P.h.D. course
Universitat Autonoma de Barcelona, 2022

(**S**ocial) **N**etwork **A**nalysis

Giulio Rossetti

Knowledge Discovery and Data Mining Laboratory (KDD) @ ISTI-CNR

giulio.rossetti@isti.cnr.it

@GiulioRossetti

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

Lecture 2

Characterizing by contraposition: Real Networks and Synthetic Models

Chapter 2

Random Networks

Summary

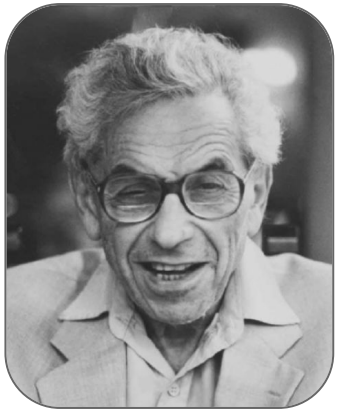
- Random Graphs
- Erdos-Renyi model
- Paths, Connectedness & Density

Reading

- Chapter 3 of Barabasi's book

Random Graphs

The Erdős-Rényi
Random Graph model (ER)



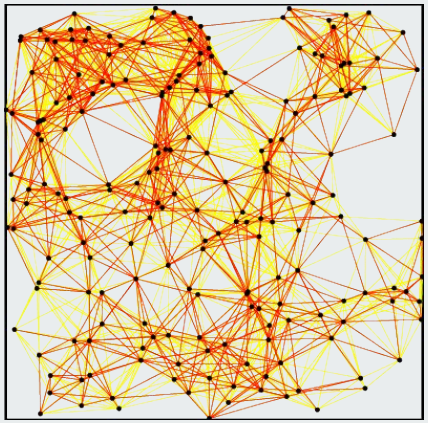
Pál Erdős
(1913-1996)

Alfréd Rényi
(1921-1970)

*"If we do not know anything else than the number **N** of nodes and the number **L** of links, the simplest thing to do is to put the links at random (no correlations)"*

[1] P. Erdős and A. Rényi.
On random graphs, I. Publicationes Mathematicae. 1959.
[2] P. Erdős and A. Rényi.
On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci., 1960.

Why using Random Graph models?



- Study some properties in a “controlled environment”
  How does property X behaves when increasing property Y ?
- Compare an observed network with a randomized version
  Is observed property X “exceptional”, or any similar network with same property Y and Z ?
- Explain a given phenomenon
  Such simple mechanism can reproduce property X and Y
- Generate synthetic datasets
  Testing an algorithm on 100 variations of the same network

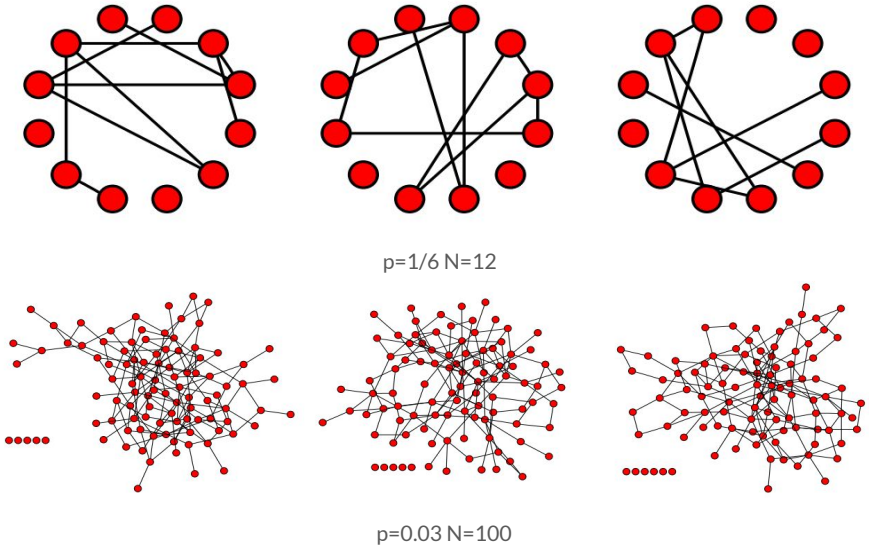
ER model

(General) Definition.
A random graph is a graph of **N** nodes where each pair of nodes is connected by probability **p**.

The **G(n,L)** definition.
1. Take **N** disconnected nodes
2. Add **L** edges uniformly at random

The **G(n,p)** definition.
1. Take **N** disconnected nodes
2. Add an edge between any of the nodes independently with probability **p**

In the G(n,p) variant, the number of edges may vary



Random Graphs

P(L): probability to have exactly **L** links in a network of **N** nodes and probability **p** (binomial distribution)

The maximum number of links in a network of N nodes.

$$P(L) = \binom{\binom{N}{2}}{L} p^L (1-p)^{\frac{N(N-1)}{2}-L}$$

Number of different ways we can choose L links among all potential links.

Reminder (Binomial Coefficient)
Number of ways, disregarding order, that k objects can be chosen from among n objects

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

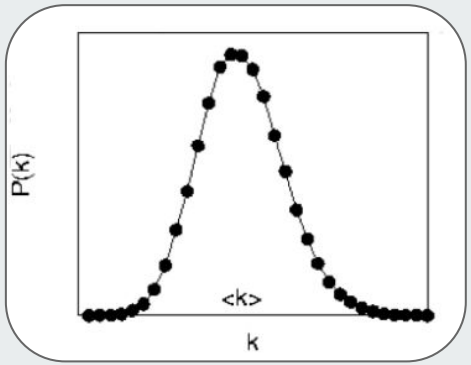
<L>: The average number of links in a random graph
<k>: The average degree (and its variance)

$$< L >= \sum_{l=0}^{\frac{N(N-1)}{2}} LP(L) = p\frac{N(N-1)}{2} \qquad < k >= \frac{2L}{N} = p(N-1)$$

$$\sigma^2 = \rho(1-\rho)\frac{N(N-1)}{2}$$

Degree Distribution

For each node, independent probabilities to take each neighbor => Binomial distribution



probability of having k edges

$$P(k) = \binom{N-1}{k} p^k (1-p)^{(N-1)-k}$$

Select k nodes from N-1

probability of missing N-1-k edges

As the network size increases, the distribution becomes increasingly narrow—we are increasingly confident that the degree of a node is in the vicinity of <k>.

$$\frac{\sigma_k}{< k >} = \left[\frac{1-p}{p}\frac{1}{(N-1)}\right]^{1/2} \approx \frac{1}{(N-1)^{1/2}}$$

Characteristics:

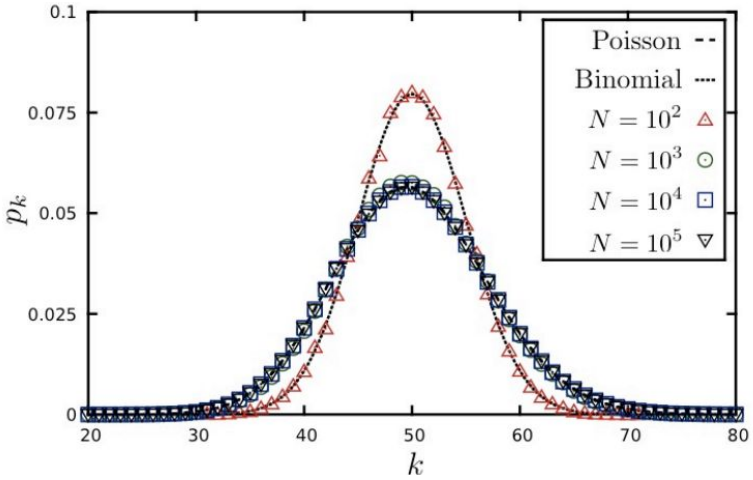
$$< k >= p(N-1)$$

$$\sigma_k^2 = p(1-p)(N-1)$$

Visual Simulation
http://www.networkpages.nl/CustomMedia/Animations/RandomGraph/ERRG/DegreeDistribution.html

Degree Distribution

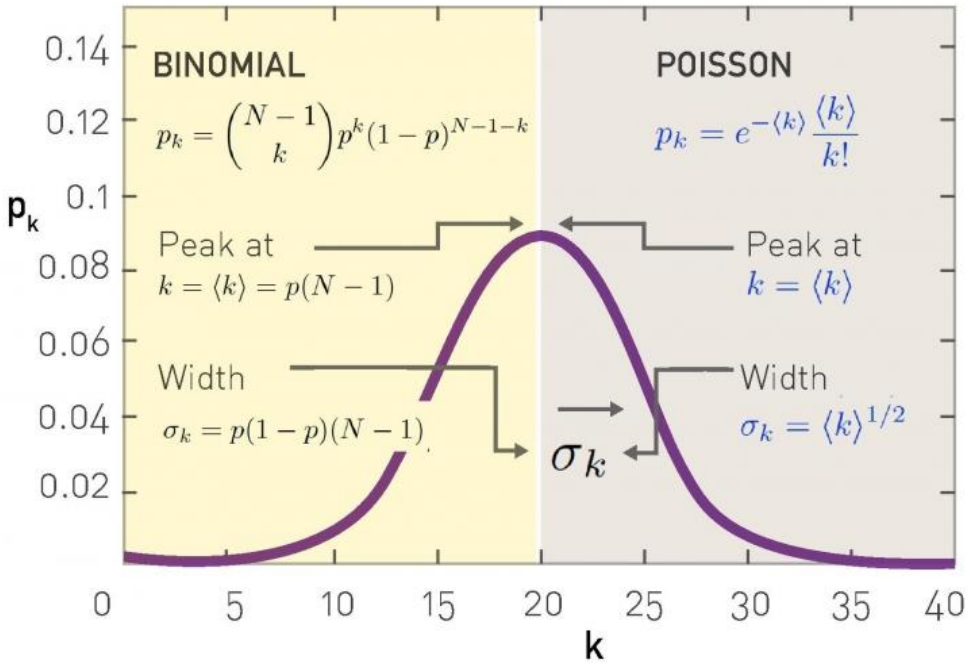
For large **N** and small **k** (p,L), we can approximate the degree distribution using a poisson distribution of parameter (mean) **λ = < k >**



Poisson distribution $P(K) = \frac{\lambda^K e^{-\lambda}}{K!}$

Distribution of degrees $P(k) = \frac{< k >^k e^{-<k>}}{k!}$

Standard deviation $\sigma = \sqrt{< k >}$

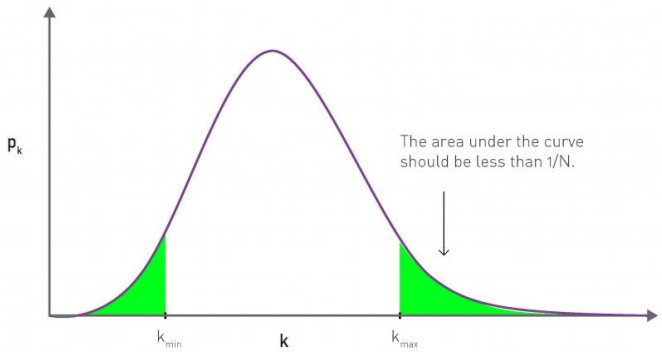


Exact Result - Binomial distribution

Large N limit - Poisson distribution

Real Networks are not Poisson

Maximum & Minimum Degree



Let's assume <k> =1,000, $N=10^9$.
We can derive the max and min degrees as follows:

$$N[1 - P(k_{\max})] \approx 1$$

$$\mathrm{I} - P(k_{\max}) = 1 - \mathrm{e}^{-<k>} \sum_{k=0}^{k_{\max}} \frac{\langle k \rangle^k}{k!} = \mathrm{e}^{-<k>} \sum_{k=k_{\max}+1}^{\infty} \frac{\langle k \rangle^k}{k!} \approx \mathrm{e}^{-<k>} \frac{\langle k \rangle \langle k \rangle^{k_{\max}}}{(k_{\max}+1)!}$$

k_{max} = 1,185

$$NP(k_{\min}) \approx 1$$

$$P(k_{\min}) = e^{-\langle k \rangle} \sum_{k=0}^{k_{\min}} \frac{\langle k \rangle^k}{k!}$$

k_{min} = 816

$$< \mathrm{K} > \pm \sigma_k \quad \sigma_k = < k >^{1/2}$$

$$\sigma_k = 31.62$$

No Outliers in a Random Society!

The most connected individual has degree $k_{max} \sim 1{,}185$

The least connected individual has degree $k_{min} \sim 816$

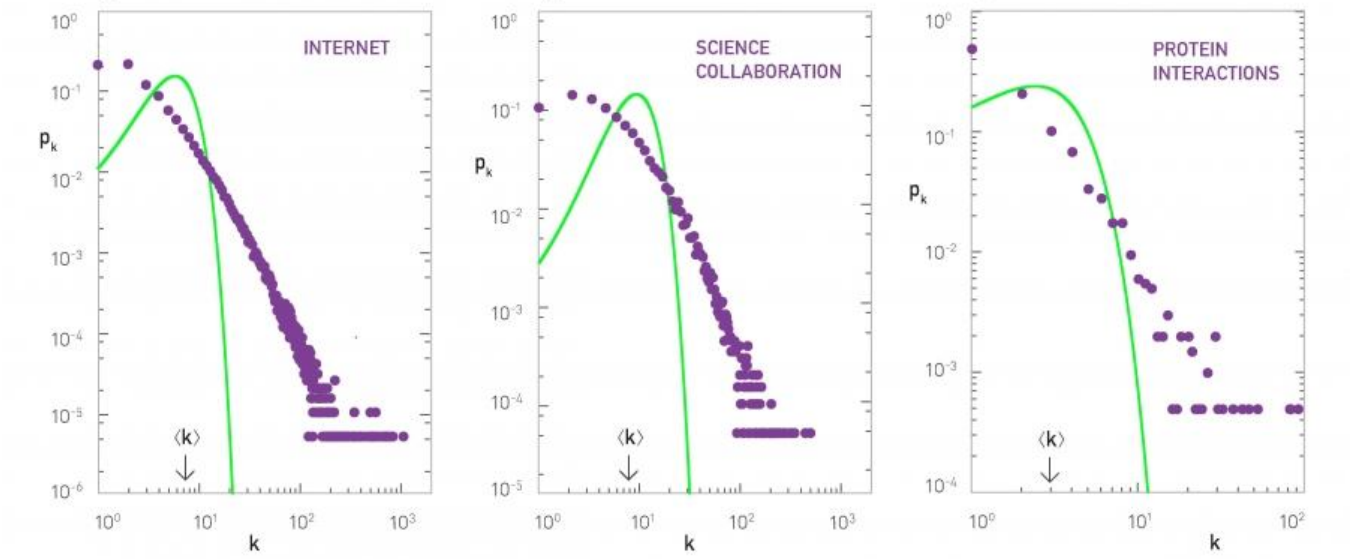
The probability to find an individual with degree $k > 2{,}000$ is 10^{-27}.

Hence the chance of finding an individual with 2,000 acquaintances is so tiny that such nodes are virtually nonexistent in a random society.

- A random society would consist of mainly average individuals, with everyone with roughly the same number of friends.
- It would lack outliers, individuals that are either highly popular or recluse.

Facing Reality:

Degree distribution of real networks



Clustering & Distance

Clustering in Random Graphs

For fixed average degree, **C** is decreasing as **N** goes large

- Low clustering coefficient
- It is vanishing with the system size

Reminder (clustering coeff.)

$$C_i \equiv \frac{2n_i}{k_i(k_i-1)}$$

where $\mathbf{n_i}$ is the number of links between the neighbours of node **i**

We know that $p = \frac{< k >}{n-1}$

thus,

$$C_i = \frac{2 < k >}{n-1} \frac{k_i(k_i-1)}{2} \frac{1}{k_i(k_i-1)} = \frac{< k >}{n-1} = p$$

Clustering

ER Graphs vs Real-World

ER

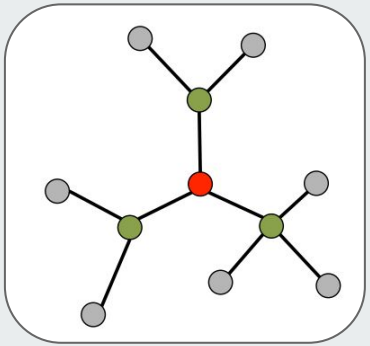
Expected Small Clustering Coefficient

$$C_i \equiv \frac{1}{N} < k >= p$$

Real-World 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 *et al.*, 2001a, Pastor-Satorras *et al.*, 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 *et al.*, 2001
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási *et al.*, 2001
E. coli, substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000
E. coli, 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 *et al.*, 2001b
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998
C. Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998

Distance in Random Graphs



Low Clustering coefficient

➔ Random graphs tend to have a tree-like topology with almost constant node degrees.

nr. of first neighbors:	$N(u)_1 =< k >$
nr. of second neighbors:	$N(u)_2 =< k >^2$
nr. of neighbours at distance d:	$N(u)_d =< k >^d$

Intuition: At which distance are all nodes reached?

$$n =< k >^d \Rightarrow \log_{<k>} n = d \Rightarrow d = \frac{\log n}{\log < k >}$$

Diameter, avg. distance is O(log n)

Distance

ER Graphs vs Real-World

ER

Logarithmically short distance among nodes

$$d = \frac{\log n}{\log \langle k \rangle}$$

Real-World Networks

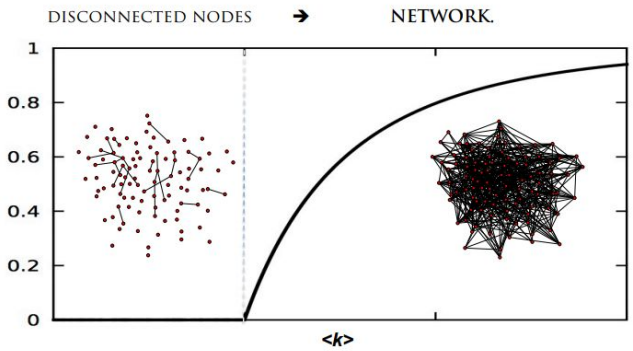
Network	Size	$\langle k \rangle$	ℓ	ℓ_{rand}	C	C_{rand}	Reference
WWW, site level, undir.	153 127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999
Internet, domain level	3015–6209	3.52–4.11	3.7–3.76	6.36–6.18	0.18–0.3	0.001	Yook *et al.*, 2001a, Pastor-Satorras *et al.*, 2001
Movie actors	225 226	61	3.65	2.99	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 *et al.*, 2001
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási *et al.*, 2001
E. coli, substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000
E. coli, 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 *et al.*, 2001b
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998
C. Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998

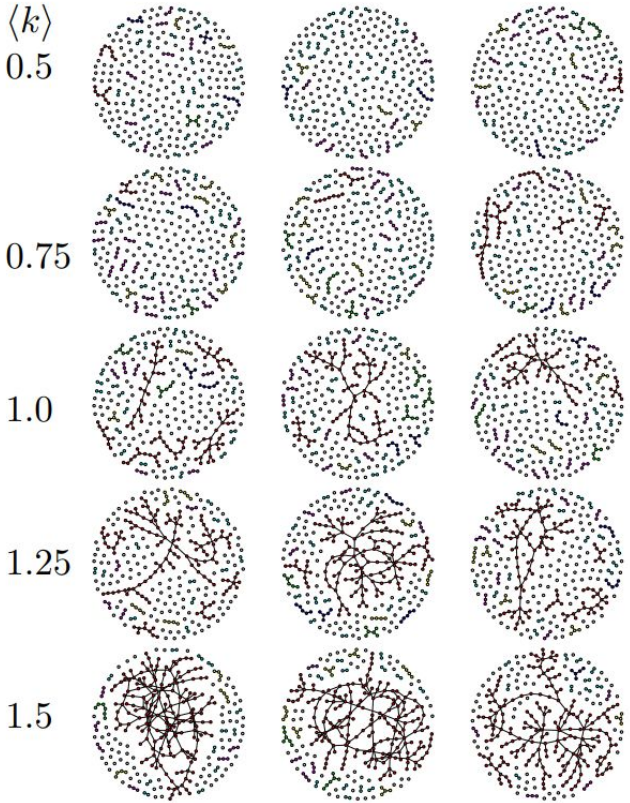
Connected Components

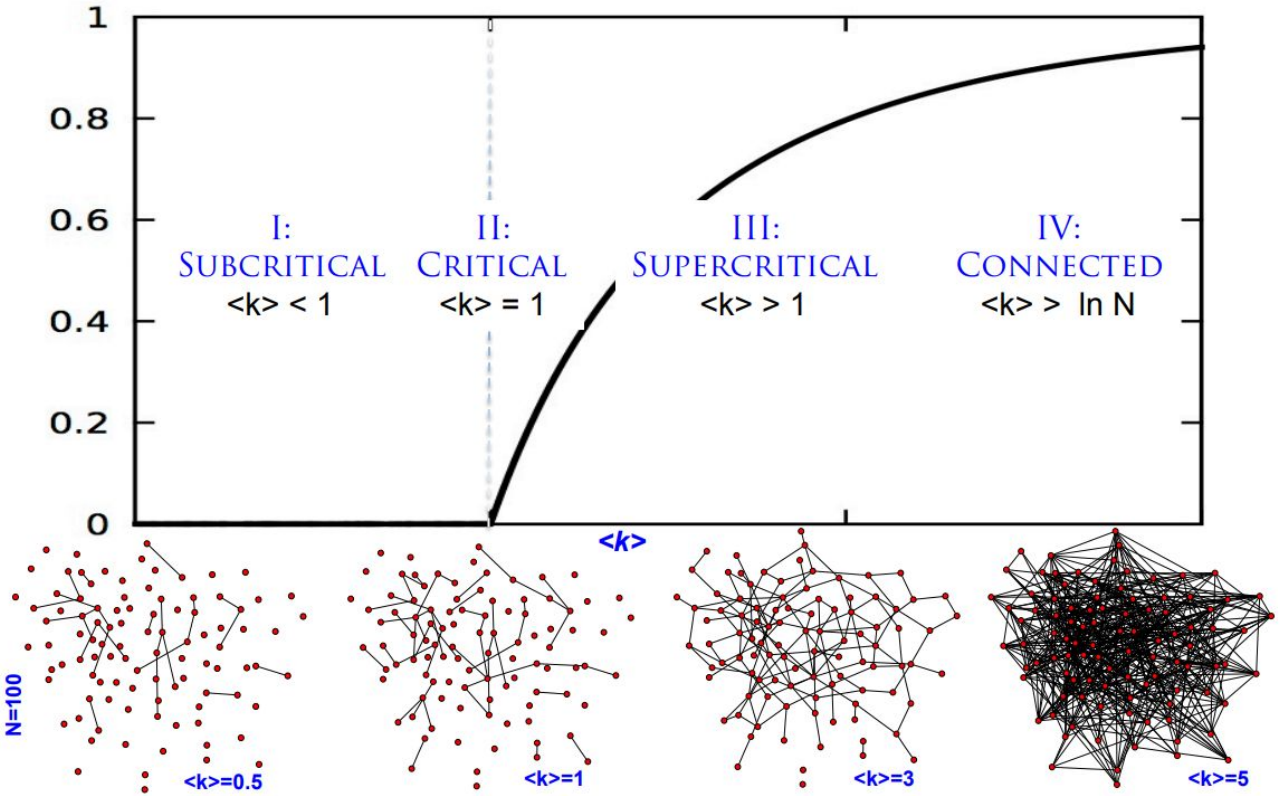
Random Graphs

Connected Components

Network structure goes through a transition.
How and when does this transition happen?







Structural (percolation) phase transition at <k> = 1 (or equivalently when p=1/N)

Network Regimes

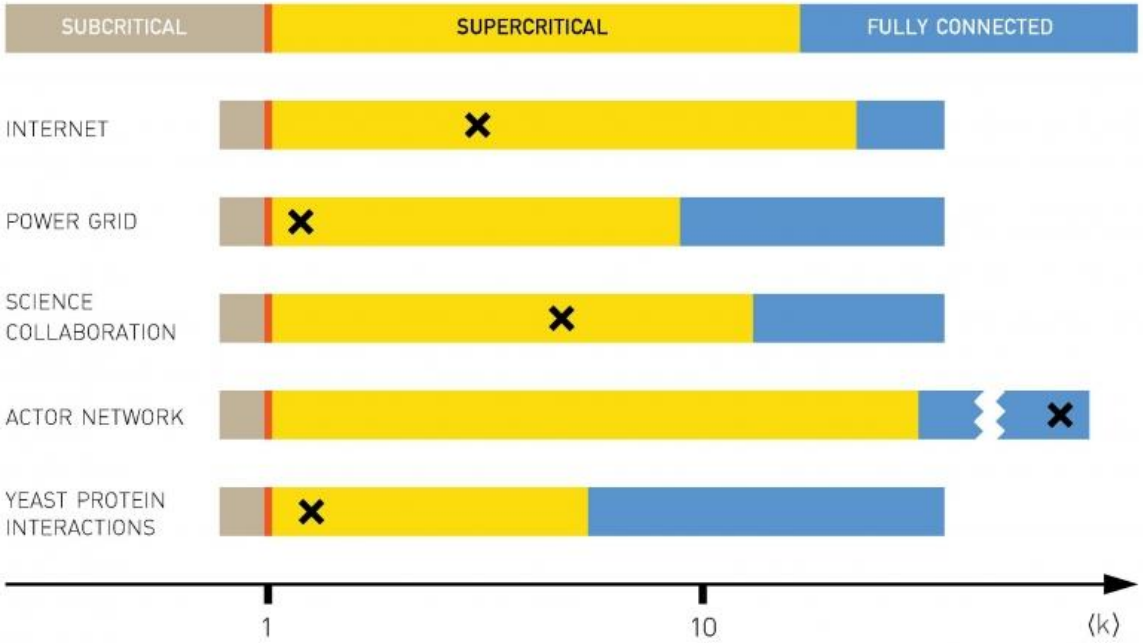
Subcritical ($\langle k \rangle < 1$, $p < p_c = 1/N$)
- No giant component;
- N-L isolated clusters, cluster size distribution is exponential;
- The largest cluster is a tree, its size ~ ln N.

Supercritical ($\langle k \rangle > 1$, $p > p_c = 1/N$)
- Unique giant component: $NG \sim (p-p_c)N$;
- GC has loops;
- Cluster size distribution: exponential.

Critical ($\langle k \rangle = 1$, $p = p_c = 1/N$)
- Unique giant component: $NG \sim N^{2/3}$
- Contains a vanishing fraction of all nodes, $NG/N \sim N^{-1/3}$
- Small components are trees, GC has loops.

Connected ($\langle k \rangle > \ln N$, $p > (\ln N)/N$)
- Only one cluster: NG=N;
- GC is dense;
- Cluster size distribution: None.

Visual Simulation
http://www.networkpages.nl/CustomMedia/Animations/RandomGraph/ERRG/AddoneEdgepATime.html



Network	N	L	‹k›	‹d›	d_{max}	ln*N*/ln‹*k*›
Internet	192,244	609,066	6.34	6.98	26	6.58
WWW	325,729	1,497,134	4.60	11.27	93	8.31
Power Grid	4,941	6,594	2.67	18.99	46	8.66
Mobile-Phone Calls	36,595	91,826	2.51	11.72	39	11.42

Real Network are Supercritical

Summarizing...

Random Networks
in a Nutshell

Degree Distribution
(Poisson for large N)

$$p_k = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle}$$

Clustering
(vanishing for large size)

$$C_i = \frac{< k >}{n-1} = p$$

Path length
(distance with logarithmic relation to nodes)

$$\mathcal{O}(\log n)$$

More on distances in **Chapter 4**!

Network	Degree Distribution	Path Length	Clustering Coefficient
Real-world networks	Broad	Short	Large
ER graphs	Poissonian	Short	Small

ER model is not capturing the properties of any real system but it serves as a reference system for any other network model

Chapter 2

Conclusion

Take Away Messages

1. ER model generates random graphs
2. In ER different values of *p* reflects different network regimes
3. Configuration models allow the generation of random graphs having heterogeneous degree distributions

Suggested Readings

- Chapter 3 of Barabasi's book

What's Next

Chapter 3:
It's a Small World!

Chapter 3

It's a Small World

Summary

- Six degrees of separation
- Watts-Strogatz model

Reading

- Chapter 20 of Kleinberg's book.

History of

Six Degrees

Karinthy, Frigyes

1929:
Minden másképpen van (Everything is Different)
Láncszemek (Chains)

"Look, Selma Lagerlöf just won the Nobel Prize for Literature, thus she is bound to know King Gustav of Sweden, after all he is the one who handed her the Prize, as required by tradition. King Gustav, to be sure, is a passionate tennis player, who always participates in international tournaments. He is known to have played Mr. Kehrling, whom he must therefore know for sure, and as it happens I myself know Mr. Kehrling quite well."

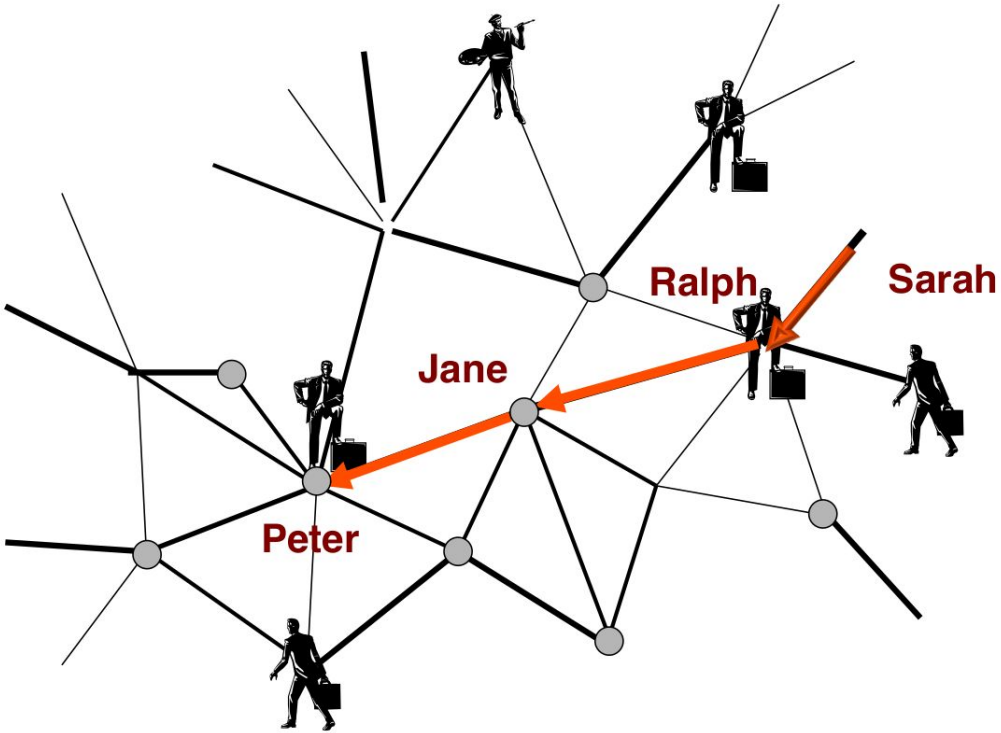
History of

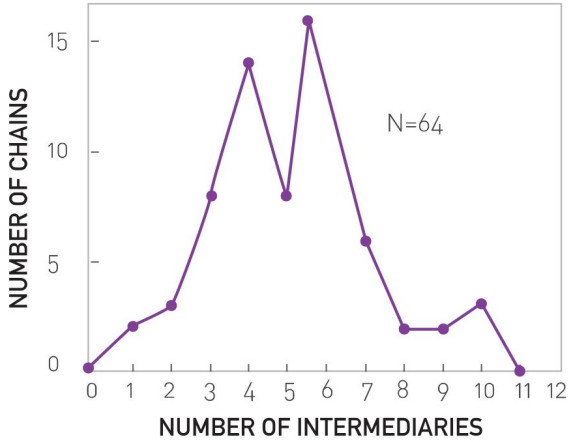
Six Degrees

1967: ***Stanley Milgram***

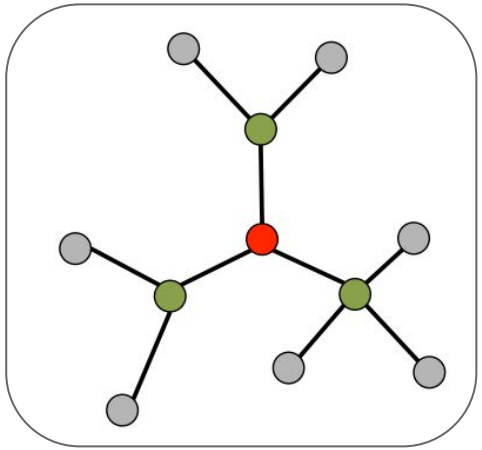
HOW TO TAKE PART IN THIS STUDY

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 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 (POST CARDS AND ALL) TO A PERSONAL ACQUAINTANCE WHO IS MORE LIKELY THAN YOU TO KNOW THE TARGET PERSON. You may send the folder to a friend, relative or acquaintance, but it must be someone you know on a first name basis.*





Milgram Experiment



nr. of nodes at distance one (d=1)	$N(u)_1 =< k >$
nr. of nodes at distance two (d=2)	$N(u)_2 =< k >^2$
nr. of nodes at distance d (d=d)	$N(u)_d =< k >^d$

$$N = 1 + \langle k \rangle + \langle k \rangle^2 + \ldots + \langle k \rangle^{d_{\max}} = \frac{\langle k \rangle^{d_{\max}+1} - 1}{\langle k \rangle - 1} \approx \langle k \rangle^{d_{\max}}$$

$$d_{\max} = \frac{\log N}{\log \langle k \rangle}$$

$$< d >= \frac{\log N}{\log \langle k \rangle}$$

We will call the small world phenomena the property that the average path length or the diameter depends logarithmically on the system size.

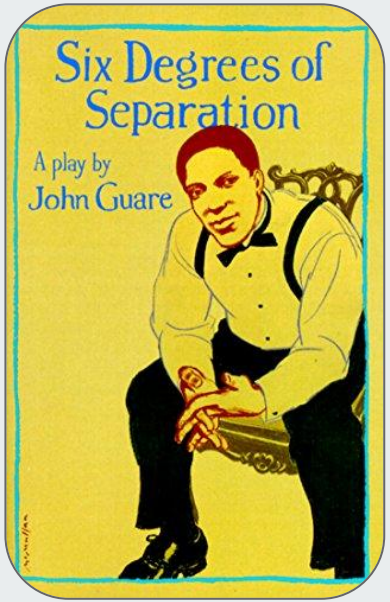
”Small” means that ⟨d⟩ is proportional to log N, rather than N.

The 1/log⟨k⟩ term implies that denser the network, the smaller will be the distance between the nodes.

Small World Phenomena

History of

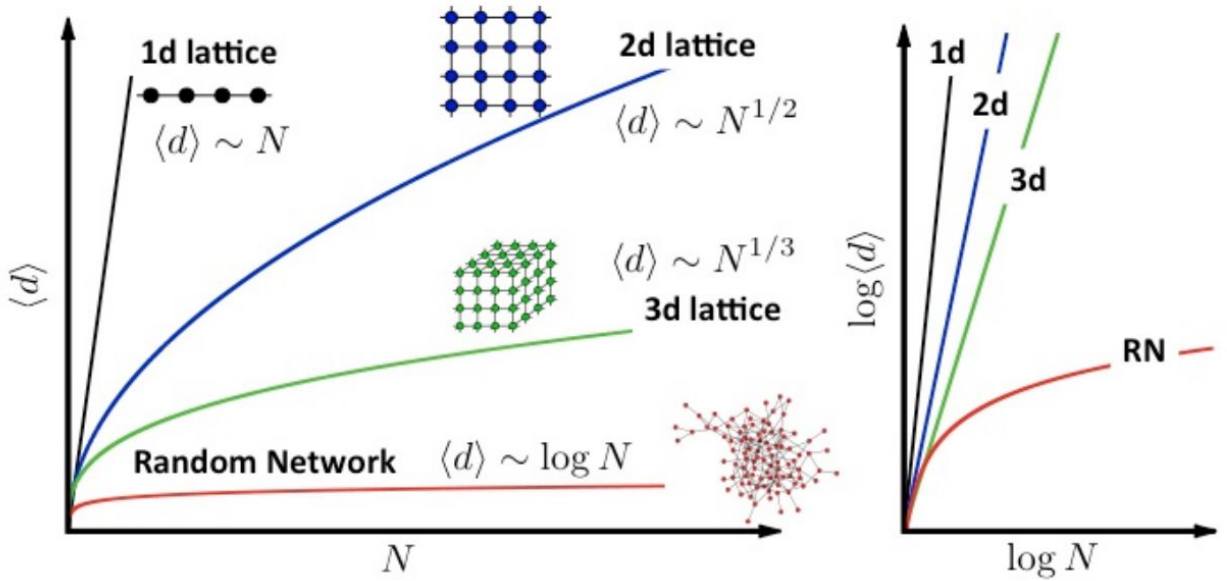
Six Degrees



"Everybody on this planet is separated by only six other people.

Six degrees of separation.
Between us and everybody else on this planet.
The president of the United States.
A gondolier in Venice.... It's not just the big names.
It's anyone.
A native in a rain forest.
A Tierra del Fuegan.
An Eskimo.
I am bound to everyone on this planet by a trail of six people.

It's a profound thought.
How every person is a new door, opening up into other worlds."



Why are small worlds surprising? Surprising compared to what?!

Watts-Strogatz Model

A model for the Small-World phenomena

One of the first paper on network science…

Real world network observations lead to a contradiction w.r.t. ER graphs:

- High clustering coefficient and
- Short distances

Duncan Watts

Steve Strogatz

NATURE | VOL 393 | 4 JUNE 1998

Collective dynamics of 'small-world' networks

Duncan J. Watts* & Steven H. Strogatz

Department of Theoretical and Applied Mechanics, Kimball Hall, Cornell University, Ithaca, New York 14853, USA

Networks of coupled dynamical systems have been used to model biological oscillators[1–4], Josephson junction arrays[5,6], excitable media[7], neural networks[8–10], spatial games[11], genetic control networks[12] and many other self-organizing systems. Ordinarily, the connection topology is assumed to be either completely regular or completely random. But many biological, technological and social networks lie somewhere between these two extremes.

Table 1 Empirical examples of small-world networks

	L_{actual}	L_{random}	C_{actual}	C_{random}	N
Film actors	3.65	2.99	0.79	0.00027	22500
Power grid	18.7	12.4	0.080	0.005	4941
C. elegans	2.65	2.25	0.28	0.05	282

Clustering vs. Interconnectedness

Random networks:

- Logarithmically short distance among nodes

$$d = \frac{\log N}{\log\langle k\rangle}$$

- Vanishing clustering coefficient for large size

$$C_i \equiv \frac{1}{N}\langle k\rangle = p$$

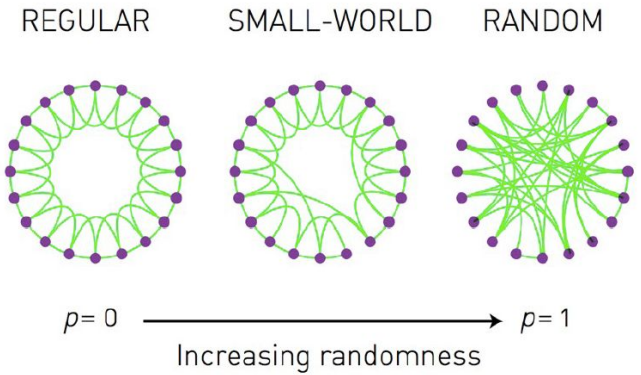
Real Networks:

High Clustering & Short Distances

Network	Size	$\langle k\rangle$	ℓ	ℓ_{rand}	C	C_{rand}
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

From Regular Lattices to Random Networks

A model to capture
large clustering coefficient and
short distances observed in real networks.

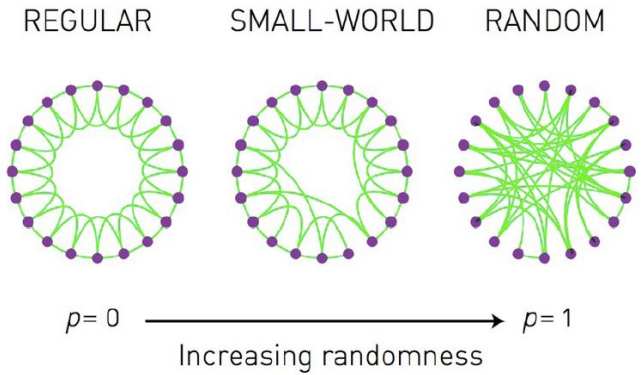


Fixed parameters:
- n - system size
- K - initial coordination number

Variable parameters:
- p - rewiring probability

From Regular Lattices to Random Networks

A model to capture
large clustering coefficient and
short distances observed in real networks.

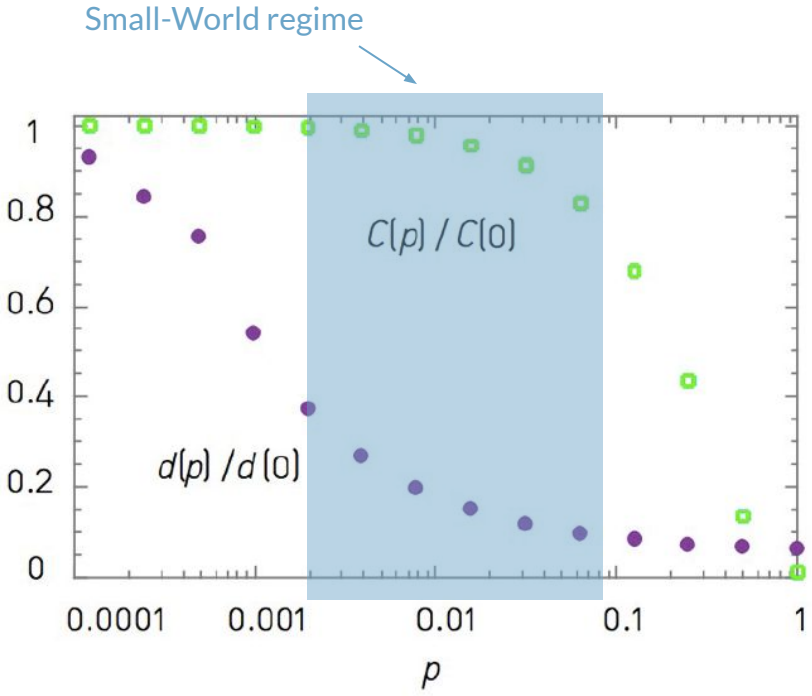


Algorithm:

1. Start with a ring lattice with n nodes in which every node is connected to its first K neighbors (K/2 on either side)
2. Randomly rewire each edge of the lattice with probability p such that self-connections are excluded.

From Regular Lattices to Random Networks

By varying p the network can be transformed from a completely ordered (p=0) to a completely random (p=1) structure



n and K are chosen:

$$n >> K >> ln(n) >> 1$$

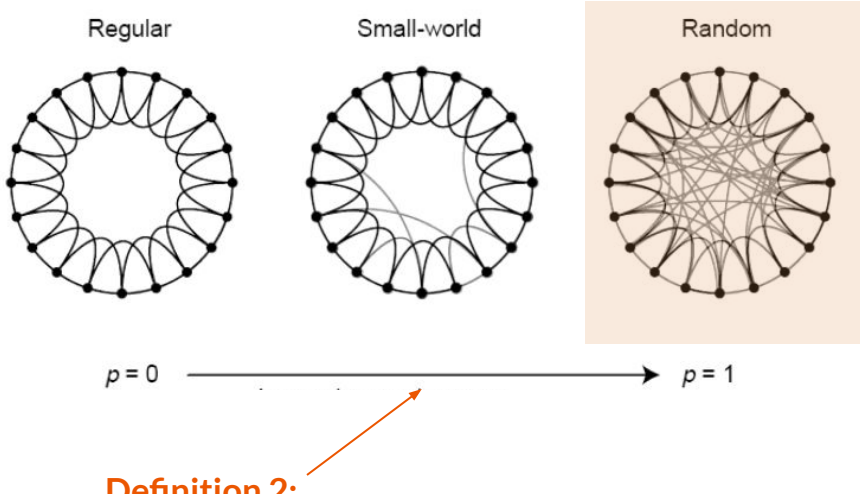
thus the random graph remains connected

$$K >> ln(n)$$

Measuring Watts-Strogatz Graphs

Watts-Strogatz model

Alternative definitions



Definition 1:

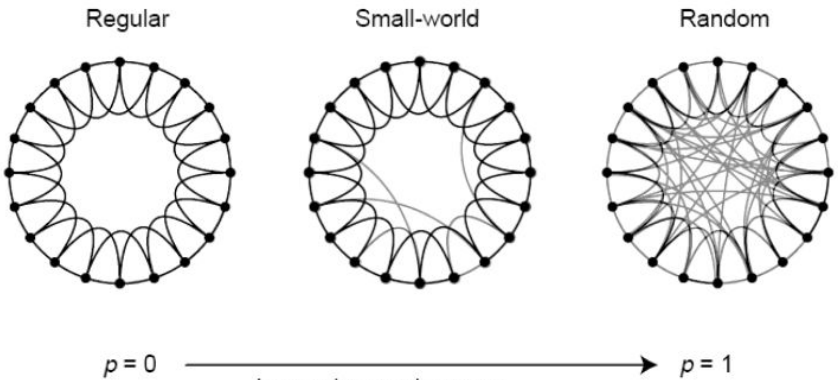
1. Start with a ring lattice with N nodes in which every node is connected to its first K neighbours (K/2 on either side).
2. Randomly rewire each edge of the lattice with probability p such that self-connections and duplicate edges are excluded.

Definition 2:

1. Start with a ring lattice with N nodes in which every node is connected to its first K neighbours (K/2 on either side).
2. For every edge in the network add an additional edge with independent probability p, connecting two nodes selected uniformly at random .

Definition 2

Global Clustering Coefficient



p=0

regular ring with constant clustering:

$$C = \frac{3(K-2)}{4(K-1)}$$

- $0 \leqq C \leqq 3/4$
- independent of n

p>0

we can count triangles and tuples

$$C = \frac{\frac{1}{4}NK\left(\frac{1}{2}K-1\right)\times 3}{\frac{1}{2}NK(K-1)+NK^2p+\frac{1}{2}NK^2p^2} = \frac{3(K-2)}{4(K-1)+8Kp+4Kp^2}$$

- Independent of n
- if $p \to 0$ it recovers the ring value
- if $p \to 1$ it well approximates 1

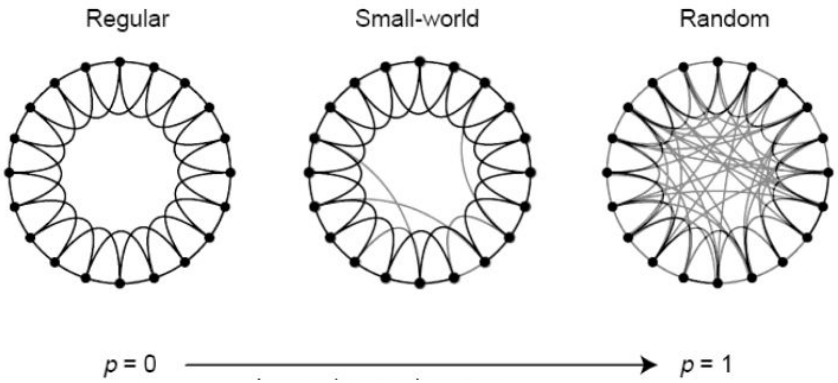
Definition 2

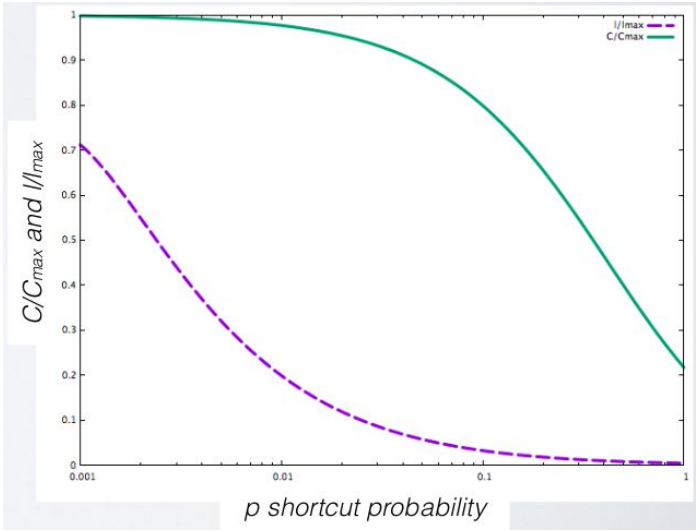
Average path length

No closed form solution

From numerical simulations we can approximate it as:

$$l = \frac{\ln(nKp)}{K^2 p}$$



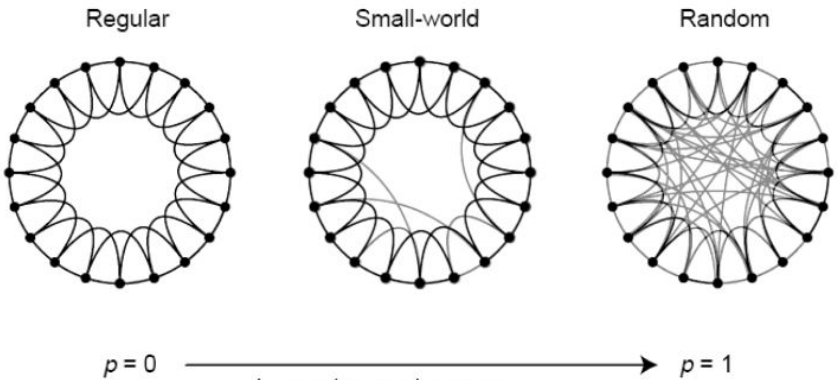


Definition 2

Degree Distribution

p=0
each node has the same degree K
(Dirac delta function)

p>0
approximates a Poisson distribution just like a random network



p>0

- each node has degree K + shortcut links
- Number of shortcut edges:

$$s = \frac{1}{2}NK \times p$$

- Each node will have on average Kp number of shortcuts
- The degree distribution is

$$P(k) = e^{-Kp}\frac{(Kp)^{(k-K)}}{(k-K)!}$$

if k≥K and P(k)=0 if k<K

Summarizing...

W-S Networks
in a Nutshell

Degree Distribution	$e^{-Kp}\frac{(Kp)^{(k-K)}}{(k-K)!}$
Clustering	$\frac{3(K-2)}{4(K-1)+8Kp+4Kp^2}$
Path length	$\frac{\ln(nKp)}{K^2p}$

Network	Degree Distribution	Path Length	Clustering Coefficient
Real-world networks	Broad	Short	Large
ER graphs	Poissonian	Short	Small
Configuration model	Custom, can be broad	Short	Small
Watts & Strogatz (in SW regime)	Poissonian	Short	Large

Chapter 3

Conclusion

Take Away Messages

1. Small diameters and high clustering coefficient deeply characterize social networks topologies

Suggested Readings

- Chapter 20 of Kleinberg's book

What's Next

Chapter 4:
Scale Free Networks

Chapter 4

Scale Free Networks

Summary

- Scale Free Networks
- Power Law degree distribution
- Barabasi-Albert model

Reading

- Chapters 4 & 5 of Barabasi's book.

Example

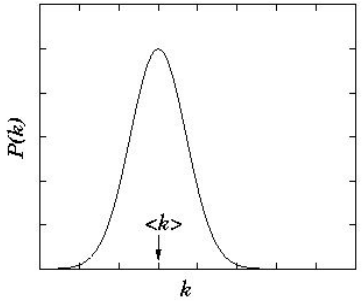
World Wide Web

Nodes: WWW documents
Links: URL links

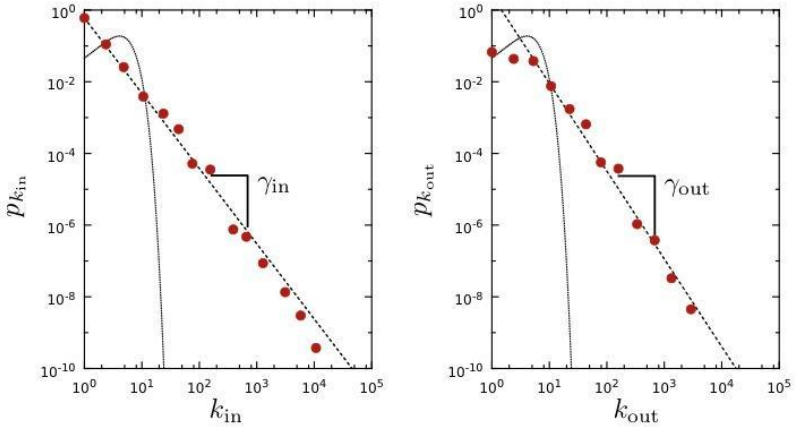
Over 3 billion documents

Data Collection:
web crawler collected all URL's found in a document and followed them recursively

R. Albert, H. Jeong, A-L Barabasi, Nature, 401 130 (1999).



Expected



Observed

Scale-Free

A network is called Scale-free when its degree distribution follows *(to some extent)* a Power-Law distribution:

$$P(k) \sim Ck^{-\gamma} = C\frac{1}{k^{\gamma}}$$

with γ called the exponent of the distribution

Discrete Formalism:

As node degrees are always positive integers, the discrete formalism captures the probability that a node has exactly k links:

$$p_k = Ck^{-\gamma}. \quad \sum_{k=1}^{\infty} p_k = 1$$

$$C\sum_{k=1}^{\infty} k^{-\gamma} = 1$$

$$C = \frac{1}{\sum_{k=1}^{\infty} k^{-\gamma}} = \frac{1}{\zeta(\gamma)}, \qquad p_k = \frac{k^{-\gamma}}{\zeta(\gamma)}$$

Interpretation p_k

Continuum Formalism:

In analytical calculations it is often convenient to assume that the degrees can take up any positive real value:

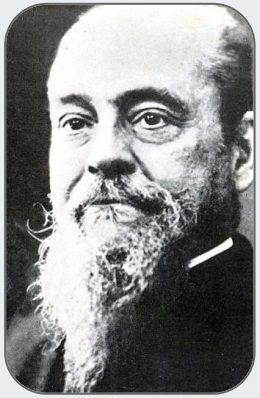
$$p(k) = Ck^{-\gamma} \qquad \int_{k_{\min}}^{\infty} p(k)dk = 1$$

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

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

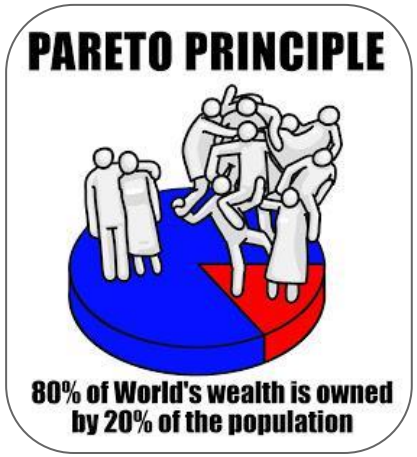
Interpretation $\int_{k_1}^{k_2} p(k)dk$

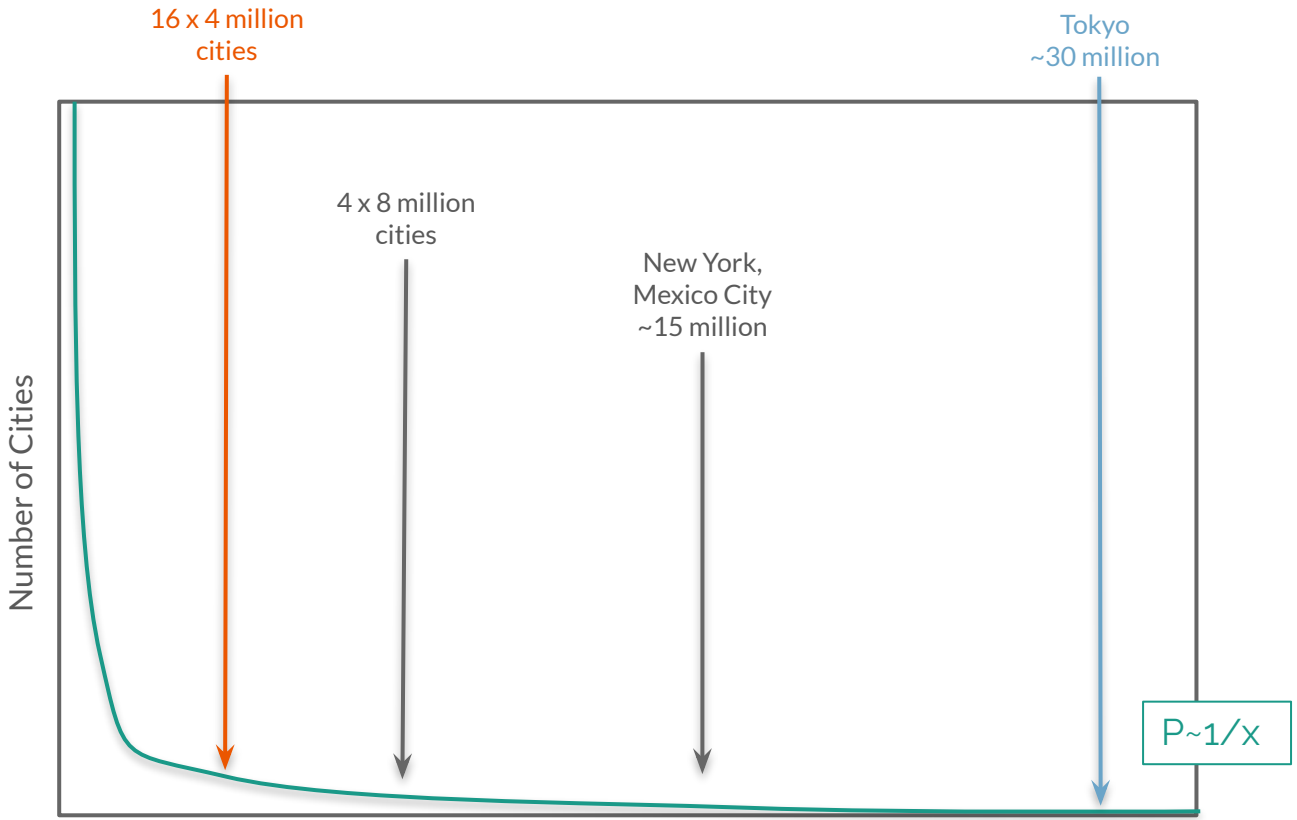
80/20 Rule



Vilfredo Federico Damaso Pareto (1848 – 1923)

Italian economist, political scientist and philosopher, who had important contributions to our understanding of income distribution and to the analysis of individuals choices.
A number of fundamental principles are named after him, like Pareto efficiency, Pareto distribution *(another name for a power-law distribution)*, the Pareto principle (or 80/20 law).





Sizes of Cities:
there is an equivalent number of people living in cities of all sizes!

Power-Law

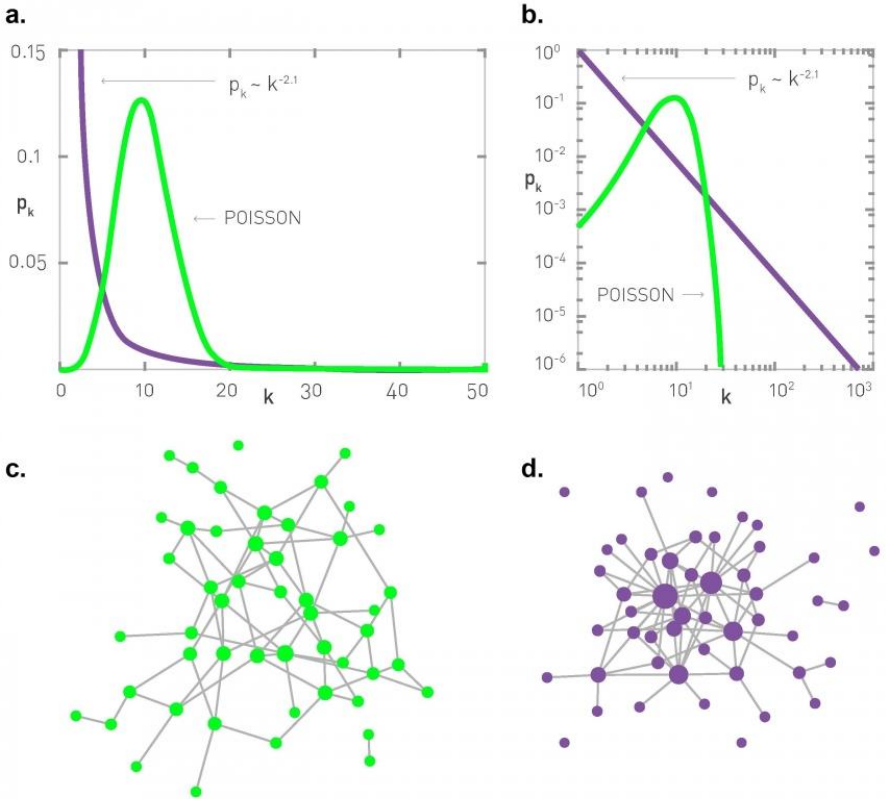
Hubs

The main difference between a random and a scale-free network comes in the *tail* of the degree distribution, representing the high-*k* region of p_k

For small *k* the power law is above the Poisson function, indicating that a scale-free network has a large number of small degree nodes, most of which are absent in a random network.

For *k* in the vicinity of <*k*> the Poisson distribution is above the power law, indicating that in a random network there is an excess of nodes with degree $k \approx \langle k \rangle$

For large k the power law is above the Poisson curve, indicating that the probability of observing a high-degree node, or *hub*, is several orders of magnitude higher in a scale-free than in a random network



Example

Hubs

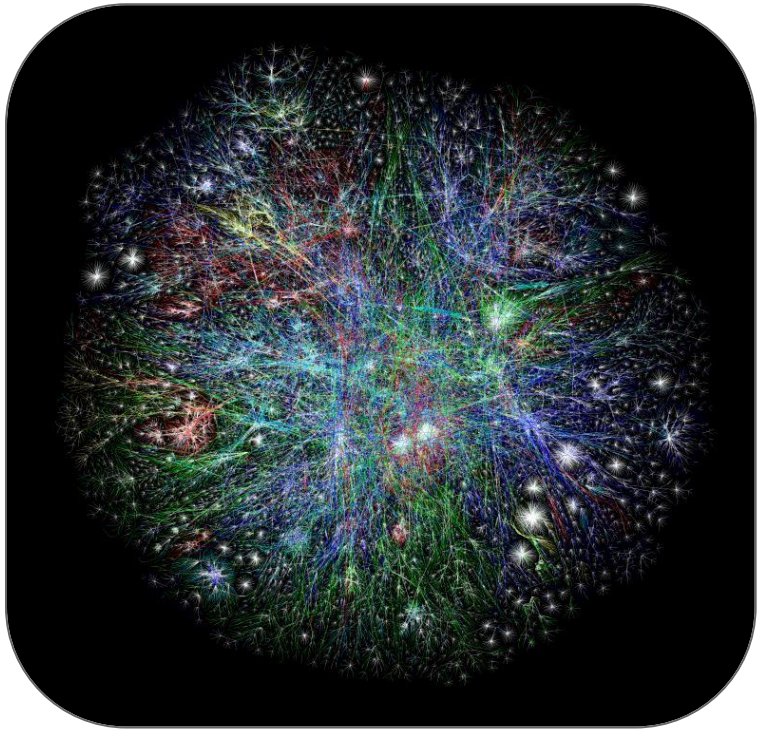
Let us use the WWW to illustrate the properties of the high-k regime.

The probability to have a node with k~100 is

- About $p_{100} \simeq 10^{-30}$ in a Poisson distribution
- About $p_{100} \simeq 10^{-4}$ if p_k follows a power law

Consequently, if the WWW were to be a random network, according to the Poisson prediction we would expect 10^{-18} k>100 degree nodes, or none.

For a power law degree distribution, we expect about $N_{k>100} = 10^{8}$ k>100 degree nodes

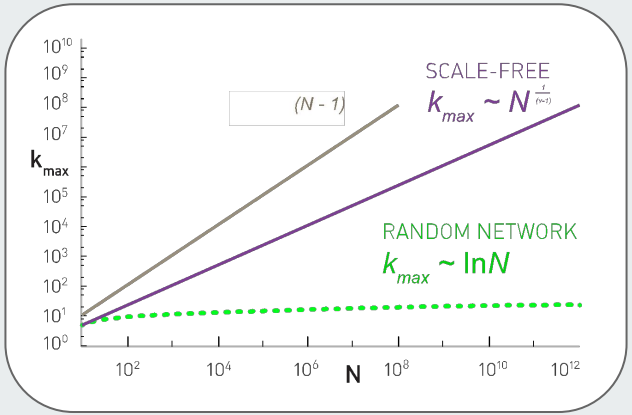


The biggest Hub

All real networks are finite

↓

We have an expected maximum degree, k_{max}



Estimating k_{max}

$$\int_{k_{\max}}^{\infty} P(k)\,dk \approx \frac{1}{N}$$

the probability to have a node larger than k_{max} should not exceed the prob. to have one node, i.e. *1/N* fraction of all nodes

$$\int_{k_{\max}}^{\infty} P(k)\,dk = (\gamma-1)k_{\min}^{\gamma-1}\int_{k_{\max}}^{\infty} k^{-\gamma}\,dk$$

$$= \frac{(\gamma-1)}{(-\gamma+1)}k_{\min}^{\gamma-1}\left[k^{-\gamma+1}\right]_{k_{\max}}^{\infty} = \frac{k_{\min}^{\gamma-1}}{k_{\max}^{\gamma-1}} \approx \frac{1}{N}$$

therefore,

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

Ultra Small World

$$< l > \sim \begin{cases} const. & \gamma = 2 \\ \dfrac{\ln\ln N}{\ln(\gamma - 1)} & 2 < \gamma < 3 \\ \dfrac{\ln N}{\ln\ln N} & \gamma = 3 \end{cases}$$

Size of the biggest hub is of order O(N).
Most nodes can be connected within two layers of it, thus the average path length will be independent of the system size.

The average path length increases slower than logarithmically.
In a random network all nodes have comparable degree, thus most paths will have comparable length. In a scale-free network the vast majority of the path go through the few high degree hubs, reducing the distances between nodes.

Some key models produce γ=3, so the result is of particular importance for them.
This was first derived by Bollobas and collaborators for the network diameter in the context of a dynamical model, but it holds for the average path length as well.

Small World

$$\ln N \qquad \gamma > 3$$

The second moment of the distribution is finite, thus in many ways the network behaves as a random network. Hence the average path length follows the result that we derived for the random network model earlier.

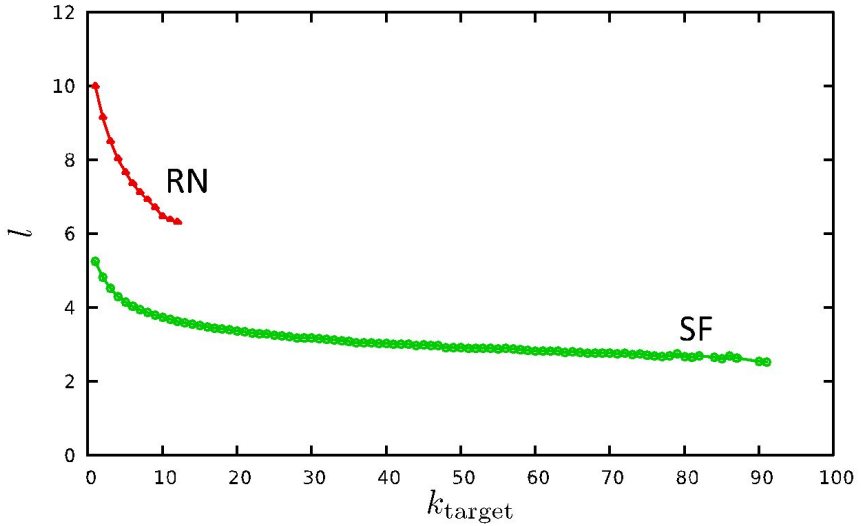
Small-World in Scale-Free networks

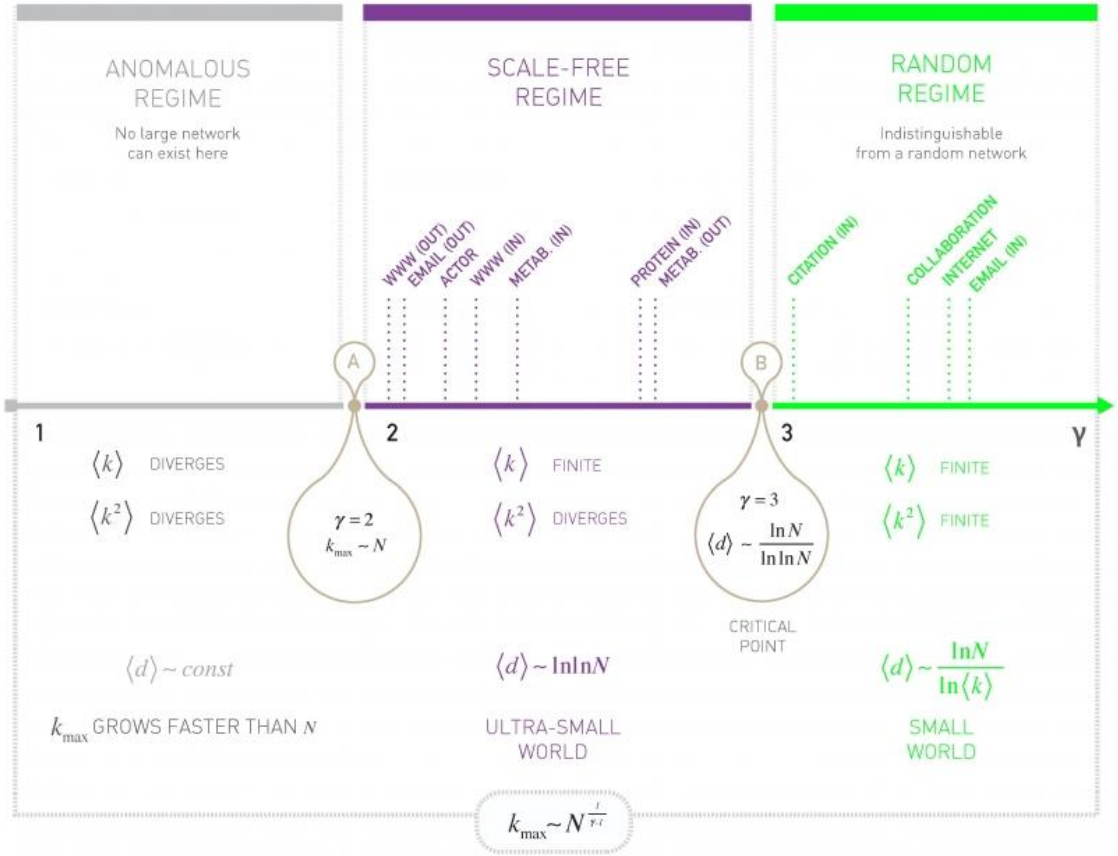
Bollobas, (1985), Newman (2001), Dorogovtsev et al (2002), Chung and Lu (2002), Bollobas (2002), Cohen (2003)

We are always close to the Hubs

"It's always easier to find someone who knows a famous or popular figure than some run-the-mill, insignificant person."

(Frigyes Karinthy, 1929)





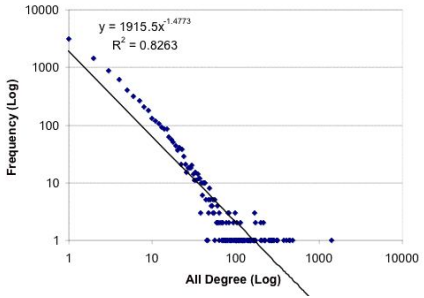
Behavior of Scale-Free networks

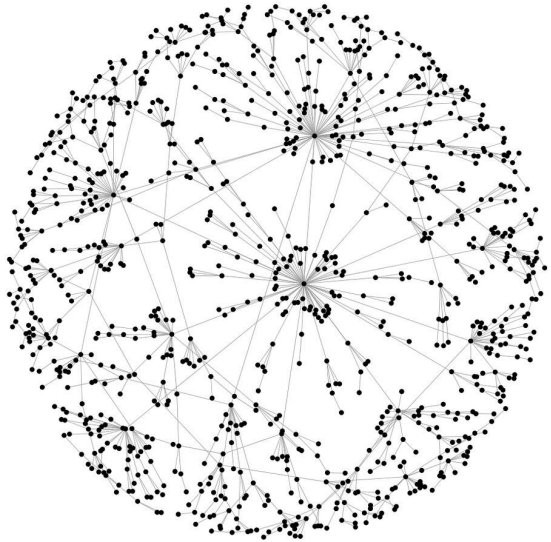
The Barabási-Albert model

Modeling Scale-Free Networks

1. Why does the random network model of Erdős and Rényi fail to reproduce the hubs and the power laws observed in many real networks?
2. Why do so different systems as the WWW or the cell converge to a similar scale-free architecture?

Hubs represent the most striking difference between a random and a scale-free network.



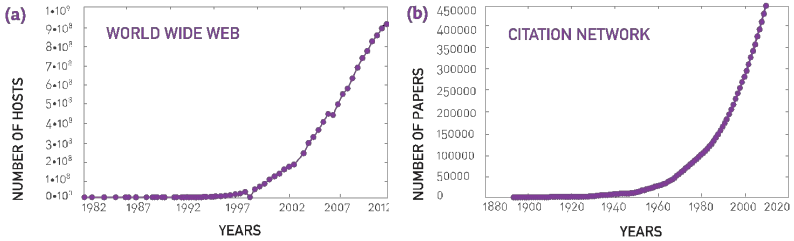


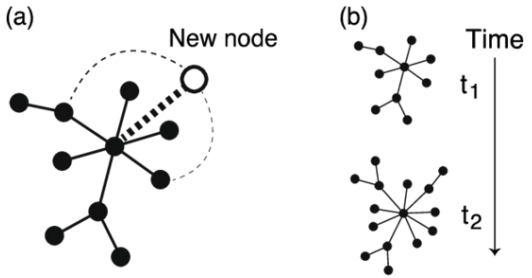
Growth and Preferential Attachment

The random network model differs from real networks in two important characteristics:

Growth:
While the random network model assumes that the number of nodes is fixed (time invariant), real networks are the result of a growth process that continuously increases.

Preferential Attachment:
While nodes in random networks randomly choose their interaction partner, in real networks new nodes prefer to link to the more connected nodes.





BA model

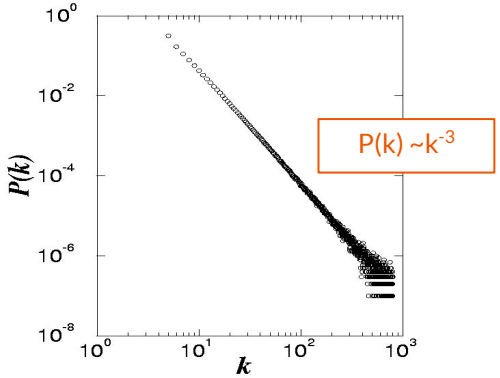
1. Networks continuously expand by the addition of new nodes
   WWW : addition of new documents
2. New nodes prefer to link to highly connected nodes.
   WWW : linking to well known sites

*Barabási & Albert, Science **286,** 509 (1999)*

1. Start with m_0 connected nodes
2. At each timestep add a new node with m links that connect it to nodes already in the network
3. The probability $\Pi(k)$ that on of the links connects to node i depends on the degree k_i of i

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

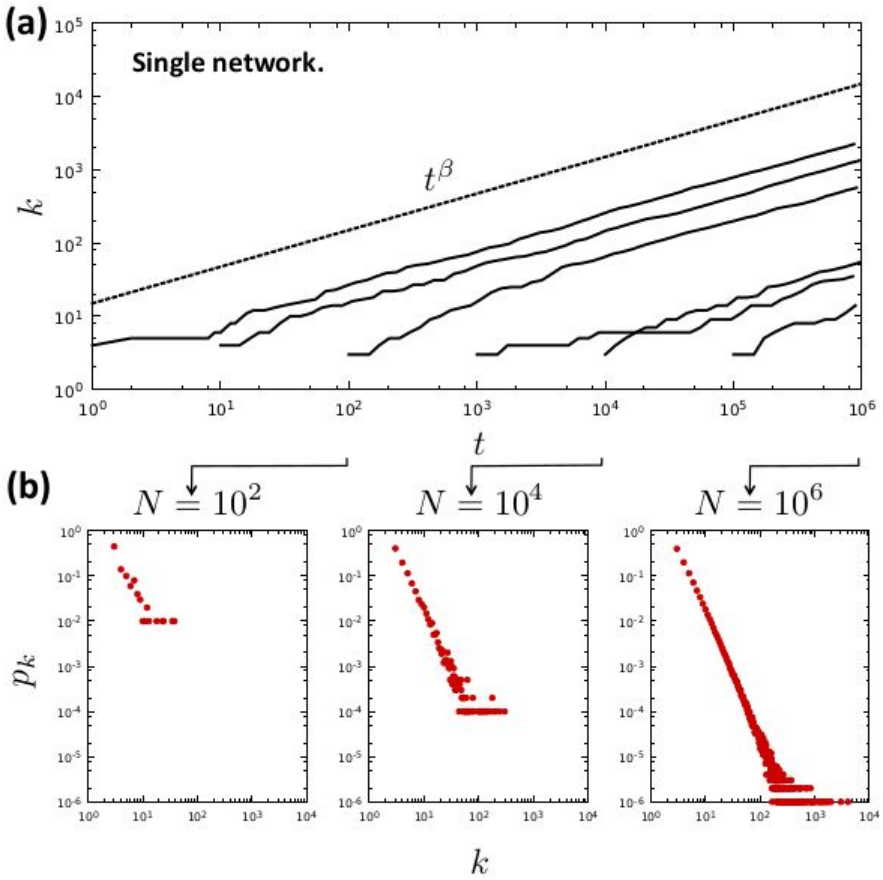
The emerging network will be scale-free with degree exponent $\gamma=3$ independently from the choice of m



BA model

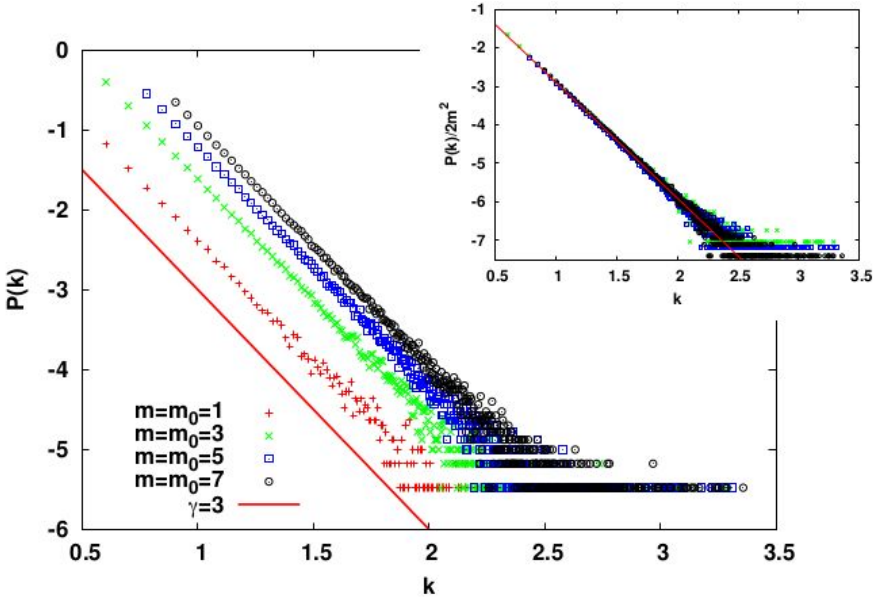
- The degree of each node increases as power-law with exponent ½
- The earlier a node was added the larger its degree
  (due to its arrival time, not because of faster growth)

Barabási & Albert, Science ***286,*** *509 (1999)*



BA model

- The degree exponent is independent of m
- The degree exponent is stationary in time and the degree distribution is time independent
- The exponent is compatible to the exponents of real networks



*Barabási & Albert, Science **286,** 509 (1999)*

Ultra Small World

$$< l > \sim \begin{cases} const. & \gamma = 2 \\ \dfrac{\ln \ln N}{\ln(\gamma - 1)} & 2 < \gamma < 3 \\ \dfrac{\ln N}{\ln \ln N} & \gamma = 3 \end{cases}$$

Small World

$$\ln N \qquad \gamma > 3$$

Size of the biggest hub is of order O(N).
Most nodes can be connected within two layers of it, thus the average path length will be independent of the system size.

The average path length increases slower than logarithmically.
In a random network all nodes have comparable degree, thus most paths will have comparable length. In a scale-free network the vast majority of the path go through the few high degree hubs, reducing the distances between nodes.

Some key models produce γ=3, so the result is of particular importance for them.
This was first derived by Bollobas and collaborators for the network diameter in the context of a dynamical model, but it holds for the average path length as well.

The second moment of the distribution is finite, thus in many ways the network behaves as a random network. Hence the average path length follows the result that we derived for the random network model earlier.

BA model: Path Length

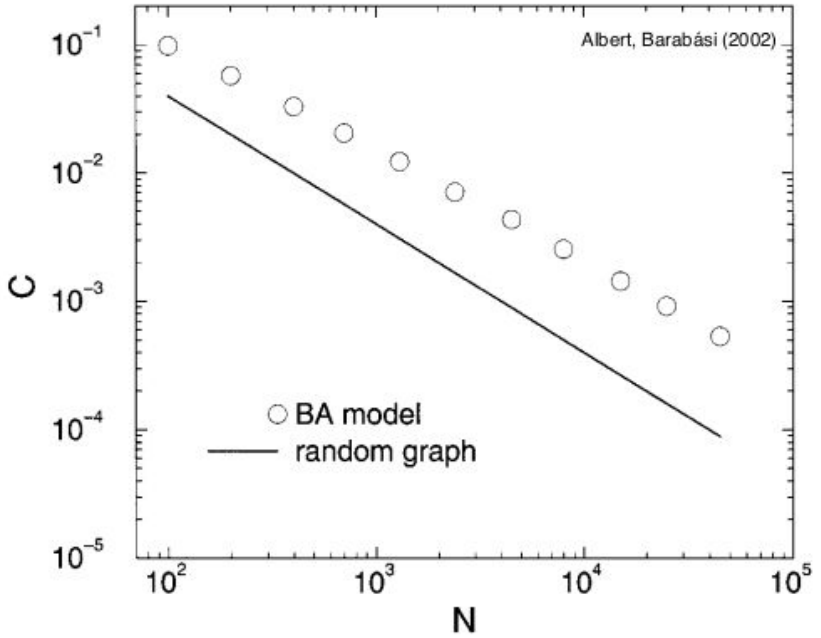
Bollobas, (1985), Newman (2001), Dorogovtsev et al (2002), Chung and Lu (2002), Bollobas (2002), Cohen (2003)

BA model

Clustering Coefficient

The clustering coeff. decreases with the system size as

$$C = \frac{m}{4}\frac{(\ln N)^2}{N}$$



Due to its definition the BA model induces non-trivial degree correlation

$$n_{kl} \simeq k^{-2} l^{-2}$$

Summarizing...

BA Networks
in a Nutshell

Number of nodes	$N = t$
Number of links	$N = mt$
Average degree	$\langle k \rangle = 2m$
Degree Distribution	$P(k) \sim Ck^{-\gamma}$
Clustering	$\frac{m}{4}\frac{(\ln N)^2}{N}$
Path length	$\frac{\ln N}{\ln \ln N}$

Network	Degree Distribution	Path Length	Clustering Coefficient
Real-world networks	Broad	Short	Large
ER graphs	Poissonian	Short	Small
Watts & Strogatz (in SW regime)	Poissonian	Short	Large
Barabasi Albert (Scale-Free)	Power-Law	Short	Rather Small

Chapter 4

Conclusion

Take Away Messages

1. Real world networks have heavy tailed degree distributions
2. Scale-Free networks
3. Ultra Small-world phenomena
4. BA models scale-free with $\gamma=3$
5. Additional models explains local behaviours, clustering coeff., ...

Suggested Readings

- Chapters 4 & 5 of Barabasi's book
- Chapter 18 of Kleinberg's book

What's Next

Lecture 3:
Micro, Meso & Macro:
Different perspectives