Introduction to Network Science

I. Makarov & L.E. Zhukov

BigData Academy MADE from VK

Social Network Analysis and Machine Learning on Graphs

Class Details

- Instructor: Ilya Makarov
- Tutor: Vitaly Pozdnyakov
- Course length: 10 lectures/classes
- Invited lecturers: Andrey Kuznetsov, Nikolay Anokhin, Dmitrii Kiselev + connected course on Large Scale RecSys
- Lecturer's Telegram: @iamakarov (urgent)
- Tutor's Telegram: @pozdnyakov_vitaliy (urgent)
- Discord (questions, deadlines, grading)
- Programming: Python, iPython notebooks, Anaconda distribution
- Python libraries: NetworkX, pyG, DGL
- Visualization: yEd, Gephi

Prerequisites

- Discrete Mathematics
- Linear Algebra
- Algorithms and Data Structures
- Probability Theory
- Differential Equations
- Programming in Python

Textbooks

- "Network Science", Albert-Laszlo Barabasi, Cambridge University Press, 2016. http://networksciencebook.com
- "Networks: An Introduction". Mark Newman. Oxford University Press, 2010.
- "Social Network Analysis. Methods and Applications". Stanley Wasserman and Katherine Faust, Cambridge University Press, 1994
- "Networks, Crowds, and Markets: Reasoning About a Highly Connected World". David Easley and John Kleinberg, Cambridge University Press 2010.

Topics

- Statistical properties and network modeling
- Network structure and dynamics
- Processes on networks
- Predictions on networks (ML)
- Network embeddings
- Graph neural networks
- Knowledge graph retrieval and completion
- Spark and BigData for relational Data
- Distributed ML on large graphs

Descriptive Analysis

1. Introduction to network science
2. Power law and scale-free networks
3. Random graphs
4. Small world and dynamical growth models
5. Centrality measures
6. Pagerank and link analysis
7. Structural similarity in networks
8. Network cores and community structure
9. Graph partitioning algorithms
10. Community detection

Network science

- Sociology (SNA)
- Mathematics (Graphs)
- Computer Science (Graphs)
- Statistical Physics (Complex networks)
- Economics (Networks)
- Bioinformatics (Networks)

Research Journals

- IEEE Transactions on Network Science and Engineering [IEEE, 2014]
- Network Science [Cambridge University Press, 2013]
- Journal of Complex Networks [Oxford Academic, 2013]
- Social Network Analysis and Mining [Springer, 2011]
- Social Networks [Elsevier, 1979]

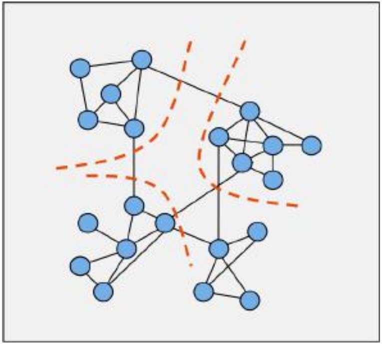
- Applied Network Science (ANS) [SpringerOpen, 2017]
- Int. Journal of Network Science, [InderScience Publishers, 2016]
- Computational Social Networks [SpringerOpen, 2014]
- Journal of Complex Networks [Oxford Academic, 2013]
- Social Networking [Scientific Research publisher, 2012]
- International Journal of Social Network Mining (IJSNM) [InderScience Publishers, 2012]

Conferences

- International School and Conference on Network Science, NetSci, NetSci-X
- International Workshop on Complex Networks and their Applications, Complenet
- SIAM Workshop on Network Science
- International Conference on Computational Social Science, IC^2S^2
- International Conference on Social Network Analysis, INSNA
- StatPhys Satellite Conference Complex Networks
- ACM Conference on Online Social Networks
- Conference on Complex Systems

Terminology

- network = graph
- nodes = vertices, actors
- links = edges, relations
- clusters = communities



Graph theory

- **Network** is represented by a graph $G(V, E)$, comprising a set of vertices V and a set of edges E, connecting those vertices.
- **Graph** can be represented by an adjacency matrix A, where A_{ij} - availability of an edge between nodes i and j
- In an **unweighted graph** A_{ij} is binary $\{0, 1\}$, in a **weighted graph** an edge can carry a weight, A- non-binary.
- **Undirected graph** is a graph where edges have no orientation, edges are defined by unordered pairs of vertices, $A_{ij} = A_{ji}$
- **Directed graph** is a graph where edges have a direction associated with them, edges are defined by ordered pairs of vertices, $A_{ij} \neq A_{ji}$

Graph theory

- A **path** between nodes i and j is a sequence of edges connecting vertices, starting at i and ending at j, where every vertex in the sequence is distinct
- **Distance** between two vertices in a graph is the number of edges in a shortest path (graph geodesic) connecting them.
- The **diameter** of a network is the largest shortest paths (distance between any two nodes) in the network
- **Average path length** is bounded from above by the diameter; in some cases, it can be much shorter than the diameter

Graph theory

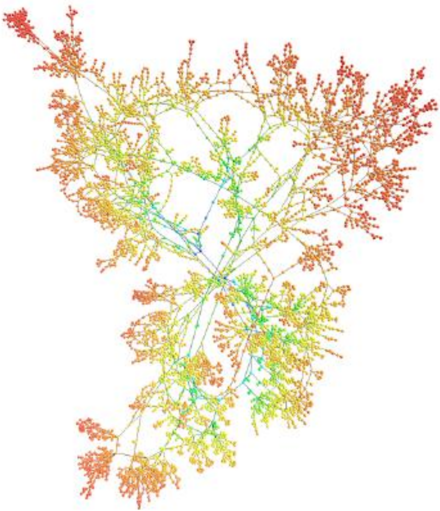
- A graph is **connected** when there is a path between every pair of vertices.
- A **connected** component is a maximal connected subgraph of the graph. Each vertex belongs to exactly one connected component, as does each edge.
- A directed graph is called **weakly connected** if replacing all of its directed edges with undirected edges produces a connected (undirected) graph.
- A directed graph is **strongly connected** if it contains a directed path between every pair of vertices. A directed graph can be connected but not strongly connected.

Graph theory

- The **degree** of a vertex of a graph is the number of edges incident to the vertex
- A vertex with degree 0 is called an **isolated vertex**.
- In a directed graph the number of head ends adjacent to a vertex is called the **in-degree** of the vertex and the number of tail ends adjacent to a vertex is its **out-degree**
- A vertex with in-degree=0 is called a **source vertex**, with out-degree=0 is a **sync vertex**

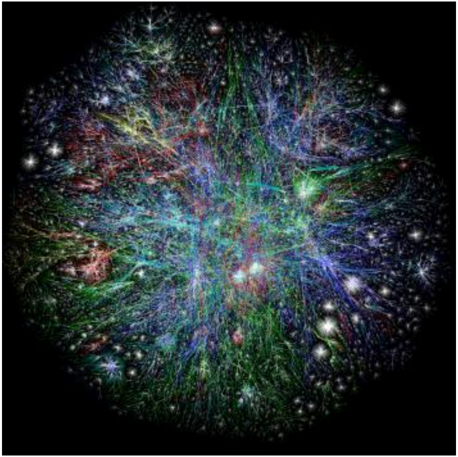
Complex networks

- not regular, but not random
- non-trivial topology
- scale-free networks
- universal properties
- everywhere
- complex systems



Examples: Internet

Internet traffic routing (BGP)



Barret Lyon, 2003

Examples: Social network - Facebook friendship

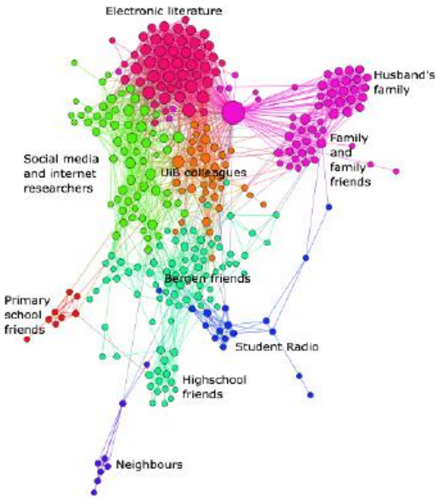


image from Jill Walker Rettberg, jiltttxt.net

Examples: Political blogs

red-conservative blogs, blue -liberal, orange links from liberal to conservative, purple from conservative to liberal

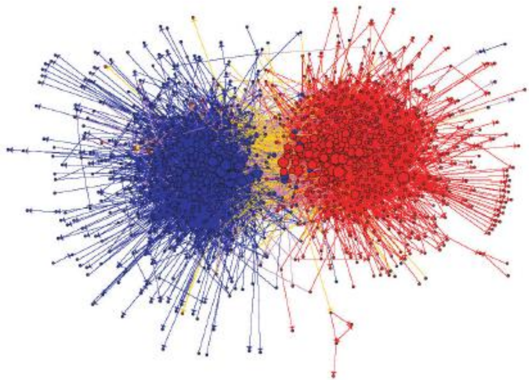


image from L. Adamic, N. Glance, 2005

Examples: Communications

Enron emails

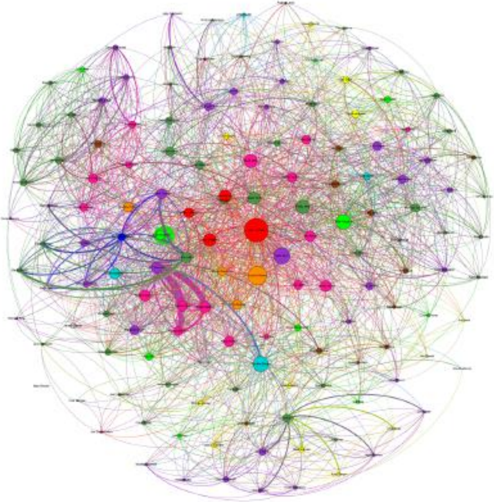
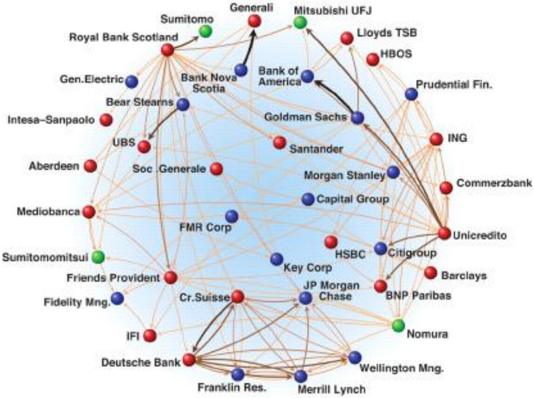


image from www.philipstarritt.com/

Examples: Finance

existing relations between financial institutions



F. Schweitzer, 2009

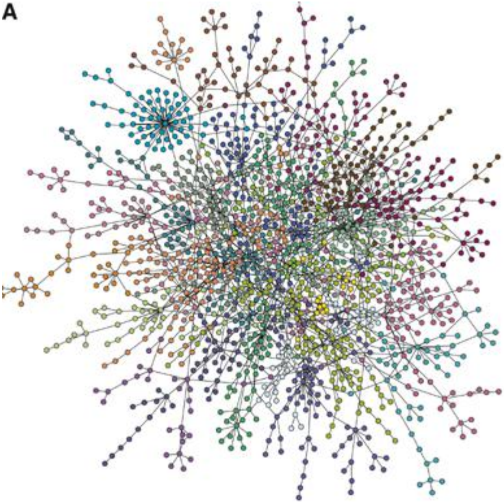
Examples: Transportation

Zurich public transportation map

image from http://www.visualcomplexity.com

Examples: Biology

Yeast protein interaction network



H. Jeong et.al., 2001

Facebook

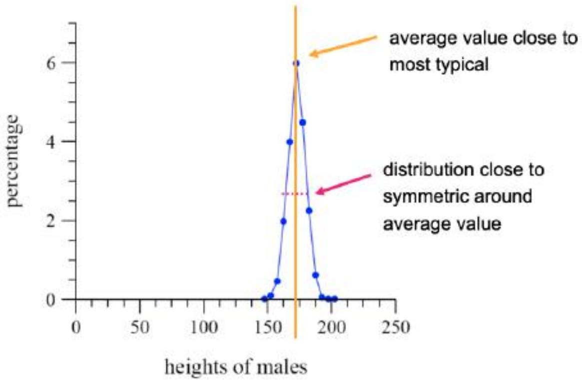
Friendship graph 500 mln people

image by Paul Butler, 2010

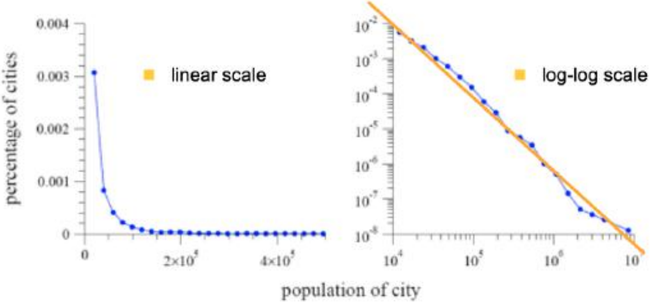
Complex networks

1. Power law node degree distribution: "scale-free" networks
2. Small diameter and average path length: "small world" networks
3. Hight clustering coefficient: transitivity

Typical normal distributions



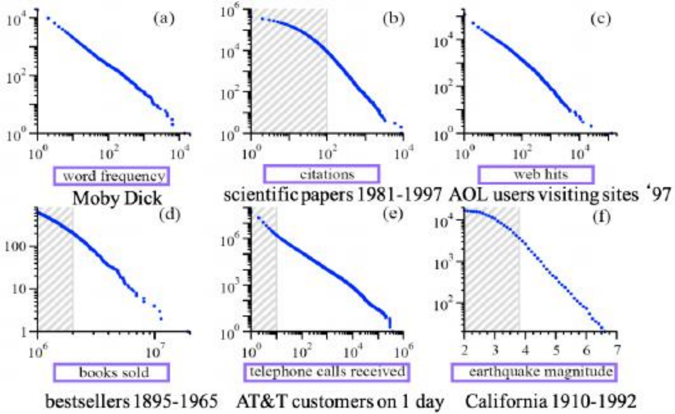
Power law distributions



$$f(k) = \frac{C}{k^{\gamma}} = Ck^{-\gamma}$$

$$\log f(k) = log\ C - \gamma log\ k$$

Power law distributions



Power law

Quantity of interest - frequency distribution of node degrees

$$f(k) \sim \frac{1}{k^\gamma}$$

- "A study of large sociogram", Anatol Rapoport and William Horrah, 1961
- "Networks of Scientific Papers", Derec J. de Solla Price, 1965
- "Diameter of the World-Wide Web", Reka Albert, Hawoong Jeong, Albert-Laszlo Barabasi, 1999
- "The Web as a graph: Measurements, models and methods", Jon Kleinberg et. al, 1999

Power law

Citation of scientific papers for 1961

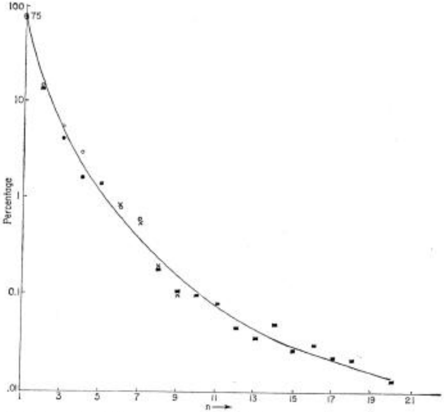
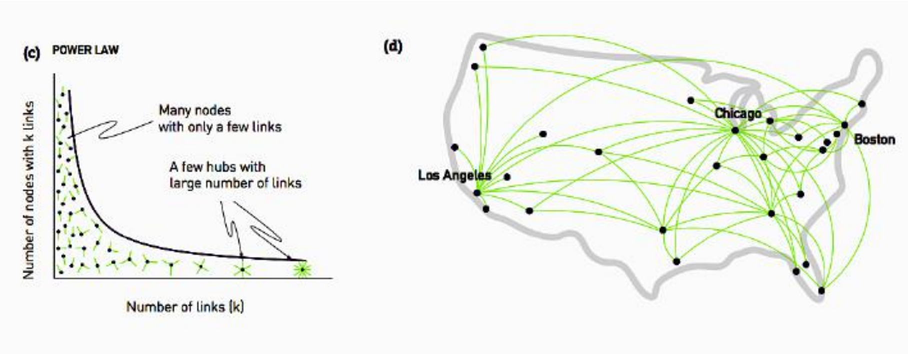


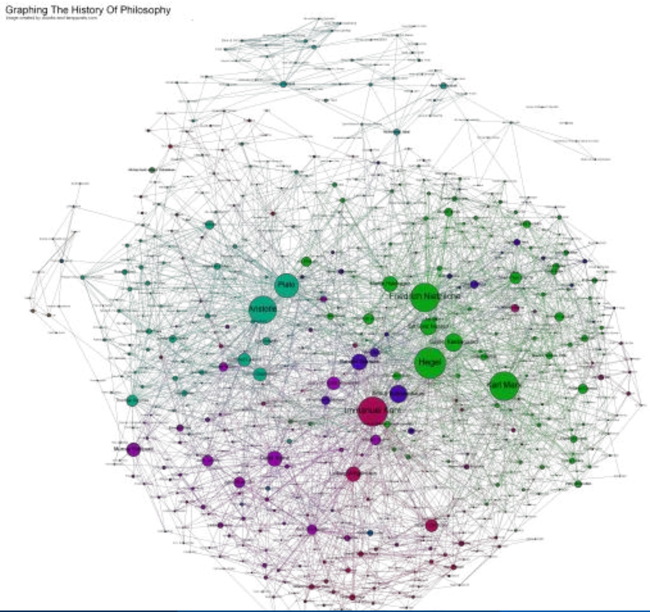
Fig. 2. Percentages (relative to total number of cited papers) of papers cited various numbers (n) of times, for a single year (1961). The data are from Garfield's 1961

from D.Price, 1965

Power law degree distribution



Power law



Power law

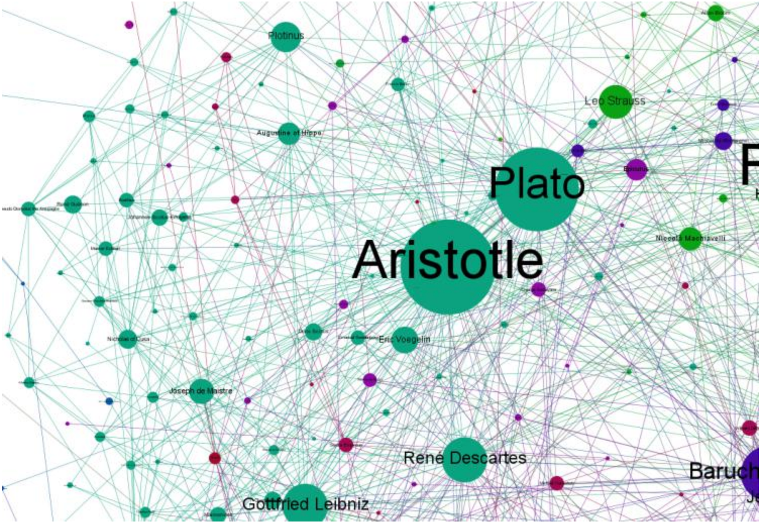


image from http://www.coppelia.io

Triads

The Strength of Weak Ties[1]

Mark S. Granovetter
Johns Hopkins University

Analysis of social networks is suggested as a tool for linking micro and macro levels of sociological theory. The procedure is illustrated by elaboration of the macro implications of one aspect of small-scale interaction: the strength of dyadic ties. It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another. The impact of this principle on diffusion of influence and information, mobility opportunity, and community organization is explored. Stress is laid on the cohesive power of weak ties. Most network models deal, implicitly, with strong ties, thus confining their applicability to small, well-defined groups. Emphasis on weak ties lends itself to discussion of relations *between* groups and to analysis of segments of social structure not easily defined in terms of primary groups.

- "The Strength of Weak Ties", Mark Grannoveter, 1973
- "Spread of Information through a Population with Socio-Structural Bias. Assumption of Transitivity", Anatol Rapoport, 1953

Triadic closure

- strength of a tie
- high transitivity
- high clustering coefficient

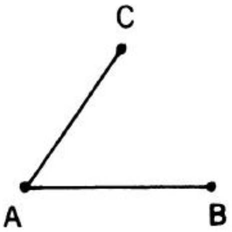


FIG. 1.—Forbidden triad

If A and B and B and C are strongly linked, the the tie between B and C is always present

Grannoveter, 1973

High clustering

Facebook friendship

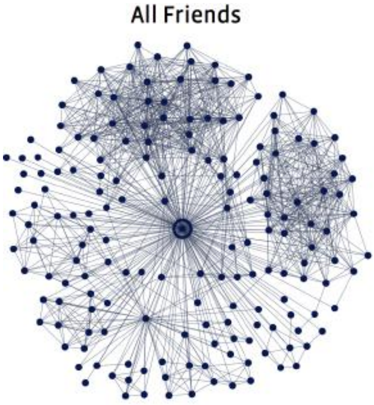
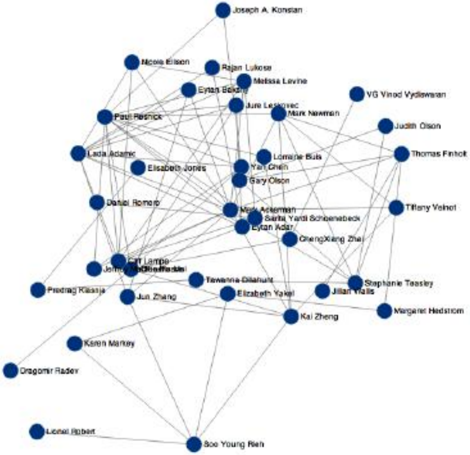


image from Cameron Marlow, Facebook

Co-author network



Small world: six degrees of separation

© Al Satterwhite

An Experimental Study of the Small World Problem*

JEFFREY TRAVERS

Harvard University

AND

STANLEY MILGRAM

The City University of New York

Arbitrarily selected individuals (N=296) in Nebraska and Boston are asked to generate acquaintance chains to a target person in Massachusetts, employing "the small world method" (Milgram, 1967). Sixty-four chains reach the target person. Within this group the mean number of intermediaries between starters and targets is 5.2. Boston starting chains reach the target person with fewer intermediaries than those starting in Nebraska; subpopulations in the Nebraska group do not differ among themselves. The funneling of chains through sociometric "stars" is noted, with 48 per cent of the chains passing through three persons before reaching the target. Applications of the method to studies of large scale social structure are discussed.

- "The small-world problem". Stanley Milgram, 1967
- "An experimental study of the small world problem", Jeffrey Travers, Stanley Milgram, 1969

Stanley Milgram's 1967 experiment

HOW TO TAKE PART IN THIS STUDY

1. ADD YOUR NAME TO THE ROSTER AT THE BOTTOM OF THIS SHEET, so that the next person who receives this letter will know who it came from.

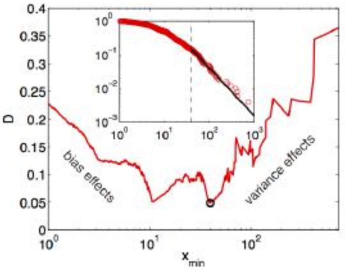
2. DETACH ONE POSTCARD. FILL IT OUT AND RETURN IT TO HARVARD UNIVERSITY. No stamp is needed. The postcard is very important. It allows us to keep track of the progress of the folder as it moves toward the target person.

3. IF YOU KNOW THE TARGET PERSON ON A PERSONAL BASIS, MAIL THIS FOLDER DIRECTLY TO HIM (HER). Do this only if you have previously met the target person and know each other on a first name basis.

4. IF YOU DO NOT KNOW THE TARGET PERSON ON A PERSONAL BASIS, DO NOT TRY TO CONTACT HIM DIRECTLY. INSTEAD, MAIL THIS FOLDER (POSTCARDS AND ALL) TO A PERSONAL ACQUAINTANCE WHO IS MORE LIKELY THAN YOU TO KNOW THE TARGET PERSON. You may send the folder

Stanley Milgram's 1967 experiment

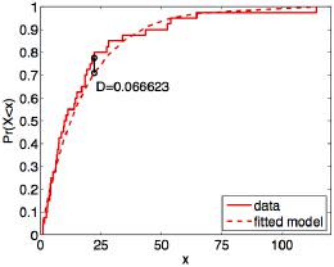
- Starting persons:
  - 296 volunteers, 217 sent
  - 196 in Nebraska
  - 100 in Boston
- Target person - Boston stockbroker
- Information given: target name, address, occupation, place of employment, college, hometown



J. Travers, S. Milgram, 1969

Stanley Milgram's 1967 experiment

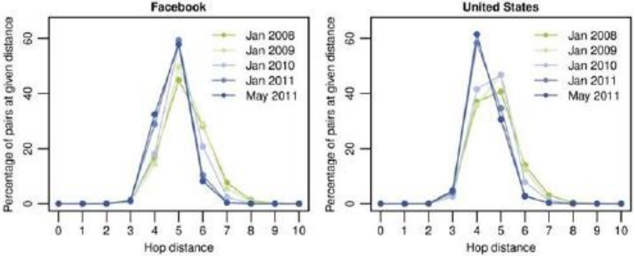
- Reached the target $N = 64(29\%)$
- Average chain length $\langle L \rangle = 5.2$
- Channels:
  - hometown $\langle L \rangle = 6.1$
  - business contacts $\langle L \rangle = 4.6$
  - from Boston $\langle L \rangle = 4.4$
  - from Nebraska $\langle L \rangle = 5.7$



J. Travers, S. Milgram, 1969

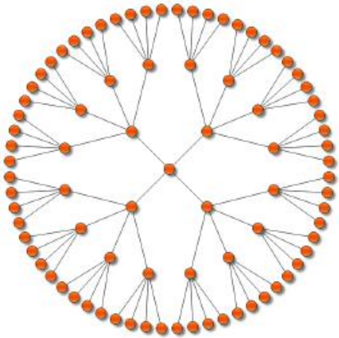
Small world

- Email graph:
  D. Watts (2001), 48,000 senders, $\langle L \rangle \approx 6$
- MSN Messenger graph:
  J. Leskovec et al (2007), 240mln users, $\langle L \rangle \approx 6.6$
- Facebook graph:
  L. Backstrom et al (2012), 721 mln users, $\langle L \rangle \approx 4.74$



figures from L.Backstrom, 2012

Simple model



An estimate: $z^d = N$, $d = \log N / \log z$
$N \approx 6.7$ bln, $z = 50$ friends, $d \approx 5.8$.

Network Data

- The Colorado Index of Complex Networks (ICON)
  http://icon.colorado.edu
- Stanford Large Network Dataset Collection
  http://snap.stanford.edu/data/index.html
- UCI Network Data Repository
  http://networkdata.ics.uci.edu

References

- Scale free networks. A.-L. Barabasi, E. Bonabeau, Scientific American 288, 50-59 (2003)
- Scale-Free Networks: A Decade and Beyond. A.-L. Barabasi, Science 325, 412-413 (2009)
- The Physics of Networks. Mark Newman, Physics Today, November 2008, pp. 33:38.

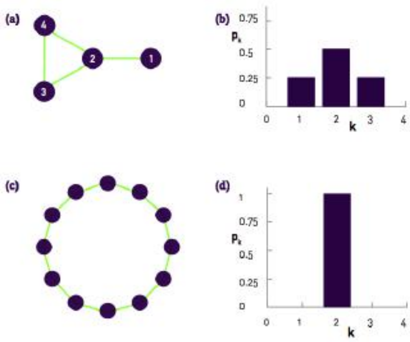
References

- The Small-World Problem. Stanley Milgram. Psychology Today, Vol 1, No 1, pp 61-67, 1967
- An Experimental Study of the Small World Problem. J. Travers and S. Milgram. . Sociometry, vol 32, No 4, pp 425-433, 1969
- Planetary-Scale Views on a Large Instant-Messaging Network. J. Leskovec and E. Horvitz. , Procs WWW 2008
- Four Degrees of Separation. L. Backstrom, P. Boldi, M. Rosa, J. Ugander, S. Vigna, WebSci '12 Procs. 4th ACM Web Science Conference, 2012 pp 33-42

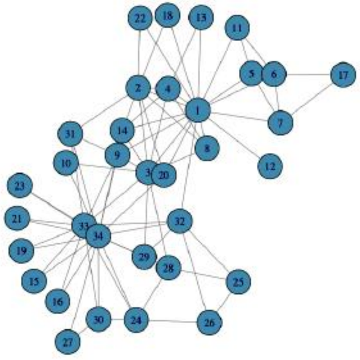
Node degree distribution

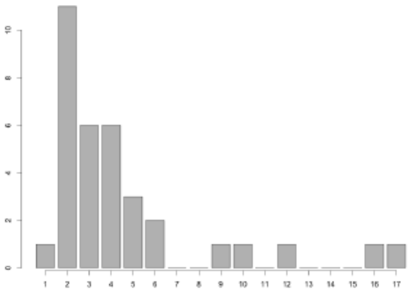
- k_i - node degree, i.e. number of nearest neighbors, $k_i = 1, 2, ...k_{\max}$
- n_k - number of nodes with degree k, $n_k = \sum_i \mathcal{I}(k_i == k)$
- total number of nodes $N = \sum_k n_k$
- Degree distribution is a fraction of the nodes with degree k

$$P(k_i = k) = P_k = \frac{n_k}{\sum_k n_k} = \frac{n_k}{N}$$



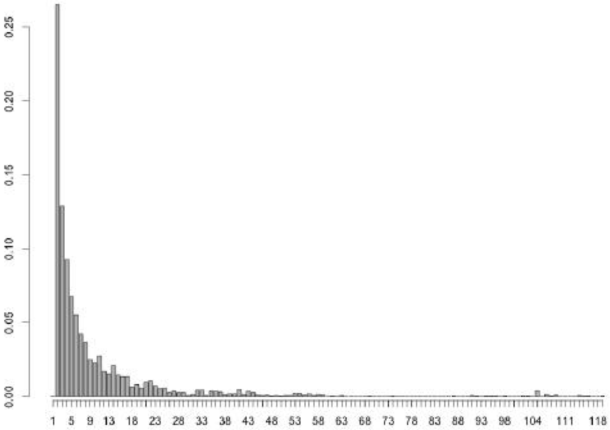
Node degree distribution



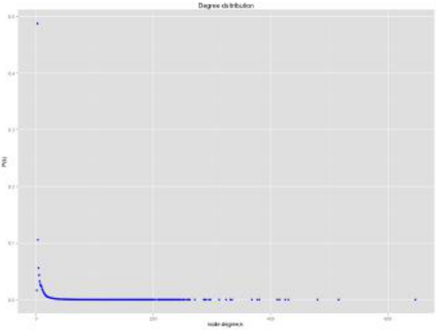


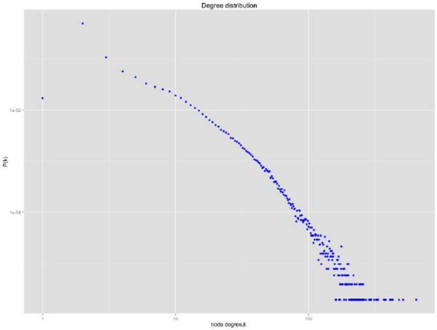
Degree distribution

Degree distribution

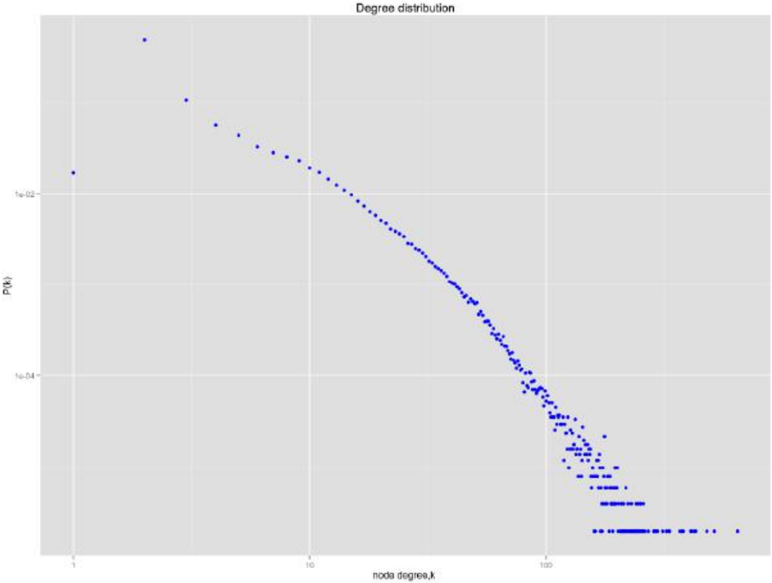


Power law degree distribution





Power law degree distribution



Discrete power law distribution

- Power law distribution, $k \in \mathbb{N}$, $\gamma \in \mathbb{R} > 0$

$$P_k = Ck^{-\gamma} = \frac{C}{k^{\gamma}}$$

- Log-log coordinates

$$\log P_k = -\gamma \log k + \log C$$

- Normalization

$$\sum_{k=1}^{\infty} P_k = C \sum_{k=1}^{\infty} k^{-\gamma} = C\zeta(\gamma) = 1; \quad C = \frac{1}{\zeta(\gamma)}$$

- Riemann zeta function, $\gamma > 1$

$$P_k = \frac{k^{-\gamma}}{\zeta(\gamma)}$$

Power law continuous approximation

- Power law, $k \in \mathbb{R}$, $\gamma \in \mathbb{R} > 0$

$$p(k) = Ck^{-\gamma} = \frac{C}{k^{\gamma}}, \quad \text{for } k \geq k_{min}$$

- Normalization ($\gamma > 1$)

$$1 = \int_{k_{min}}^{\infty} p(k)dk = C \int_{k_{\min}}^{\infty} \frac{dk}{k^{\gamma}} = \frac{C}{\gamma - 1} k_{\min}^{-\gamma+1}$$

$$C = (\gamma - 1) k_{\min}^{\gamma-1}$$

- Power law normalized PDF

$$p(k) = (\gamma - 1) k_{\min}^{\gamma-1} k^{-\gamma} = \frac{\gamma - 1}{k_{min}} \left(\frac{k}{k_{min}} \right)^{-\gamma}$$

Moments

- Power law PDF, $\gamma > 1$:

$$p(k) = \frac{C}{k^\gamma}, \;\; k \geq k_{min}; C = (\gamma - 1)k_{\text{min}}^{\gamma-1}$$

- First moment (mean value), $\gamma > 2$:

$$\langle k \rangle = \int_{k_{\text{min}}}^{\infty} kp(k)dk = C\int_{k_{\text{min}}}^{\infty} \frac{dk}{k^{\gamma-1}} = \frac{\gamma - 1}{\gamma - 2}k_{\text{min}}$$

- Second moment, $\gamma > 3$:

$$\langle k^2 \rangle = \int_{k_{\text{min}}}^{\infty} k^2 p(k)dk = C\int_{k_{\text{min}}}^{\infty} \frac{dk}{k^{\gamma-2}} = \frac{\gamma - 1}{\gamma - 3}k_{\text{min}}^2$$

- m-th moment, $\gamma > m + 1$:

$$\langle k^m \rangle = \int_{k_{\text{min}}}^{k_{\text{max}}} k^m p(k)dk = C\frac{k_{\text{max}}^{m+1-\gamma} - k_{\text{min}}^{m+1-\gamma}}{m + 1 - \gamma}$$

Hubs in networks

- How does the network size affect the size of its hubs(natural cutoff)?
- Probability of observing a single node with degree $k > k_{\max}$:

$$Pr(k \geq k_{\max}) = \int_{k_{\max}}^{\infty} p(k)dk$$

- Expected number of nodes with degree $k \geq k_{\max}$:

$$N \cdot Pr(k \geq k_{\max}) = 1$$

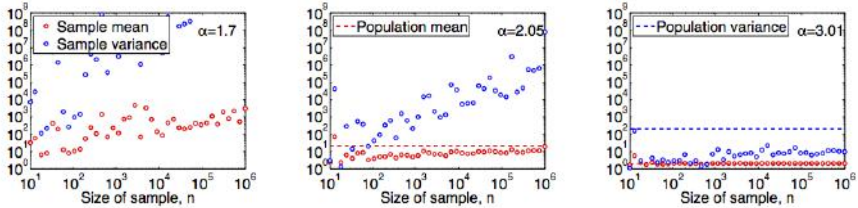
- Expected largest node degree in exponential network $p(k) = Ce^{-\lambda k}$

$$k_{max} = k_{min} + \frac{\ln N}{\lambda}$$

- Expected largest node degree in power law network $p(k) = Ck^{-\gamma}$

$$k_{max} = k_{min} \; N^{\frac{1}{\gamma - 1}}$$

Moments



$$\langle k \rangle = C \frac{k_{\max}^{2-\gamma} - k_{\min}^{2-\gamma}}{2-\gamma}, \qquad \langle k^2 \rangle = C \frac{k_{\max}^{3-\gamma} - k_{\min}^{3-\gamma}}{3-\gamma}$$

$$\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$$

Clauset et.al, 2009

Scale free network

Degree of a randomly chosen node:

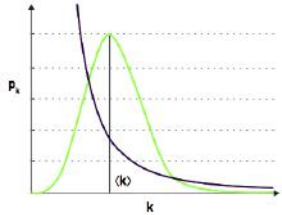
$$k = \langle k \rangle \pm \sigma_k, \quad \sigma_k^2 = \langle k^2 \rangle - \langle k \rangle^2$$

Poisson degree distribution (random network) has a scale $\langle k \rangle$:

$$k = \langle k \rangle \pm \sqrt{\langle k \rangle}$$

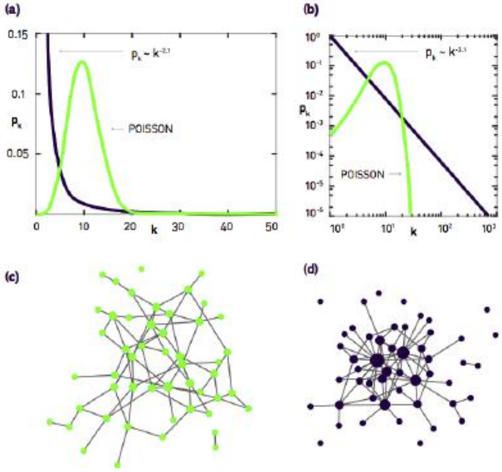
Power law network with $2 < \gamma < 3$ is scale free:

$$k = \langle k \rangle \pm \infty$$



from A.-L., Barabasi, 2016

Hubs in networks

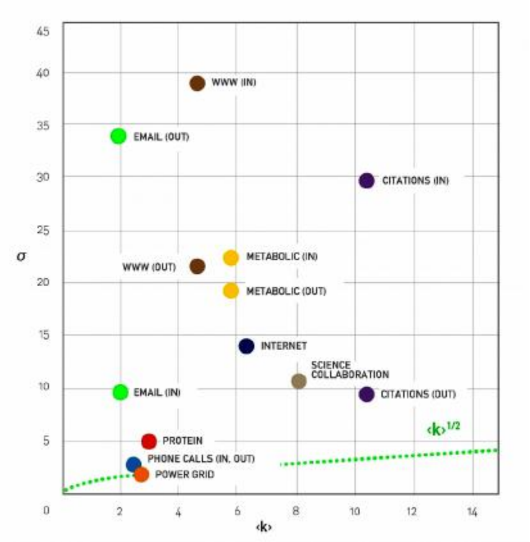


from A.-L., Barabasi, 2016

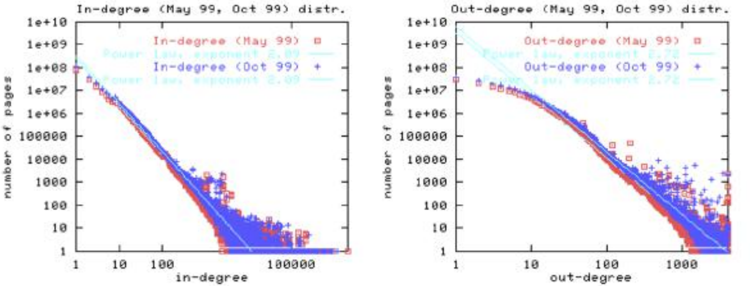
Degree fluctuation in real networks

Network	N	L	$\langle k \rangle$	$\langle k_{in}^2 \rangle$	$\langle k_{out}^2 \rangle$	$\langle k^2 \rangle$	γ_{in}	γ_{out}	γ
Internet	192,244	609,066	6.34	-	-	240.1	-	-	3.42*
WWW	325,729	1,497,134	4.60	1546.0	482.4	-	2.00	2.31	-
Power Grid	4,941	6,594	2.67	-	-	10.3	-	-	Exp.
Mobile-Phone Calls	36,595	91,826	2.51	12.0	11.7	-	4.69*	5.01*	-
Email	57,194	103,731	1.81	94.7	1163.9	-	3.43*	2.03*	-
Science Collaboration	23,133	93,437	8.08	-	-	178.2	-	-	3.35*
Actor Network	702,388	29,397,908	83.71	-	-	47,353.7	-	-	2.12*
Citation Network	449,673	4,689,479	10.43	971.5	198.8	-	3.03*	4.00*	-
E. Coli Metabolism	1,039	5,802	5.58	535.7	396.7	-	2.43*	2.90*	-
Protein Interactions	2,018	2,930	2.90	-	-	32.3	-	-	2.89*-

Degree fluctuation in real networks



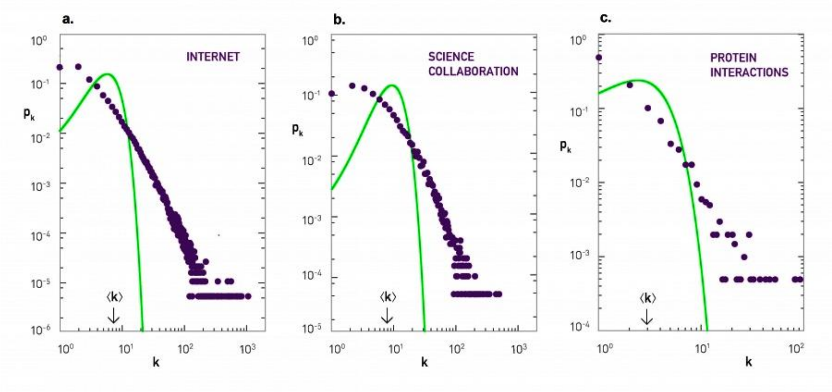
Scale-free networks



In- and out- degrees of WWW crawl 1999

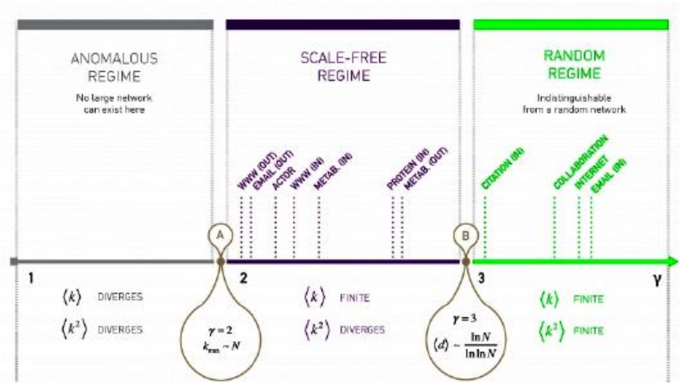
Broder et.al, 1999

Scale-free networks



from A.-L. Barabasi, 2016

Properties of scale free networks



from A.-L., Barabasi, 2016

Plotting power Laws

- Power law PDF

$$p(k) = Ck^{-\gamma}; \quad \log p(k) = \log C - \gamma \log k$$

- Cumulative distribution function (CDF)

$$F(k) = Pr(k_i \leq k) = \int_0^k p(k)dk$$

- Complimentary cumulative distribution function cCDF

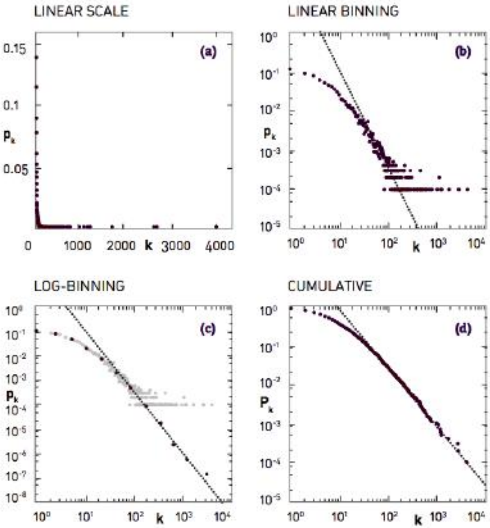
$$\bar{F}(k) = Pr(k_i > k) = 1 - F(k) = \int_k^\infty p(k)dk$$

- Power law cCDF

$$\bar{F}(k) = \frac{C}{\gamma - 1} k^{-(\gamma-1)}$$

$$\log \bar{F}(k) = \log \frac{C}{\gamma - 1} - (\gamma - 1) \log k$$

Plotting power laws



from A.-L., Barabasi, 2016

Parameter estimation: γ

Maximum likelihood estimation of parameter γ

- Let $\{x_i\}$ be a set of n observations (points) independently sampled from the distribution

$$P(x_i) = \frac{\gamma - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\gamma}$$

- Probability of the sample sequence

$$P(\{x_i\}|\gamma) = \prod_i^n \frac{\gamma - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\gamma}$$

Maximum likelihood

- log-likelihood

$$\mathcal{L} = \ln P(\gamma|\{x_i\}) = n\ln(\gamma - 1) - n\ln x_{\min} - \gamma \sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}$$

- maximization $\frac{\partial \mathcal{L}}{\partial \gamma} = 0$

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

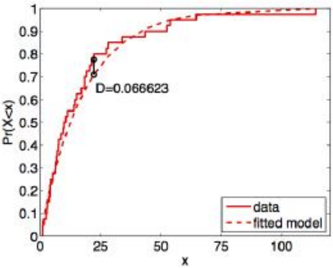
- error estimate

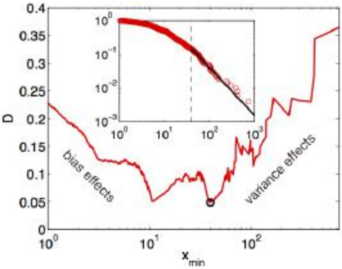
$$\sigma = \sqrt{n}\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}\right]^{-1} = \frac{\gamma - 1}{\sqrt{n}}$$

Parameter estimation: k_{min}

- Kolmogorov-Smirnov test (compare model and experimental CDF)

$$D = \max_x |F(x|\gamma, x_{min}) - F_{exp}(x)|$$





- find

$$x^*_{min} = argmin_{x_{min}} D$$

Clauset et.al, 2009

References

- Power laws, Pareto distributions and Zipf's law, M. E. J. Newman, Contemporary Physics, pages 323–351, 2005.
- Power-Law Distribution in Empirical Data, A. Clauset, C.R. Shalizi, M.E.J. Newman, SIAM Review, Vol 51, No 4, pp. 661-703, 2009.
- A Brief History of Generative Models for Power Law and Lognormal Distributions, M. Mitzenmacher, Internet Mathematics Vol 1, No 2, pp 226-251.