gesis Leibniz Institute for the Social Sciences

39. Methodenseminar: Big Data Module II

Introduction to Social Network Science with Python

Macro-Scale Analysis

Fariba Karimi, Haiko Lietz, & Marcos Oliveira

July 17, 2019

Leibniz Association

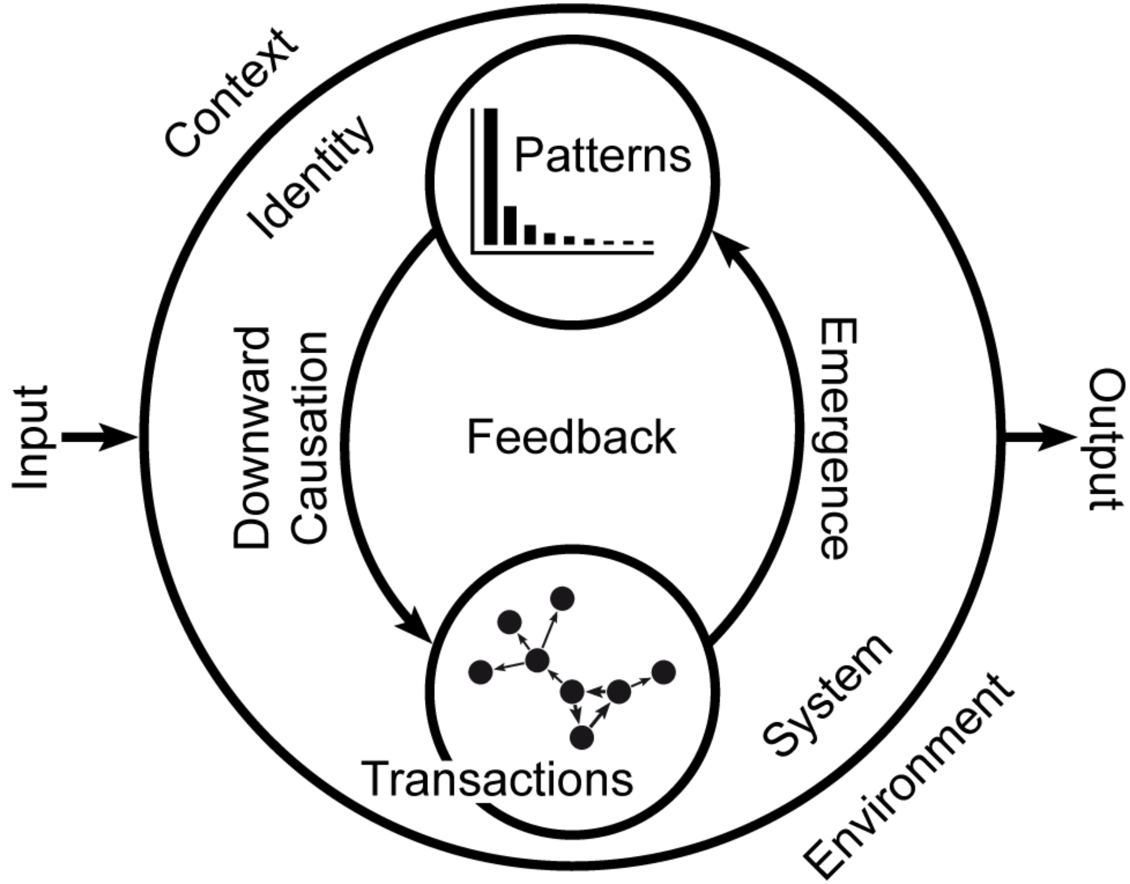
Leibniz Institute
for the Social Sciences

Macro-Scale Analysis

Complex System

A system that consists of many **parts**, the **interactions** of which give rise to a **whole** that is **more than the sum of the parts**

Why Macro Matters: Collective Control

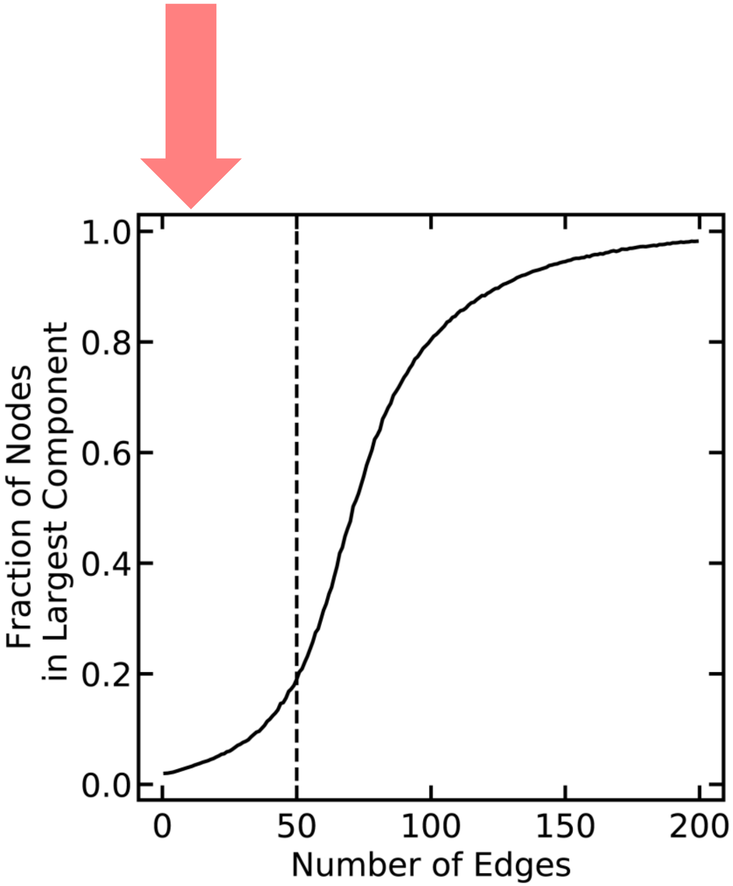


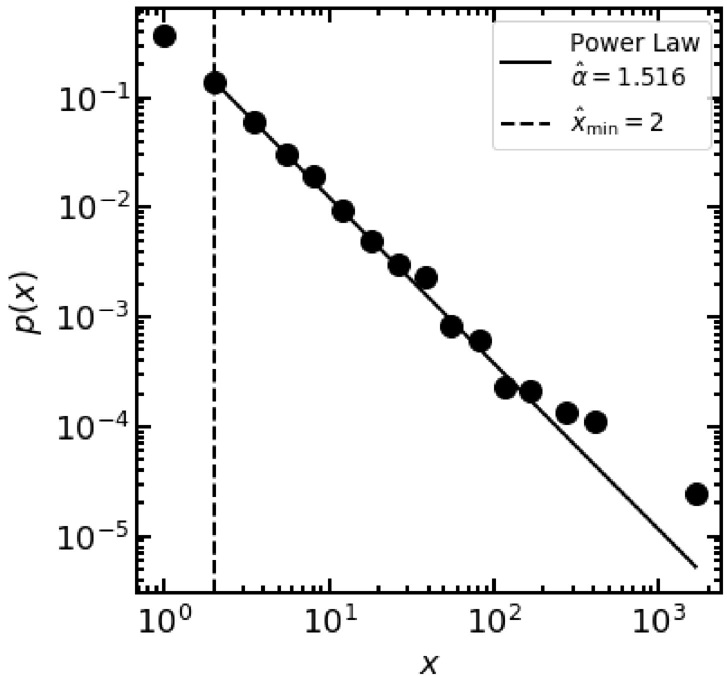
[1] Mohr (1998). *Annual Review of Sociology* 24:345–370.
[2] Fuhse (2009). *Sociological Theory* 27:51–73.
[3] Page (2015). *Annual Review of Sociology* 41:21–41.

Macro-Scale Analysis

Small-World Networks

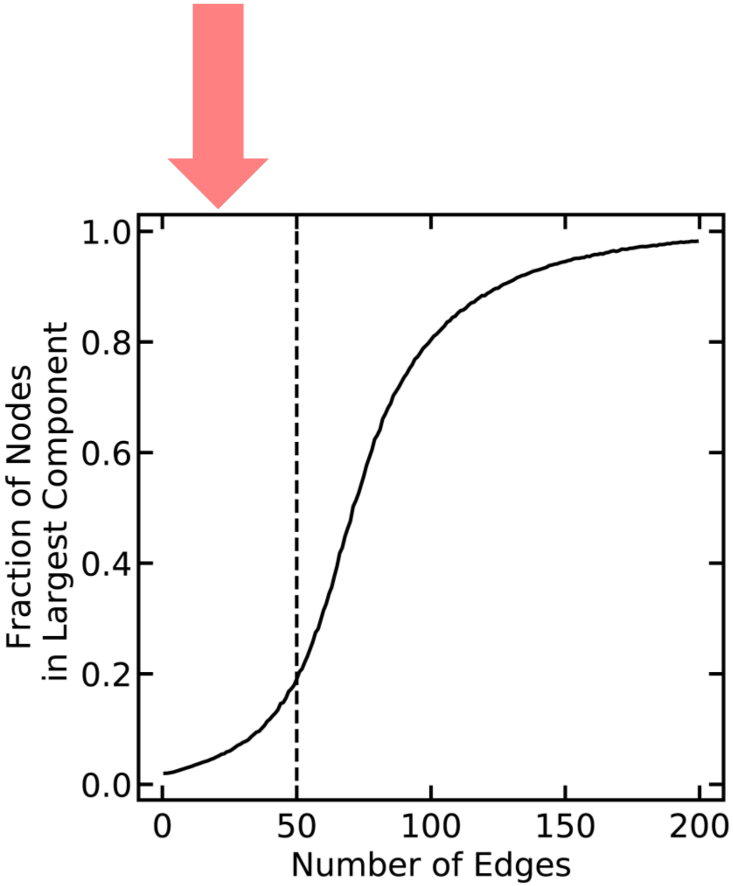
Percolation In Erdős-Rényi Graph

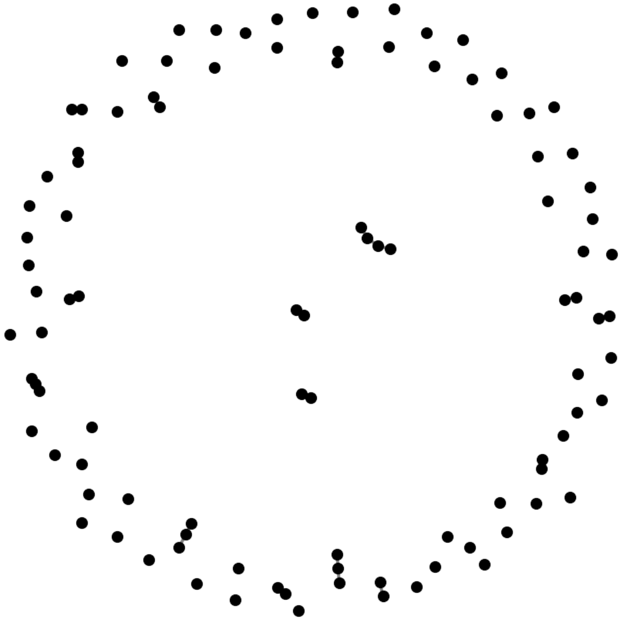




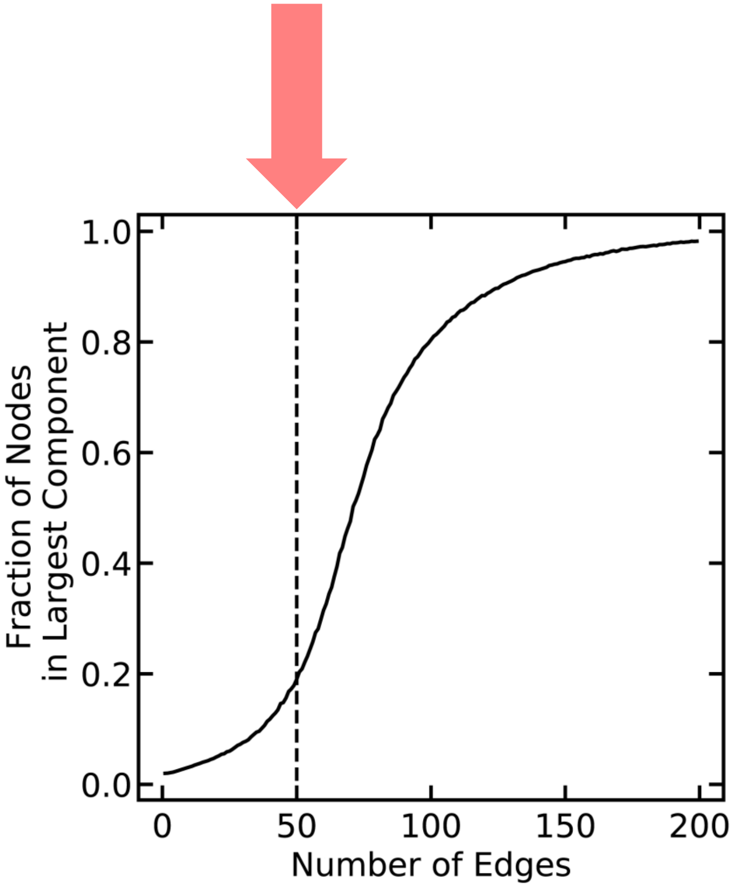
gesis Leibniz Institute for the Social Sciences

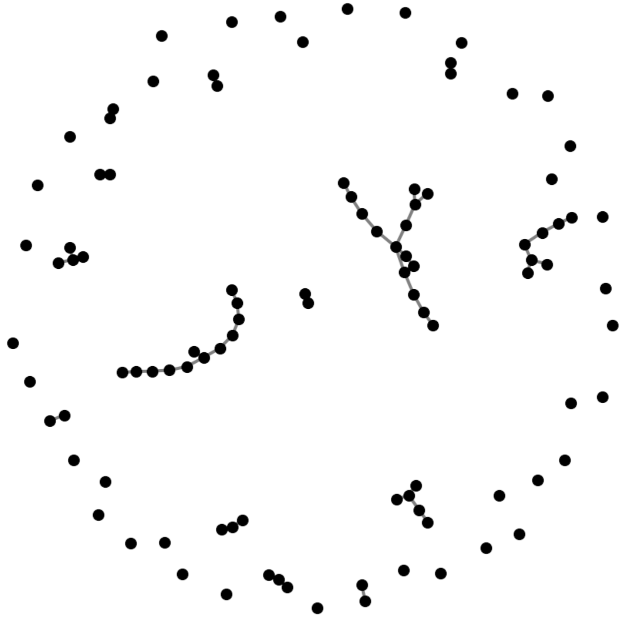
Percolation In Erdős-Rényi Graph



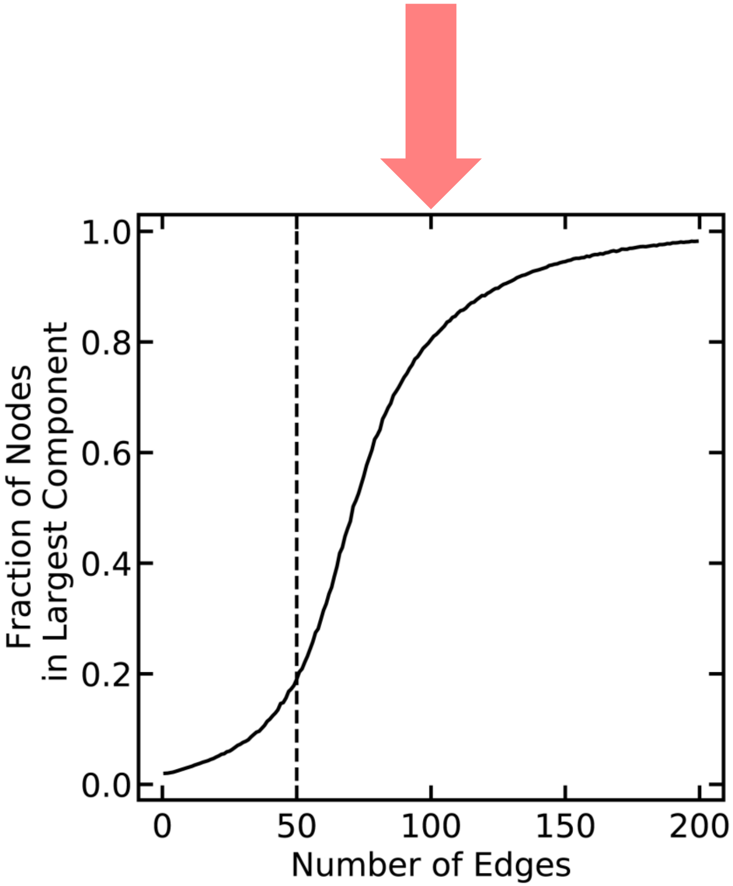


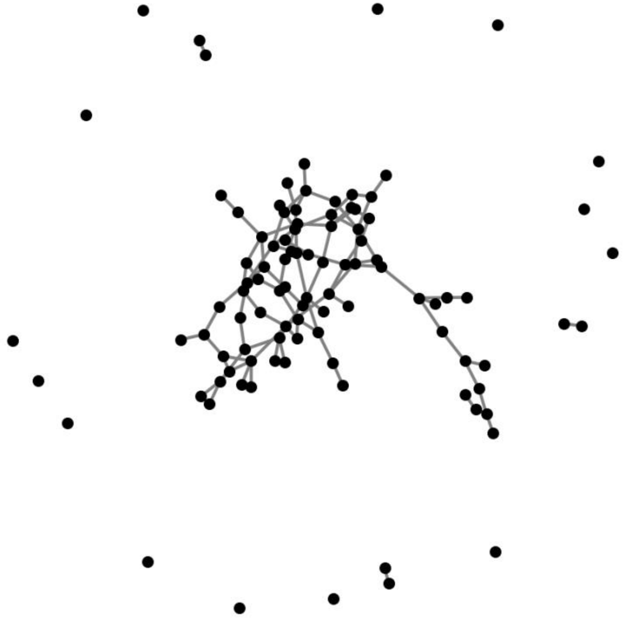
Percolation In Erdős-Rényi Graph



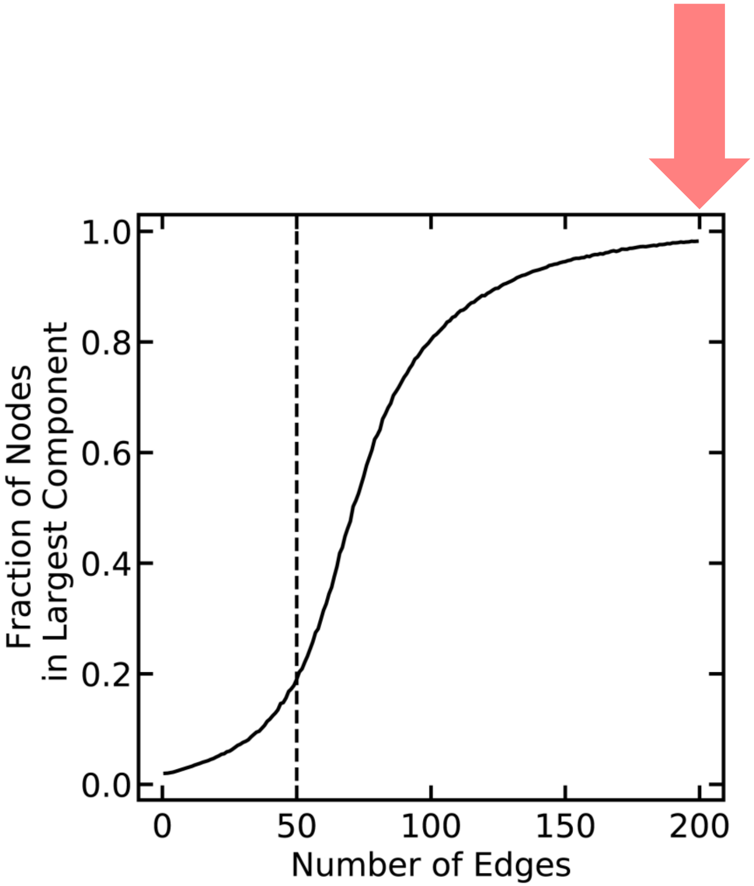


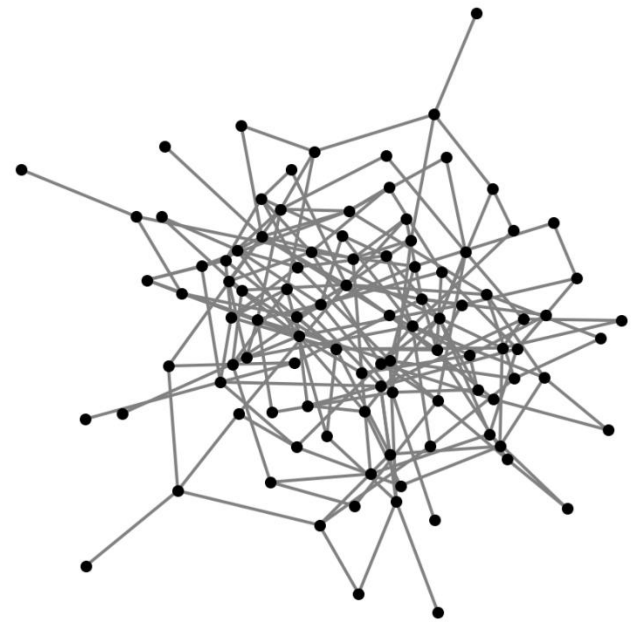
Percolation In Erdős-Rényi Graph



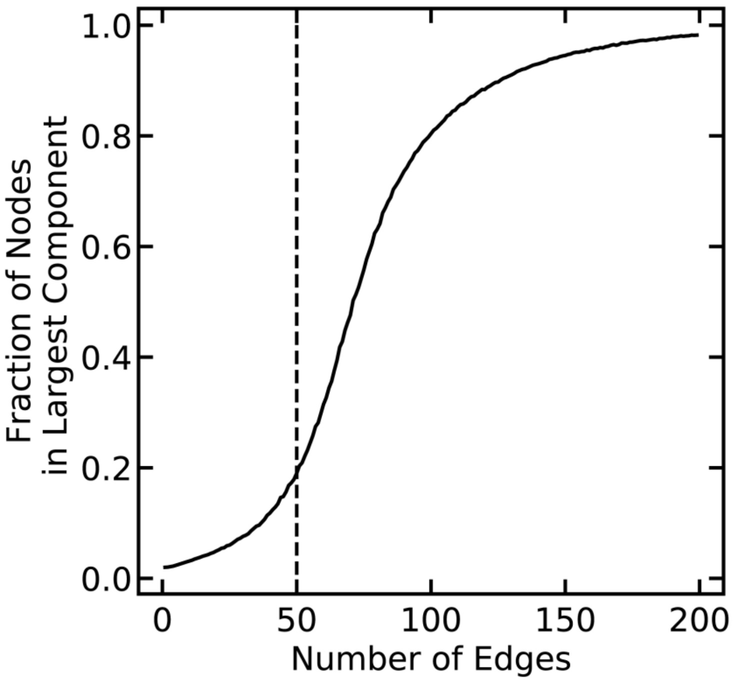


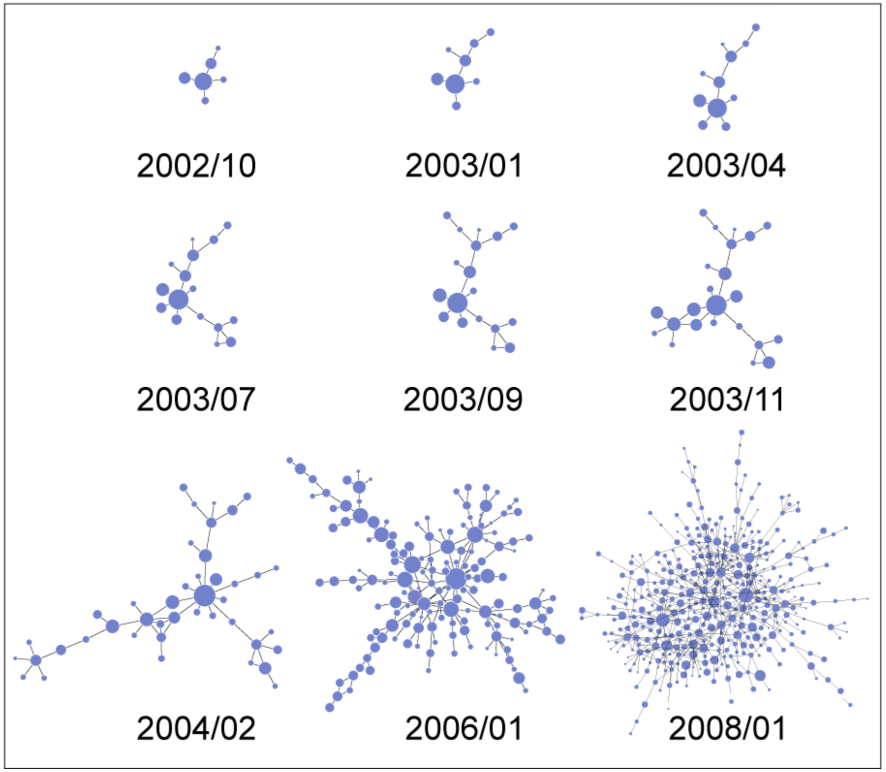
Percolation In Erdős-Rényi Graph





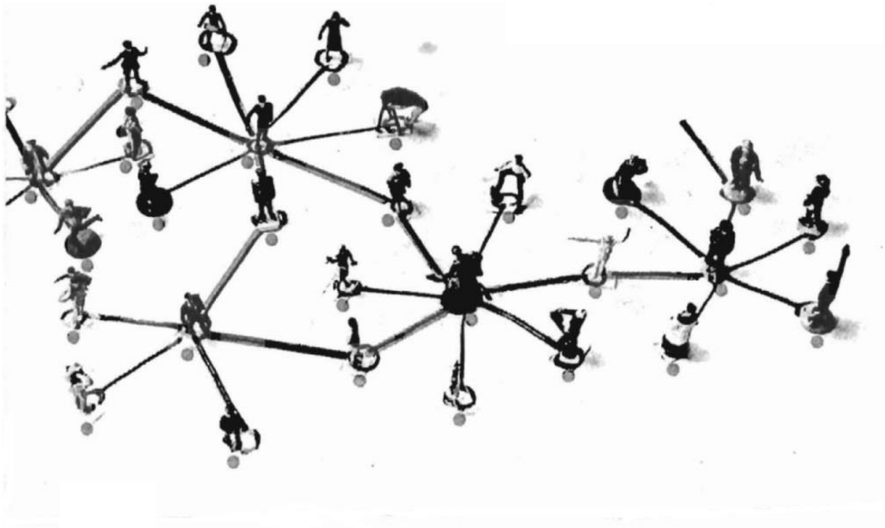
Evidence for Social Percolation





[1] Lee, D.K. *et al.* 2010. *Physical Review E* 82:026112.

The Small-World Phenomenon

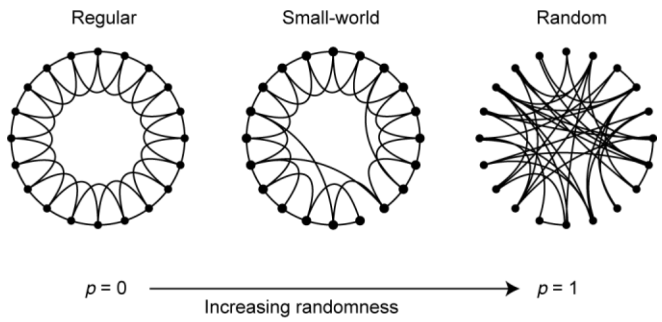


- **Small-World Experiment of 1967**: Randomly selected persons were asked to contact an unknown but roughly descried target person by writing a postcard to a broker, who acts accordingly, and so on…
- **Result**: 3 of 60 chain latters reached the target person through five brokers, on average (**six degrees of separatrion**)

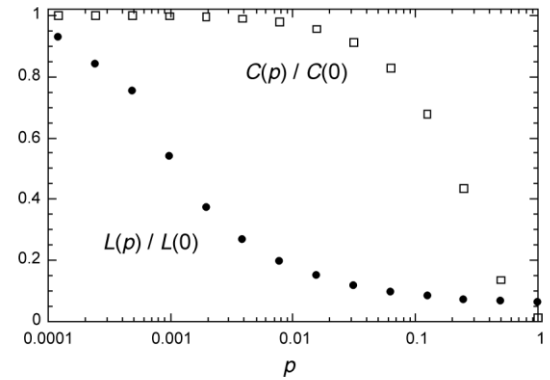
[1] Milgram, S. 1967. *Psychology Today* 2:60–67.
[2] Travers, J. & Milgram, S. 1969. *Sociometry* 32, 425–443.

Small-World Networks: Topological

- **Small-world networks** [1]: Networks with two realistic properties (high average clustering coefficient and short average path length) from rewiring lattice with probability p



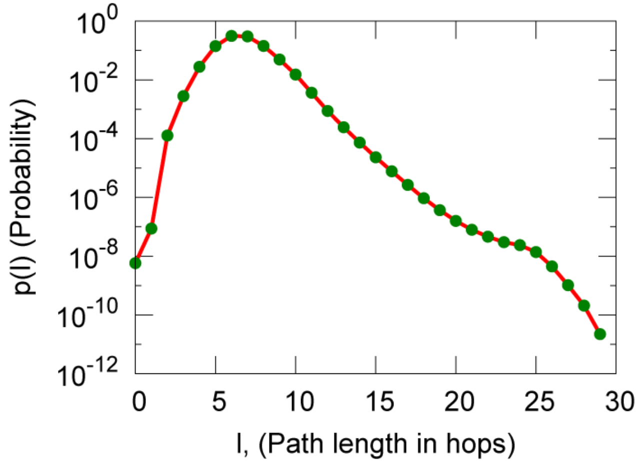
- **Small-world-ness** of a graph [2]
  - $Q = \frac{CC_{\text{norm}}}{L_{\text{norm}}}$, with
    - $CC_{\text{norm}} = \frac{CC_{\text{actual}}}{CC_{\text{random}}}$ (CC: average clustering coefficient)
    - $L_{\text{norm}} = \frac{L_{\text{actual}}}{L_{\text{random}}}$ (L: average shortest path length)

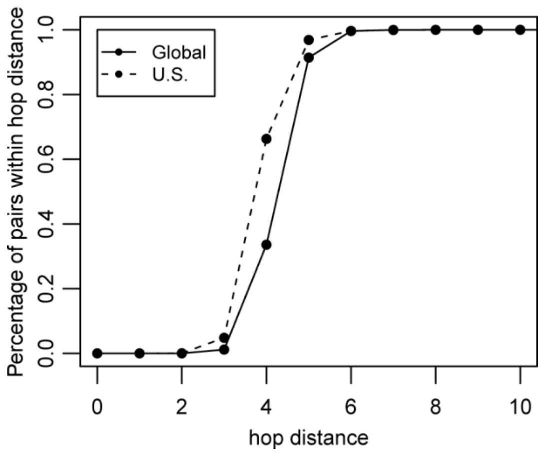


[1] Watts, D.J. & Strogatz, S.H. 1998. *Nature* 393:440–442.
[2] Humphries, M.D. & Gurney, K. 2008. *PLoS ONE* 3:e0002051.

The Small-World Phenomenon Confirmed

- **Microsoft Messenger** [1]
  - Used friendship network of 180M nodes, 1.342B edges
  - 99.9% in largest connected component
  - Randomly sampled 1000 nodes
  - $L = 6.6$ (seven degrees of separation)
- **Facebook** [2]
  - Friendship network of 721M nodes and 68.7B edges
  - 99.9% in largest connected component
  - 92% within six degrees, 99.6% within six degrees
  - $L = 4.7$ (five degrees of separation)





[1] Leskovec, J. & Horvitz, E. 2008. *Proc. WWW 2008*. 915–924.
[2] Ugander, J. *et al.* 2011. arXiv:1111.4503v1

The World Is Small

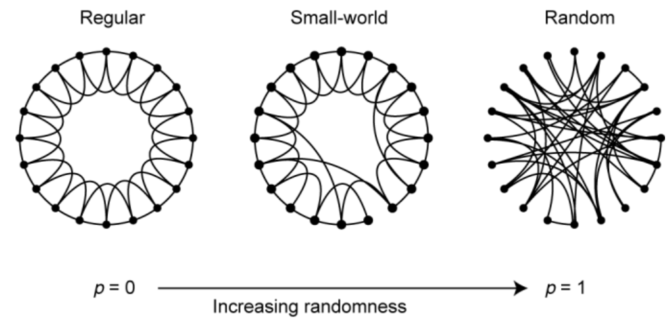
Authors	Network	Period	N	k	L Actual	L Random	CC Actual	CC Random	Lr	CCr	Q
Organizations											
Kogut and Walker (2001)	German firms	1993–1997	291	2.02	5.64	3.01	0.84	0.022	1.87	38.18	20.38
Baum *et al.* (2003)	Canadian I-banks	1952–1957	53	1.36	3.21	4.556	0.023	0.027	0.70	0.85	1.21
		1969–1974	41	2.22	2.82	3.176	0.283	0.054	0.89	5.24	5.90
		1985–1990	142	3.83	2.95	3.144	0.273	0.027	0.94	10.11	10.78
Davis *et al.* (2003)	US Co. interlocks	1982	195	6.8	3.15	2.7	0.24	0.039	1.17	6.15	5.27
		1999	195	7.2	2.98	2.64	0.2	0.039	1.13	5.13	4.54
Verspagen and Duyster (2004)	Strategic alliances*	1980–1996	5504	5.29	4.2	5.25	0.34	0.0008	0.80	425.00	531.25
Schilling and Phelps, (forthcoming)	US alliances in 11 2-digit SIC codes**	1992–2000	171 (157)	3.11 (1.42)	20.39 (18.69)	5.62 (3.01)	0.26 (0.18)	0.04 (0.039)	3.85 (2.84)	10.44 (7.53)	2.71 (2.65)
Persons											
Davis *et al.* (2003)	US Director interlocks	1982	2366	19.1	4.03	2.61	0.91	0.009	1.54	101.11	65.48
		1990	2078	17.4	3.98	2.65	0.89	0.009	1.50	98.89	65.84
		1999	1916	16.3	3.86	2.69	0.88	0.009	1.43	97.78	68.14
Fleming *et al.* (forthcoming)	US patenting inventors***	1986–1990	7069	4.73	2.73	1.14	0.736	0.0452	2.394737	16.28	6.80
Kogut and Walker (2001)	German Co. ownership	1993–1997	429	3.56	6.09	5.16	0.83	0.008	1.18	103.75	87.91
Newman (2004)	Biology co-authorship	1995–1999	1,520,251	18.1	4.6		0.066				
	Physics co-authorship	1995–1999	52,909	9.7	5.9		0.43				
	Mathematics co-authorship	1940–2006	253,339	3.9	7.6		0.15				
Moody, 2004	Sociologists co-authorship	1963–1999	128,151		9.81	7.57	0.194	0.207	1.30	0.94	0.72
		1989–1999	87,731		11.53	8.24	0.266	0.302	1.40	0.88	0.63
Goyal *et al.*	Economists co-authorship	1980–1989	48,608	1.244			0.182				
		1990–1999	81,217	1.672			0.157				
Watts (1999)	Hollywood Film actors	1898–1997	226,000	61	3.65	2.99	0.79	0.00027	1.22	2925.93	2396.85
Smith (2006)	U.S. Rappers		5533		3.9		0.18				
	U.S. Jazz musicians		1275		2.79		0.33				
	Brazilian pop		5834		2.3		0.84				
Technology											
Watts (1999)	Power grids		4941	2.94	18.7	12.4	0.08	0.005	1.51	16.00	10.61
Vazquez *et al.* (2002)	Internet	1997	3112	3.5	3.8		0.18				
		1998	3834	3.6	3.8		0.21				
		1999	5287	3.8	3.7		0.24				

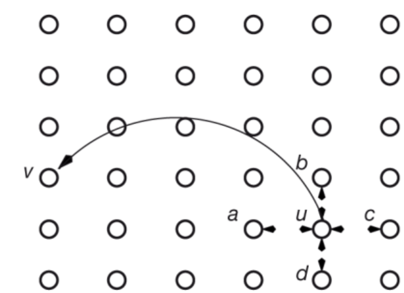
* Chemicals and Electronics Industries, ** average across industries for analysis of separate industries, see Schilling and Phelps, forthcoming.
*** Path length for giant component, **** average for biology, physics, and mathematics.
Empty cells appear when small world statistics were not included in the original article.

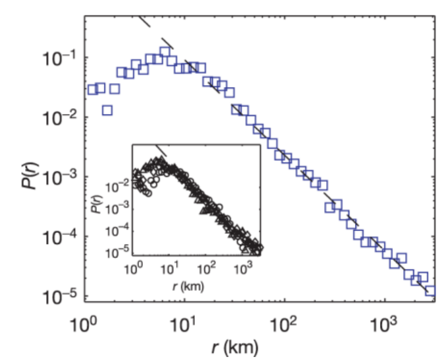
[1] Uzzi, B. *et al.* 2007. *European Management Review* 4:77–91.

Small-World Networks: Navigational

- How can the Small World be searched?
- By mostly searching locally but occasionally searching globally!
- Search is most efficient if **search distance decays as a power law** with the exponent β equal to the dimension of the grid [1]
- Such behavior is known from **human travel** ($\beta \approx 1.6$) [2]







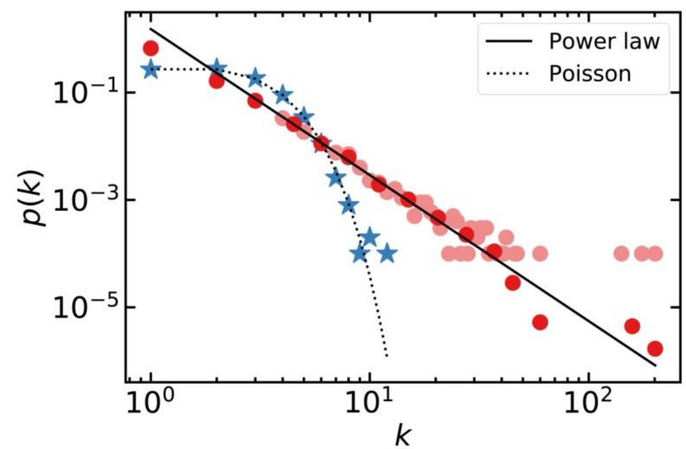
[1] Kleinberg, J.M. 2000. *Nature* 406:845.
[2] Brockmann, D. *et al.* 2006. *Nature* 439:462–465.

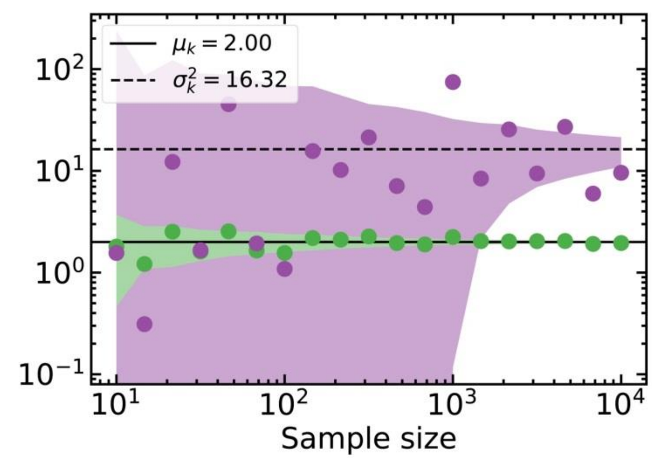
Macro-Scale Analysis

Scale-Free Networks

Scale-Free Networks

- **Scale-free networks** [1]: Generative model for networks without a characteristic scale
- **Scale-free-ness**
  - Characterized by **power law distribution** of degree k, $p(k) \propto k^{-\alpha}$, exponent α being the only shape parameter
  - **Scale invariance**: Proportionate scaling of k does not change the shape of the distribution
  - Typically one is interested in $\alpha \leq 3$
  - **Inifinite variance** ($\alpha \leq 3$): Sampling is not reliable as Central Limit Theorem fails

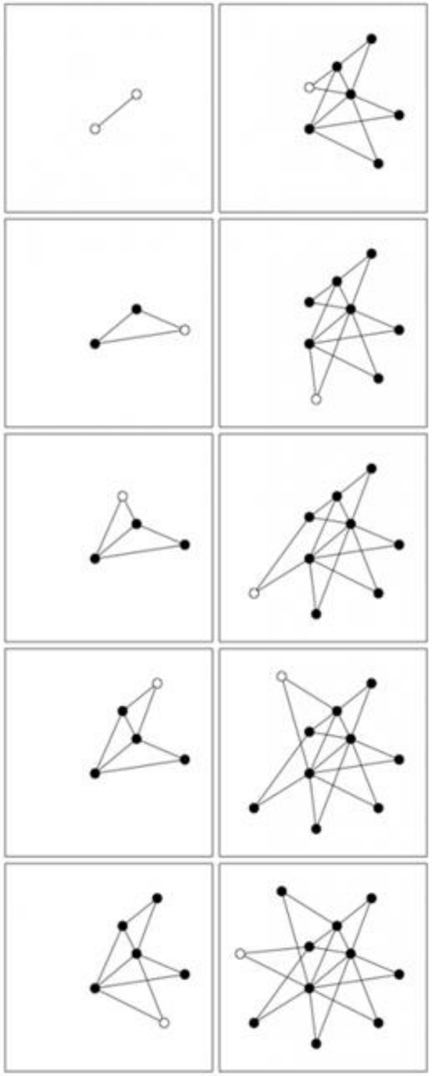




[1] Barabási, A.-L. & R. Albert. 1999. *Science* 286:510–512.

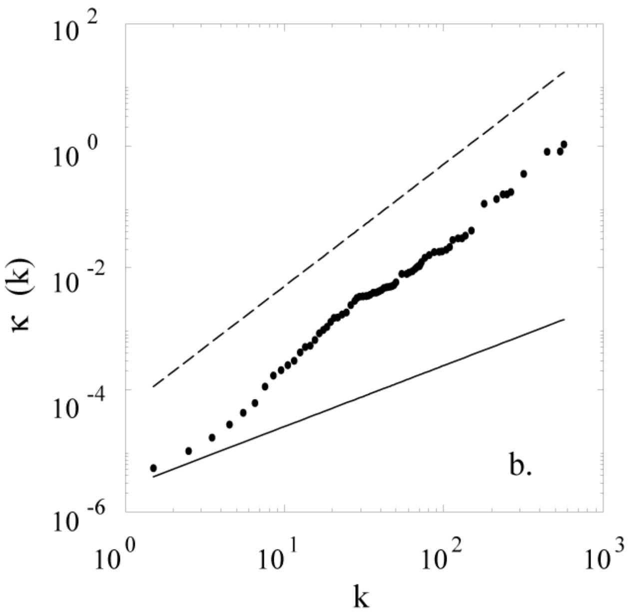
Scale-Free Networks

- **Scale-free networks** [1]: Generative model for networks without a characteristic scale
- **Generative model**
  - Combination of **growth** and **preferential attachment**: Probability of new vertex to attach to existing vertex i is $\prod_i = \frac{k_i}{\sum_j k_j}$, j being all vertices



[1] Barabási, A.-L. & R. Albert. 1999. *Science* 286:510–512.

- **Scale-free networks** [1]: Generative model for networks without a characteristic scale
- **Generative model**
  - Combination of **growth** and **preferential attachment**: Probability of new vertex to attach to existing vertex i is $\prod_i = \frac{k_i}{\sum_j k_j}$, j being all vertices
- **Measuring preferential attachment** [2] as the returns of selections at t_1 to those at t_0



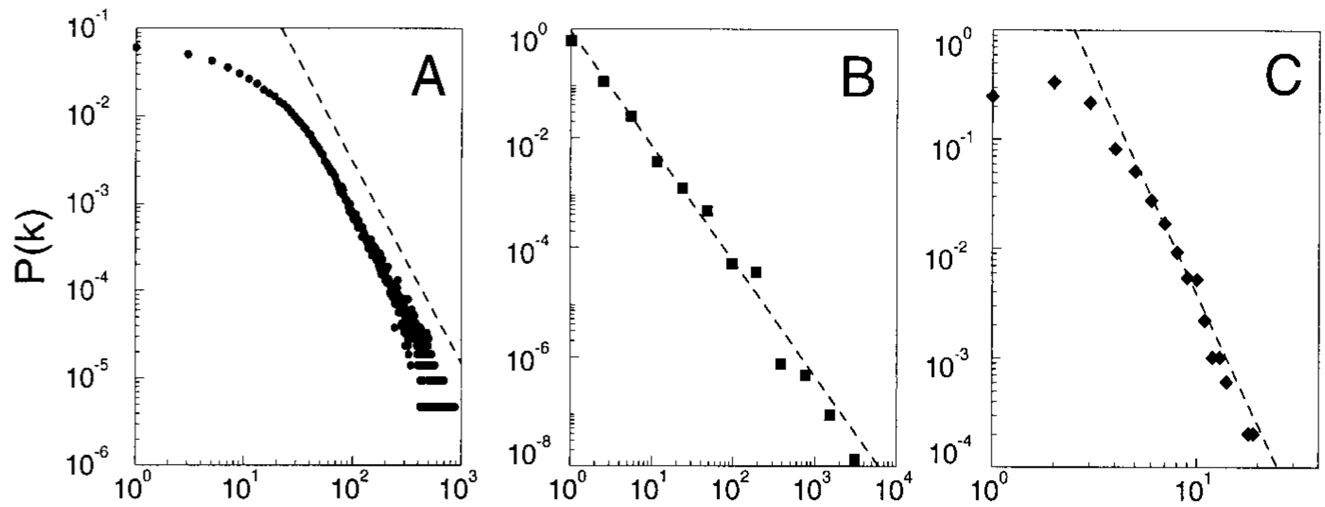
[1] Barabási, A.-L. & R. Albert. 1999. *Science* 286:510–512.
[2] Jeong, H. *et al.* 2003. *Europhysics Letters* 61: 567–572.

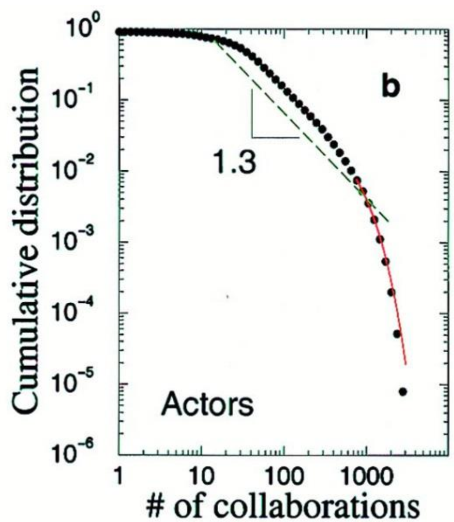
Distributions Point At Mechanisms

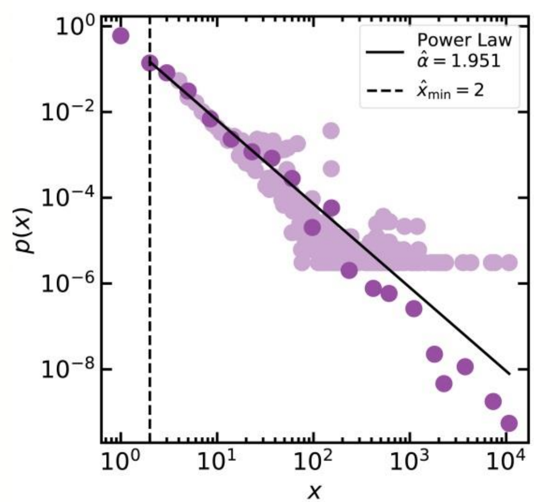
	Function	Mechanism
Exponential	$p(k) \propto \mathrm{e}^{-\lambda k}$	Random process
Lognormal	$p(k) \propto \frac{1}{k}\mathrm{e}^{-\frac{(\log(k)-\mu)^2}{2\sigma^2}}$	Multiplicative growth [1]
Stretched Exponential	$p(k) \propto (\lambda k)^{\beta-1}\mathrm{e}^{-(\lambda k)^{\beta}}$	Sublinear preferential growth [2]
Power Law	$p(k) \propto \mathrm{k}^{-\alpha}$	Linear preferential growth [3]
Truncated Power Law	$p(k) \propto \mathrm{k}^{-\alpha}\mathrm{e}^{-\lambda k}$	Linear preferential growth with cutoff effects [4]

[1] Mitzenmacher, M. 2011. *Internet Mathematics* 1:226–251.
[2] Krapivsky, P. L. *et al.* 2000. *Physical Review Letters* 85:4629–4632.
[3] Barabási, A.-L. & R. Albert. 1999. *Science* 286:510–512.
[4] Amaral, L.A.N. *et al.* 2000. *Proc. Nat. Acad. Sci. USA* 97:11149–11152.

Fitting Power Laws Is Tricky







[1] Barabási, A.-L. & R. Albert. 1999. *Science* 286:510–512.
[2] Amaral, L.A.N. *et al.* 2000. *Proc. Nat. Acad. Sci. USA* 97:11149–11152.

Maximum Likelihood Method

- Estimation the **parameters** α (scaling exponent) and $x_{\min}$ (lower cutoff)
  - Minimizing the maximum Kolmogorov-Smirnov distance D between the observed data and the model
- Computing the **plausibility** p of a power-law fit
  - Synthesizing n power laws with $\hat{\alpha}$ and $\hat{x}_{\min}$ (estimated from the data) and counting the fraction in which D between the observed data and the model is smaller than D between the synthetic data and the model
  - Power-law hypothesis is confirmed if $p \geq 0.1$
- Identifying the **best fitting function**
  - Comparing the power law to other functions and checking the sign of the log-likelihood ratio R
  - Sign is meaningful if $p_R < 0.1$

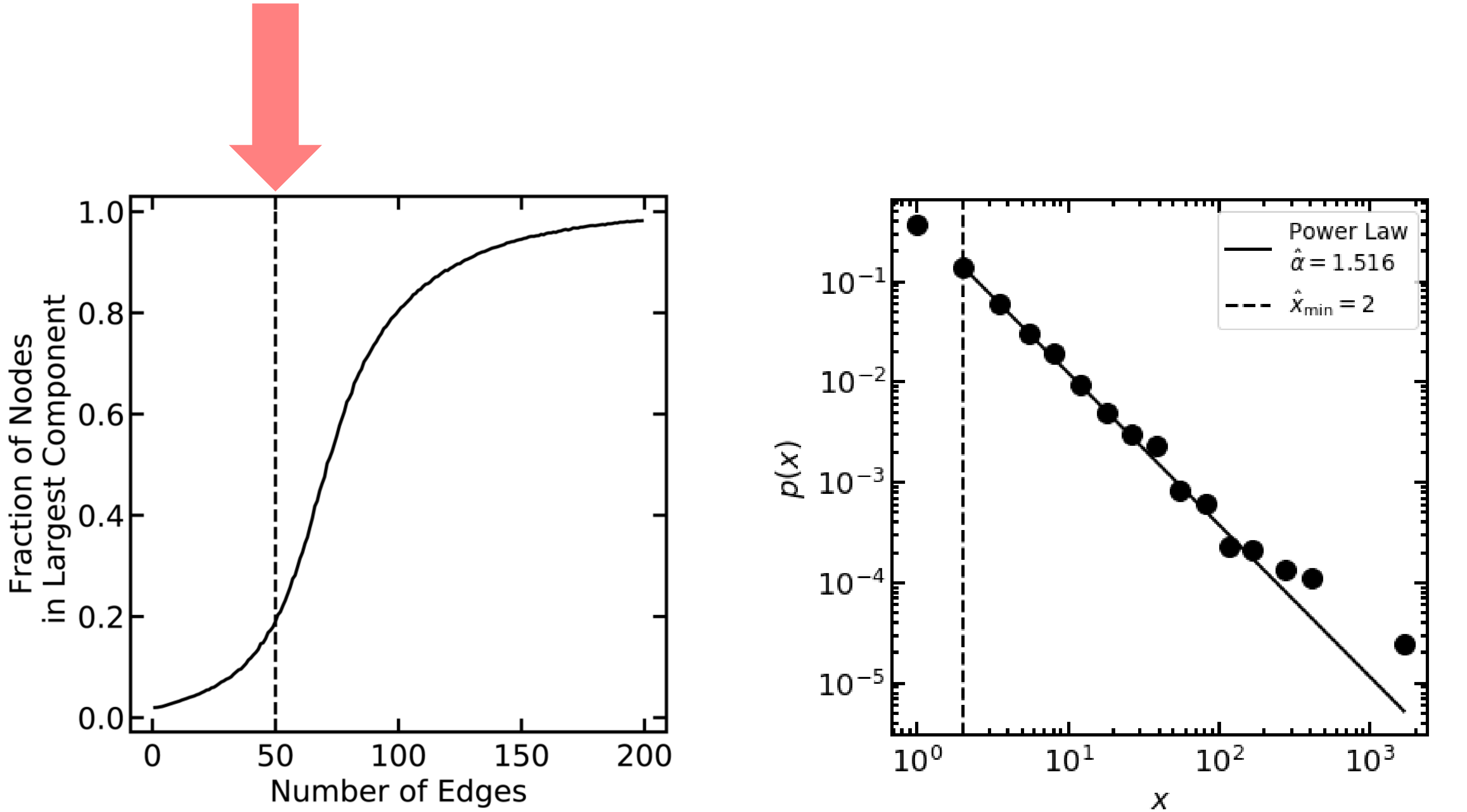
[1] Clauset, A. *et al.* 2009. *SIAM Review* 51:661–703.

Many Systems Are Scale-Free

Quantity	n	$\langle x \rangle$	σ	$x_{\max}$	$\hat{x}_{\min}$	$\hat{\alpha}$	n_{tail}	p
count of word use	18 855	11.14	148.33	14 086	7 ± 2	1.95(2)	2958 ± 987	**0.49**
protein interaction degree	1846	2.34	3.05	56	5 ± 2	3.1(3)	204 ± 263	**0.31**
metabolic degree	1641	5.68	17.81	468	4 ± 1	2.8(1)	748 ± 136	0.00
Internet degree	22 688	5.63	37.83	2583	21 ± 9	2.12(9)	770 ± 1124	**0.29**
telephone calls received	51 360 423	3.88	179.09	375 746	120 ± 49	2.09(1)	$102\,592 \pm 210\,147$	**0.63**
intensity of wars	115	15.70	49.97	382	2.1 ± 3.5	1.7(2)	70 ± 14	**0.20**
terrorist attack severity	9101	4.35	31.58	2749	12 ± 4	2.4(2)	547 ± 1663	**0.68**
HTTP size (kilobytes)	226 386	7.36	57.94	10 971	36.25 ± 22.74	2.48(5)	6794 ± 2232	0.00
species per genus	509	5.59	6.94	56	4 ± 2	2.4(2)	233 ± 138	**0.10**
bird species sightings	591	3384.36	10 952.34	138 705	6679 ± 2463	2.1(2)	66 ± 41	**0.55**
blackouts ($\times 10^3$)	211	253.87	610.31	7500	230 ± 90	2.3(3)	59 ± 35	**0.62**
sales of books ($\times 10^3$)	633	1986.67	1396.60	19 077	2400 ± 430	3.7(3)	139 ± 115	**0.66**
population of cities ($\times 10^3$)	19 447	9.00	77.83	8 009	52.46 ± 11.88	2.37(8)	580 ± 177	**0.76**
email address books size	4581	12.45	21.49	333	57 ± 21	3.5(6)	196 ± 449	**0.16**
forest fire size (acres)	203 785	0.90	20.99	4121	6324 ± 3487	2.2(3)	521 ± 6801	0.05
solar flare intensity	12 773	689.41	6520.59	231 300	323 ± 89	1.79(2)	1711 ± 384	**1.00**
quake intensity ($\times 10^3$)	19 302	24.54	563.83	63 096	0.794 ± 80.198	1.64(4)	$11\,697 \pm 2159$	0.00
religious followers ($\times 10^6$)	103	27.36	136.64	1050	3.85 ± 1.60	1.8(1)	39 ± 26	**0.42**
freq. of surnames ($\times 10^3$)	2753	50.59	113.99	2502	111.92 ± 40.67	2.5(2)	239 ± 215	**0.20**
net worth (mil. USD)	400	2388.69	4 167.35	46 000	900 ± 364	2.3(1)	302 ± 77	0.00
citations to papers	415 229	16.17	44.02	8904	160 ± 35	3.16(6)	3455 ± 1859	**0.20**
papers authored	401 445	7.21	16.52	1416	133 ± 13	4.3(1)	988 ± 377	**0.90**
hits to web sites	119 724	9.83	392.52	129 641	2 ± 13	1.81(8)	$50\,981 \pm 16\,898$	0.00
links to web sites	241 428 853	9.15	106 871.65	1 199 466	3684 ± 151	2.336(9)	$28\,986 \pm 1560$	0.00

[1] Clauset, A. *et al.* 2009. *SIAM Review* 51:661–703.

Component Size Distribution In Erdős-Rényi Graph At Phase Transition



Scaling Analysis

- Two quantities x and y are in a **scaling relationship** if they are related mathematically as $y \propto x^{\beta}$
  - $\beta < 1$: Decreasing returns of the average ${}^{y}/_{x}$ to scale
  - $\beta = 1$: Constant returns of the average ${}^{y}/_{x}$ to scale
  - $\beta > 1$: Increasing returns of the average ${}^{y}/_{x}$ to scale
- **Allometry**: Study of relationship of body size to, *e.g.*, metabolic rate, shape, anatomy, physiology, and behavior
- **Methodological caveat**: Account for noise in x and y, if present, when measuring residuals [2]

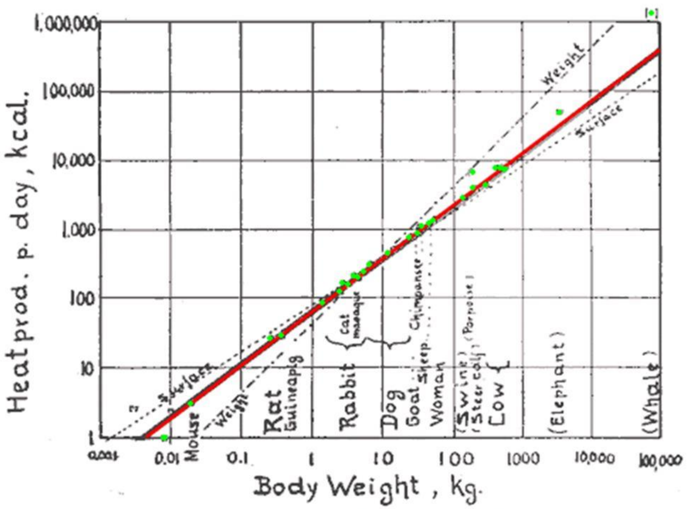
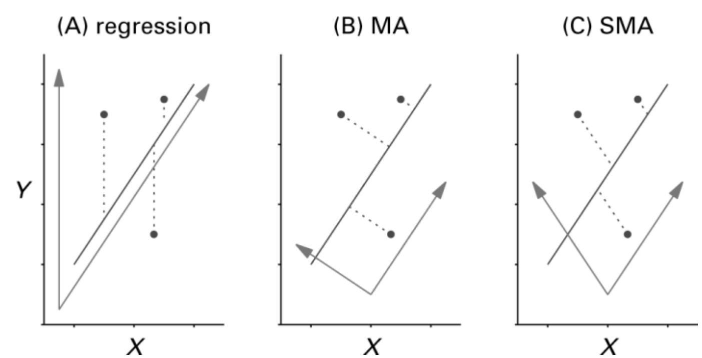
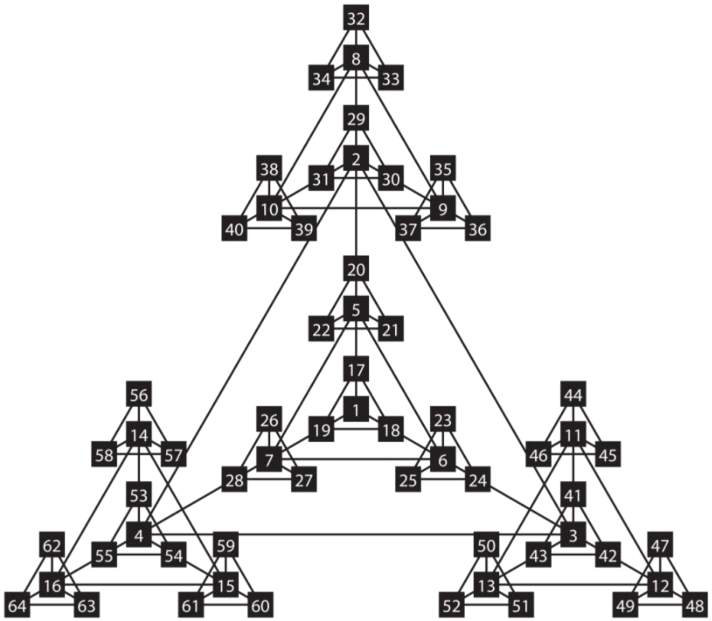


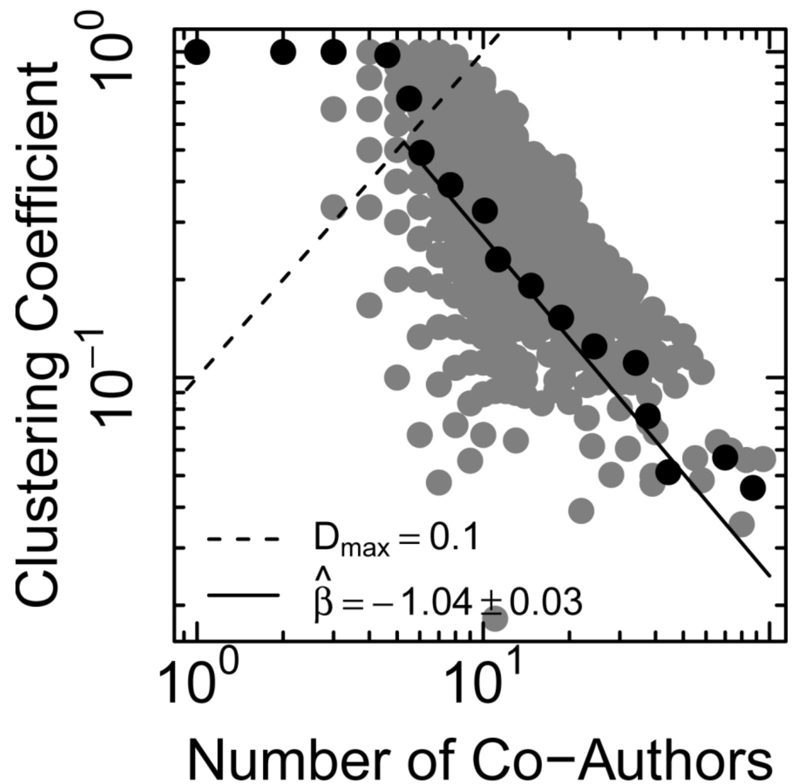
Fig. 1. Log. metabol. rate/log body weight



[1] Kleiber, M. 1947. *Physiological Reviews* 27:511–41.
[2] Warton, D.I. *et al.* 2006. *Biological Reviews* 81:259–291.

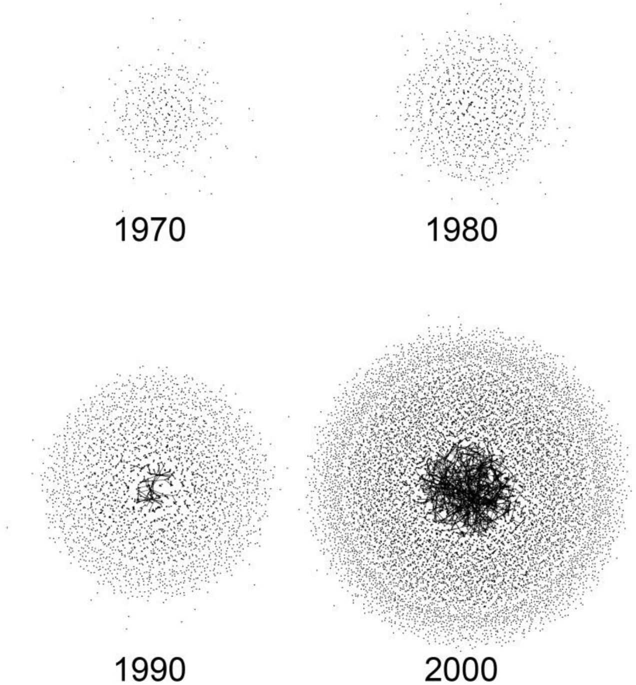
Social Scaling: Hierarchical Modularity

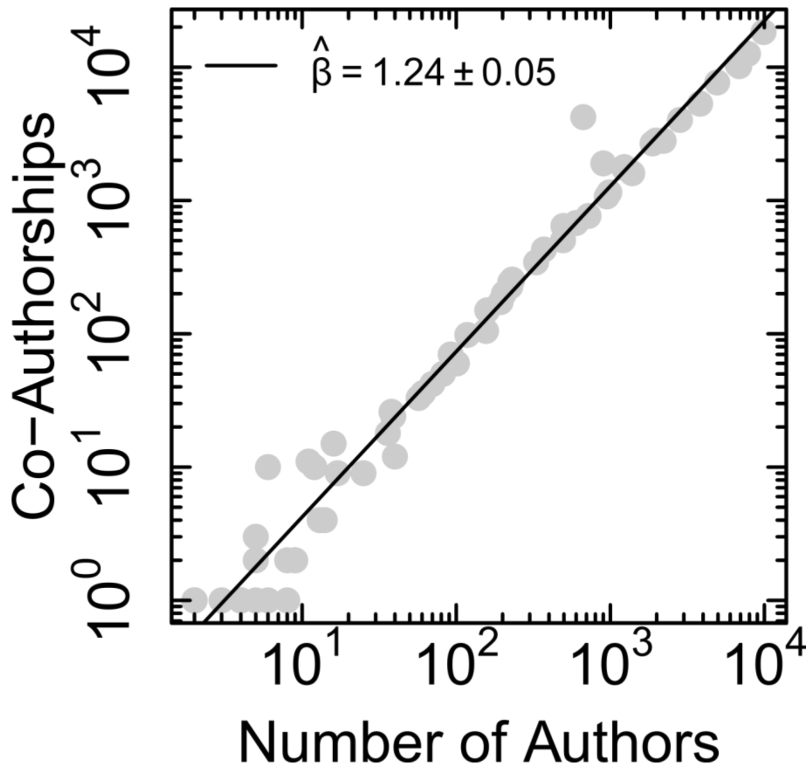




[1] Ravasz, E. & Barabási, A.-L. 2003. *Physical Review E* 67:026112.
[2] Lietz, H. 2016. *Scale-Free Identity*. Dissertation, University of Duisburg-Essen.

Social Scaling: Network Densification





[1] Ravasz, E. & Barabási, A.-L. 2003. *Physical Review E* 67:026112.
[2] Lietz, H. 2016. *Scale-Free Identity*. Dissertation, University of Duisburg-Essen.

Next: Small-World Networks (Demo)

Leibniz Institute for the Social Sciences

Leibniz Association