

Complex Networks

Mark Jelasity

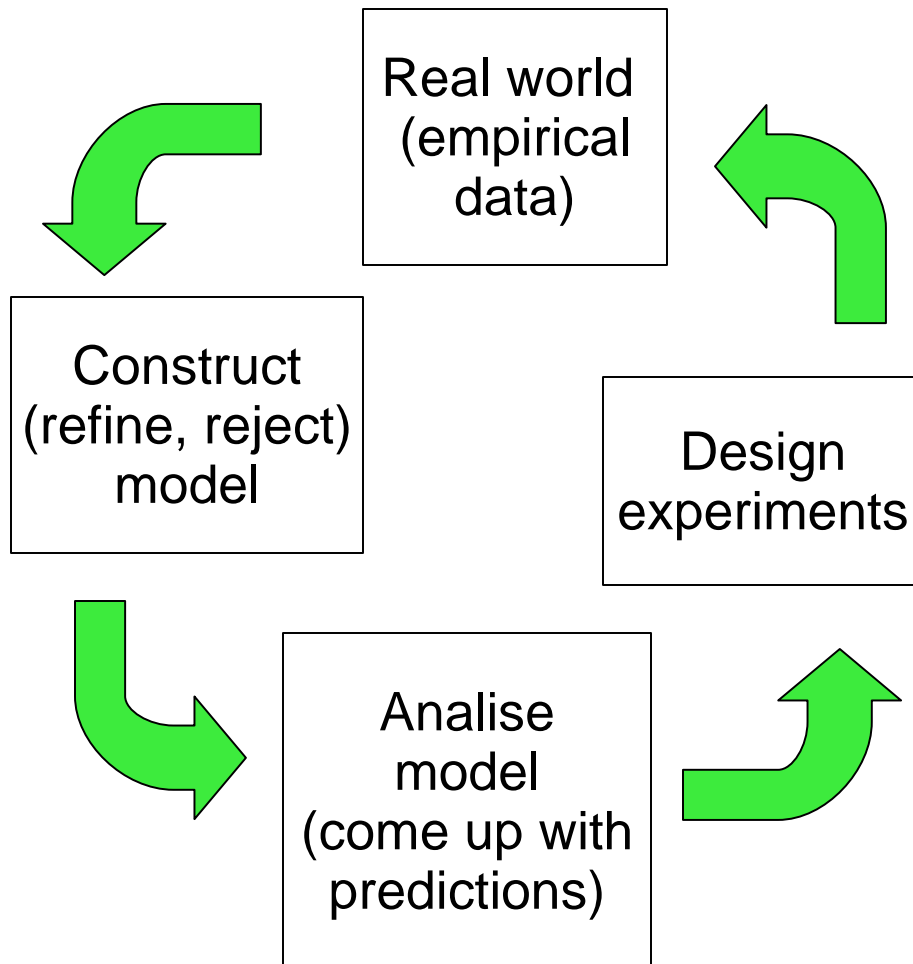
Motivation

- Where are the networks?
 - Some example computer systems
 - **WWW, Internet routers, software components**
 - Large decentralized systems
 - **Communication topology is always a non-trivial network**
 - Other networks
 - **Social relationships, food web, chemical reactions (DNA), etc**
- Complex self-managing systems will inevitably have to deal with complex networks

Motivation

- Some networks are actually important (not only interesting)
 - WWW, Internet, food web, metabolic nets, etc
- Some common aspects must be understood for most networks
 - Robustness
 - Epidemics (spreading of info, etc)
 - Efficiency
 - **function of network depending on its structure**
 - Design and engineering
 - **need to understand emergent properties**

This is empirical science



- Complex networks is a branch of physics

- Empirical: loop of modeling and observation

- Models capture only selected aspects

Outline

- Basic concepts recap (graphs, probability)
- Graph models
 - Erdős-Rényi
 - General degree distribution
 - Watts-Strogatz
 - Barabási-Albert
 - [motifs]

Graph theoretical concepts

- Node, edge
- Graph
 - Directed, undirected, simple
- Paths
 - Length, average length, diameter
- Connected graph
 - Strongly, weakly
- Node degree
 - In-, out-, average, distribution

Probability

- Discrete distribution, random variable
- Expectation value, variance
- What is a random graph?
 - Probability space of graphs

The model

- Simple undirected graph $G_{N,p}$
- Parameters
 - N : number of nodes
 - p : probability of connecting any pairs of nodes
- Algorithm
 - Start with empty graph of N nodes
 - Draw all $N(N-1)/2$ possible edges with probability p
- Stats of degree of a fixed node i
 - $\langle k_i \rangle = p(N-1)$, k_i has binomial distr, approx Poisson

Probabilistic properties

- Usual question: $P(Q)$ over a probability space of graphs
 - Q can be eg “connected”, or “contains a triangle”, etc
- Usually $P(Q)$ depends on N and p
- We are interested in “almost always” Q :

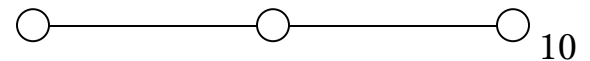
$$P_{N,p}(Q) \rightarrow 1 \quad (N \rightarrow \infty)$$

Probabilistic properties

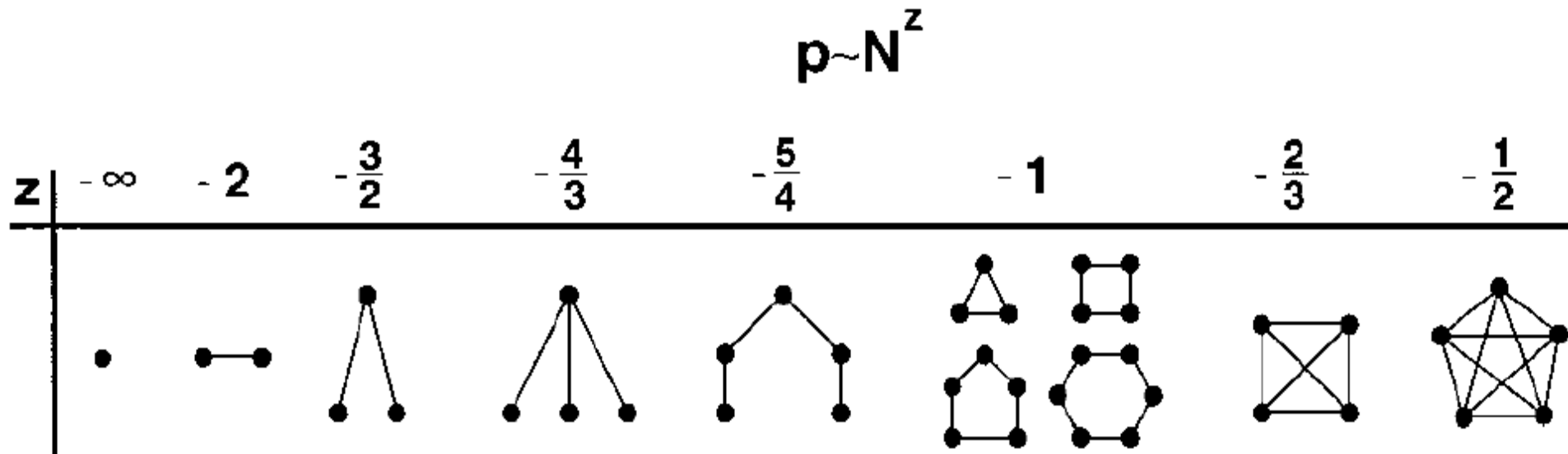
- Often there is a critical probability p_c such that

$$\lim_{N \rightarrow \infty} P_{N,p}(Q) = \begin{cases} 0 & \frac{p(N)}{p_c(N)} \rightarrow 0 \\ 1 & \frac{p(N)}{p_c(N)} \rightarrow \infty \end{cases}$$

- We are interested in p_c for different Q-s
- Example: $G_{N,p}$ has a subgraph



Critical pr. for small subgraphs



- Note the case $p \sim 1/N$ where cycles of all order appear
- Note that this is understood as N tends to ∞

Connectivity

- Let's look at connectivity as a function of p
 - AKA “graph evolution”: when we keep adding edges
- Note that if p grows slower than $1/N$, the graph is a disconnected collection of small (constant size) components
- If $p \sim 1/N$, avg node degree $\langle k \rangle$ is constant, cycles of all order have finite probability
 - What's going on if $\langle k \rangle$ is constant?

The case when $p \sim 1/N$

- $0 < \langle k \rangle < 1$
 - One cycle, otherwise trees, the largest being $O(\ln N)$ size
 - The number of clusters is $N - n$ (ie each new edge connects two clusters)
- $\langle k \rangle = 1$
 - Critical value: largest cluster is suddenly $O(N^{2/3})$, with complex structure
- $\langle k \rangle > 1$
 - The largest cluster is of size $(1 - f(\langle k \rangle))N$ nodes where f decreases exponentially
- [If $\langle k \rangle \geq \ln N$, completely connected (but here the avg degree grows with N)]

Degree distribution

- k_i the degree of fixed node
 - k_i is binomial ($\text{Bin}(N-1, p)$)
- Degree distribution: the degree of a random node from a random graph
 - x_k : number of nodes with degree k
 - $\langle x_k \rangle = NP(k_i = k)$
 - Distribution of x_k has very low variance
 - So it is a reasonable assumption to say that a random graph $G_{N,p}$ has very close to binomial degree distribution

Diameter

- The longest shortest path
- $L = \ln N / \ln \langle k \rangle = \log_{\langle k \rangle} N$
- The reason is that these graphs are locally like trees
- The average path length (l) grows also as $\log_{\langle k \rangle} N$
- Observed networks tend to have a diameter consistent with this prediction

Statistics of some networks

Network	Size	$\langle k \rangle$	ℓ	ℓ_{rand}	C	C_{rand}	Reference
WWW, site level, undir.	153 127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999
Internet, domain level	3015–6209	3.52–4.11	3.7–3.76	6.36–6.18	0.18–0.3	0.001	Yook <i>et al.</i> , 2001a, Pastor-Satorras <i>et al.</i> , 2001
Movie actors	225 226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz, 1998
LANL co-authorship	52 909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c
MEDLINE co-authorship	1 520 251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c
SPIRES co-authorship	56 627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c
NCSTRL co-authorship	11 994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c
Math. co-authorship	70 975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabási <i>et al.</i> , 2001
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási <i>et al.</i> , 2001
<i>E. coli</i> , substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000
<i>E. coli</i> , reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Solé, 2000
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Solé, 2000
Words, co-occurrence	460.902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Solé, 2001
Words, synonyms	22 311	13.48	4.5	3.84	0.7	0.0006	Yook <i>et al.</i> , 2001b
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998
<i>C. Elegans</i>	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998

Clustering coefficient

- Definition of clustering coefficient
 - Ratio of actual and possible number of edges between neighbors of a node
- In this model it is evident
 - $C = p = \langle k \rangle / N$
 - Very small
- This does NOT predict the clustering in real networks

Some other similar models

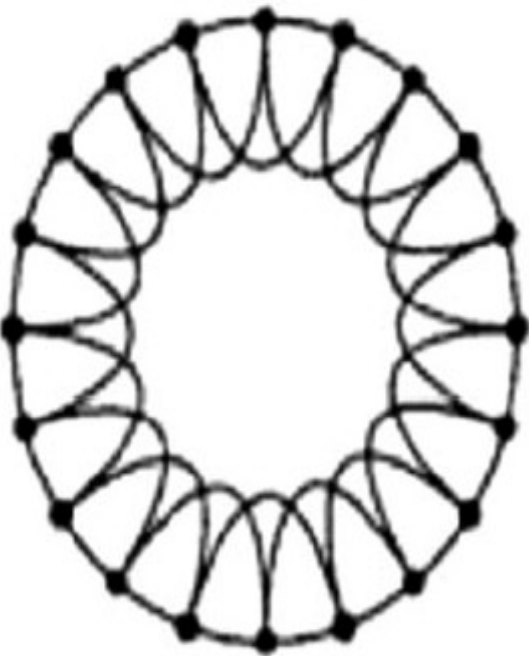
- $G_{r\text{-reg}}$: probability space is the set of r -regular graphs with equal probability
 - $G_{3\text{-reg}}$ is Hamiltonian
 - Note that $G_{3/(N-1),N}$ is not even connected
- $G_{r\text{-out}}$: we generate a random graph by adding 3 edges from all nodes
 - $G_{4\text{-out}}$ is Hamiltonian
 - It is believed that $G_{3\text{-out}}$ is also Hamiltonian
- So we need to be careful
- When there is guarantee that all nodes have some edges, things are radically different

Watts-Strogatz model

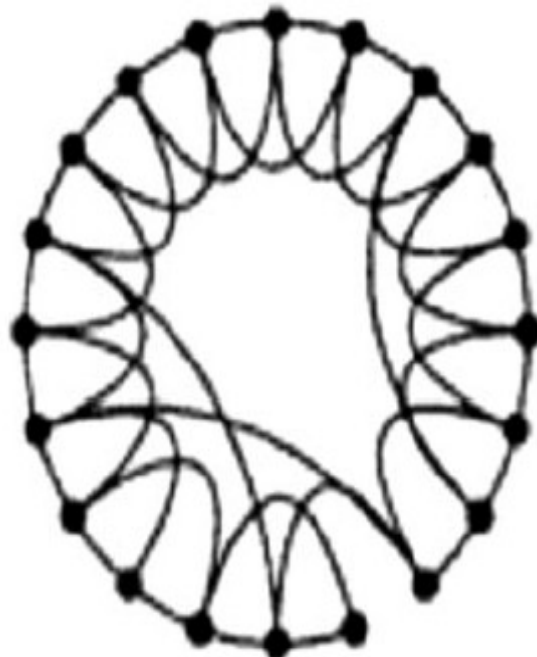
- Motivation: random graphs don't model clustering
- Local structure + randomness (“shortcuts”)
 - Ring with links to K nearest neighbors
 - Rewire each of the $K/2$ links to the left of a node with probability p ($pNK/2$ shortcuts on average)
- Clustering is $c=3(K-2)/4(K-1)$ if $p=0$
- Average path length is $O(N)$ if $p=0$
- With $p=1$ we get the $G_{k/2\text{-out}}$ model, not the Erdős-Rényi model

Watts-Strogatz model

Regular



Small-world



Random



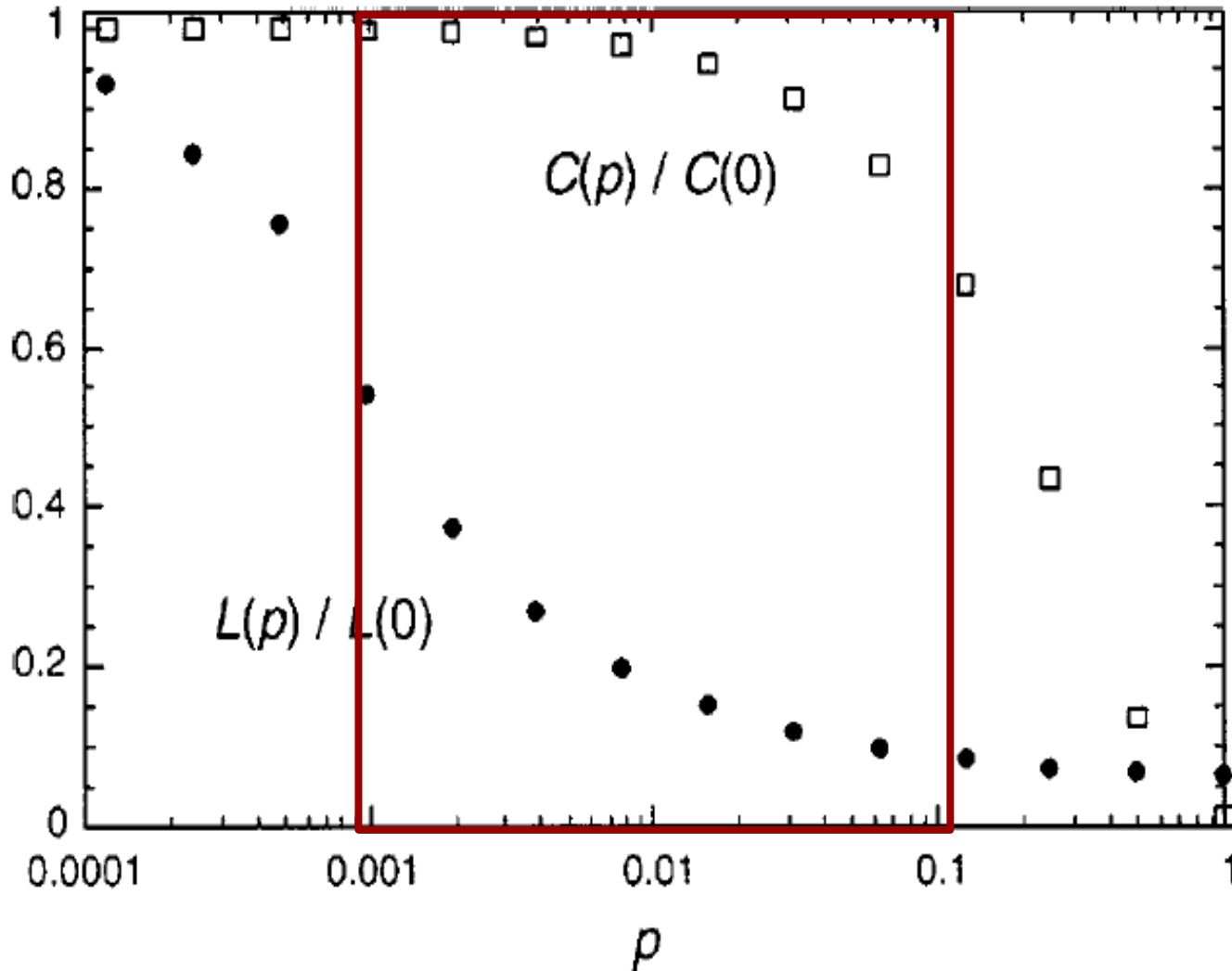
$p = 0$



$p = 1$

Increasing randomness

The small world region



In a wide region clustering is large, path length is short: small world graph

Statistical properties

- Clustering in the general case

$$C(p) \approx C(0)(1-p)^3 = \frac{3(K-2)}{4(K-1)}(1-p)^3$$

- Degree distribution

- Transition from constant (K) to Poisson(K/2)+K/2

- Path length

- Small p: linear; large p: logarithmic
- Transition: $p=2/NK$ (1 shortcut on average)

Growth models

- So far we can model clustering and path length. Is this all? No
- Degree distribution is very often heavy tail
 - $P(k) \sim k^{-\gamma}$ (often some cutoff eg $P(k) \sim k^{-\gamma} e^{-k/\gamma}$)
 - Without cutoff
 - **No expectation value (ie $\langle k \rangle = \infty$) if $\gamma \leq 2$**
 - **No variance (ie $\text{Var}(k) = \infty$) if $\gamma \leq 3$, etc**
- Called scale-free because of fractals

Observed scale free networks

Network	Size	$\langle k \rangle$	κ	γ_{out}	γ_{in}	ℓ_{real}	ℓ_{rand}	ℓ_{pow}	Reference	Nr.
WWW	325 729	4.51	900	2.45	2.1	11.2	8.32	4.77	Albert, Jeong, and Barabási 1999	1
WWW	4×10^7	7		2.38	2.1				Kumar <i>et al.</i> , 1999	2
WWW	2×10^8	7.5	4000	2.72	2.1	16	8.85	7.61	Broder <i>et al.</i> , 2000	3
WWW, site	260 000				1.94				Huberman and Adamic, 2000	4
Internet, domain*	3015–4389	3.42–3.76	30–40	2.1–2.2	2.1–2.2	4	6.3	5.2	Faloutsos, 1999	5
Internet, router*	3888	2.57	30	2.48	2.48	12.15	8.75	7.67	Faloutsos, 1999	6
Internet, router*	150 000	2.66	60	2.4	2.4	11	12.8	7.47	Govindan, 2000	7
Movie actors*	212 250	28.78	900	2.3	2.3	4.54	3.65	4.01	Barabási and Albert, 1999	8
Co-authors, SPIRES*	56 627	173	1100	1.2	1.2	4	2.12	1.95	Newman, 2001b	9
Co-authors, neuro.*	209 293	11.54	400	2.1	2.1	6	5.01	3.86	Barabási <i>et al.</i> , 2001	10
Co-authors, math.*	70 975	3.9	120	2.5	2.5	9.5	8.2	6.53	Barabási <i>et al.</i> , 2001	11
Sexual contacts*	2810			3.4	3.4				Liljeros <i>et al.</i> , 2001	12
Metabolic, <i>E. coli</i>	778	7.4	110	2.2	2.2	3.2	3.32	2.89	Jeong <i>et al.</i> , 2000	13
Protein, <i>S. cerev.</i> *	1870	2.39		2.4	2.4				Jeong, Mason, <i>et al.</i> , 2001	14
Ythan estuary*	134	8.7	35	1.05	1.05	2.43	2.26	1.71	Montoya and Solé, 2000	14
Silwood Park*	154	4.75	27	1.13	1.13	3.4	3.23	2	Montoya and Solé, 2000	16
Citation	783 339	8.57			3				Redner, 1998	17
Phone call	53×10^6	3.16		2.1	2.1				Aiello <i>et al.</i> , 2000	18
Words, co-occurrence*	460 902	70.13		2.7	2.7				Ferrer i Cancho and Solé, 2001	19
Words, synonyms*	22 311	13.48		2.8	2.8				Yook <i>et al.</i> , 2001b	20

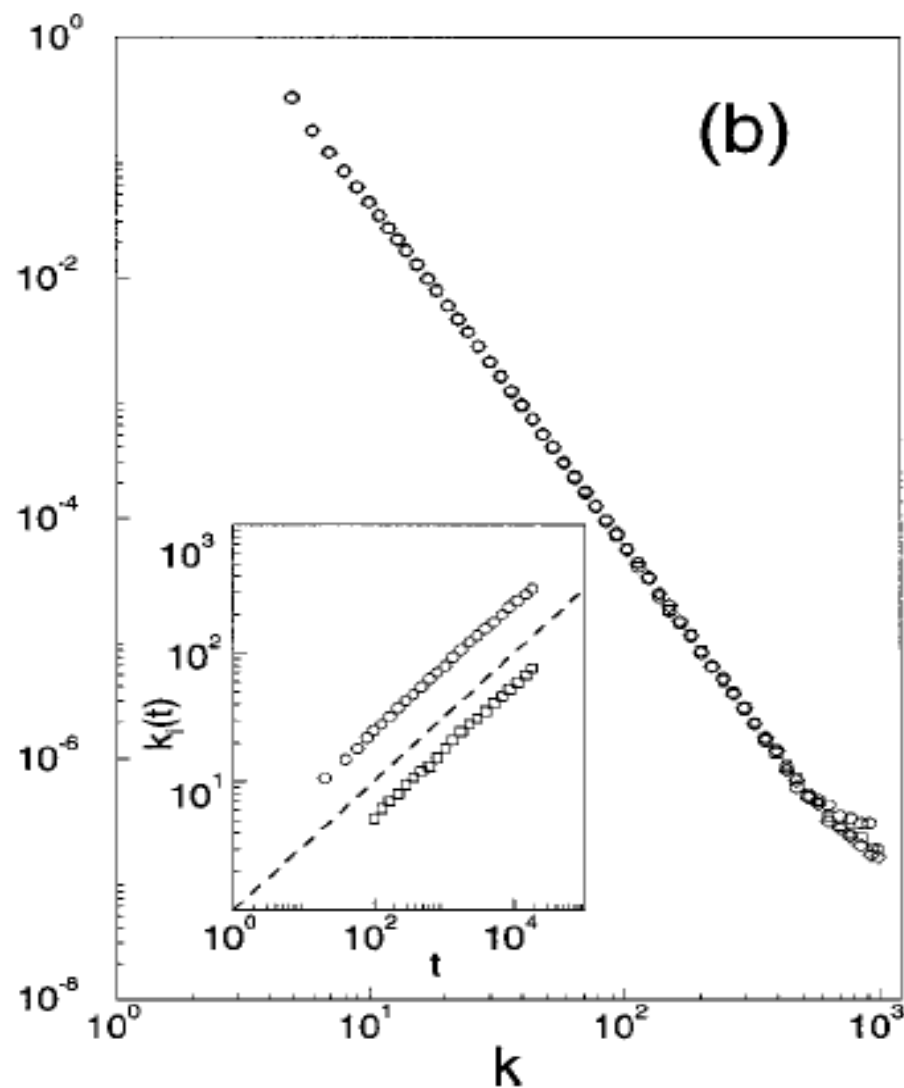
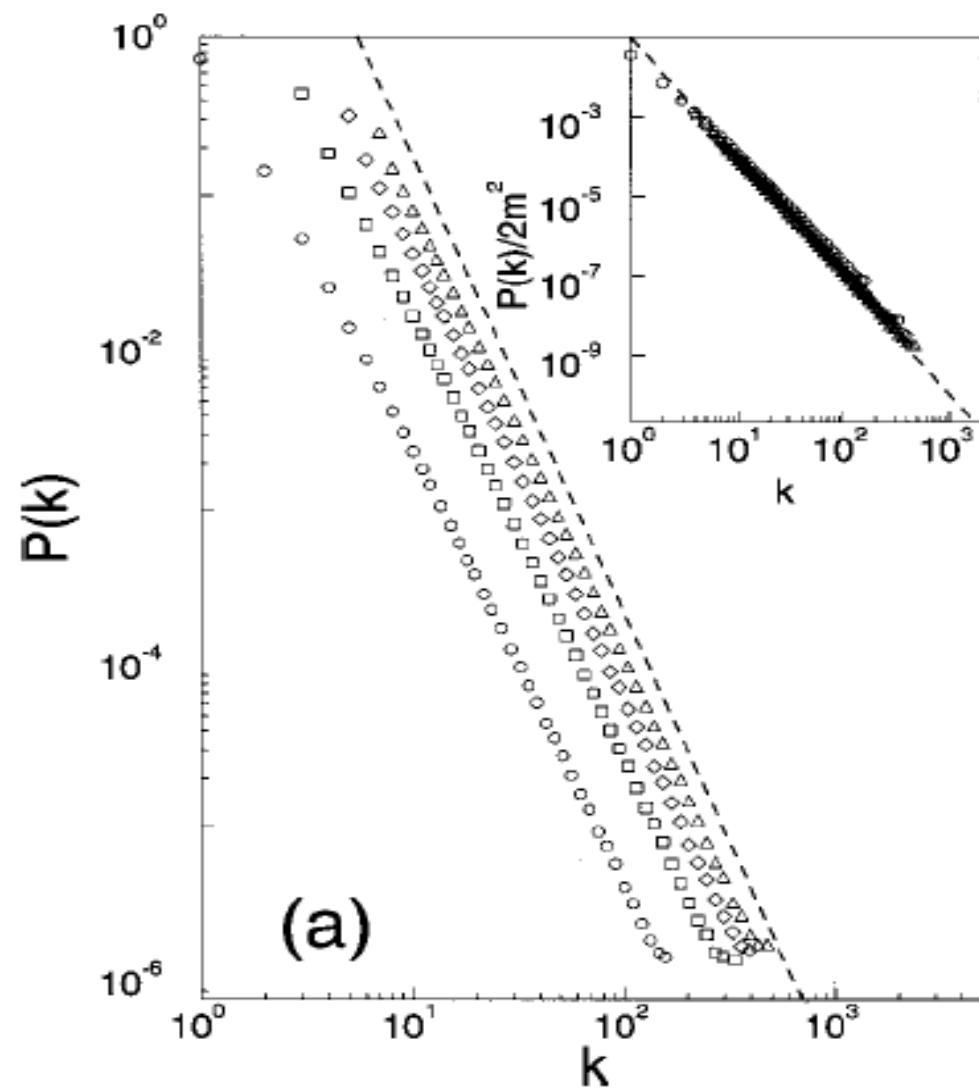
Barabási-Albert model

- Preferential attachment rule
 - Start with a small number (m_0) of nodes
 - Repeat adding a new node with $m \leq m_0$ links, where each link is linked to node i according to

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

- T time step, $t+m_0$ nodes, mt edges
- Converges to exponent $\gamma=3$

Empirical results with BA model



Some statistics

- Average path length
 - $L \sim \ln N / \ln \ln N$ (somewhat smaller than random)
- Clustering
 - $C \sim N^{-0.75}$, (recall that random was $1/N$)
- In Sum
 - Models degree distribution
 - But doesn't model clustering

General degree distribution

- BA model has another problem
 - Correlation between degree of neighbors
- General model
 - Given a sequence of degrees
 - Construct a probability space in which all graphs with the given sequence are equiprobable
 - Stubs method
 - **Problems: loop edges, multiple edges**



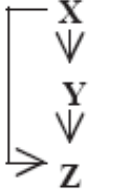

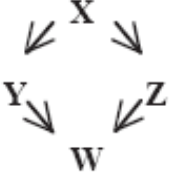

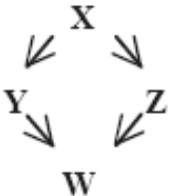
Connectivity of general model

- [Recall the ER model had $\langle k \rangle = 1$ as a tipping point for giant component]
- General rule for connectivity (critical value): $\langle k^2 \rangle - 2\langle k \rangle = 0$
- For the Poisson distribution this gives $\langle k \rangle^2 = \langle k \rangle$, that is, $\langle k \rangle = 1$

Network motifs

- Degree distribution, path length, clustering; is this all to account for?
- In a random model, small subgraphs have a theoretical distribution
- In a real network, some small subgraphs are represented more or less frequently
 - This is yet another aspect to account for in a model
 - Are motifs functional? Or just side effects?
 - In other words, should we bother?
- Z-score: $(N_{\text{real}} - N_{\text{rand}}) / \text{SD}$

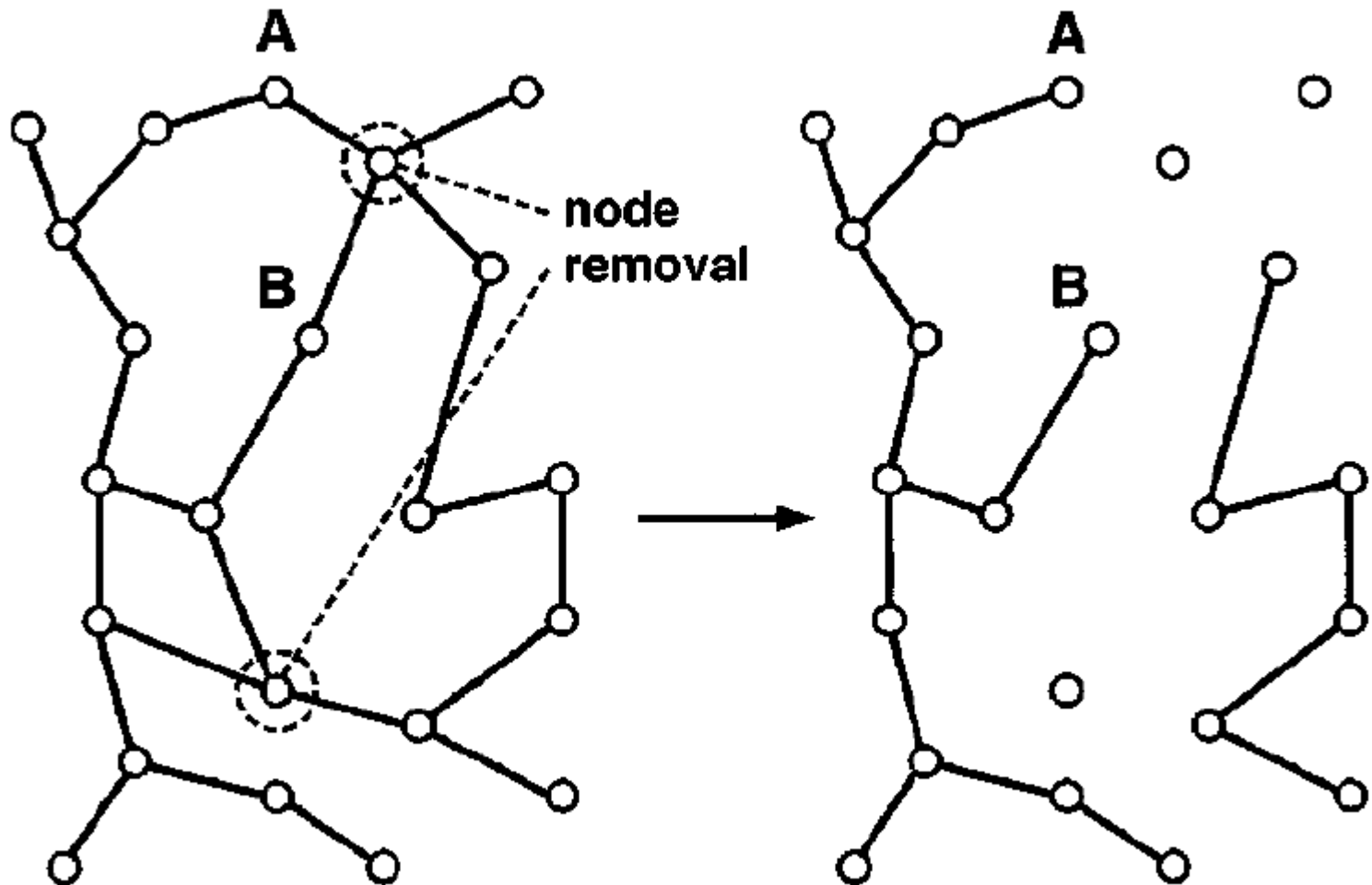
Some examples for motifs

Network	Nodes	Edges	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z score	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z score	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z score
Gene regulation (transcription)			 Feed-forward loop			 Bi-fan					
<i>E. coli</i>	424	519	40	7 ± 3	10	203	47 ± 12	13			
<i>S. cerevisiae</i> *	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons			 Feed-forward loop			 Bi-fan			 Bi-parallel		
<i>C. elegans</i> †	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs			 Three chain			 Bi-parallel					
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			

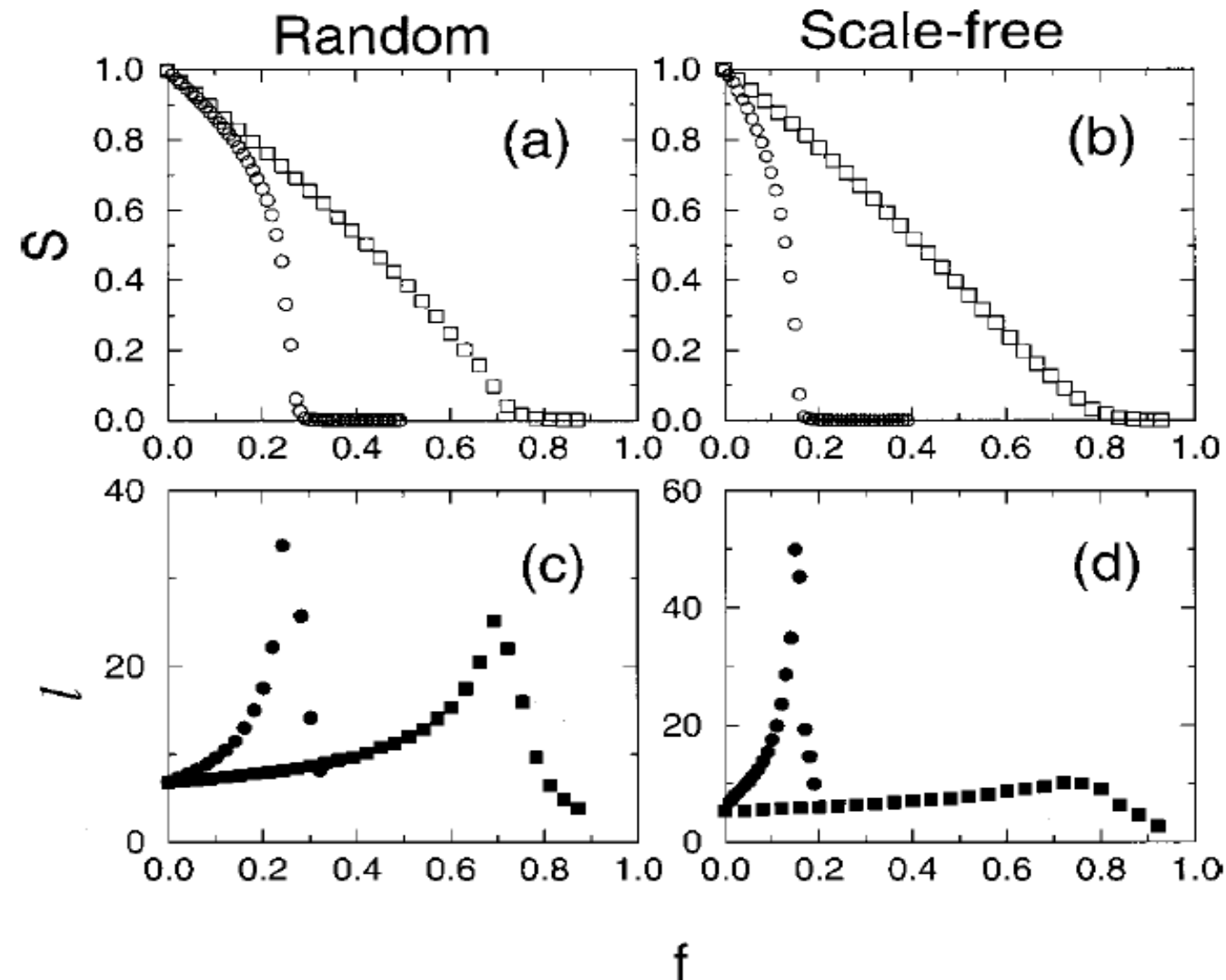
Error and attack tolerance

- We need to understand how vulnerable existing systems are
- We need to design self-healing and self-protecting systems
- Models
 - Node removal: failure
 - **A random node is removed along with all the links**
 - Node removal: attack
 - **The most connected (highest degree) nodes are removed**

Node removal



Achilles' heel

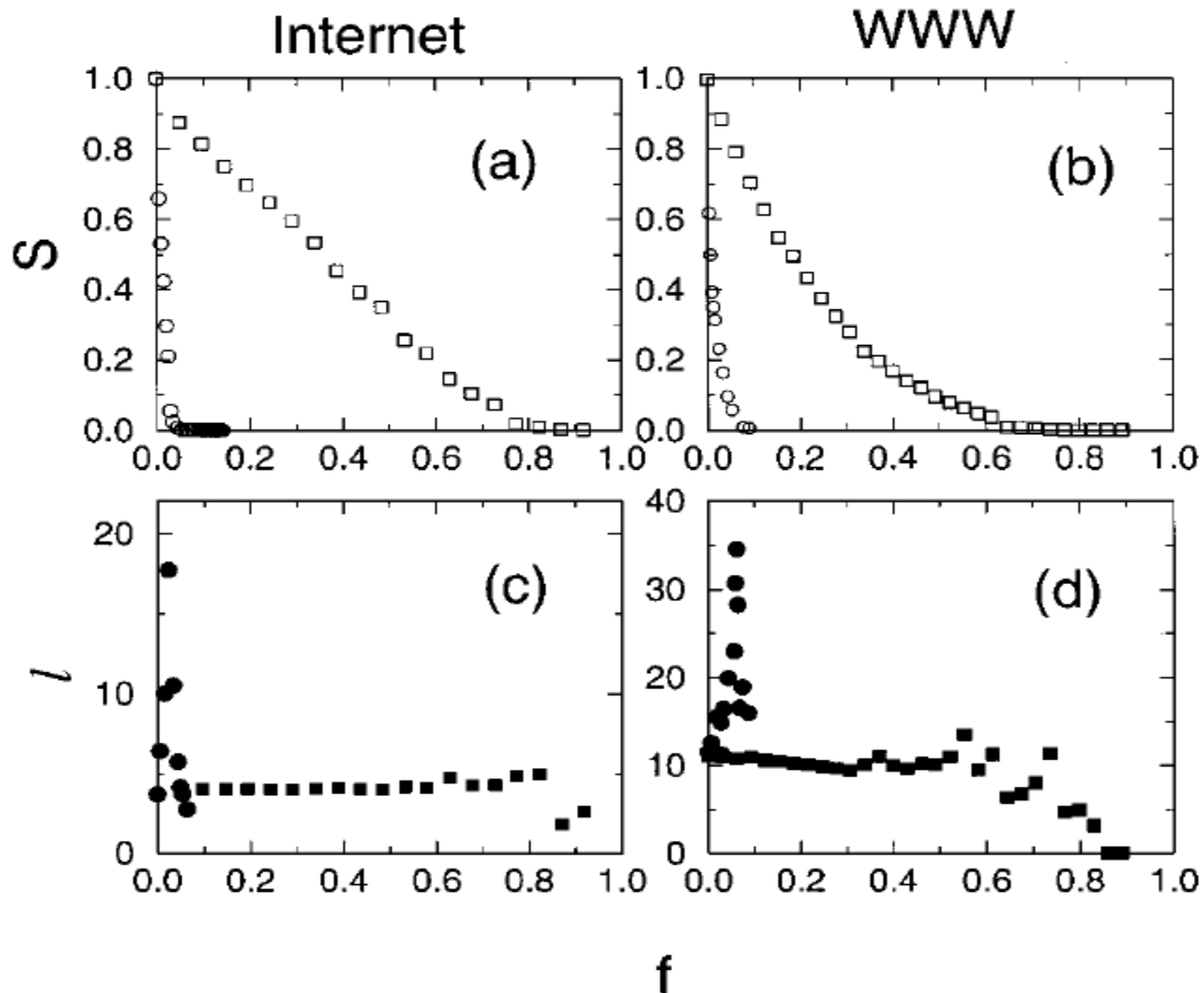


$N=10\,000$

$\langle k \rangle = 4$

ER and BA
model

Real world examples



Real world examples

- Internet and WWW
 - Extremely sensitive to attack, and extremely robust to random failure
- Cellular networks
 - 8% removal 500% increase in path length is attack, otherwise unchanged
- Ecological networks Silwood Park web
 - Error tolerance: 95% removal
 - Attack tolerance: 20% removal
 - Secondary extinctions under attack: 16% removal

Some refs

- Papers this presentation used material from
 - Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47-97, January 2002.
 - R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824-827, 2002.
 - Mark E. J. Newman. Random graphs as models of networks. In Stefan Bornholdt and Heinz G. Schuster, editors, *Handbook of Graphs and Networks: From the Genome to the Internet*, chapter 2. John Wiley, New York, NY, 2002.
- The course website
 - <http://www.inf.u-szeged.hu/~jelasity/p2p/>