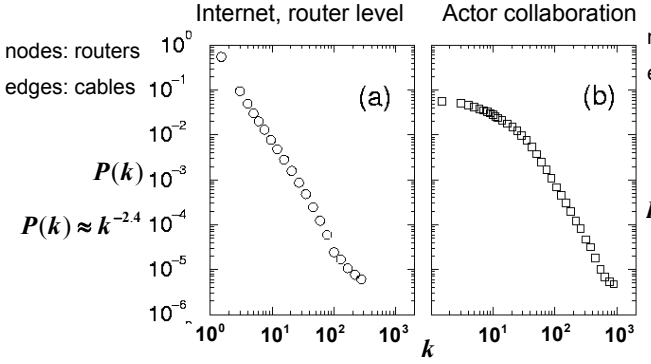
$P(k) \approx k^{-2.1}$

Power-law distributions in the wild
Web, power grid



Power-law distributions in the wild
Internet routers, Actor collaboration

nodes: routers
edges: cables

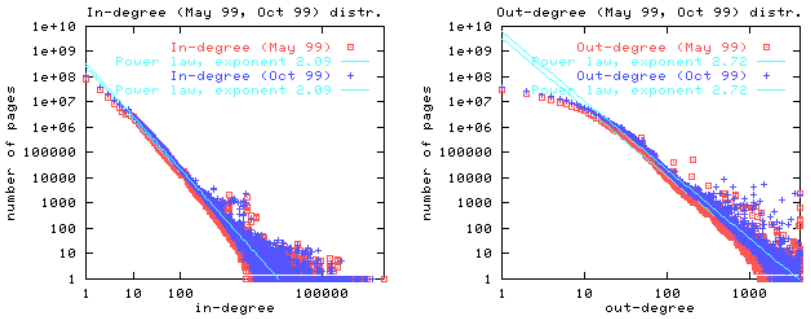


nodes: actors
edges: cast jointly

$P(k) \approx k^{-2.4}$

$P(k) \approx (31+k)^{-3}$

Power-law distributions in the wild
Web graph



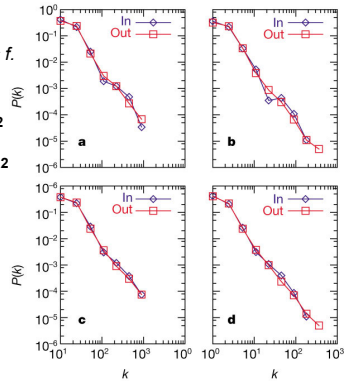
Power-law distributions in the wild
Metabolic networks

Archaeoglobus f.

$$P_{in}(k) \approx k^{-2.2}$$

$$P_{out}(k) \approx k^{-2.2}$$

C. elegans



E. coli

bipartite
nodes: metabolites, reactions
directed edges,
out: reactant (substrate)
in: product of reaction

Power-laws and popularity

- Power-laws arise in many settings other than degree distributions
- *Popularity* of actors, books, movies, songs, web pages are some examples
- Popularity is a phenomenon characterized by extreme imbalances due to network effects
- Result of positive feedback or reinforcement due to correlated decisions in a population
- The “rich-get-richer” phenomenon

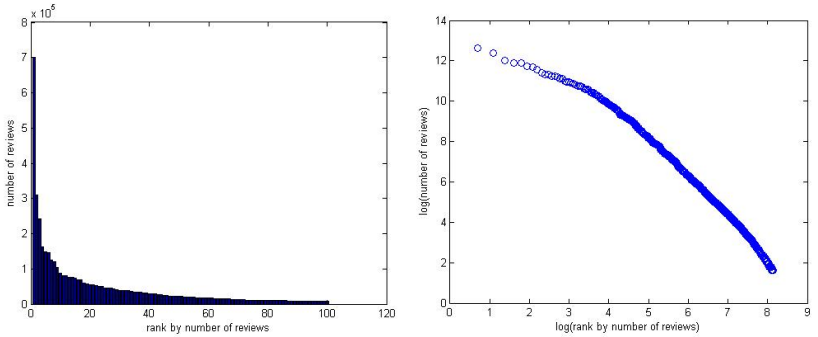
Popularity of Web pages

- Use the number of in-edges as a measure of popularity
- As a function of k, what fraction of Web pages have k in-edges?
- Supposes pages decide independently and randomly to link to other pages
- Then, the total number of in-edges at a Web page would be the sum of (many) independent random quantities — the presence of absence of a link from other pages
- By the *Central Limit Theorem*, we would expect the distribution of the number of in-edges at a page to be *normal* (“bell curve”)
- In other words, the number of Web pages with k in-edges should decay exponentially as k grows large

Rich-get-richer

- Yet, for the Web, the fraction of pages that have k in-edges follows a power-law and is approximately proportional to k^{-2}
- The fraction of cities with population k is roughly k^{-c}
- The fraction of books that have sold k copies is roughly k^{-c}
- Switching from "blockbuster" view to "niche" view, a power-law function remains power-law
  - The fraction of songs that have been downloaded k times is roughly k^{-c}
  - The number of times that the k-th most popular song has been downloaded is roughly k^{-c}

Rich-get-richer
iPhone app popularity



Rich-get-richer

- Once the rich-get-richer process gets going, the dynamics of popularity continue to enforce it
- But, how does the process get ignited in the first place?
- During the early phases, the process is very sensitive to unpredictable fluctuations
- What would happen if we could roll-back time and repeat history?
- Reasonable to expect popularity to obey power-law in each instance
- But not necessarily with the *same* ranking of popularity

Rich-get-richer

- Difficult to roll back time and repeat history
- But, we can conduct experiments to see what happens
- Salgankik, Dodds, and Watts designed such an experiment
- Created fake music download site populated with 48 obscure songs written by real groups
- Visitors could listen to the songs, see their "download count" and download them it they wanted to

Rich-get-richer

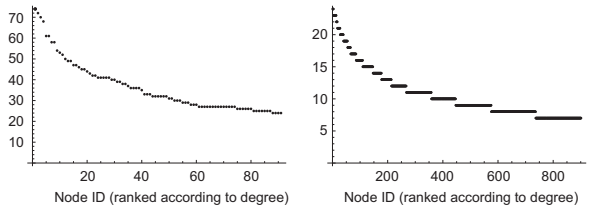
- In reality, there were 8 "parallel" copies of the site and each visitor was assigned to one at random on arrival
- The parallel sites started out in identical states with the same list of 48 songs but evolved independently
- In the end, the relative popularity of the 48 songs varied considerably among the 8 sites (although the "best" songs were never in the bottom and "worst" songs were never in the top)
- Some users were directed to a 9th site that had no "feedback" through download counters
- In this site, there was significantly less variation among the popularities of the 48 songs

Scale-free networks

- Networks with degree distributions that are described by power-laws are also called *scale-free*
- A function $f(x)$ is called *scale-free* if
  $f(bx) = C(b) \cdot f(x)$
  where $C(b)$ is some constant that depends only on b
- In other words, the overall form of the function does not change when considering values for x that are a factor b larger
- Related to *fractals* in mathematics

Scale-free networks

- Power-law distributions are scale-free
- Let $f(x)$ be a power-law function: $f(x)=cx^{-\alpha}$
  $f(bx) = c(bx)^{-\alpha} = b^{-\alpha} c\, x^{-\alpha} = C(b) \cdot f(x)$

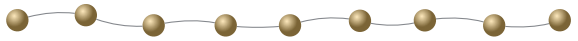


- Different but the *form* of the function remains the same

Diameters, path lengths

- Consider a *connected* network
- Recall definition of *diameter*: the longest shortest path
- Smallest diameter: 1 (independent of n)

- Largest diameter: $n-1$ (grows linearly in n)



Diameters, path lengths

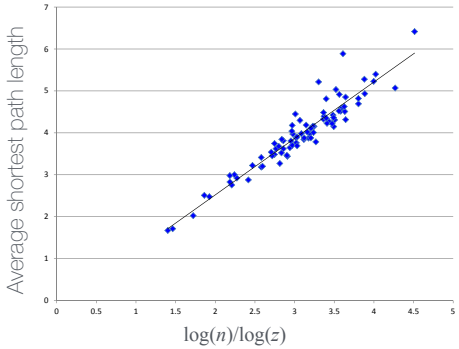
- Network exhibits *small* diameter if it is not constant but grows *sublinearly* with network size — $\log n$, $\log\log n$, etc.
- Travers and Milgram (1969)
  - diameter ~5-6, n ~200M
- Economics co-authorship (2004)
  - diameter ~9.5, n ~81,000
- Microsoft messenger (2008)
  - diameter ~6.5, n ~180M
- Facebook social graph
  - diameter ~4.74, n ~7.21M (2012)
  - diameter ~3.57, n ~1.59B (2016)
  - Facebook social graph diameter has been *shrinking*

Diameters, path lengths

- Alternative definition for diameter: expected shortest path distance between a random pair of nodes
- Thus, diameters and path lengths behave similarly
- Let z denote the average node degree
- Under some weak assumptions, it can be shown that for large n, the average shortest path length and the diameter are roughly proportional to $\log(n)/\log(z)$

Diameters, path lengths

- "Ad health" dataset from 84 high schools



Diameters, path lengths

- "Six degrees of separation" confirmation
- Take the current world adult population as 7 billion people
- Assume each person knows on the average 50 other people among friends, relatives, colleagues, etc.
- Then, $\log(n)/\log(z) = \log(7\times10^9)/\log(50) = 5.79$

Clustering coefficient

- Recall *clustering coefficient* of a node: probability that two randomly selected *friends* of it are friends themselves — probability that the "triangle" closed
- Recall *edge density* of a network: actual number of edges in proportion to the maximum possible number of edges
- Recall we consider a network to exhibit *high clustering* if the clustering coefficient is significantly greater than the edge density
- For the "Fiorentine family network", the clustering coefficient is 0.46 and the edge density is 2×20/(16×15)=0.1666, so the network is highly clustered

Real networks
Summary

Network	n	z	ℓ	$\frac{\log n}{\log z}$	CC	ρ
WWW, site level, undir.	153 127	35.21	3.1	3.35	0.1078	0.00023
Internet, domain level	3015–6209	3.52–4.11	3.7–3.76	6.36–6.18	0.18–0.3	0.001
Movie actors	225 226	61	3.65	2.99	0.79	0.00027
LANL co-authorship	52 909	9.7	5.9	4.79	0.43	1.8×10^{-4}
MEDLINE co-authorship	1 520 251	18.1	4.6	4.91	0.066	1.1×10^{-5}
SPIRES co-authorship	56 627	173	4.0	2.12	0.726	0.003
NCSTRL co-authorship	11 994	3.59	9.7	7.34	0.496	3×10^{-4}
Math. co-authorship	70 975	3.9	9.5	8.2	0.59	5.4×10^{-5}
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}
E. coli, substrate graph	282	7.35	2.9	3.04	0.32	0.026
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03
Words, co-occurrence	460.902	70.13	2.67	3.03	0.437	0.0001
Words, synonyms	22 311	13.48	4.5	3.84	0.7	0.0006
Power grid	4941	2.67	18.7	12.4	0.08	0.005
C. Elegans	282	14	2.65	2.25	0.28	0.05

z average node degree $\quad$ ℓ average path length

CC clustering coefficient $\quad$ ρ edge density

Real networks
Universal properties

- Heavy-tailed degree distribution
- Small diameter and average path length
- Highly clustered
- Very few (typically just one) connected components
- Is there a natural, simple *model* of network formation and growth that can explain how these properties arise?