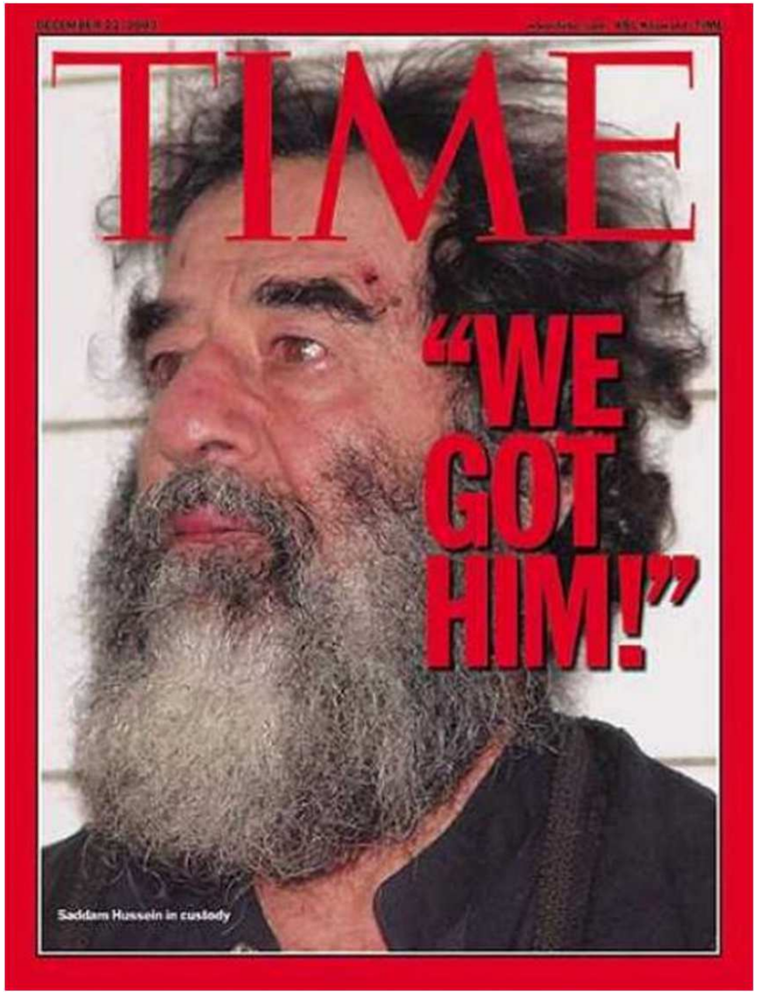
# Introduction to Network Science

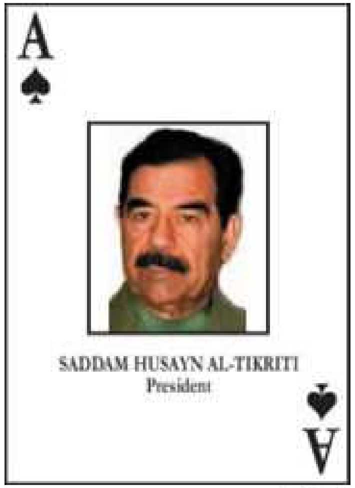
Riadh DHAOU based on the lecture of

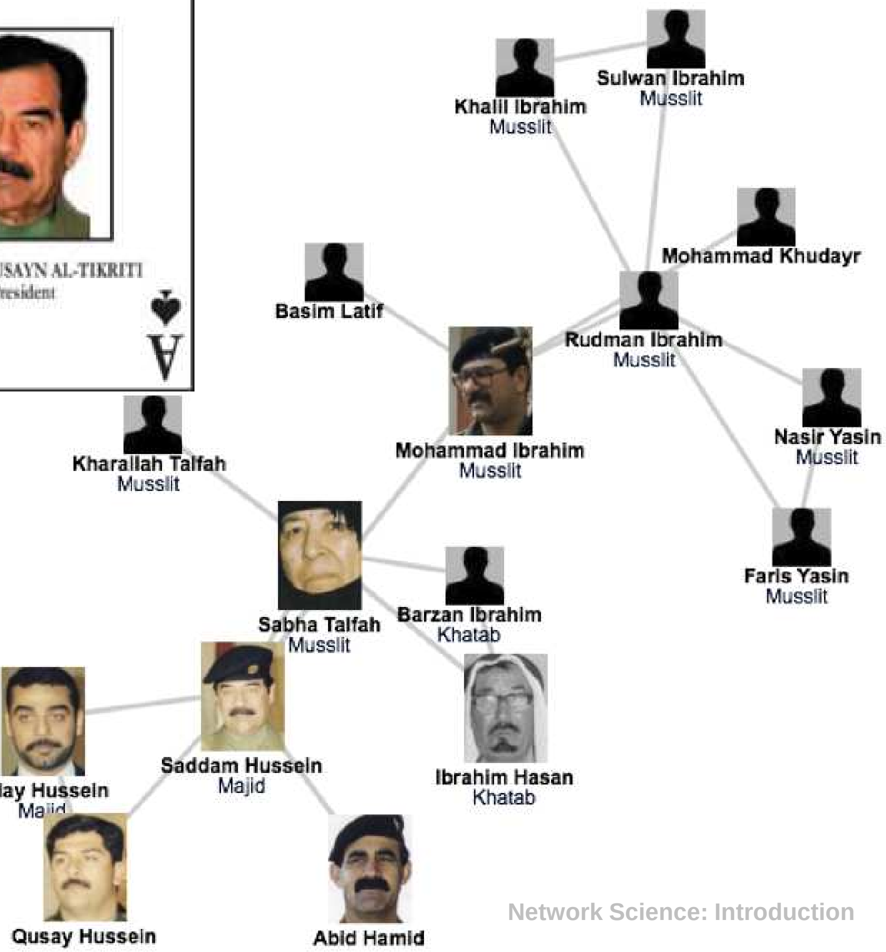
Albert-László Barabási

# FROM SADDAM HUSSEIN TO NETWORK THEORY

# A SIMPLE STORY (1) The fate of Saddam and network science



## A SIMPLE STORY (1) The fate of Saddam and network science

The capture of Saddam Hussein:

→shows the strong predictive power of networks.

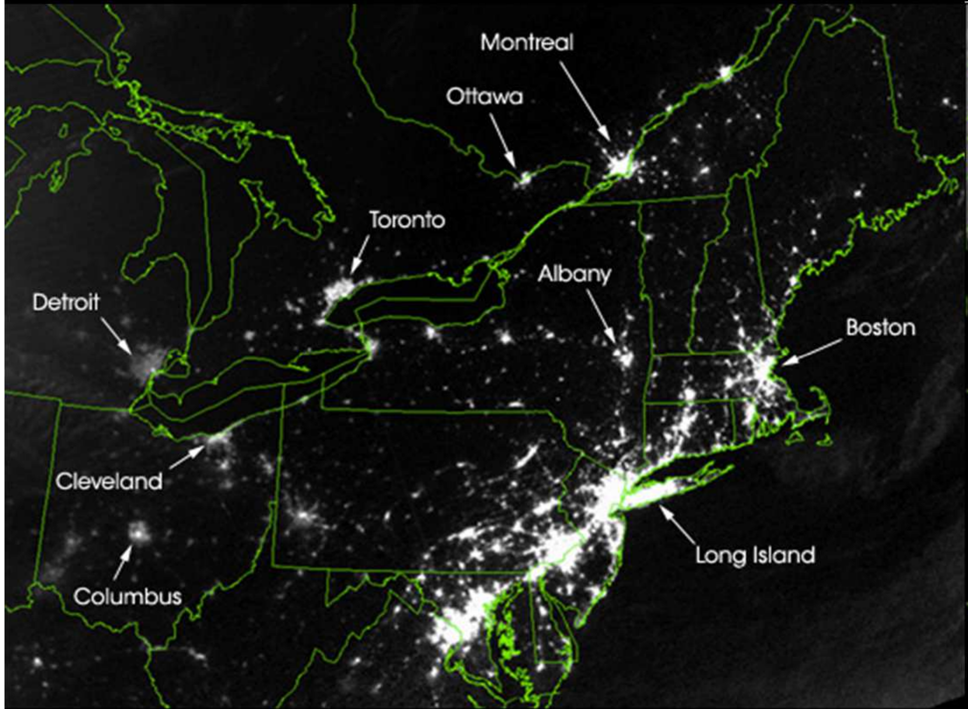
→ underlies the need to obtain accurate maps of the networks we aim to study; and the often heroic difficulties we encounter during the mapping process.

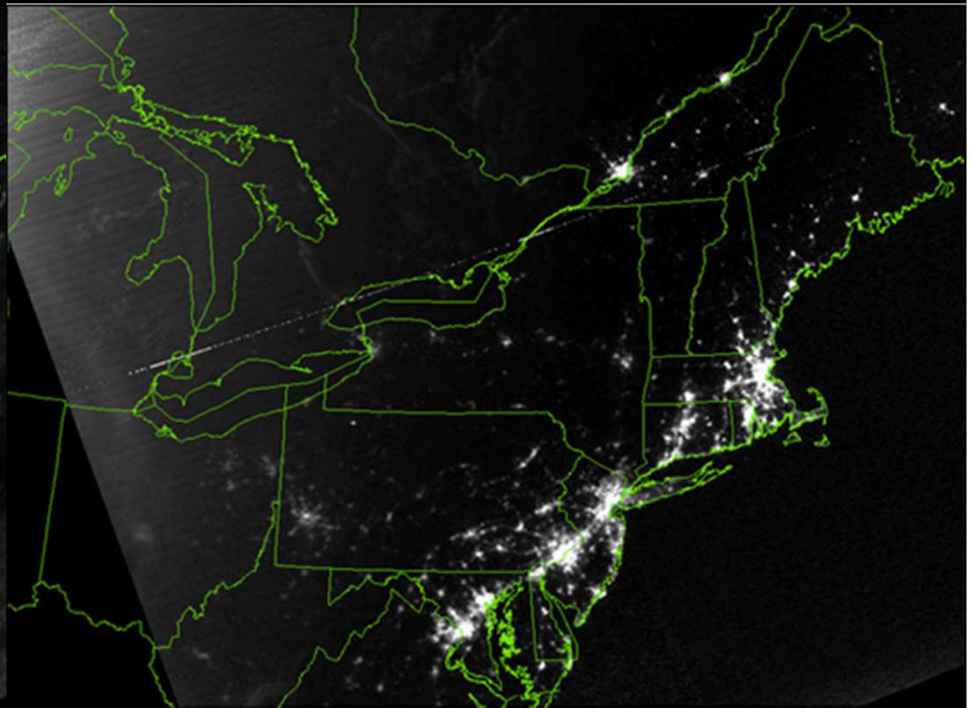
→ demonstrates the remarkable stability of these networks: The capture of Hussein was not based on fresh intelligence, but rather on his pre-invasion social links, unearthed from old photos stacked in his family album.

→ shows that the choice of network we focus on makes a huge difference: the hierarchical tree, that captured the official organization of the Iraqi government, was of no use when it came to Saddam Hussein's whereabouts.

# VULNERABILITY DUE TO INTERCONNECTIVITY

# A SIMPLE STORY (2): August 15, 2003 blackout.



August 14, 2003: 9:29pm EDT
20 hours before

August 15, 2003: 9:14pm EDT
7 hours after

# A SIMPLE STORY (2): August 15, 2003 blackout.

An important theme of this class:

→ we must understand how network structure affects the robustness of a complex system.

→ develop quantitative tools to assess the interplay between network structure and the dynamical processes on the networks, and their impact on failures.

→We will learn that failures reality failures follow reproducible laws, that can be quantified and even predicted using the tools of network science.

# NETWORKS AT THE HEART OF COMPLEX SYSTEMS

# COMPLEX SYSTEMS

## Complex

[adj., v. kuh m-pleks, kom-pleks; n. kom-pleks]
–adjective

1.
composed of many interconnected parts; compound; composite: a complex highway system.

2.
characterized by a very complicated or involved arrangement of parts, units, etc.: complex machinery.

3.
so complicated or intricate as to be hard to understand or deal with: a complex problem.

*Source: Dictionary.com*

Complexity, a **scientific theory** which asserts that some systems display behavioral phenomena that are completely inexplicable by any conventional analysis of the systems' constituent parts. These phenomena, commonly referred to as emergent behaviour, seem to occur in many complex systems involving living organisms, such as a stock market or the human brain.

*Source: John L. Casti, Encyclopædia Britannica*

## Complexity

# THE ROLE OF NETWORKS

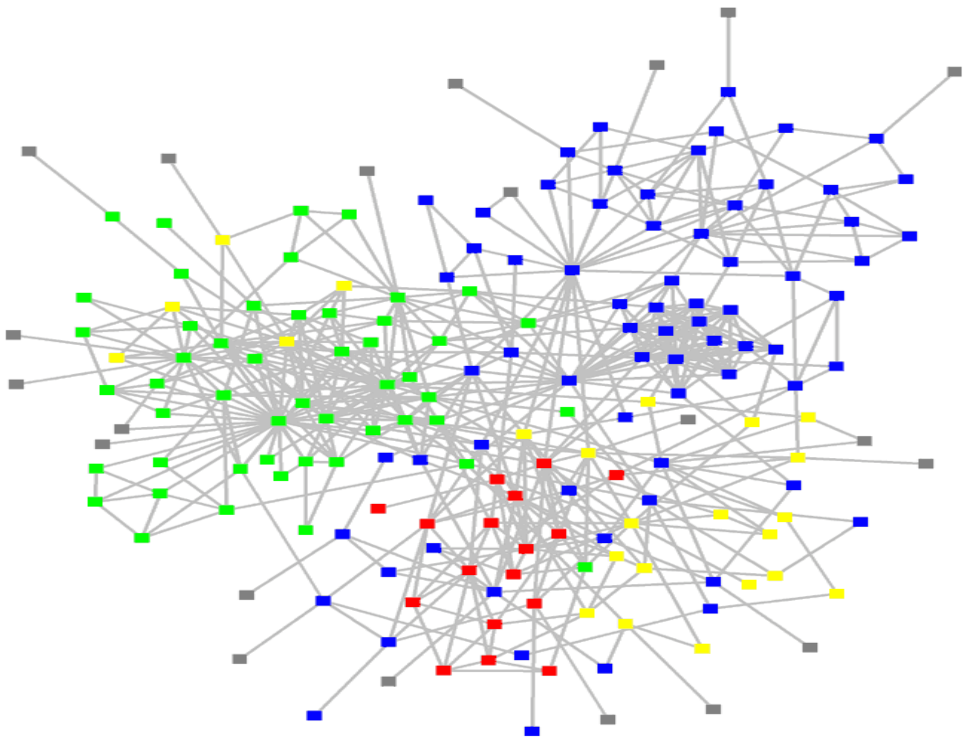
Behind each complex system there is a **network**, that defines the interactions between the component.

## SOCIETY | Factoid:

The “Social Graph” behind Facebook

*Keith Shepherd's "Sunday Best". http://baseballart.com/2010/07/shades-of-greatness-a-story-that-needed-to-be-told/*

# STRUCTURE OF AN ORGANIZATION



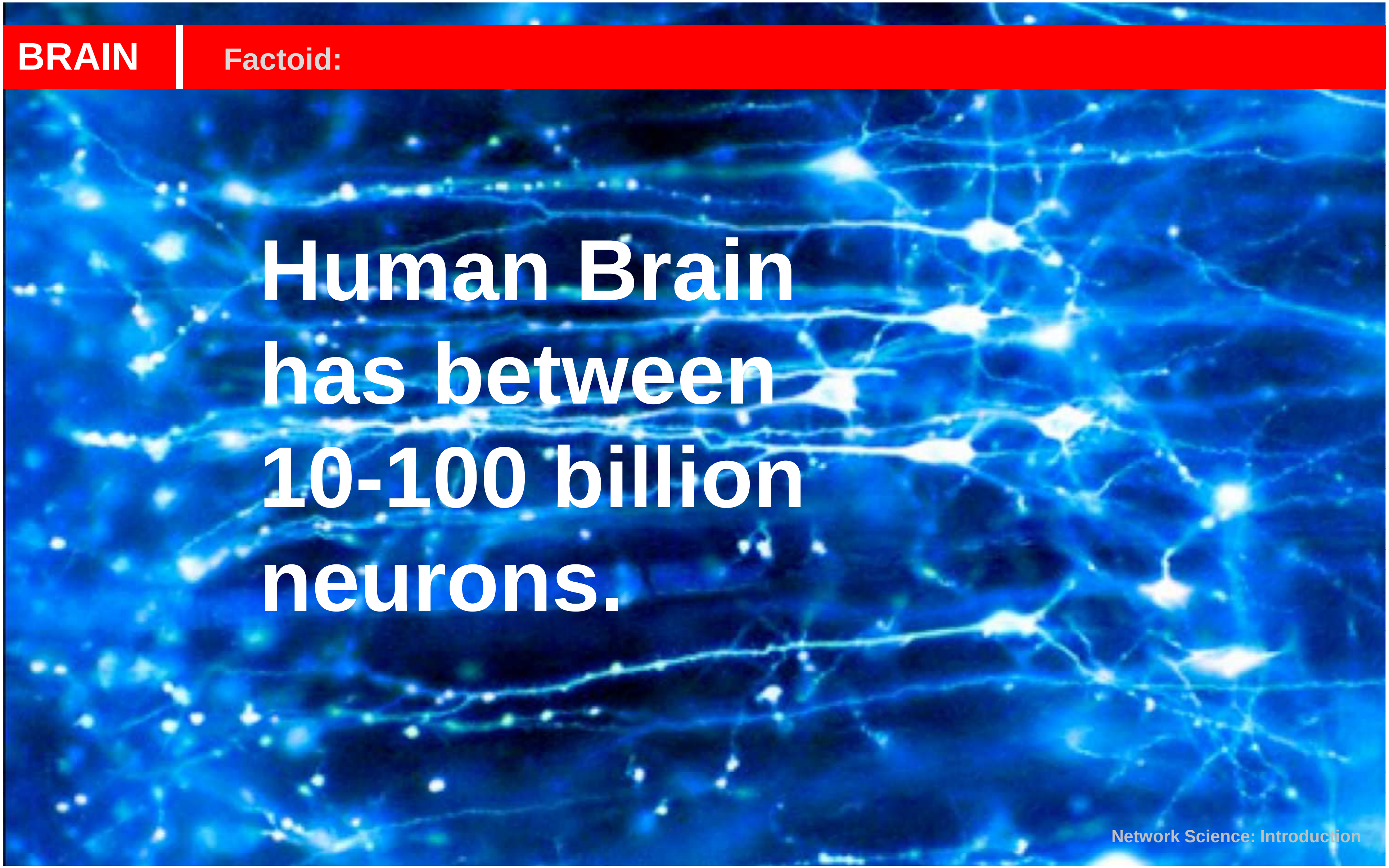
*www.orgnet.com*

: departments

: consultants

: external experts

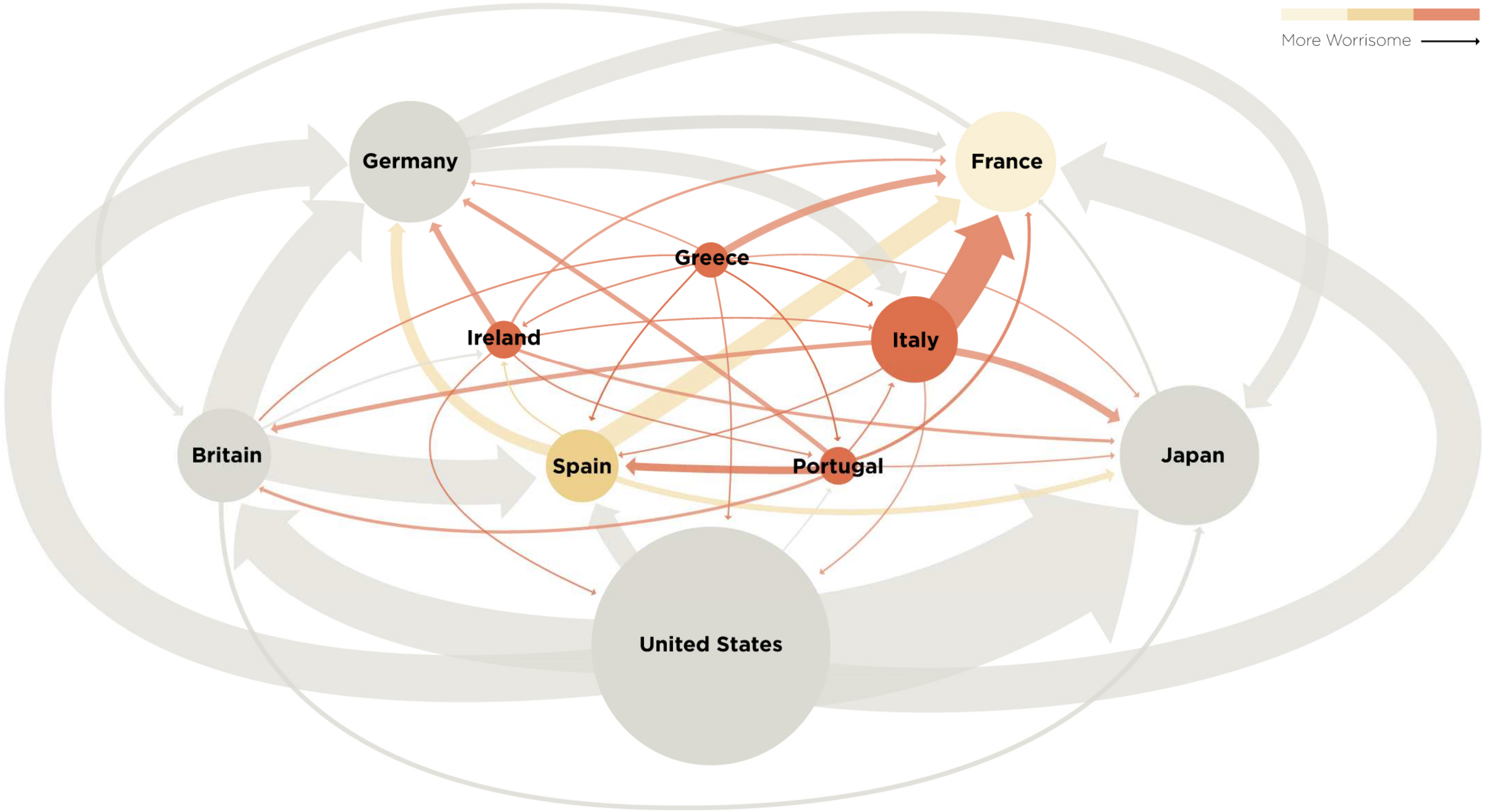
**BRAIN** | **Factoid:**



# Human Brain has between 10-100 billion neurons.

# The subtle financial networks

# The not so subtle financial networks: 2011



# BUSINESS TIES IN US BIOTECH-INDUSTRY

**Nodes:**

Companies

Investment

Pharma

Research Labs

Public

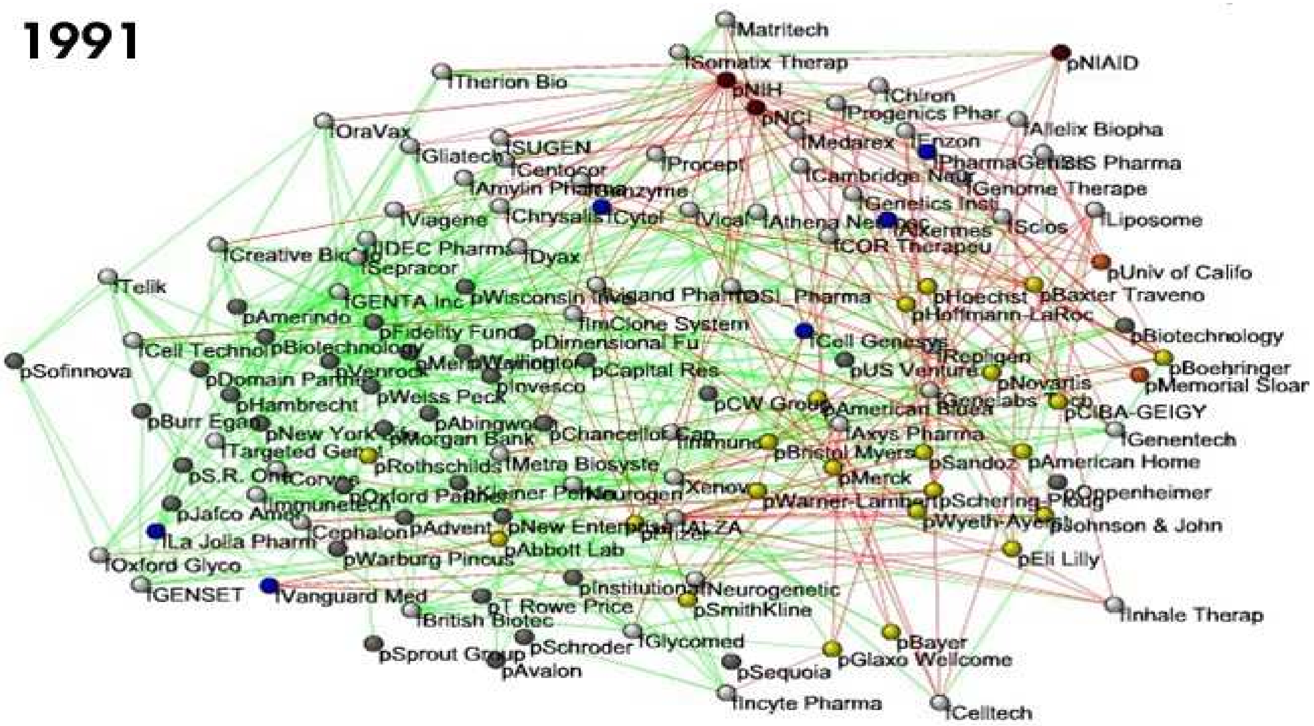
Biotechnology

**Links:**

Collaborations

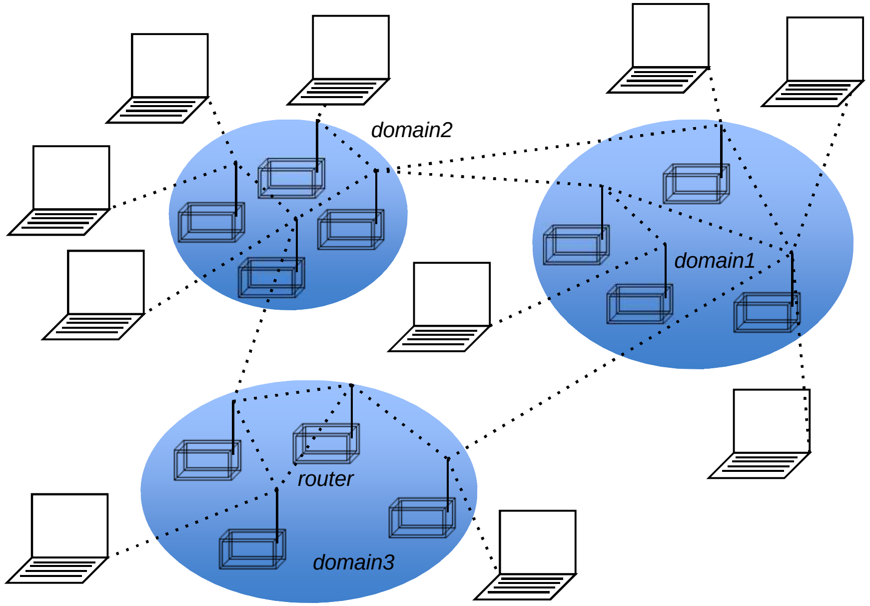
Financial

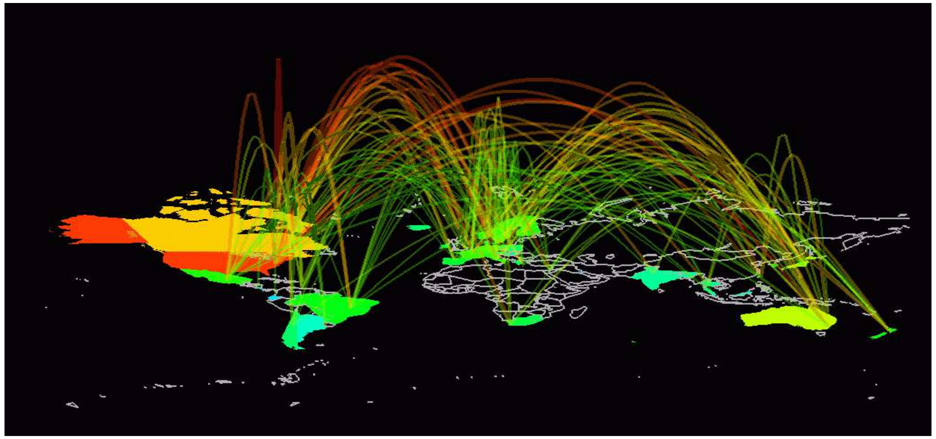
R&D

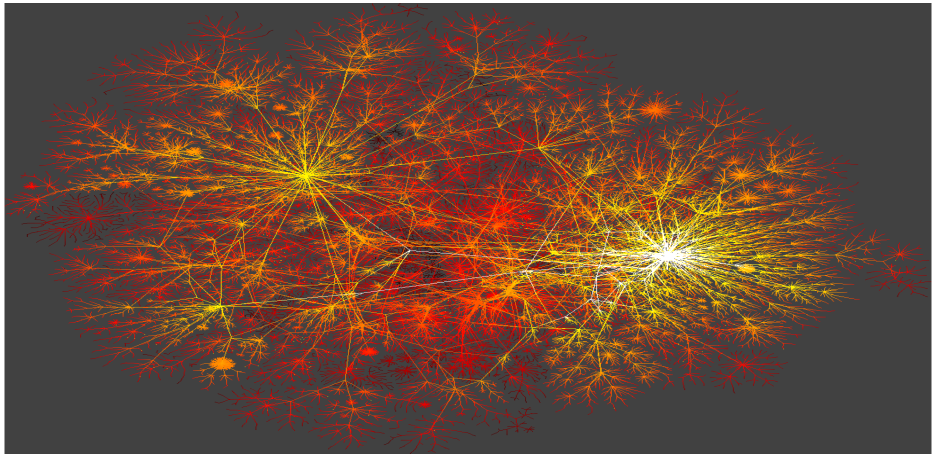



http://ecclectic.ss.uci.edu/~drwhite/Movie

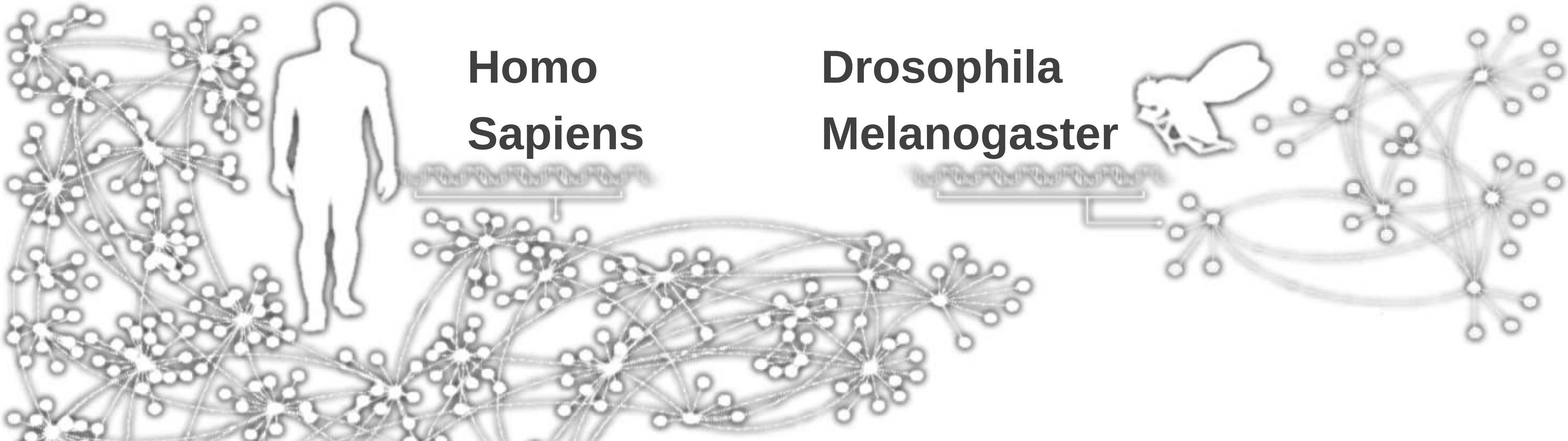
# INTERNET

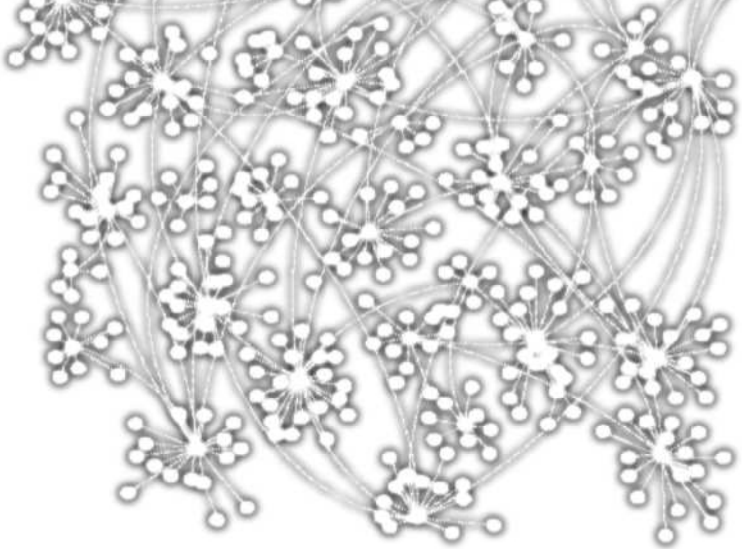




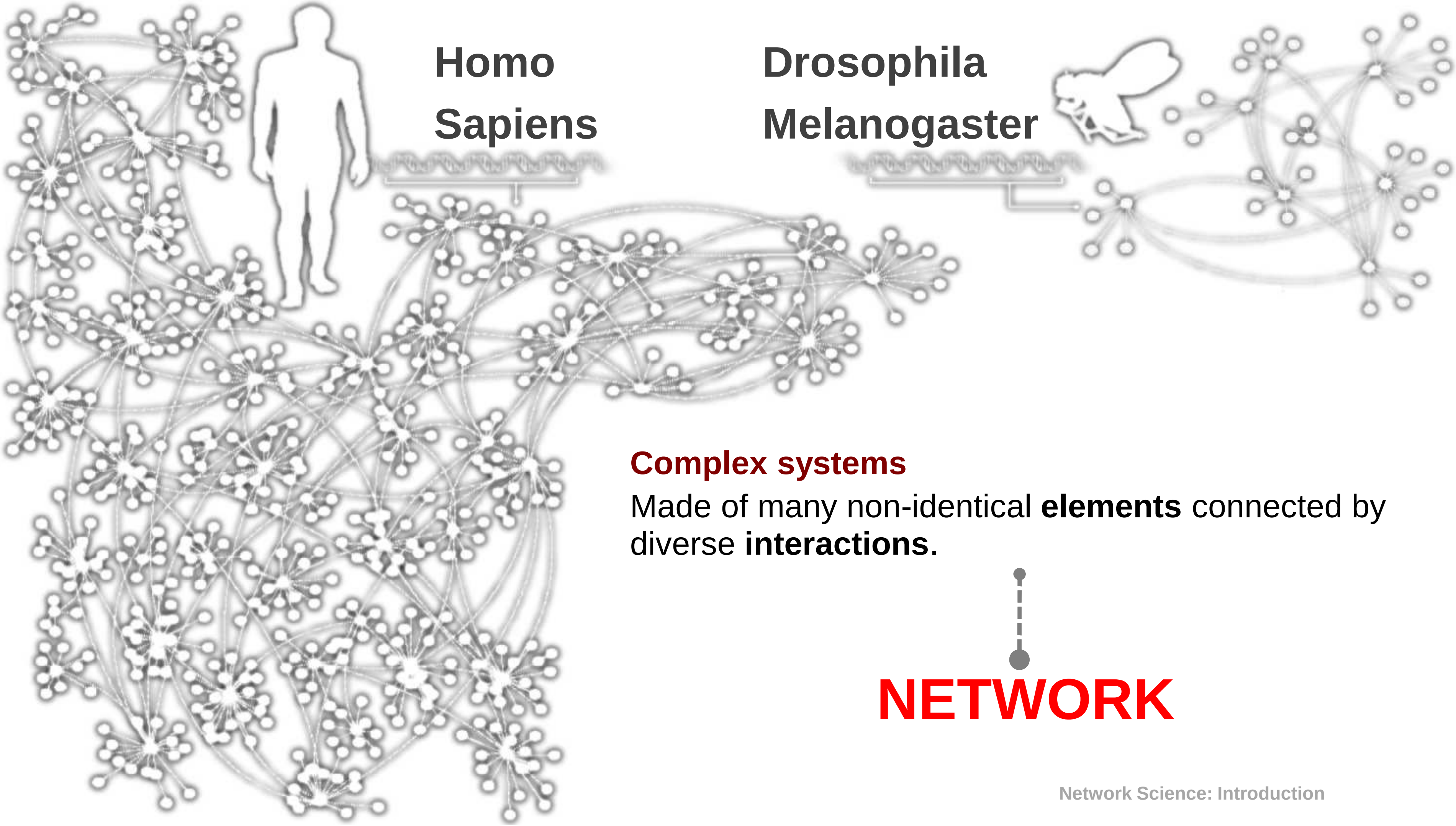
## HUMANS GENES



In the generic networks shown, the points represent the elements of each organism's genetic network, and the dotted lines show the interactions between them.

## HUMANS GENES




**Complex systems**

Made of many non-identical **elements** connected by diverse **interactions**.

# NETWORK

# THE ROLE OF NETWORKS

Behind each system studied in complexity there is an intricate wiring diagram, or a **network**, that defines the interactions between the component.

We will never understand complex system unless we map out and understand the networks behind them.

# TWO FORCES HELPED THE EMERGENCE OF NETWORK SCIENCE

## THE HISTORY OF NETWORK ANALYSIS

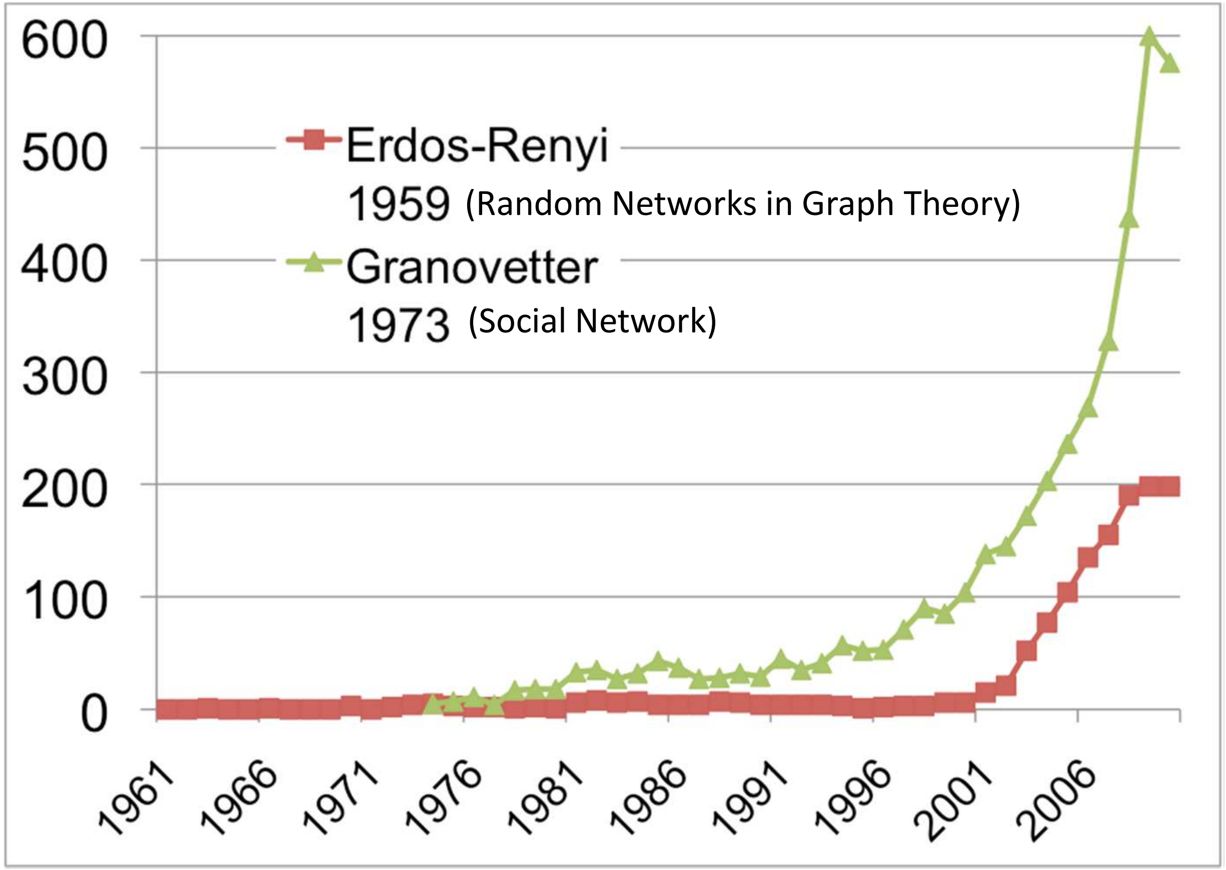
**Graph theory:** 1735, Euler

**Social Network Research:** 1930s, Moreno

**Communication networks/internet:** 1960s

**Ecological Networks:** May, 1979.

# THE HISTORY OF NETWORK ANALYSIS



# The emergence of network maps:

Movie Actor Network, 1998;
World Wide Web, 1999.
C elegans neural wiring diagram 1990
Citation Network, 1998
Metabolic Network, 2000;
Protein-protein Interaction (PPI) network, 2001

# The universality of network characteristics:

The architecture of networks emerging in various domains of science, nature, and technology are more similar to each other than one would have expected.

# THE CHARACTERISTICS OF NETWORK SCIENCE

# THE CHARACTERISTICS OF NETWORK SCIENCE

***Interdisciplinary***

***Empirical***

***Quantitative and Mathematical***

***Computational***

# THE CHARACTERISTICS OF NETWORK SCIENCE

***Interdisciplinary***

***Empirical, data driven***

***Quantitative and Mathematical***

***Computational***

# THE CHARACTERISTICS OF NETWORK SCIENCE

***Interdisciplinary***

***Empirical***

***Quantitative and Mathematical***

***Computational***

# THE CHARACTERISTICS OF NETWORK SCIENCE

*Interdisciplinary*

*Empirical*

*Quantitative and Mathematical*

***Computational***

# THE IMPACT OF NETWORK SCIENCE

# ECONOMIC IMPACT


**Google**
Market Cap(2010 Jan 1):
*$189 billion*

**Cisco Systems**
networking gear Market cap (Jan 1, 2919):
*$112 billion*

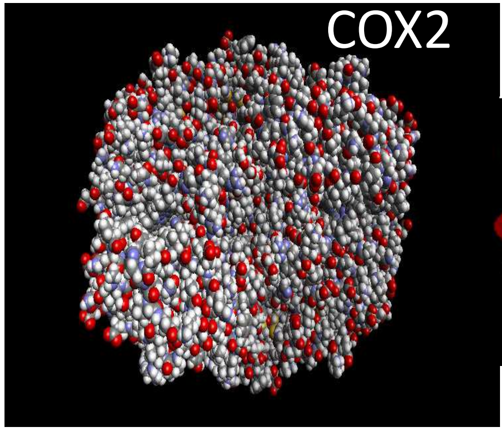
**Facebook**
market cap:
*$50 billion*

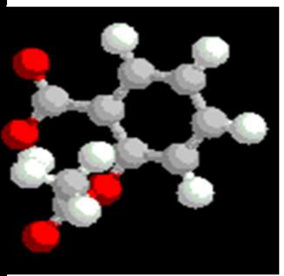
*www.bizjournals.com/austin/news/2010/11/15/facebooks... - Cached*

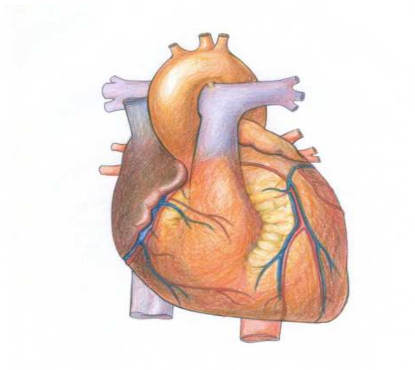
# DRUG DESIGN, METABOLIC ENGINEERING:

Reduces
Inflammation
Fever
Pain

Prevents
Heart attack
Stroke

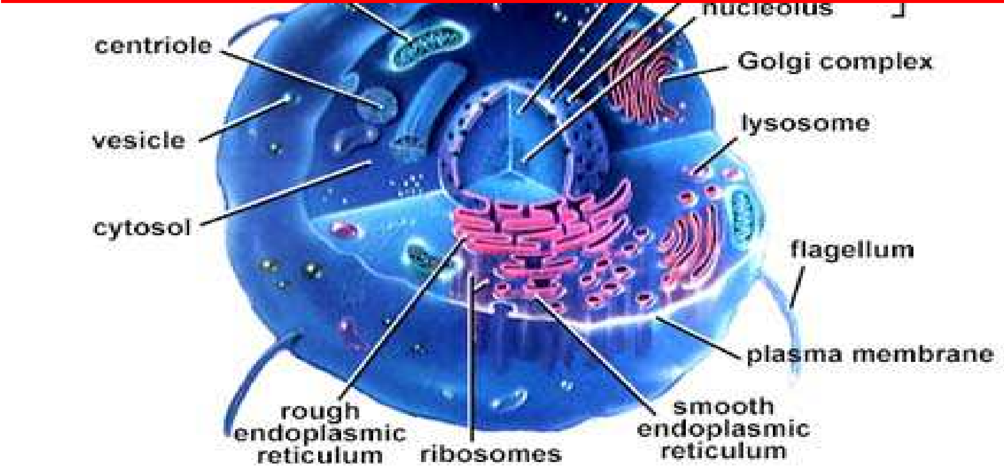


Reduces the risk of
Alzheimer's Disease

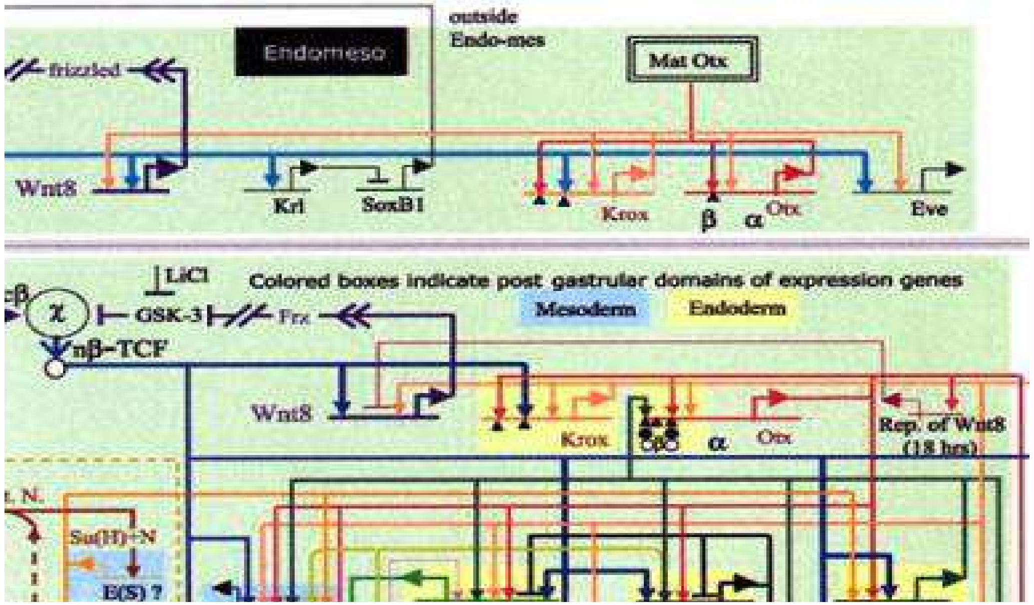
Reduces the risk of
breast cancer
ovarian cancers
colorectal cancer

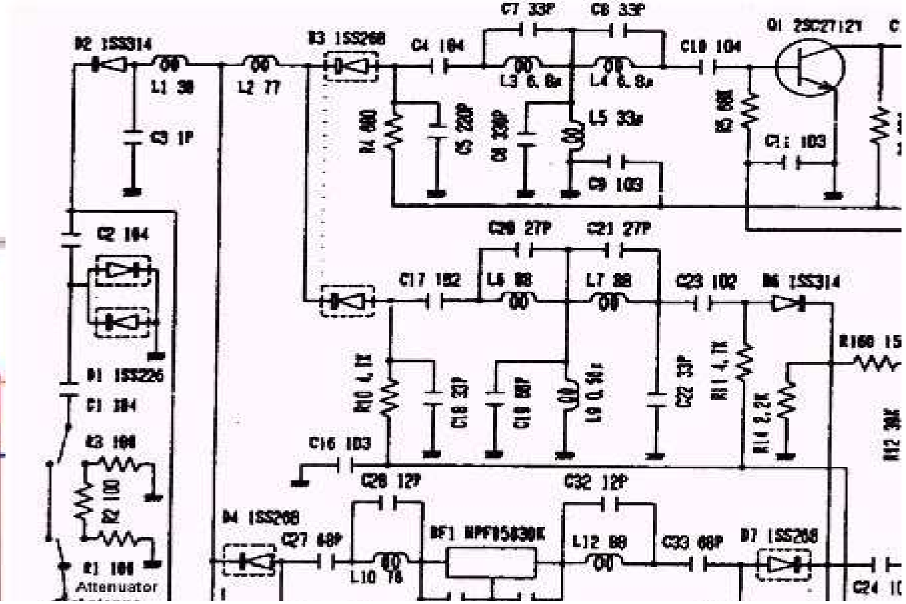
Causes
Bleeding
Ulcer

# DRUG DESIGN, METABOLIC ENGINEERING:

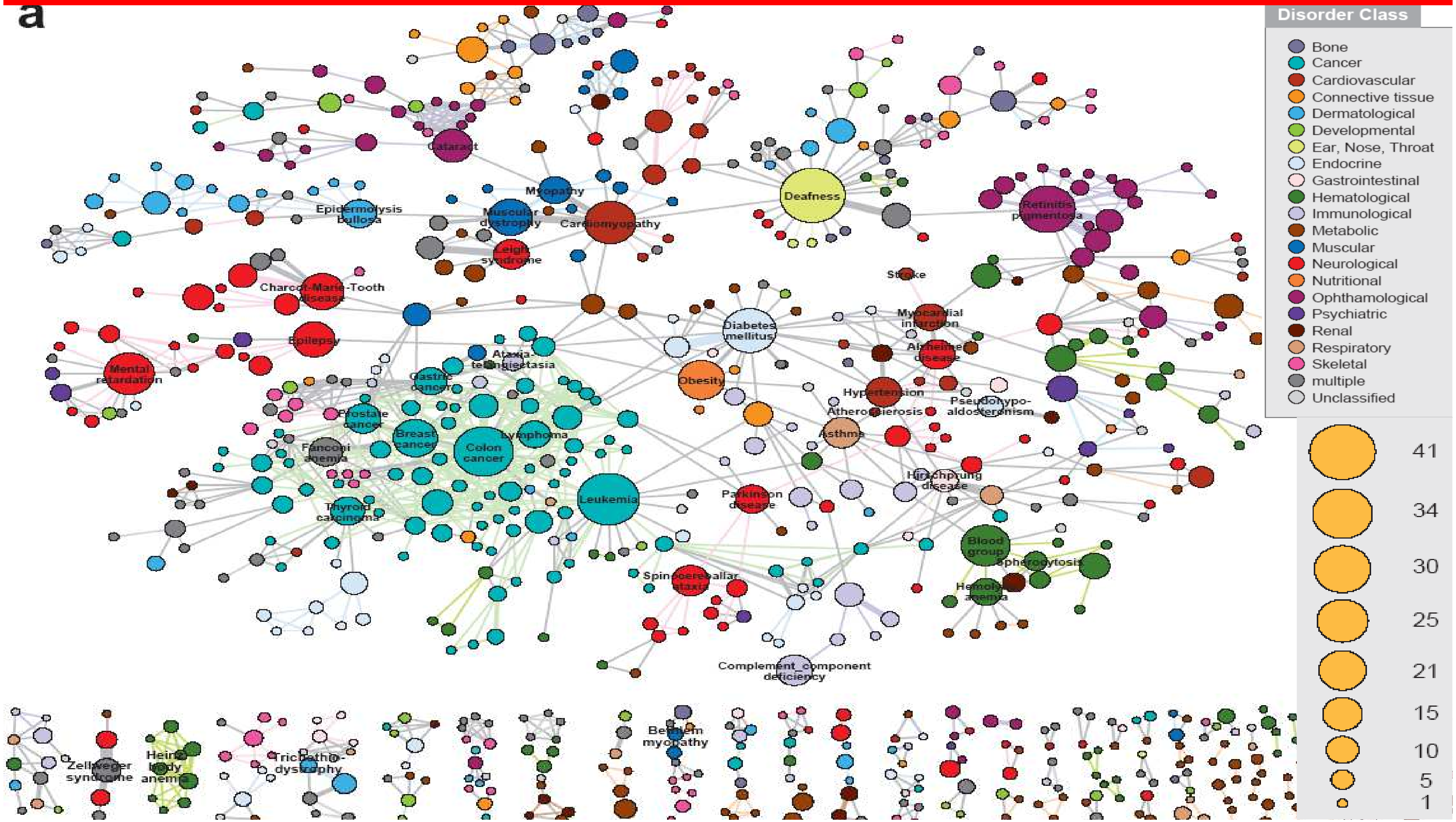




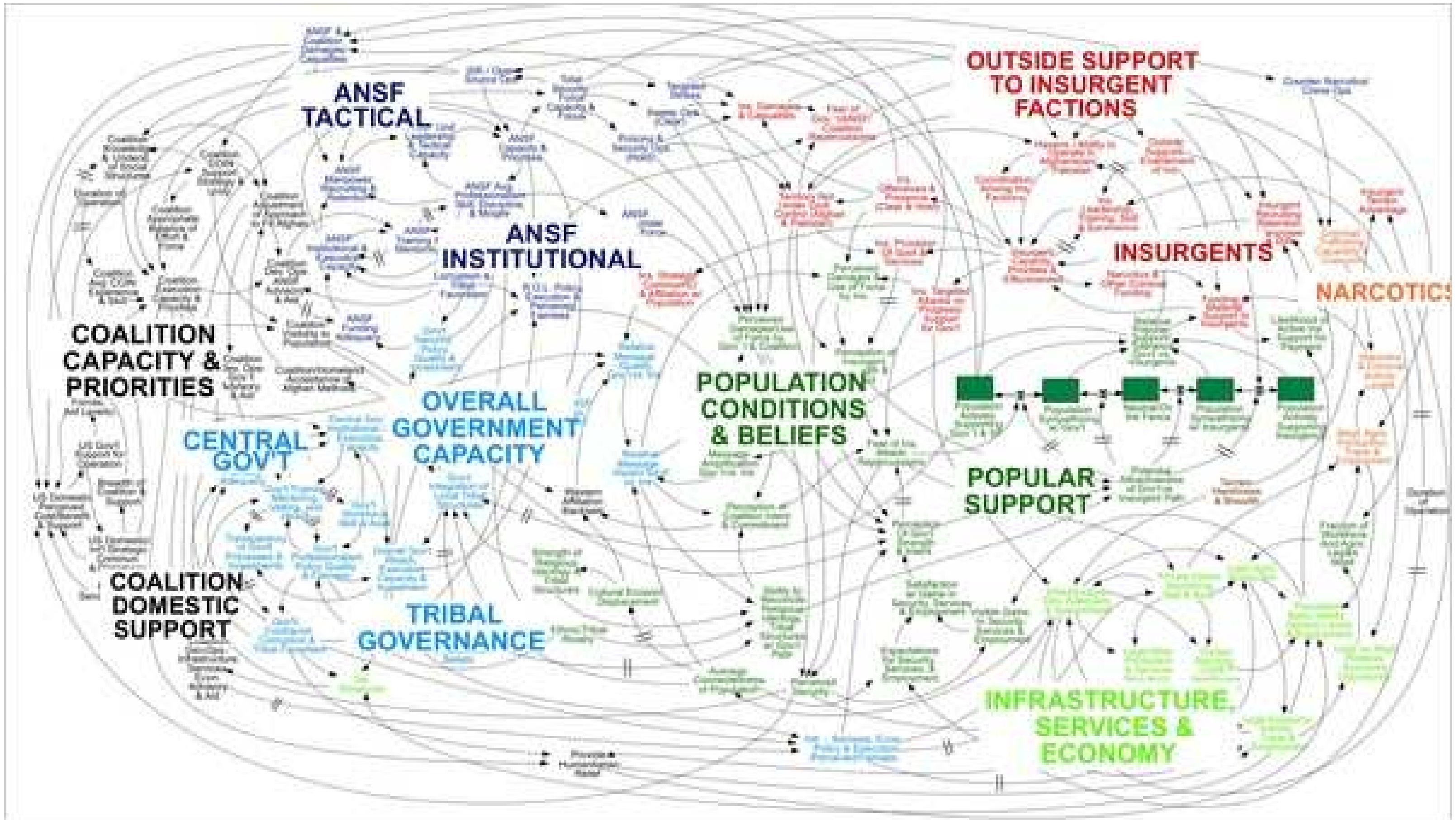


# HUMAN DISEASE NETWORK



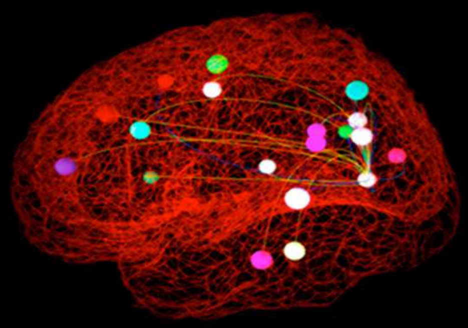
# The network behind a military engagement



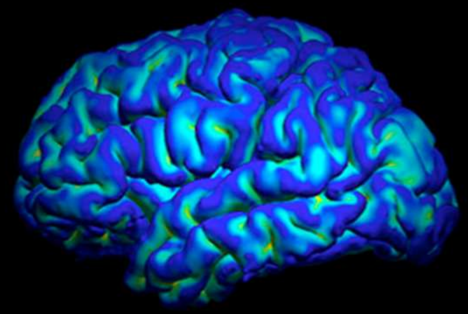
# Predicting the H1N1 pandemic

## BRAIN RESEARCH

In September 2010 the National Institutes of Health awarded $40 million to researchers at Harvard, Washington University in St. Louis, the University of Minnesota and UCLA, to develop the technologies that could systematically map out brain circuits.

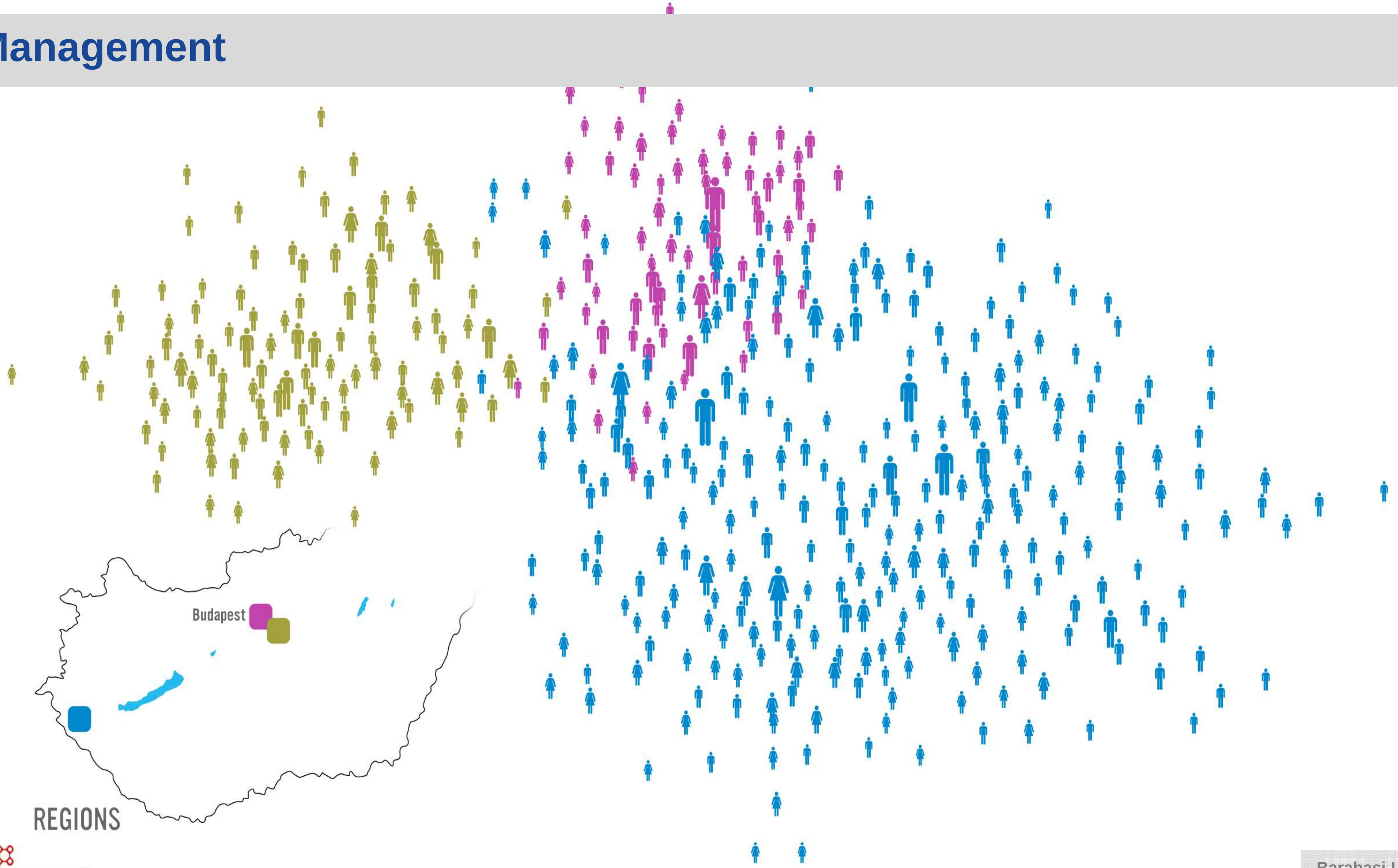


**The Human Connectome Project (HCP)** with the ambitious goal to construct a map of the complete structural and functional neural connections in vivo within and across individuals.



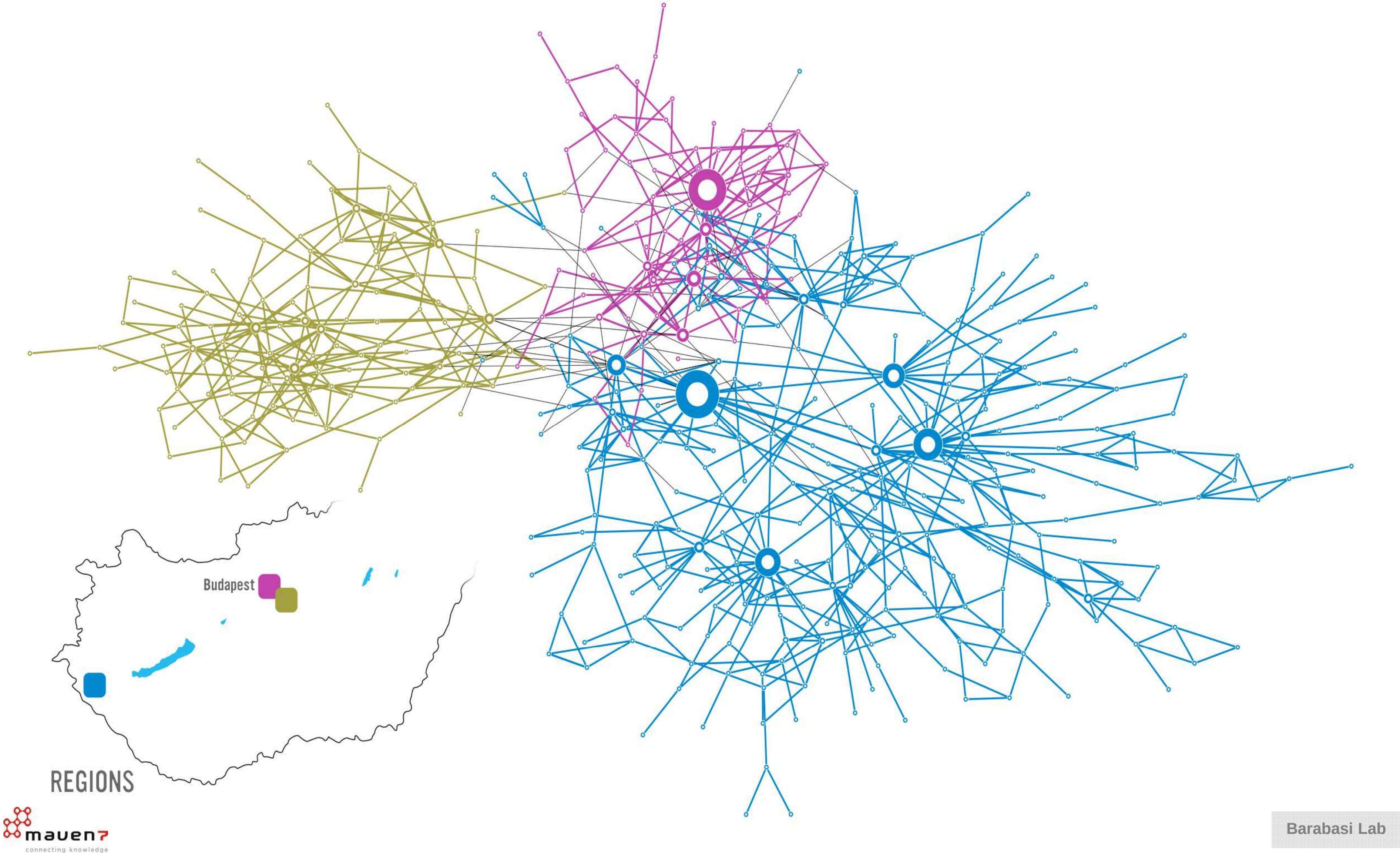
http://www.humanconnectomeproject.org/overview/

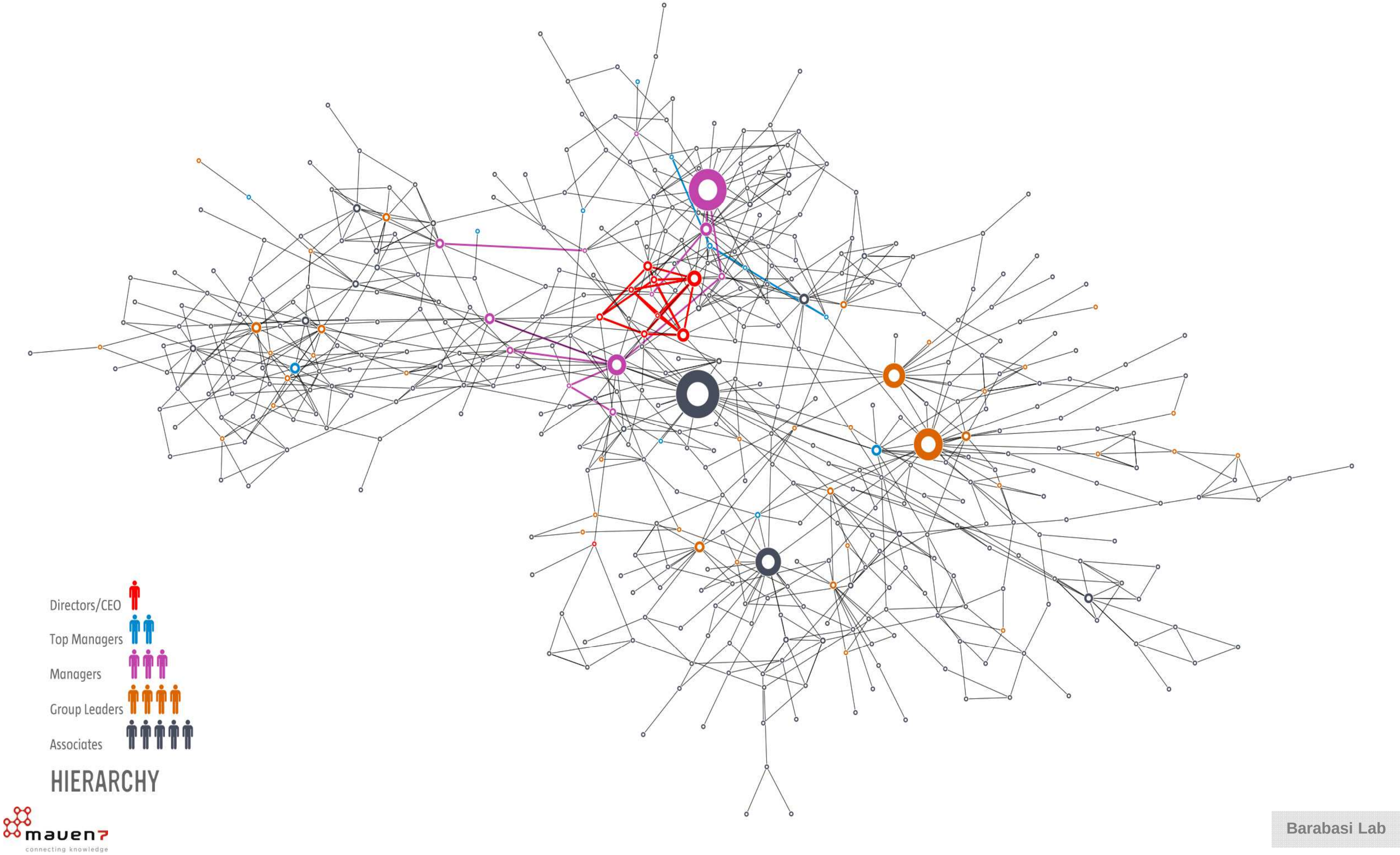
# Management



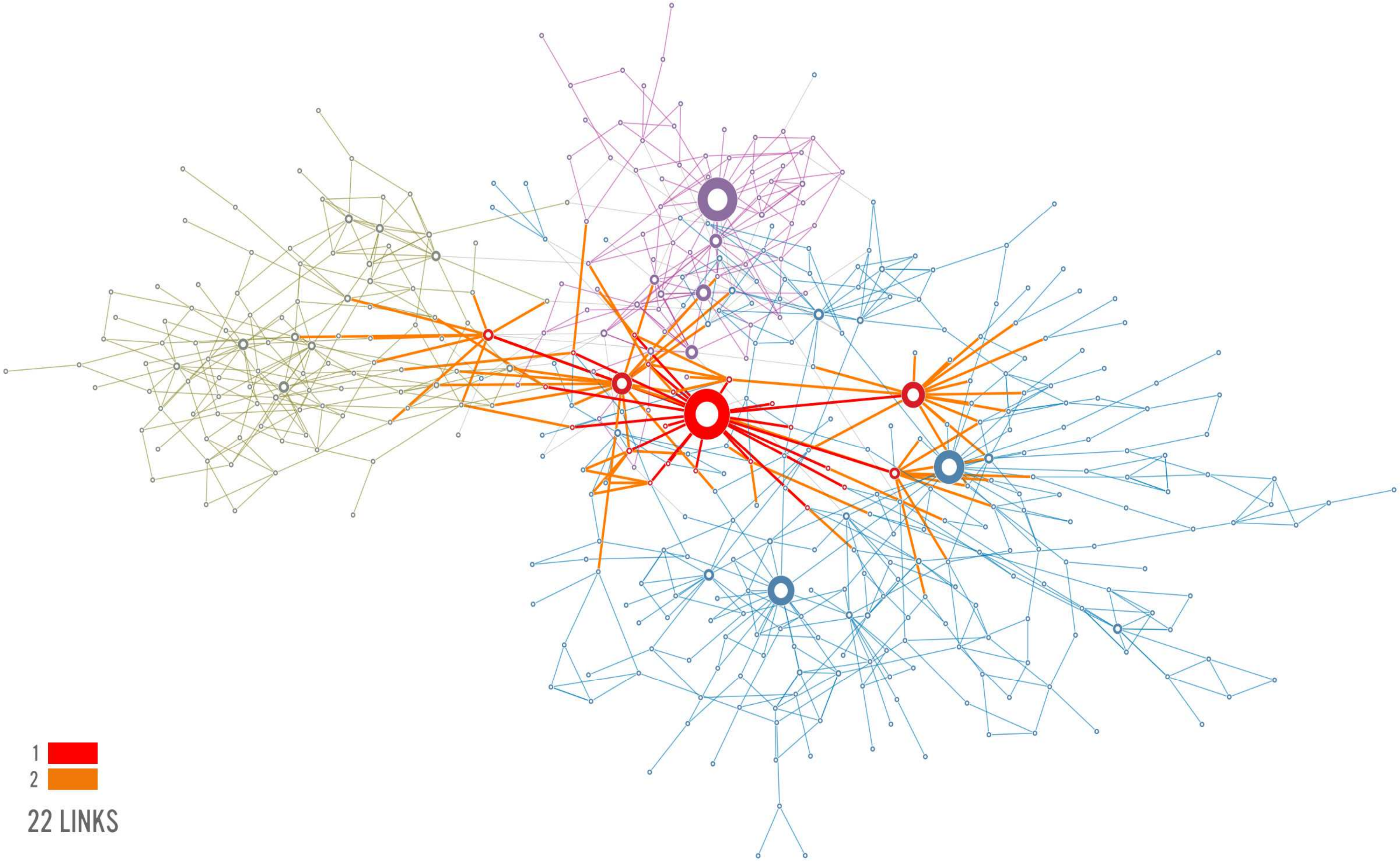

maven7 connecting knowledge

Barabasi Lab


Budapest
REGIONS
maven7
connecting knowledge
Barabasi Lab

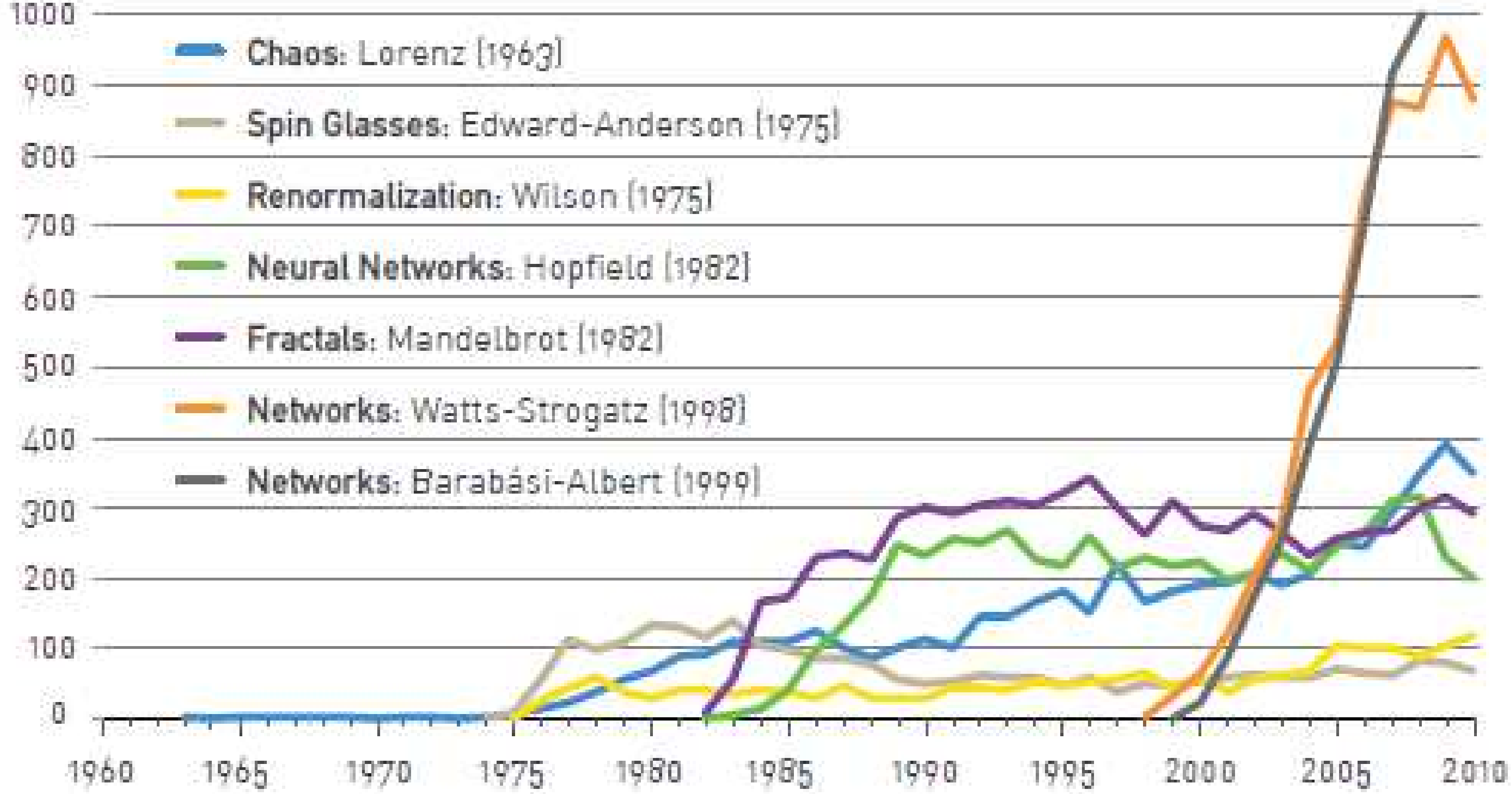


Directors/CEO
Top Managers
Managers
Group Leaders
Associates
HIERARCHY
maven7
connecting knowledge
Barabasi Lab


1
2
22 LINKS

# SCIENTIFIC IMPACT

# NETWORK SCIENCE | The science of the 21st century



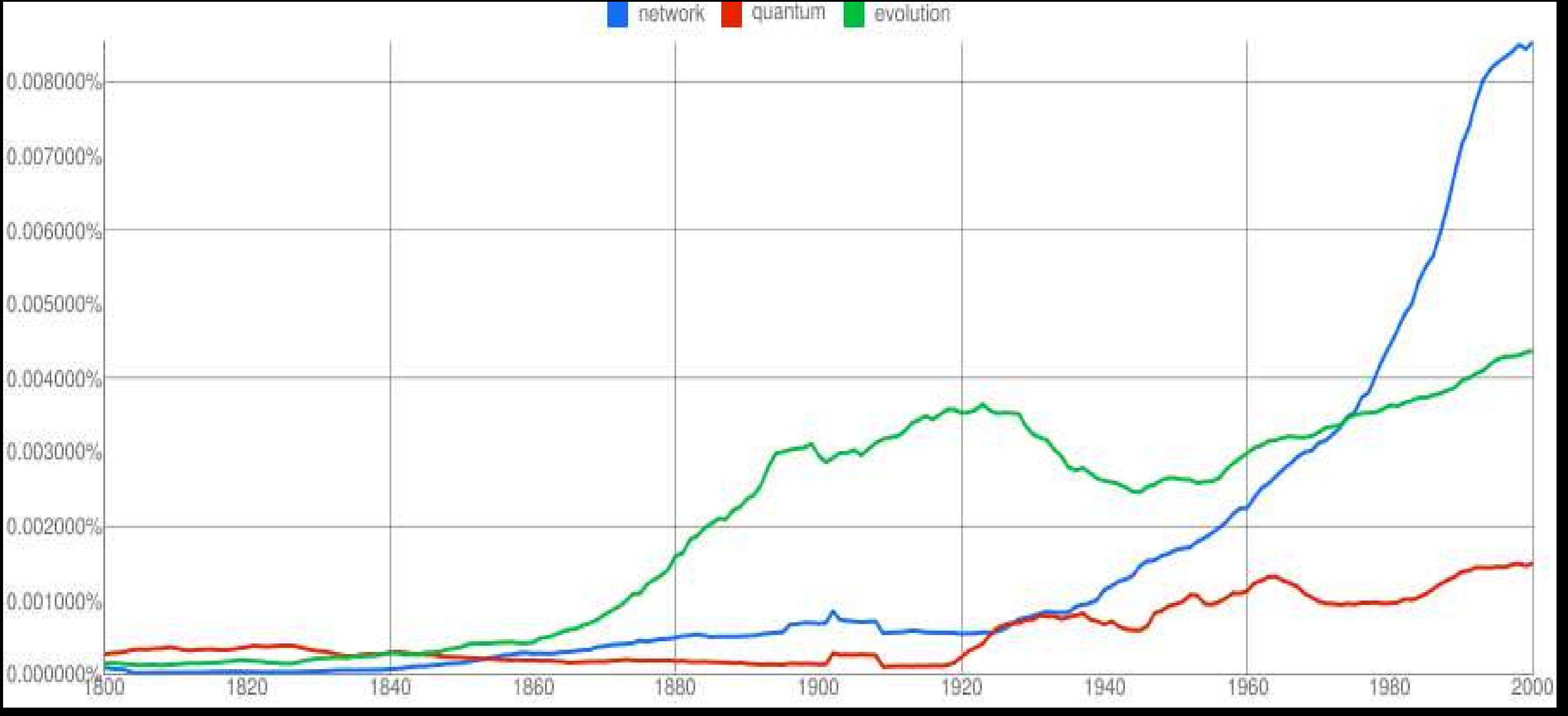

Times cited

Years

# SUMMARY

NGRAMS

# Networks Awareness



## MOST IMPORTANT | Networks Really Matter

If you were to understand the spread of diseases, **can you do it without networks?**

If you were to understand the WWW structure, searchability, etc, **hopeless without invoking the Web's topology.**

If you want to understand human diseases, **it is hopeless without considering the wiring diagram of the cell**.

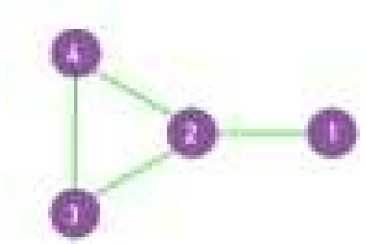
Section 10

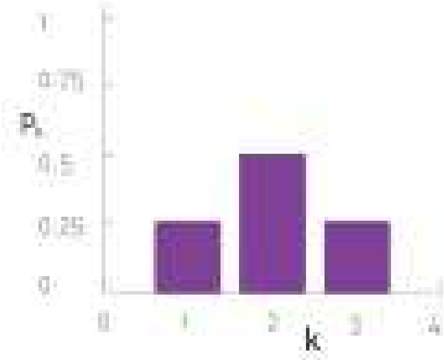
# Some Graph Properties

# DEGREE DISTRIBUTION

## Degree distribution

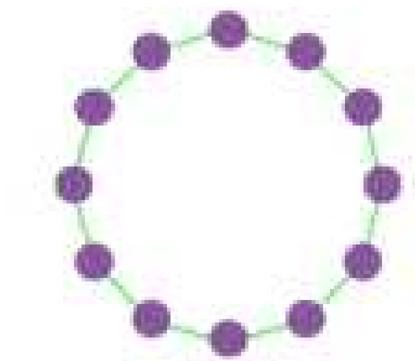
P(k): probability that a randomly chosen node has degree *k*

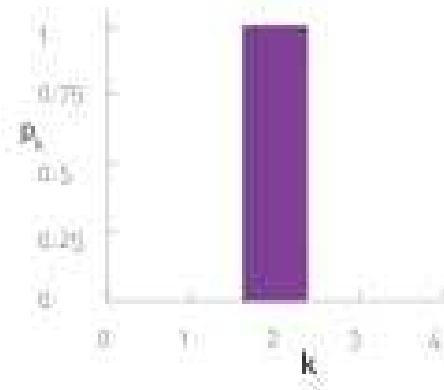



**$N_k$ = # nodes with degree k**

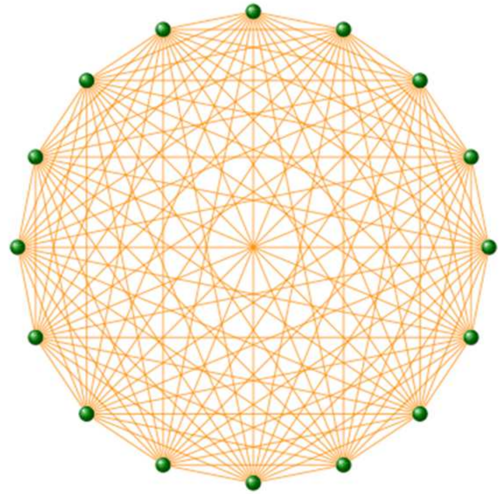
**$P(k) = N_k / N$ ❾ plot**





# COMPLETE GRAPH

The maximum number of links a network of N nodes can have is: $L_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$



A graph with degree $L=L_{max}$ is called a complete graph, and its average degree is **<k>=N-1**

## REAL NETWORKS ARE SPARSE

**Most networks observed in real systems are sparse:**

$$L \ll L_{max}$$

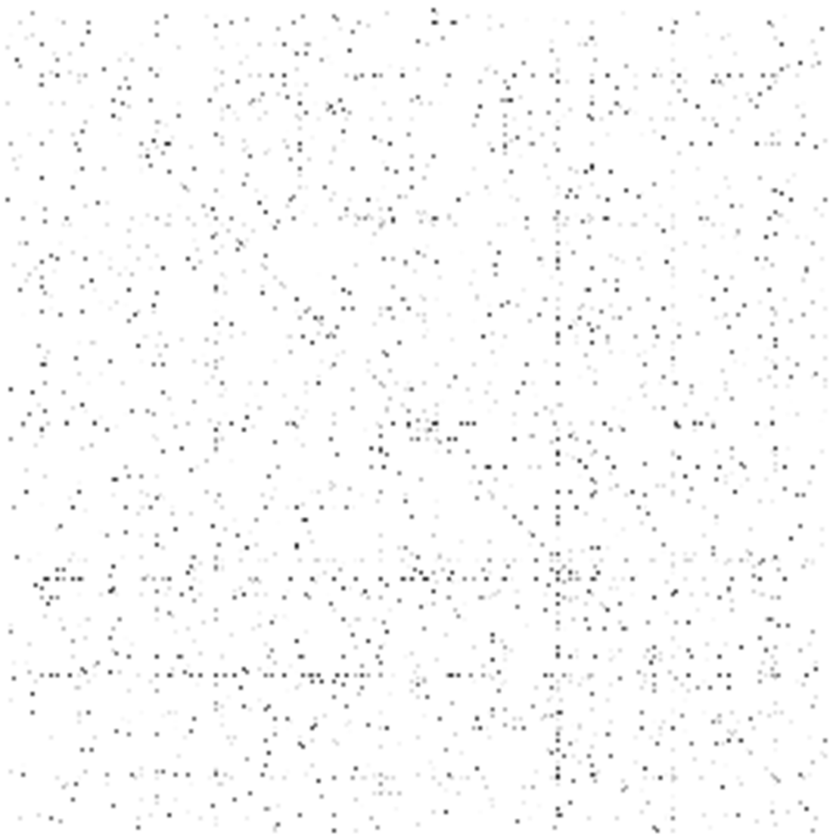
**or**

$$<k> \ll N-1.$$

| | | | | |
|---|---|---|---|---|
| WWW (ND Sample): | $N=325{,}729$; | $L=1.4\ 10^{6}$ | $L_{max}=10^{12}$ | $<k>=4.51$ |
| Protein (*S. Cerevisiae*): | $N=\ \ 1{,}870$; | $L=4{,}470$ | $L_{max}=10^{7}$ | $<k>=2.39$ |
| Coauthorship (Math): | $N=\ 70{,}975$; | $L=2\ 10^{5}$ | $L_{max}=3\ 10^{10}$ | $<k>=3.9$ |
| Movie Actors: | $N=212{,}250$; | $L=6\ 10^{6}$ | $L_{max}=1.8\ 10^{13}$ | $<k>=28.78$ |

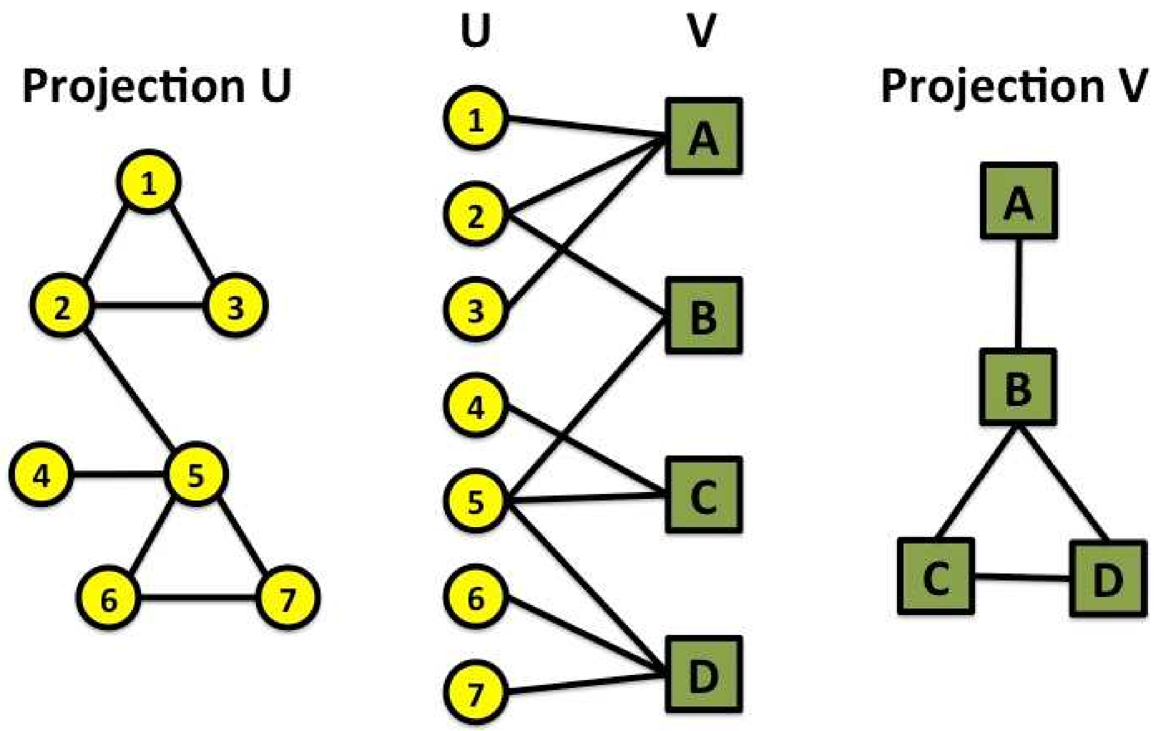
*(Source: Albert, Barabasi, RMP2002)*

# ADJACENCY MATRICES ARE SPARSE



# BIPARTITE GRAPHS

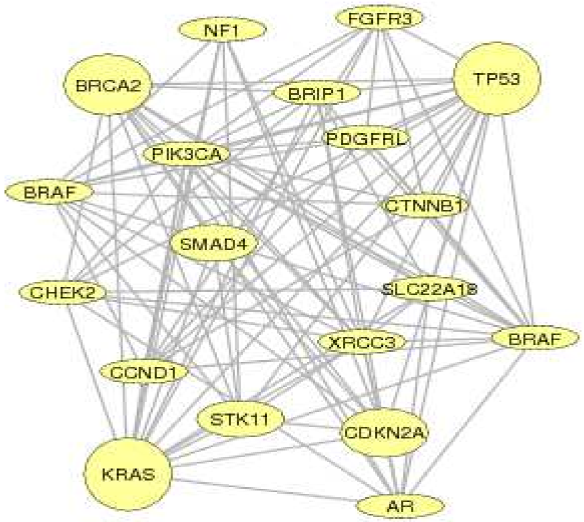
**bipartite graph** (or **bigraph**) is a graph whose nodes can be divided into two disjoint sets *U* and *V* such that every link connects a node in *U* to one in *V*; that is, *U* and *V* are independent sets.



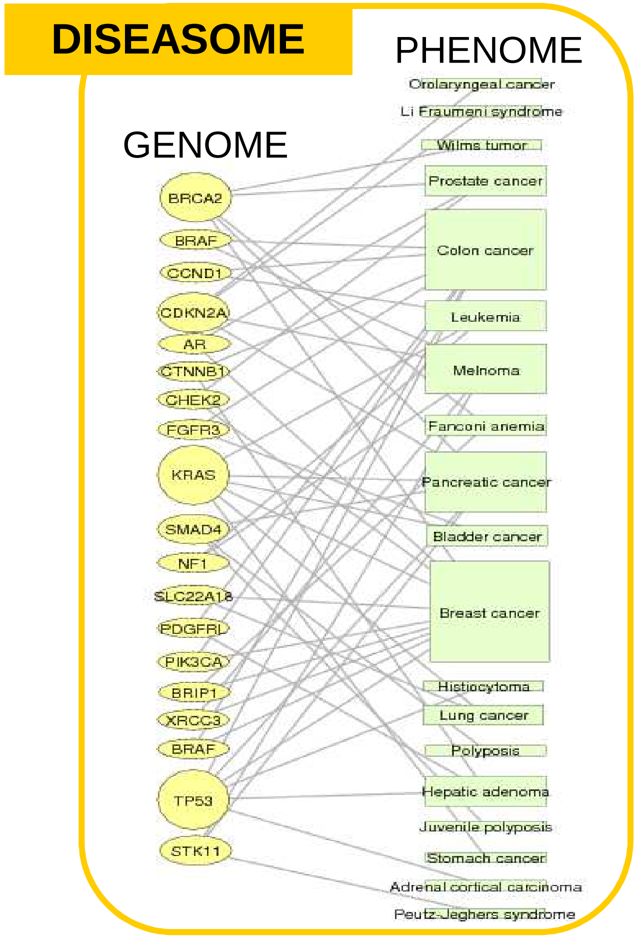

## Examples:

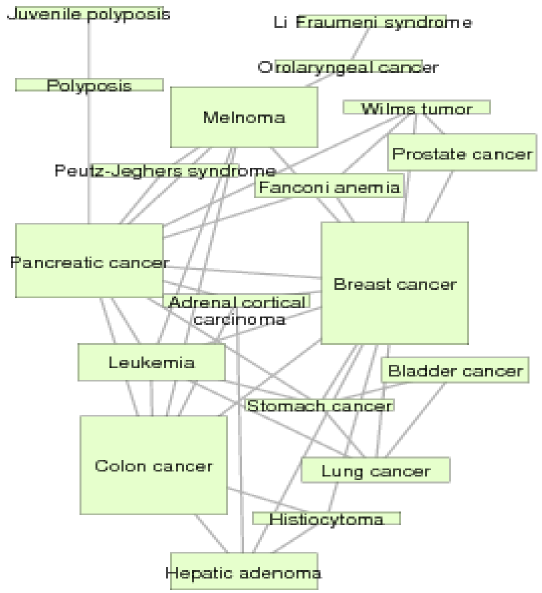
Hollywood actor network
Collaboration networks
Disease network (diseasome)

# GENE NETWORK – DISEASE NETWORK




**Gene network**

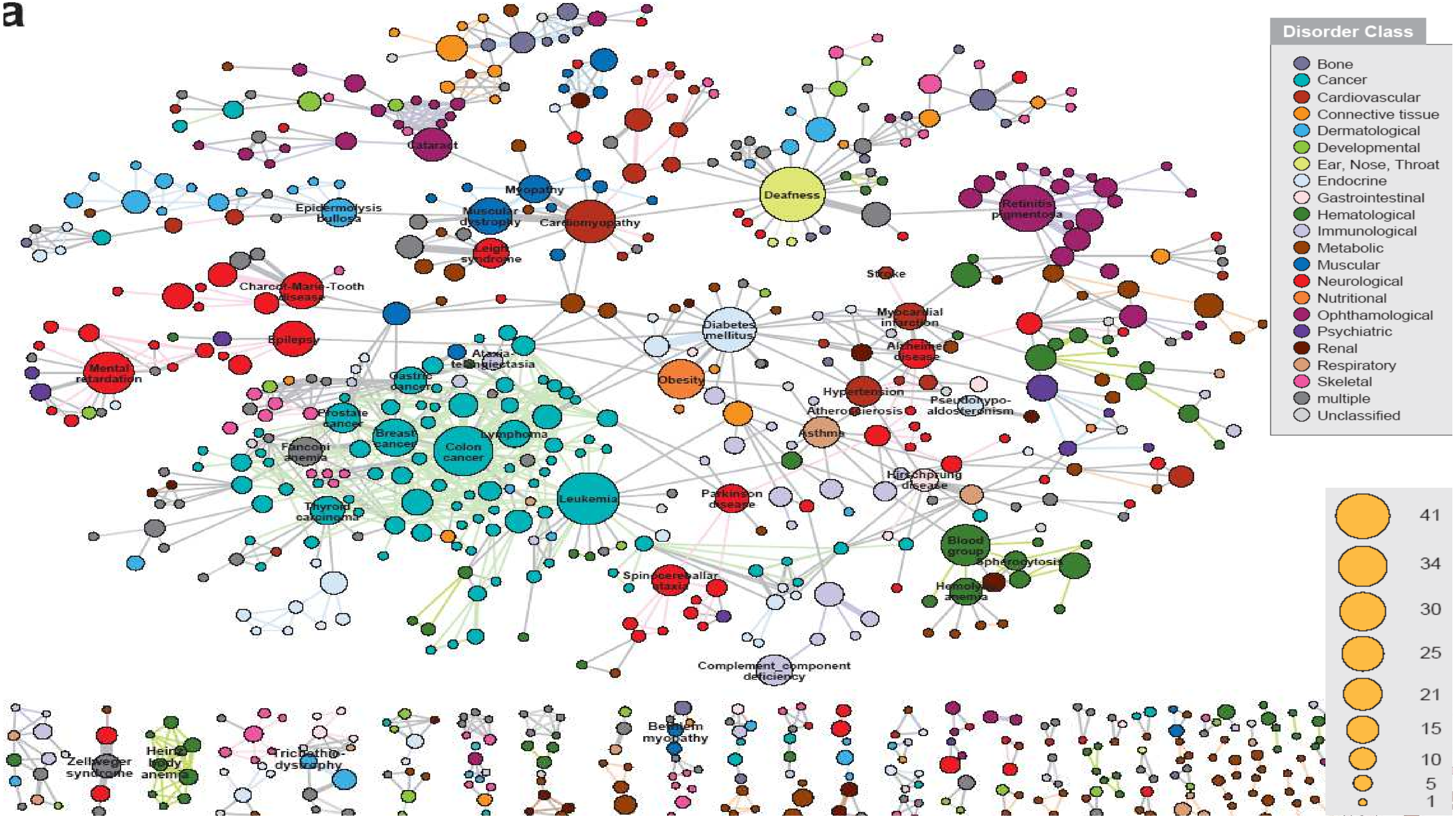


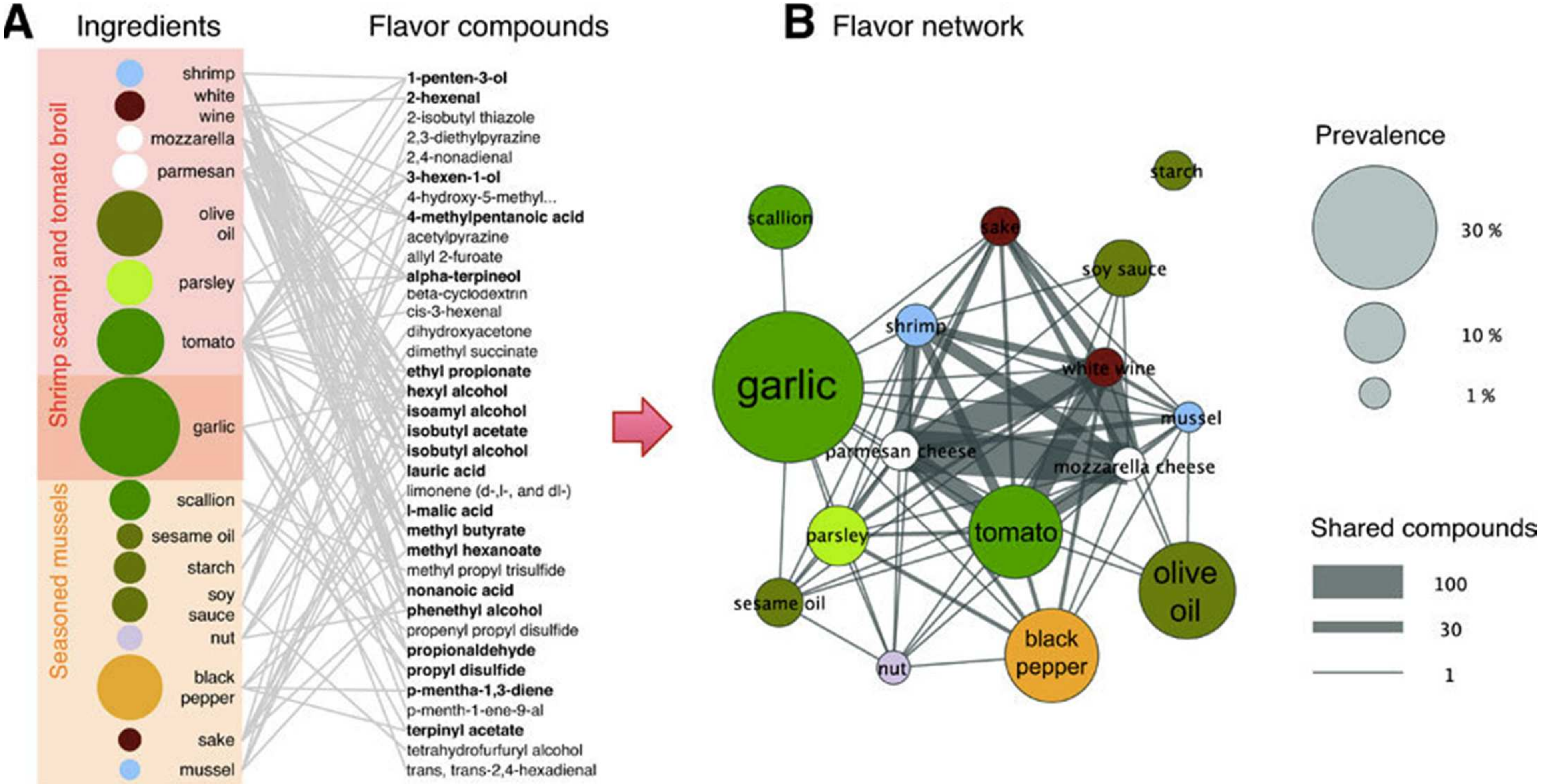

**Disease network**

*Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)*

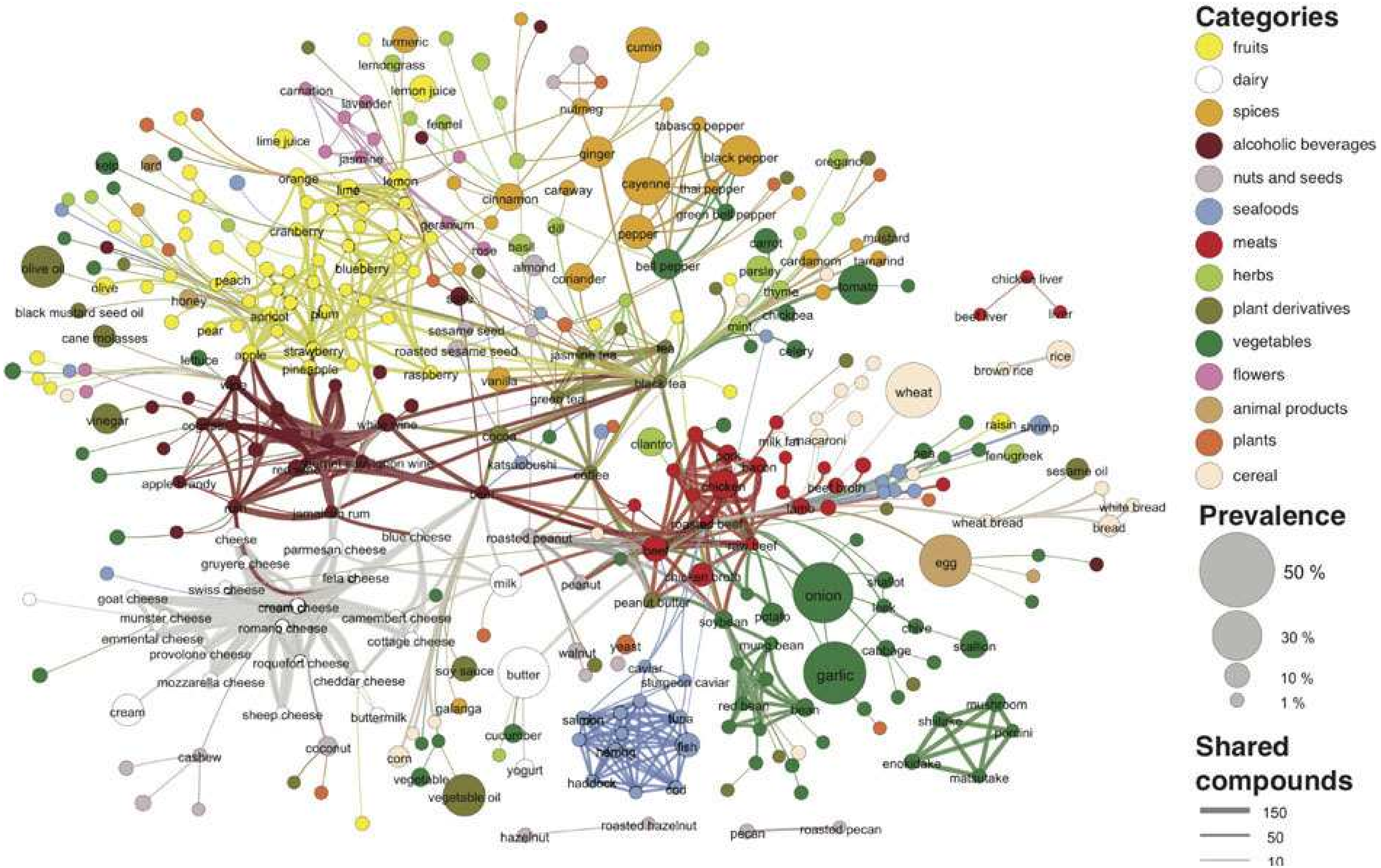
# HUMAN DISEASE NETWORK



# Ingredient-Flavor Bipartite Network




*Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, A.-L. Barabási Flavor network and the principles of food pairing , Scientific Reports 196, (2011).*



Categories
fruits
dairy
spices
alcoholic beverages
nuts and seeds
seafoods
meats
herbs
plant derivatives
vegetables
flowers
animal products
plants
cereal
Prevalence
50 %
30 %
10 %
1 %
Shared compounds
150
50
10
turmeric
cumin
lemongrass
carnation
lemon juice
lavender
nutmeg
fennel
tabasco pepper
lime juice
ginger
jasmine
black pepper
oregano
kelp
lard
orange
lime
lemon
cinnamon
caraway
cayenne
thai pepper
green bell pepper
cranberry
geranium
dill
pepper
mustard
carrot
rose
basil
bell pepper
cardamom
tamarind
olive oil
blueberry
almond
coriander
parsley
tomato
olive
peach
thyme
chicken liver
honey
black mustard seed oil
apricot
plum
chickpea
beef liver
liver
pear
sesame seed
mint
cane molasses
strawberry
roasted sesame seed
jasmine tea
tea
celery
brown rice
rice
lettuce
apple
pineapple
raspberry
vanilla
black tea
wheat
green tea
vinegar
white wine
raisin
shrimp
cognac
cocoa
cilantro
milk fat
macaroni
pea
fenugreek
katsuobushi
pork
bacon
sesame oil
apple brandy
coffee
chicken
beef broth
white bread
rum
jamaican rum
lamb
wheat bread
roasted beef
bread
cheese
blue cheese
roasted peanut
raw beef
parmesan cheese
beef
egg
gruyere cheese
feta cheese
milk
peanut
chicken broth
shallot
swiss cheese
goat cheese
cream cheese
peanut butter
onion
leek
munster cheese
romano cheese
camembert cheese
soybean
potato
chive
emmental cheese
cottage cheese
walnut
yeast
mung bean
cabbage
scallion
provolone cheese
roquefort cheese
caviar
mozzarella cheese
cheddar cheese
soy sauce
butter
sturgeon caviar
garlic
red bean
bean
mushroom
cream
galanga
salmon
tuna
shiitake
sheep cheese
buttermilk
cucumber
fish
porcini
herring
coconut
yogurt
casbew
corn
enokidake
haddock
matsutake
vegetable
cod
vegetable oil
roasted hazelnut
hazelnut
pecan
roasted pecan

# Clustering coefficient

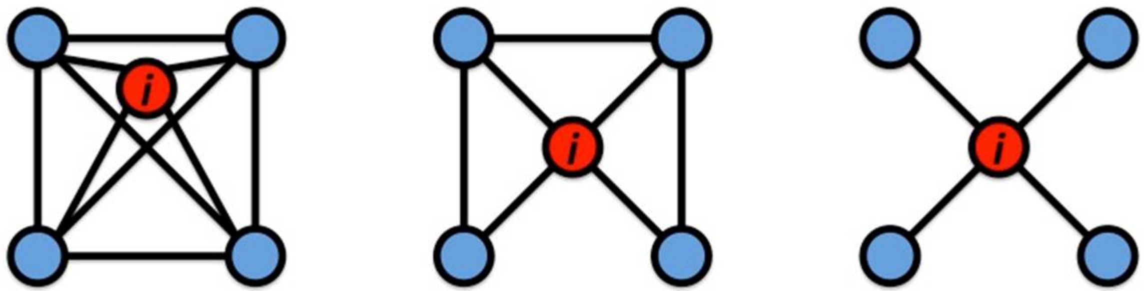
# CLUSTERING COEFFICIENT

## Clustering coefficient:

what fraction of your neighbors are connected?

- Node i with degree $k_i$
- $C_i$ in [0,1]

$$\boxed{C_i = \frac{2e_i}{k_i(k_i-1)}}$$



$C_i = 1$ $C_i = 1/2$ $C_i = 0$

**Watts & Strogatz, Nature 1998.**

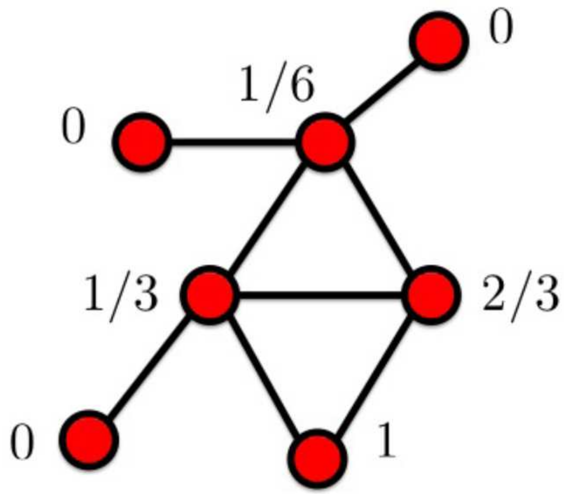
# CLUSTERING COEFFICIENT

## Clustering coefficient:

what fraction of your neighbors are connected?

- Node i with degree $k_i$
- $C_i$ in [0,1]

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$



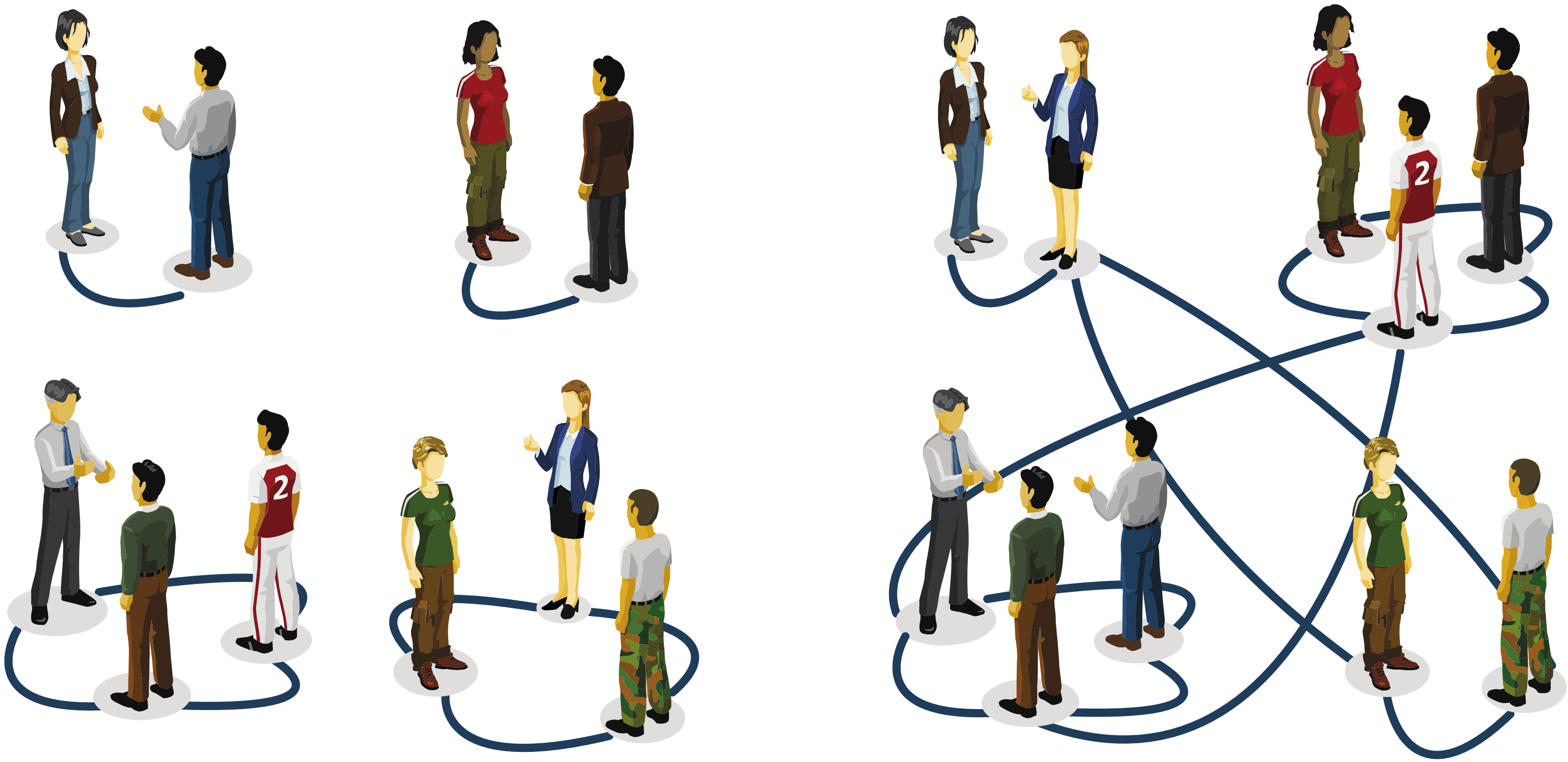

$$\langle C \rangle = \frac{13}{42} \approx 0.310$$

$$C = \frac{3}{8} = 0.375$$

**Watts & Strogatz, Nature 1998.**

# Random Networks

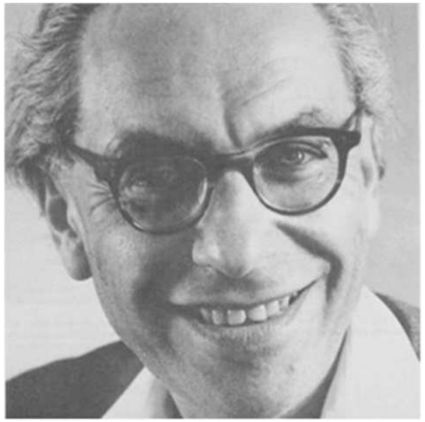
## RANDOM NETWORK MODEL



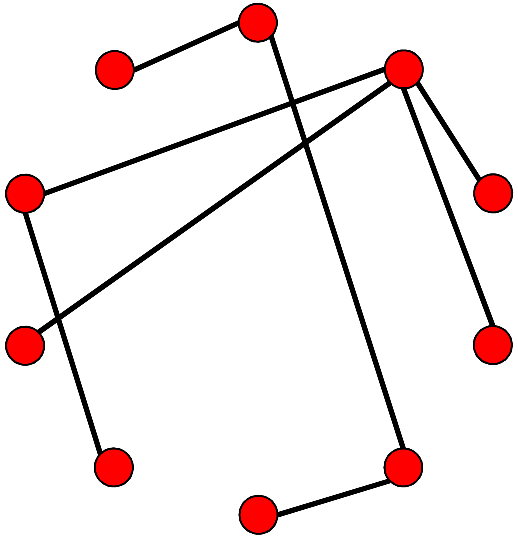
# The random network model

# RANDOM NETWORK MODEL

**Pál Erdös**
(1913-1996)



**Alfréd Rényi**
(1921-1970)



**Erdös-Rényi model (1960)**

**Connect with probability p**

p=1/6 N=10

<k> ~ 1.5

# RANDOM NETWORK MODEL

**Definition:**

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

**$G(N, L)$ Model**

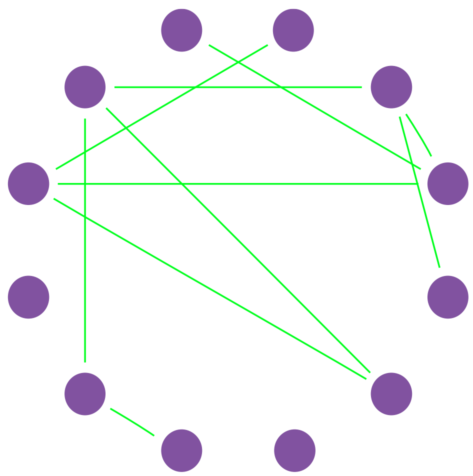
N labeled nodes are connected with $L$ randomly placed links. Erdős and Rényi used this definition in their string of papers on random networks [2-9].

**$G(N, p)$ Model**

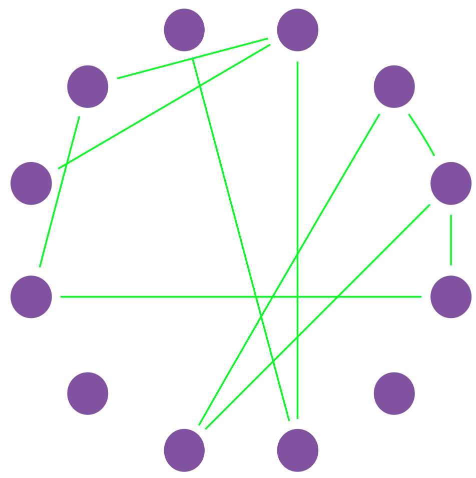
Each pair of N labeled nodes is connected with probability $p$, a model introduced by Gilbert [10].

## RANDOM NETWORK MODEL

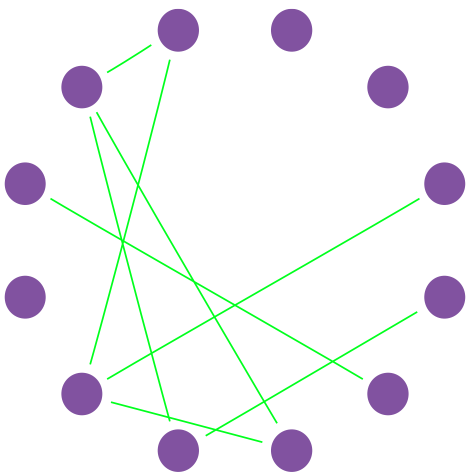
p=1/6
N=12



L=8



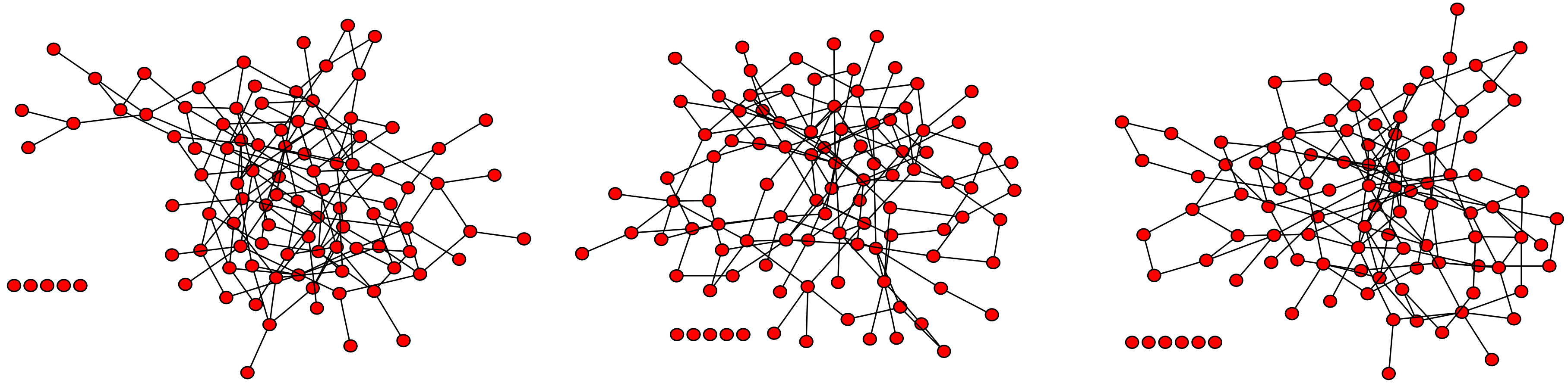
L=10



L=7

# RANDOM NETWORK MODEL

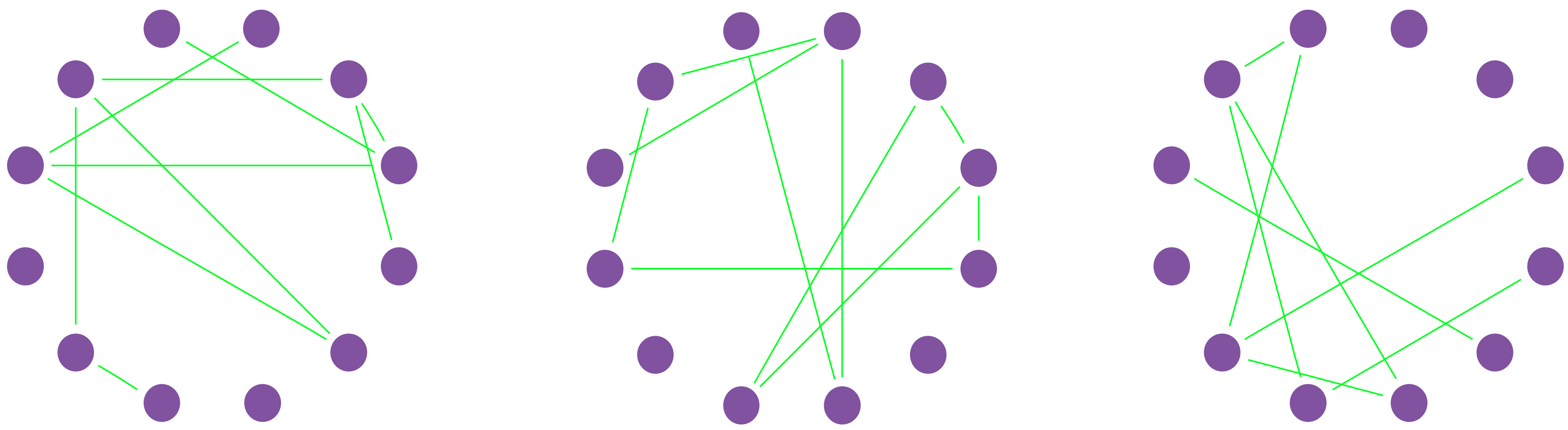
p=0.03
N=100



# The number of links is variable

# RANDOM NETWORK MODEL

p=1/6
N=12



L=8 L=10 L=7

# Number of links in a random network

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

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

$$P(L) = \underbrace{\begin{pmatrix} \begin{pmatrix} N \\ 2 \end{pmatrix} \\ L \end{pmatrix}}_{\substack{\text{Number of different ways we can choose} \\ \text{L links among all potential links.}}} p^L (1-p)^{\frac{N(N-1)}{2}-L}$$

Binomial distribution...

## MATH TUTORIAL | Binomial Distribution: The bottom line

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

$$< x >= Np$$

$$< x^2 >= p(1-p)N + p^2 N^2$$

$$\sigma_x = (< k^2 > - < k >^2)^{1/2} = [p(1-p)N]^{1/2}$$

*http://keral2008.blogspot.com/2008/10/derivation-of-mean-and-variance-of.html*

# RANDOM NETWORK MODEL

***P(L)***: the probability to have a network of exactly *L* links

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

- The average number of links <*L*> in a random graph

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

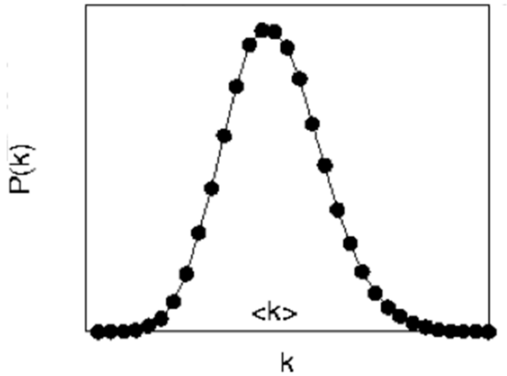
$$< k >= 2L/N = p(N-1)$$

- The standard deviation

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

# Degree distribution

# DEGREE DISTRIBUTION OF A RANDOM GRAPH




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

Select k nodes from N-1

probability of having *k* edges

probability of missing N-1-k edges

$$<k> = p(N-1) \qquad \sigma_k^2 = p(1-p)(N-1)$$

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

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>.

# DEGREE DISTRIBUTION OF A RANDOM GRAPH

$$P(k) = \binom{N-1}{k} p^k (1-p)^{(N-1)-k} \qquad\qquad <k> = p(N-1) \qquad\qquad p = \frac{<k>}{(N-1)}$$

For large N and small k, we can use the following approximations:

$$\binom{N-1}{k} = \frac{(N-1)!}{k!(N-1-k)!} = \frac{(N-1)(N-1-1)(N-1-2)...(N-1-k+1)(N-1-k)!}{k!(N-1-k)!} = \frac{(N-1)^k}{k!}$$

$$\ln[(1-p)^{(N-1)-k}] = (N-1-k)\ln(1-\frac{<k>}{N-1}) = -(N-1-k)\frac{<k>}{N-1} = -<k>(1-\frac{k}{N-1}) \cong -<k>$$

$$(1-p)^{(N-1)-k} = e^{-<k>} \qquad\qquad \ln(1+x) = \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}x^n = x - \frac{x^2}{2} + \frac{x^3}{3} - ... \quad \text{for} \quad |x| \le 1$$

$$P(k) = \binom{N-1}{k} p^k (1-p)^{(N-1)-k} = \frac{(N-1)^k}{k!} p^k e^{-<k>} = \frac{(N-1)^k}{k!}\left(\frac{<k>}{N-1}\right)^k e^{-<k>} = e^{-<k>}\frac{<k>^k}{k!}$$

# POISSON DEGREE DISTRIBUTION

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

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

$$p = \frac{<k>}{(N-1)}$$

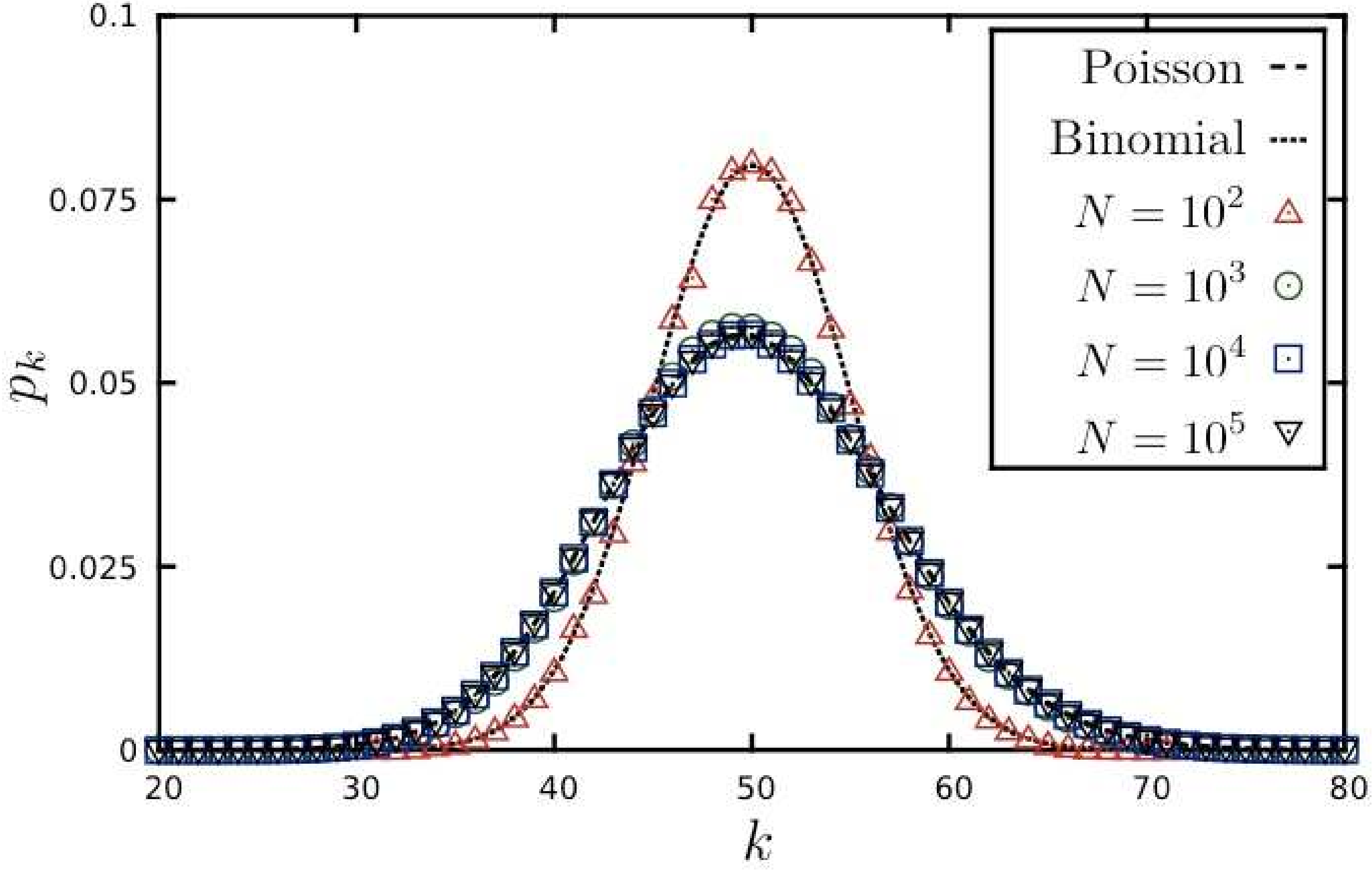
For large N and small k, we arrive to the Poisson distribution:

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

# DEGREE DISTRIBUTION OF A RANDOM GRAPH

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

<k>=50

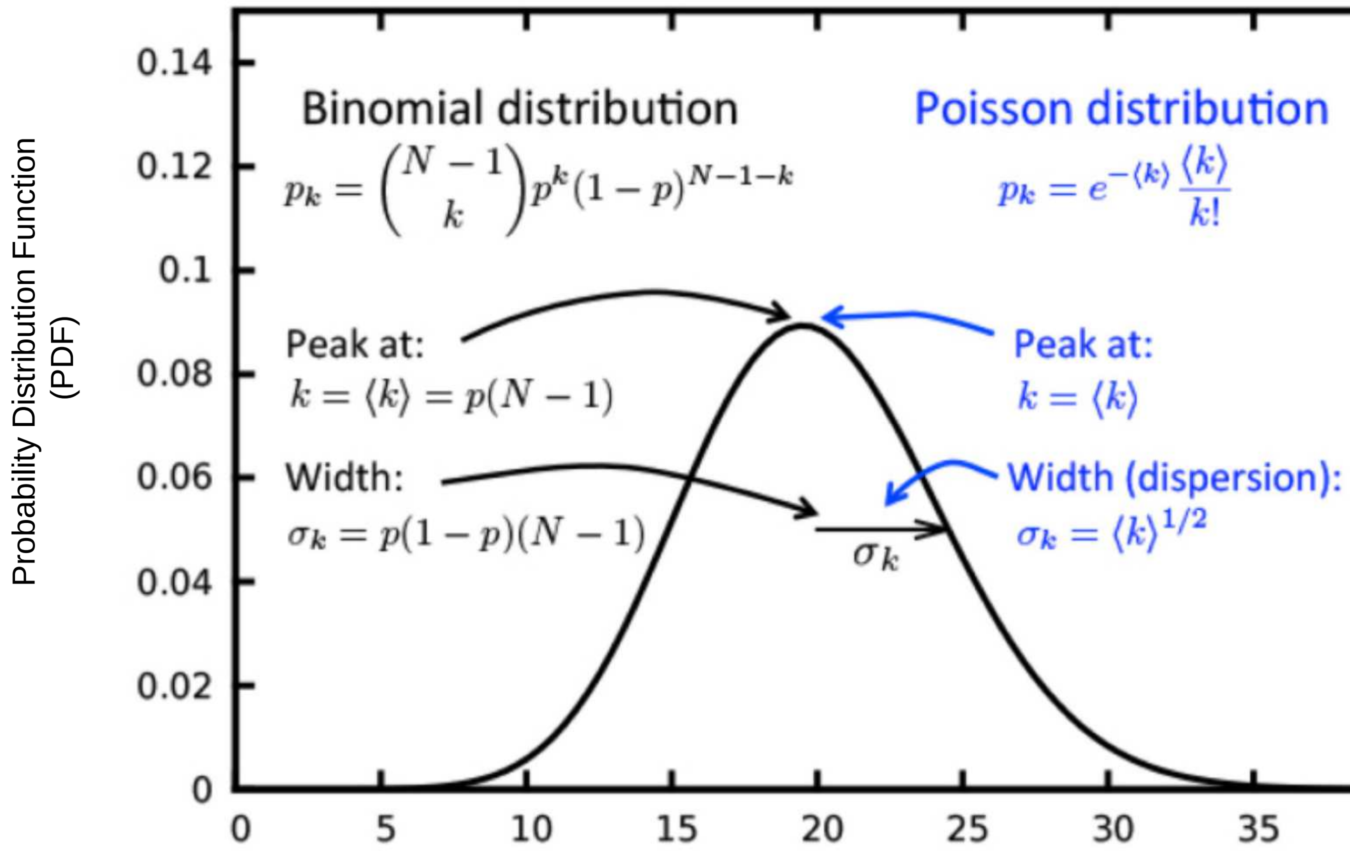



K

# DEGREE DISTRIBUTION OF A RANDOM NETWORK

**Exact Result**
-binomial distribution-

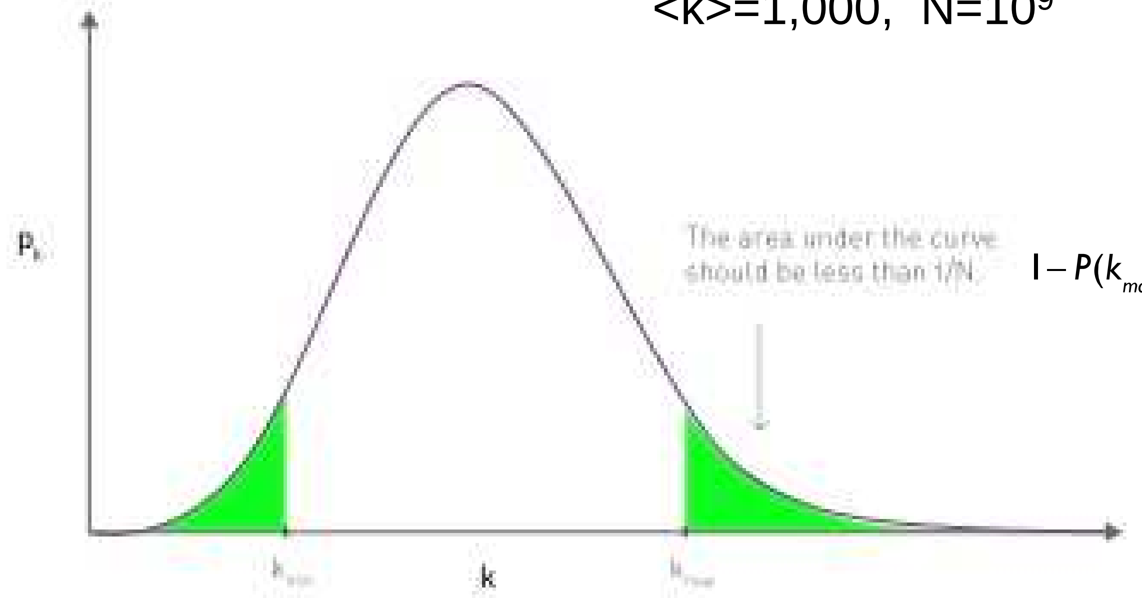
**Large N limit**
-Poisson distribution-



# Real Networks are not Poisson

## Section 12.5 Maximum and minimum degree

<k>=1,000,  N=$10^9$




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

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

<k>=1,000,  N=$10^9$

$k_{max}$=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

$$<k> \pm \sigma_k \qquad \sigma_k = <k>^{1/2}$$

$$\sigma_k = 31.62.$$

# NO OUTLIERS IN A RANDOM SOCIETY

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

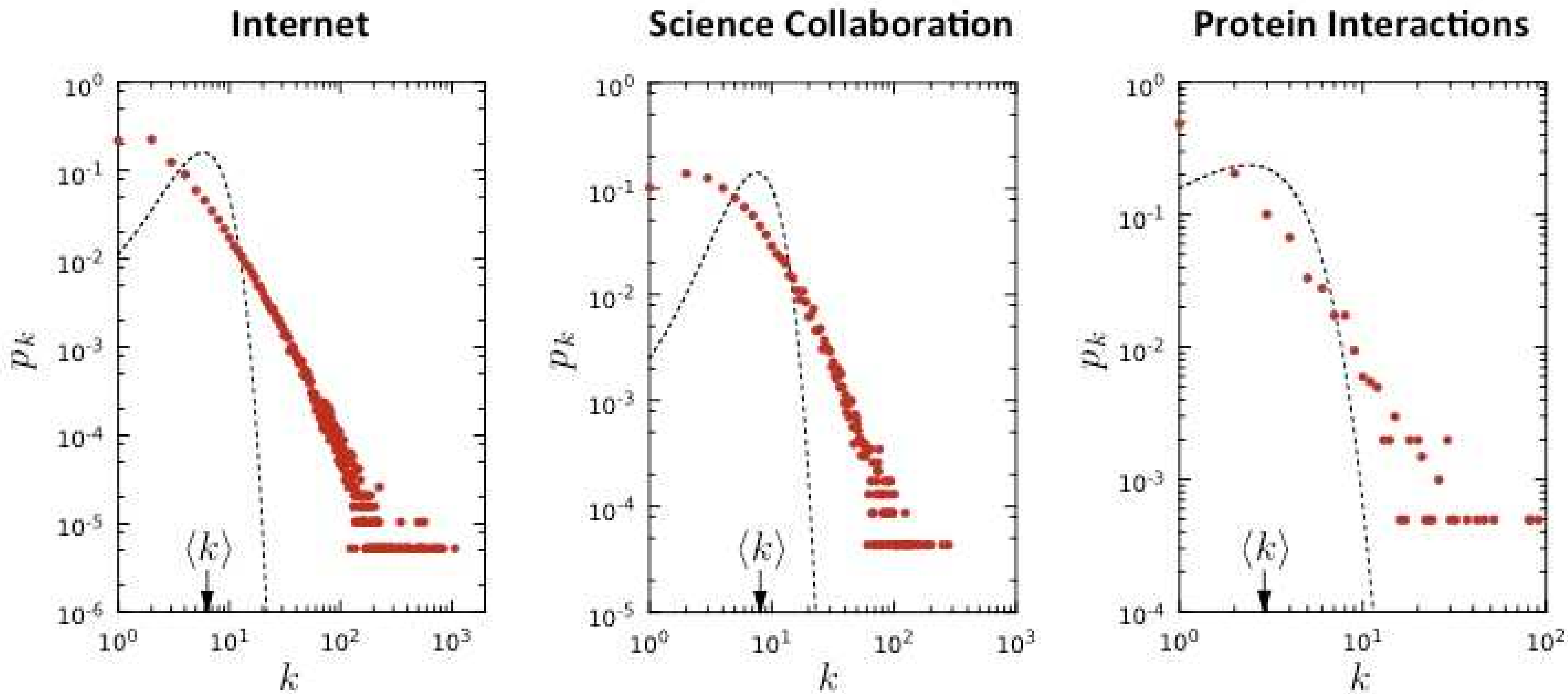
→The most connected individual has degree $k_{max}$~1,185
→The least connected individual has degree $k_{min}$ ~ 816

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 inexistent in a random society.

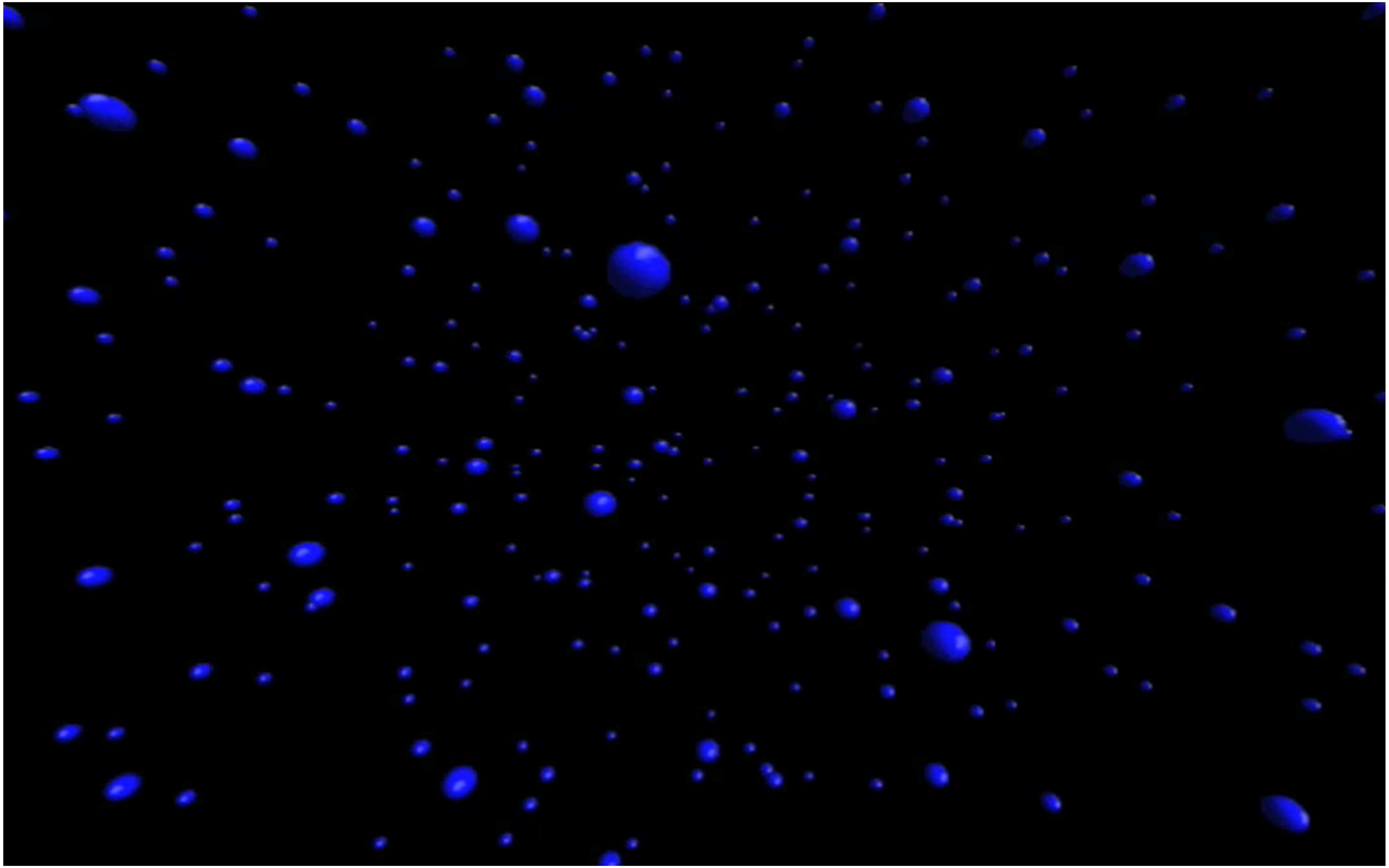
→a random society would consist of mainly average individuals, with everyone with roughly the same number of friends.
→It would lack outliers, individuals that are either highly popular or recluse.

# FACING REALITY: Degree distribution of real networks

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

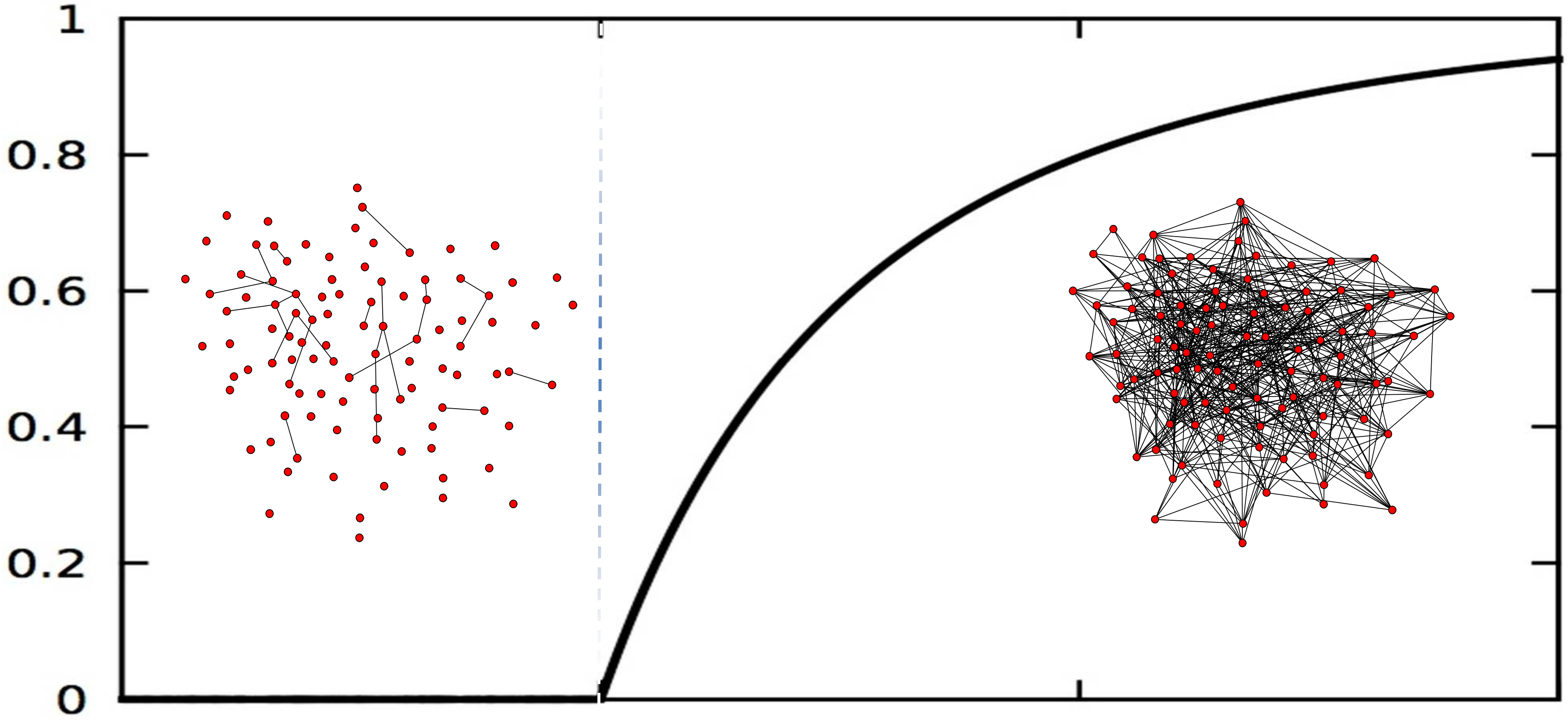


# The evolution of a random network



# EVOLUTION OF A RANDOM NETWORK

disconnected nodes ➔ **NETWORK.**




***<k>***

**How does this transition happen?**

## EVOLUTION OF A RANDOM NETWORK

disconnected nodes ➔ **NETWORK**.

***$<k_c>=1$ (Erdos and Renyi, 1959)***

The fact that at least one link per node is *necessary* to have a giant component is not unexpected. Indeed, for a giant component to exist, each of its nodes must be linked to at least one other node.

It is somewhat unexpected, however that one link is *sufficient* for the emergence of a giant component.

It is equally interesting that the emergence of the giant cluster is not gradual, but follows what physicists call a second order phase transition at $<k>=1$.

Let us denote with $u = 1 - N_G/N$ the fraction of nodes that are not in the giant component ($GC$), whose size we take to be $N_G$. If node $i$ is part of the $GC$, it must link to another node $j$, which must also be part of the $GC$. Hence if $i$ is *not* part of the $GC$, that could happen for two reasons:

- There is no link between $i$ and $j$ (probability for this is $1- p$).

- There is a link between $i$ and $j$, but $j$ is not part of the $GC$ (probability for this is $pu$).

Therefore the total probability that $i$ is not part of the $GC$ via node $j$ is $1 - p + pu$. The probability that $i$ is not linked to the $GC$ via any other node is therefore $(1 - p + pu)^{N-1}$, as there are $N - 1$ nodes that could serve as potential links to the $GC$ for node $i$. As $u$ is the fraction of nodes that do not belong to the $GC$, for any $p$ and $N$ the solution of the equation

$$u = (1 - p + pu)^{N-1} \tag{3.30}$$

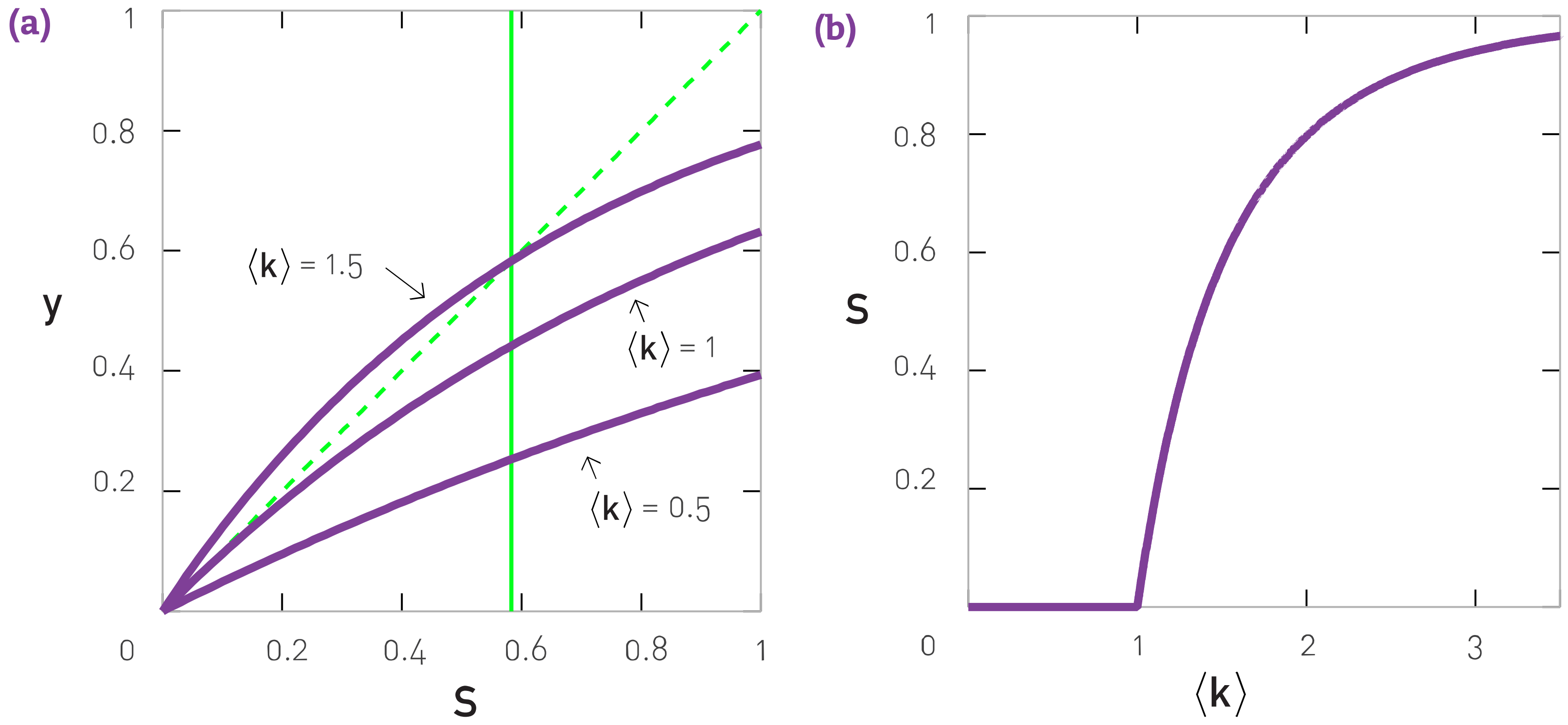
provides the size of the giant component via $N_G = N(1 - u)$. Using $p = \langle k\rangle / (N - 1)$ and taking the log of both sides, for $\langle k\rangle \ll N$ we obtain

$$\ln u \simeq (N-1)\ln\left[1 - \frac{\langle k\rangle}{N-1}(1-u)\right]. \tag{3.31}$$

Taking an exponential of both sides leads to $u = exp[- \langle k\rangle(1 - u)]$. If we denote with $S$ the fraction of nodes in the giant component, $S = N_G / N$, then $S = 1 - u$ and (3.31) results in

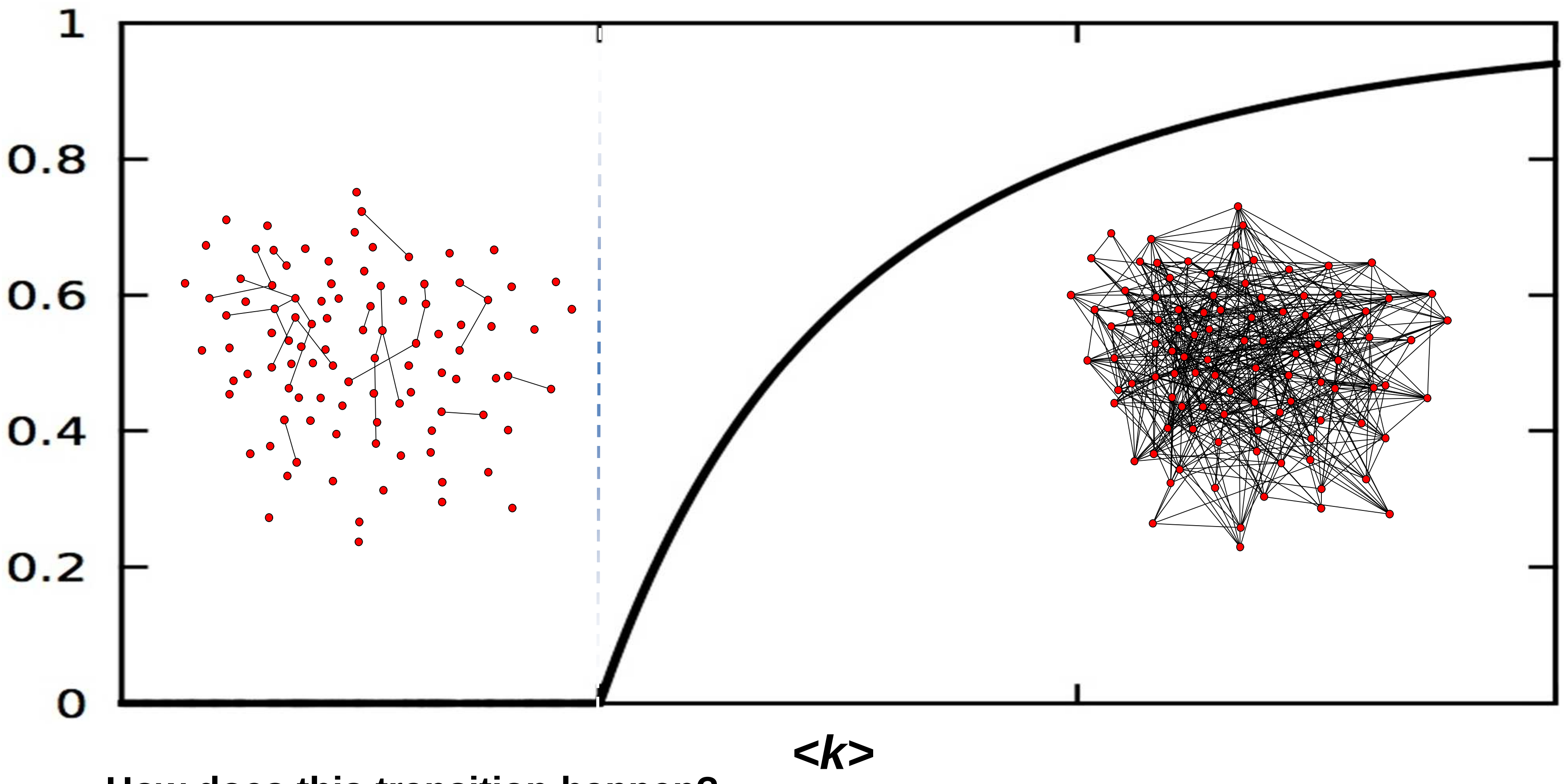
$$S = 1 - e^{-\langle k\rangle S}.$$

$$S = 1 - e^{-\langle k \rangle S}. \tag{3.32}$$



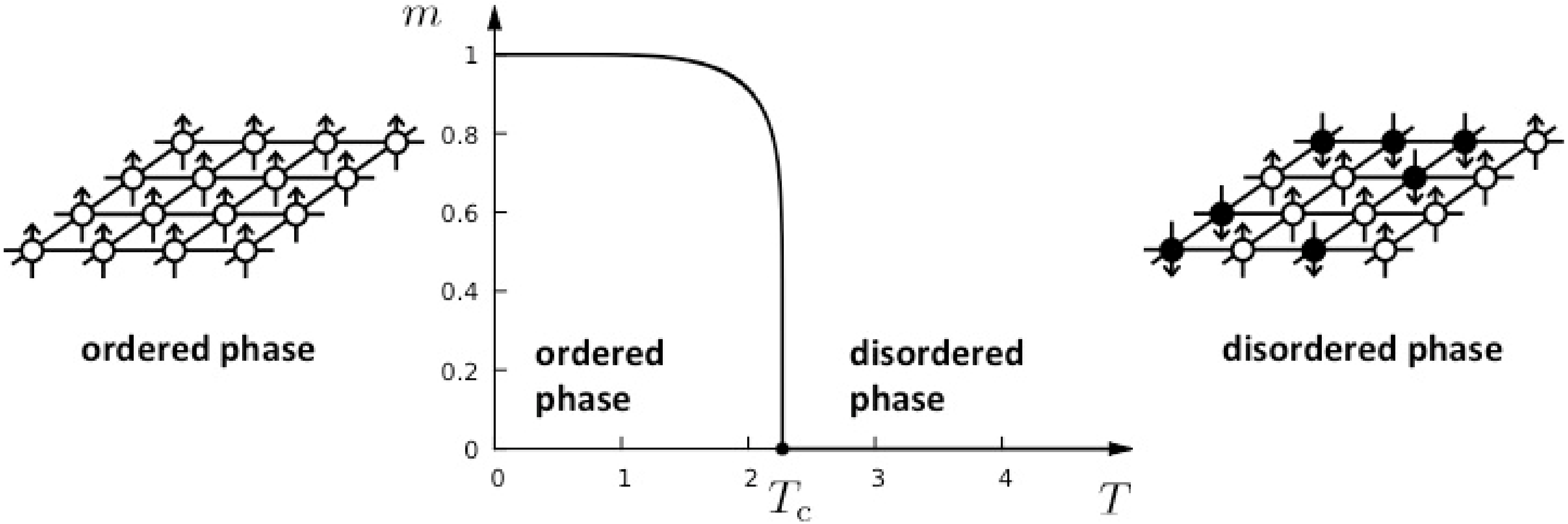
## EVOLUTION OF A RANDOM NETWORK

disconnected nodes ➔ **NETWORK.**

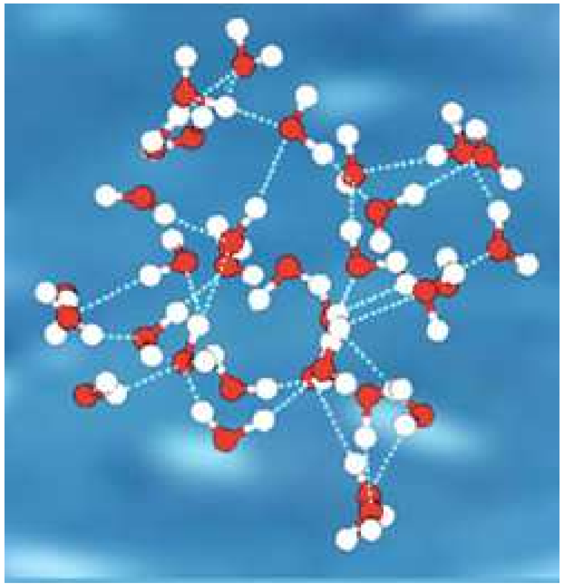



**How does this transition happen?**

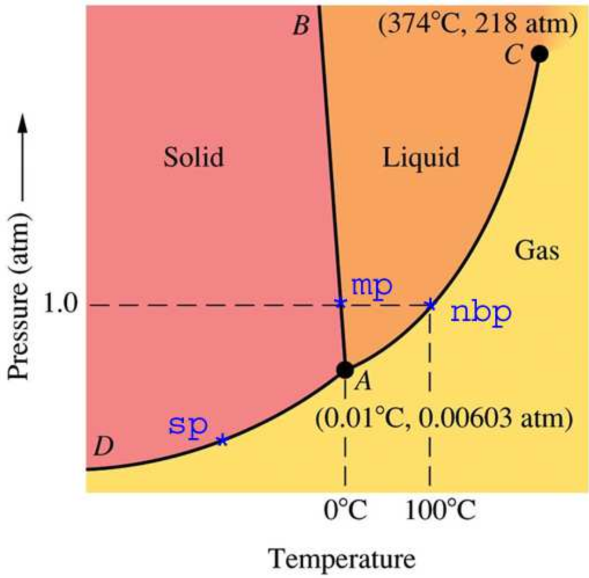
## Phase transitions in complex systems I: Magnetism

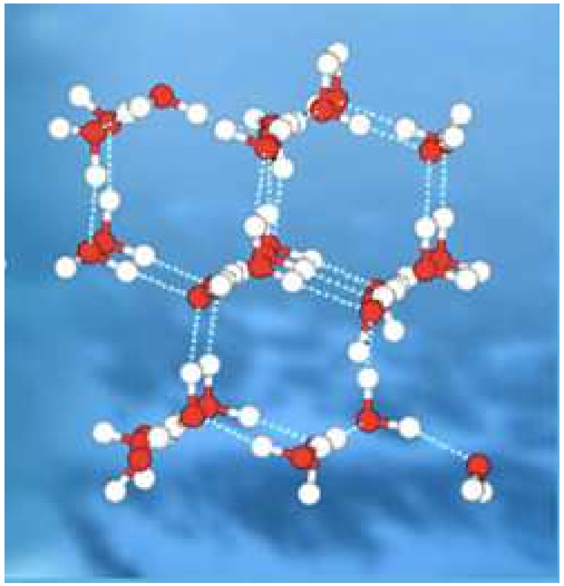


# Phase transitions in complex systems I: liquids



Water



Ice

# CLUSTER SIZE DISTRIBUTION

Probability that a randomly selected node belongs to a cluster of size *s*:

$$p(s) = \frac{e^{-<k>s}(<k>s)^{s-1}}{s!}$$

$$\langle k\rangle^{s-1} = \exp[(s-1)\ln\langle k\rangle]$$

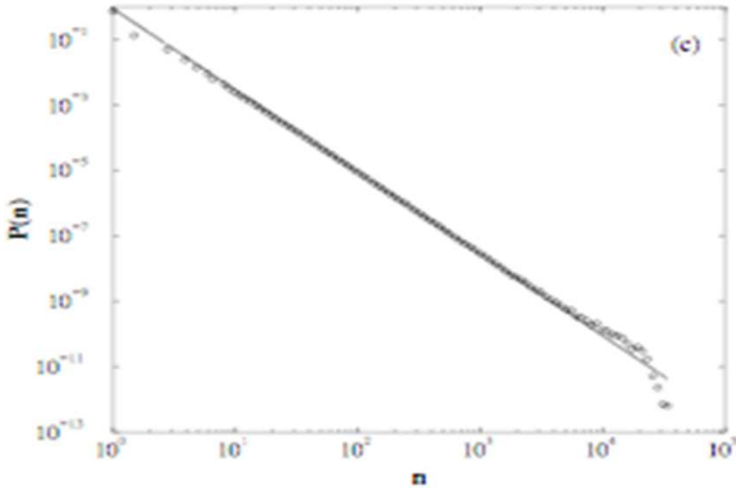
$$p(s) = \frac{s^{s-1}}{s!}e^{-\langle k\rangle s+(s-1)\ln\langle k\rangle}$$

$$s! = \sqrt{2\pi s}\left(\frac{s}{e}\right)^{s}$$

$$p(s) \sim s^{-3/2}e^{-(\langle k\rangle-1)s+(s-1)\ln\langle k\rangle}$$

At the critical point <k>=1

$$p(s) \sim s^{-3/2}$$



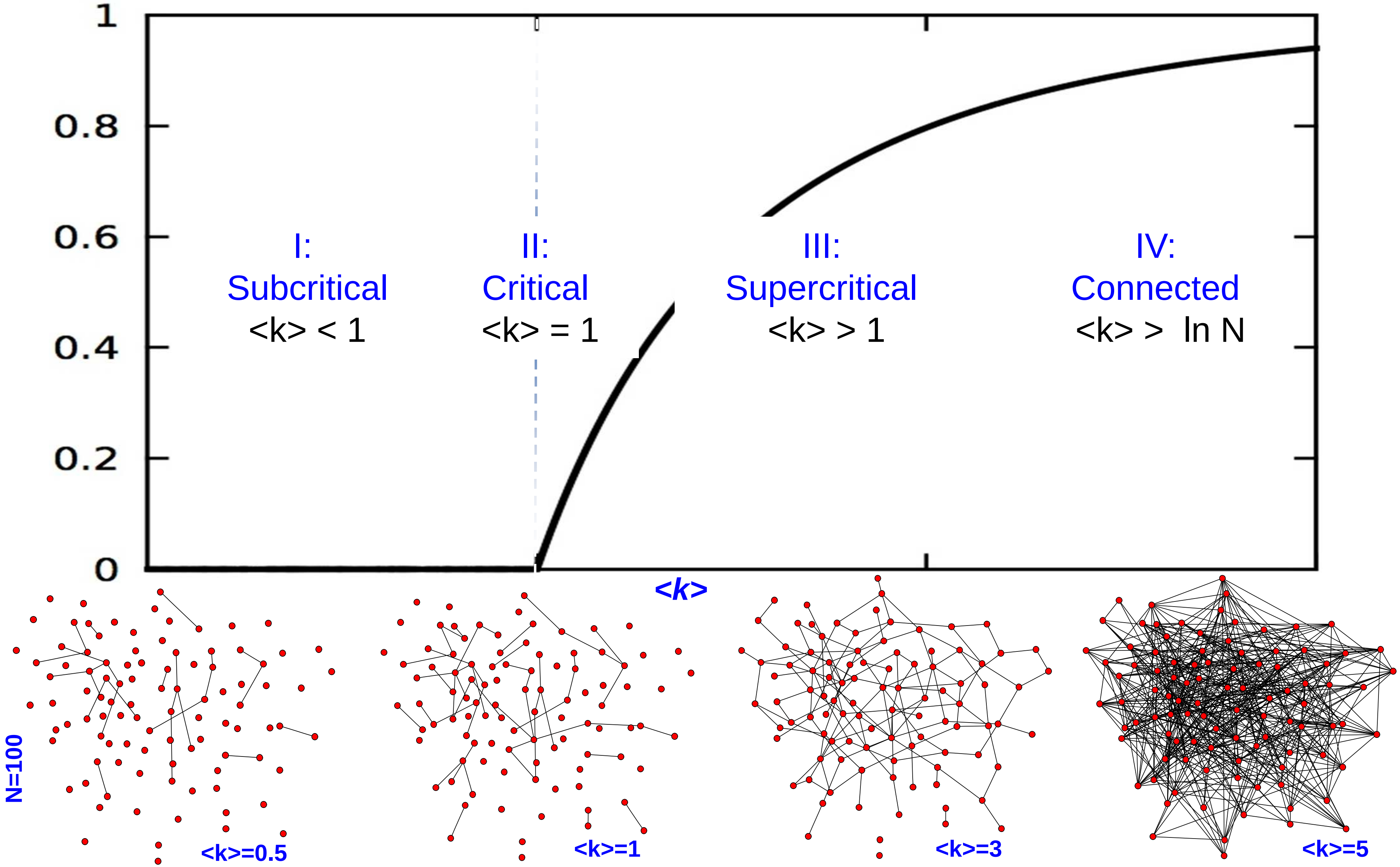

The distribution of cluster sizes at the critical point, displayed in a log-log plot. The data represent an average over 1000 systems of sizes

The dashed line has a slope of

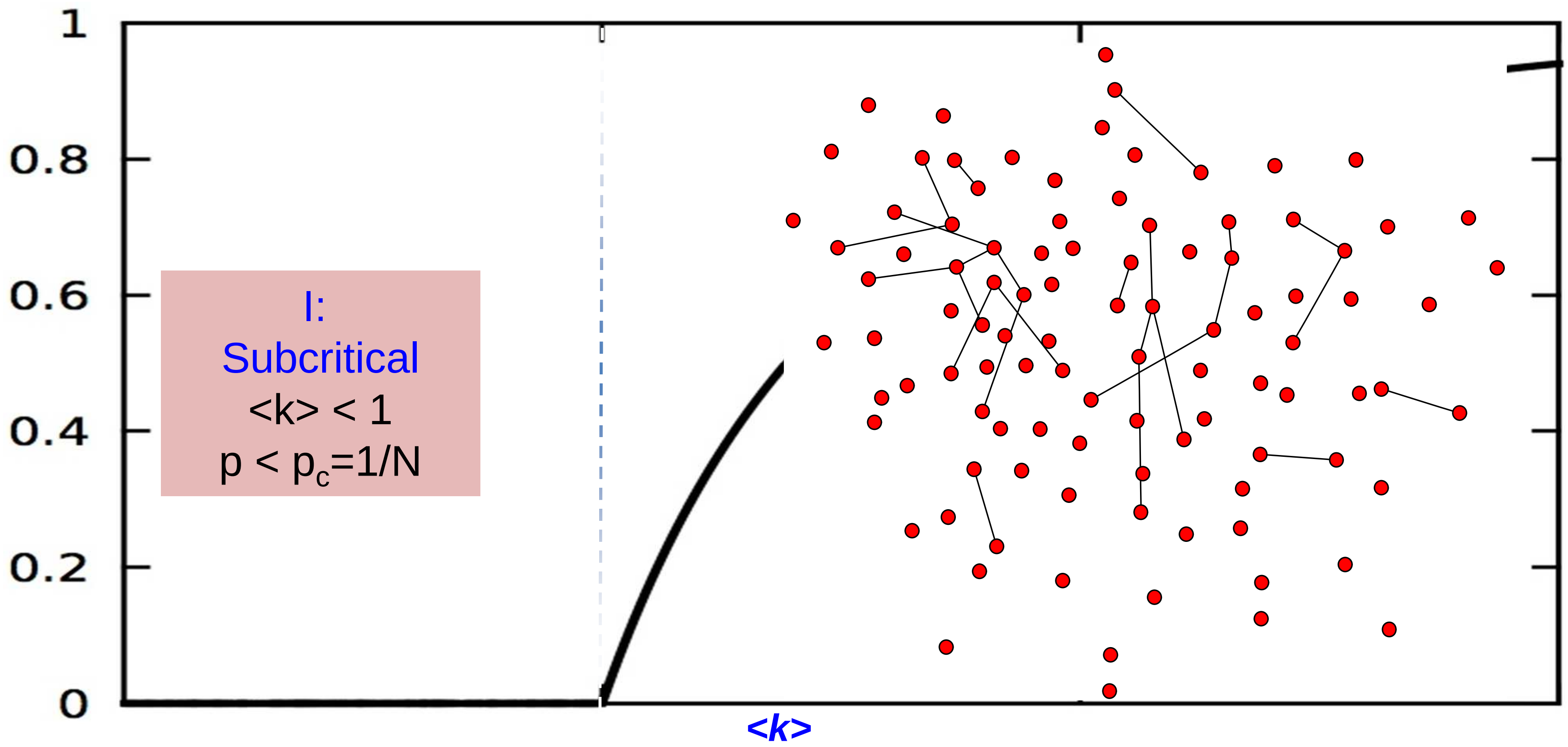
$$-\tau_n = -2.5$$



Derivation in Newman, 2010



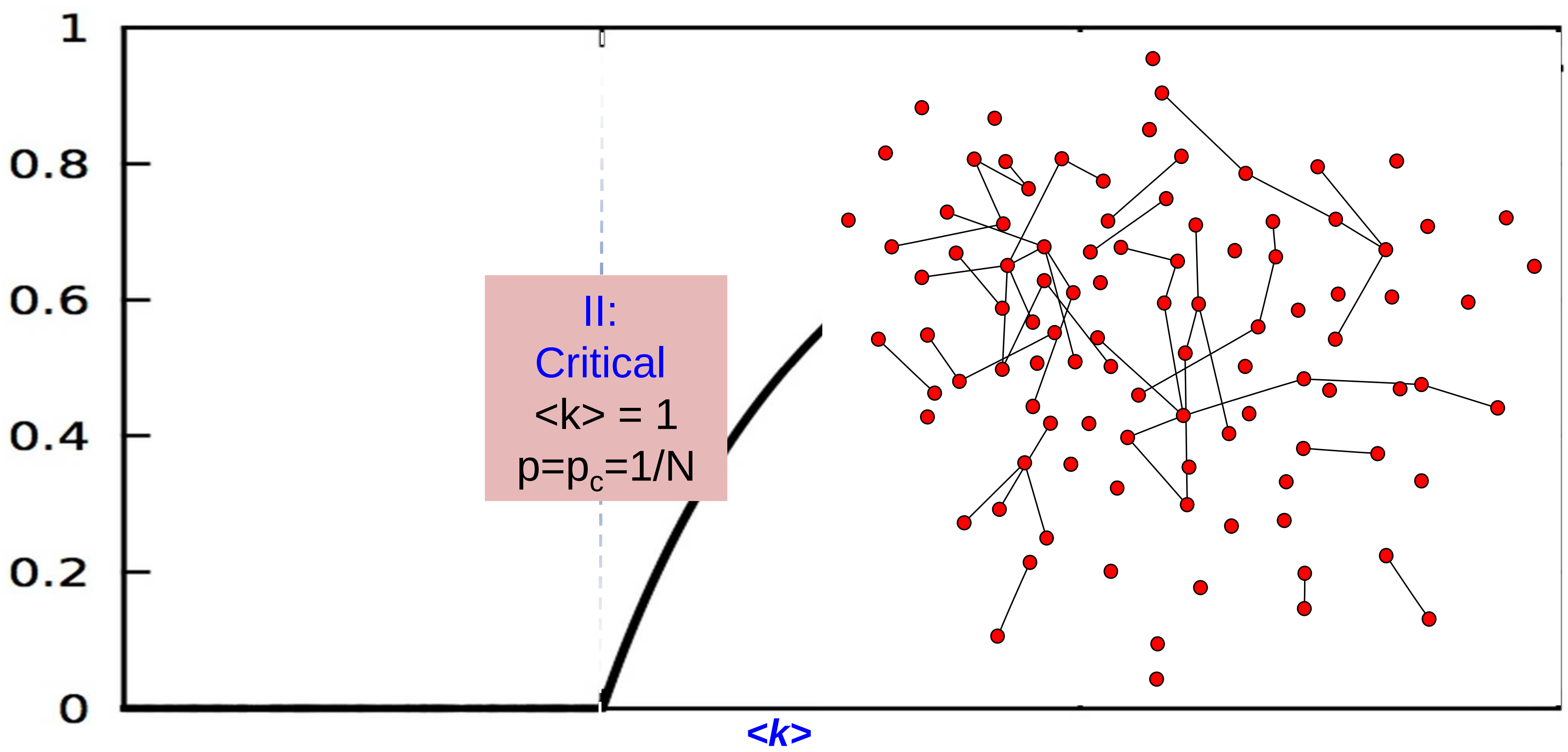
1
0.8
0.6
0.4
0.2
0
I:
Subcritical
<k> < 1
II:
Critical
<k> = 1
III:
Supercritical
<k> > 1
IV:
Connected
<k> > ln N
<k>
N=100
<k>=0.5
<k>=1
<k>=3
<k>=5




No giant component.

N-L isolated clusters, cluster size distribution is exponential $p(s) \sim s^{-3/2} e^{-(\langle k \rangle - 1)s + (s-1)\ln\langle k \rangle}$

The largest cluster is a tree, its size ~ *ln N*




Unique giant component: $N_G \sim N^{2/3}$

→contains a vanishing fraction of all nodes, $N_G/N \sim N^{-1/3}$

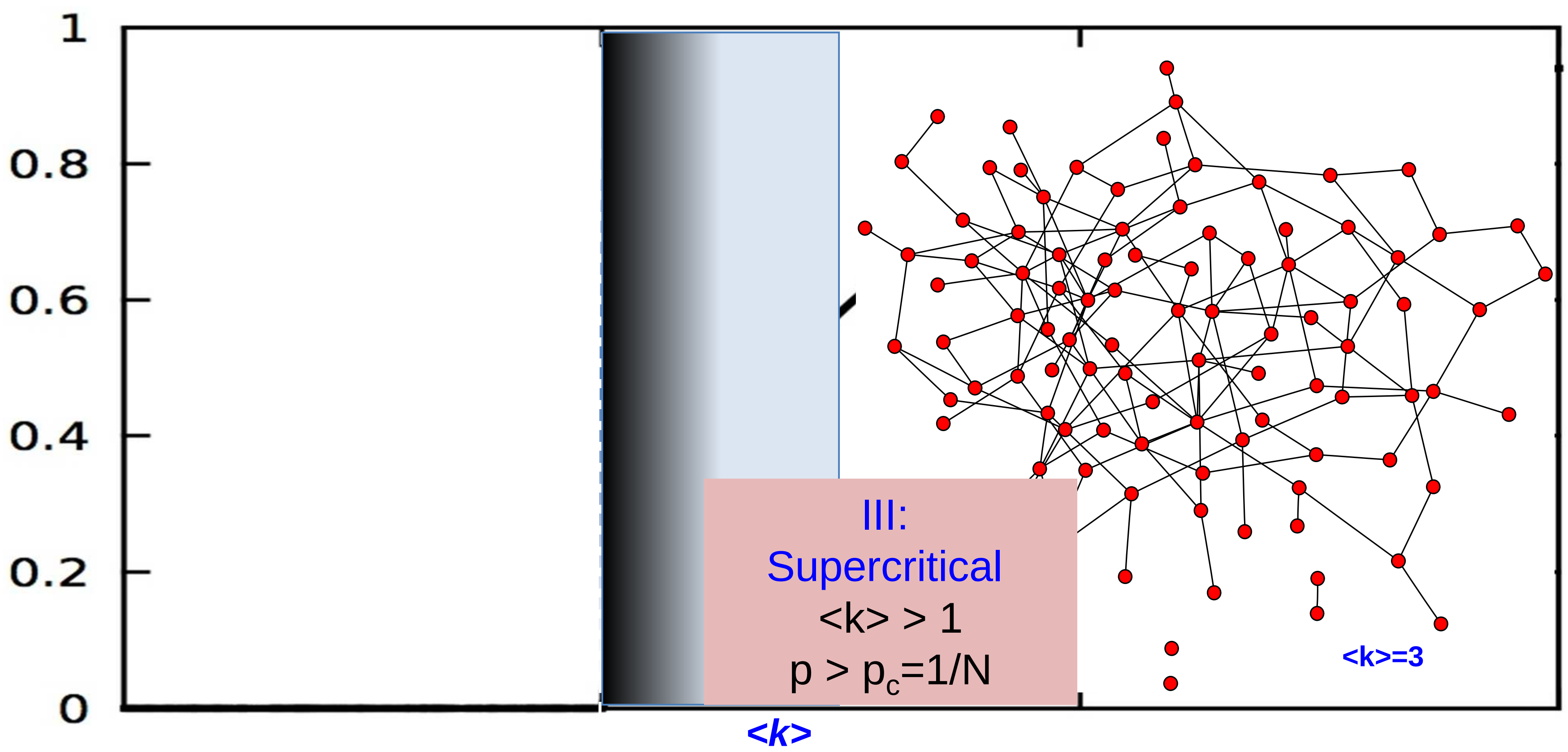
→Small components are trees, GC has loops.

Cluster size distribution: $p(s) \sim s^{-3/2}$

A jump in the cluster size:

N=1,000 → ln N~ 6.9; $N^{2/3}$~95

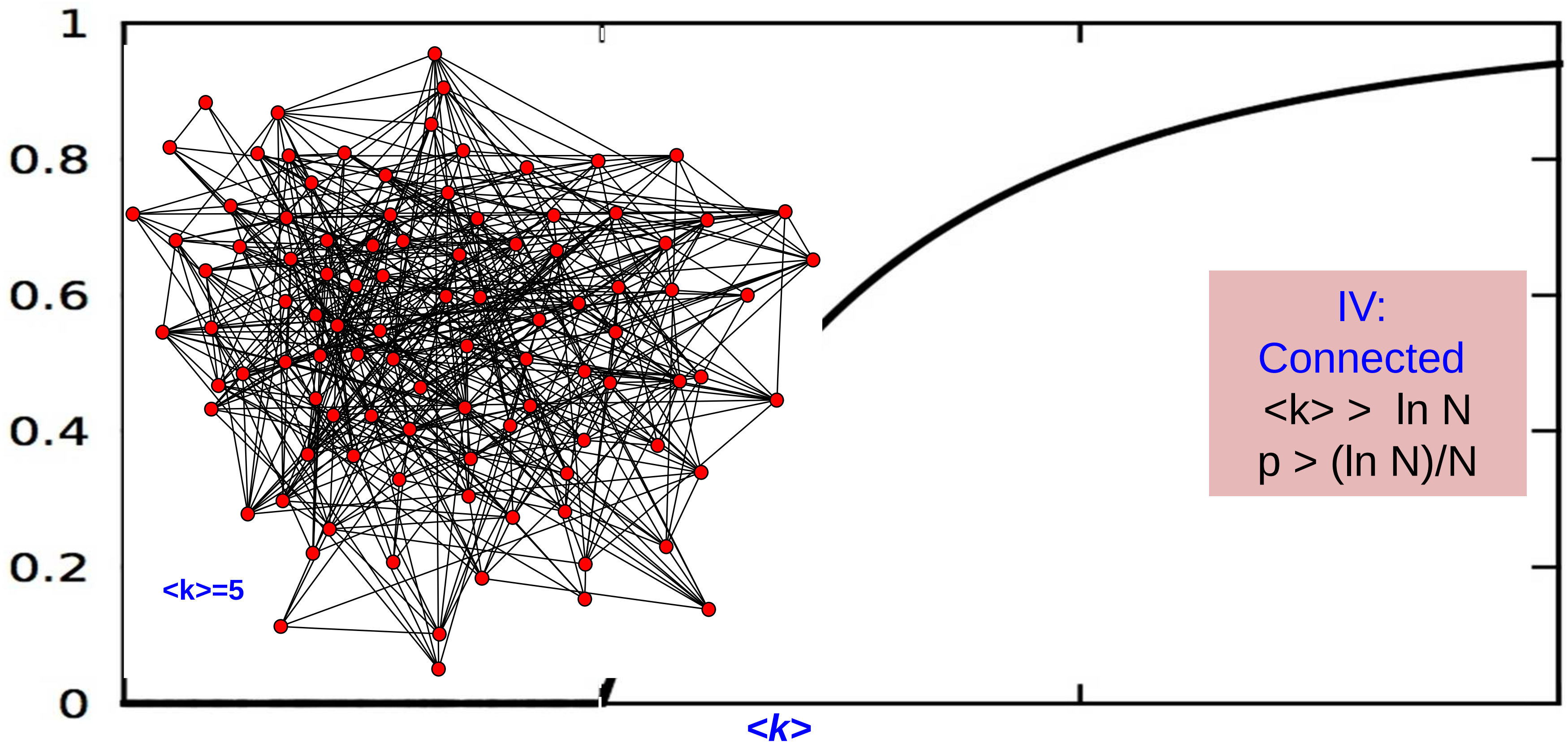
N=7 $10^9$ → ln N~ 22; $N^{2/3}$~3,659,250




Unique giant component: $N_G$~ (p-$p_c$)N

→GC has loops.

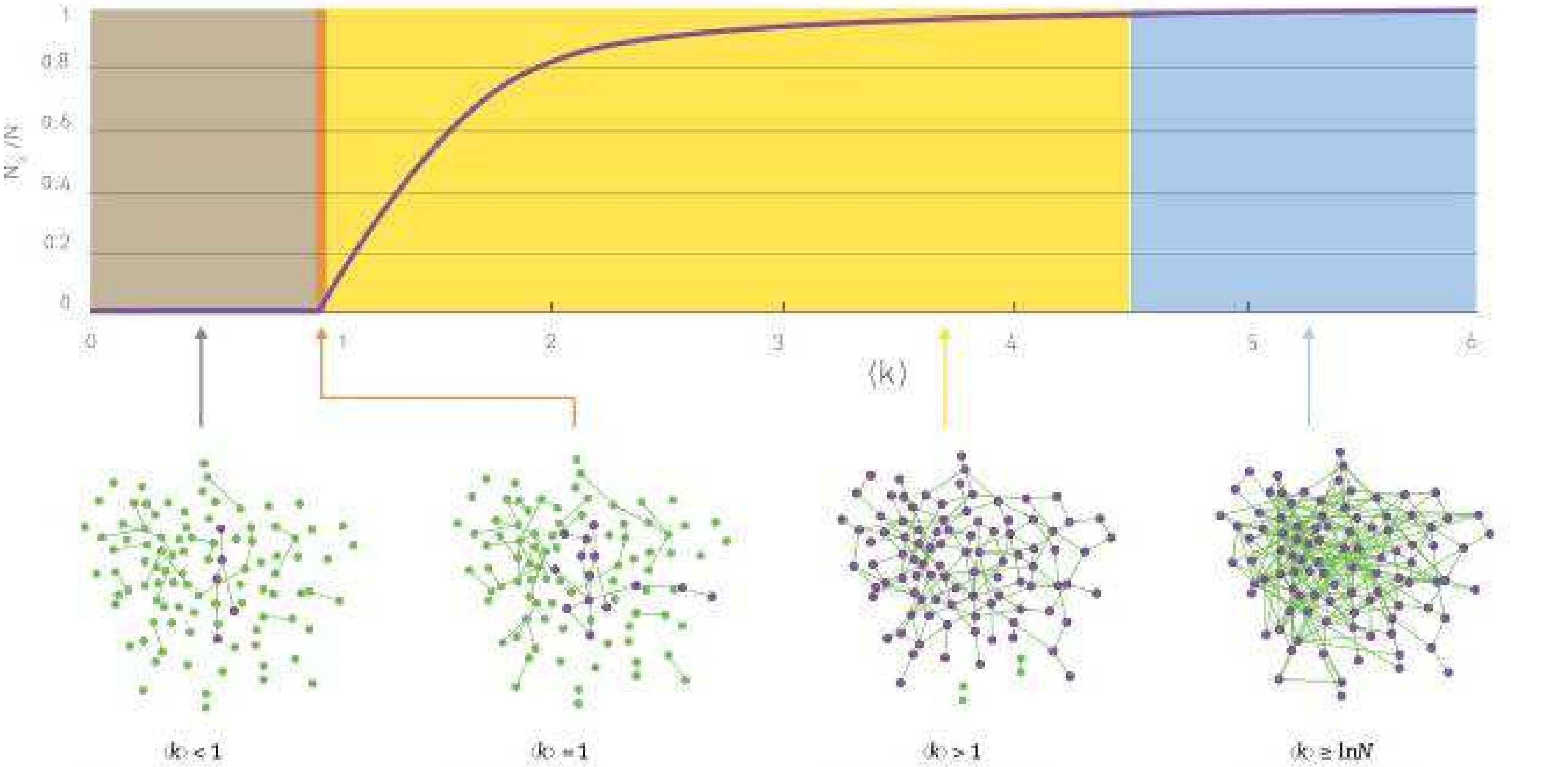
Cluster size distribution: exponential $$p(s) \sim s^{-3/2} e^{-(\langle k \rangle -1)s + (s-1)\ln\langle k \rangle}$$




Only one cluster: $N_G$=N

→GC is dense.

Cluster size distribution: None




**(b) Subcritical Regime**
- No giant component
- Cluster size distribution: $p_s \sim s^{-3/2} e^{-\alpha s}$
- Size of the largest cluster: $N_G \sim \ln N$
- The clusters are trees

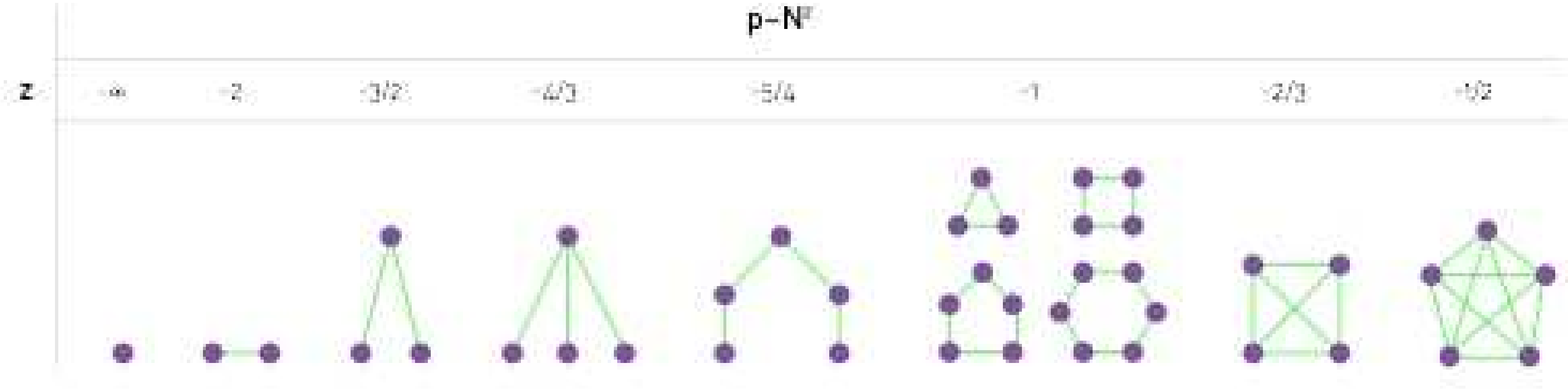
**(c) Critical Point**
- No giant component
- Cluster size distribution: $p_s \sim s^{-3/2}$
- Size of the largest cluster: $N_G \sim N^{2/3}$
- The clusters may contain loops

**(d) Supercritical Regime**
- Single giant component
- Cluster size distribution: $p_s \sim s^{-3/2} e^{-\alpha s}$
- Size of the giant component: $N_G \sim (p - p_c)N$
- The small clusters are trees
- Giant component has loops

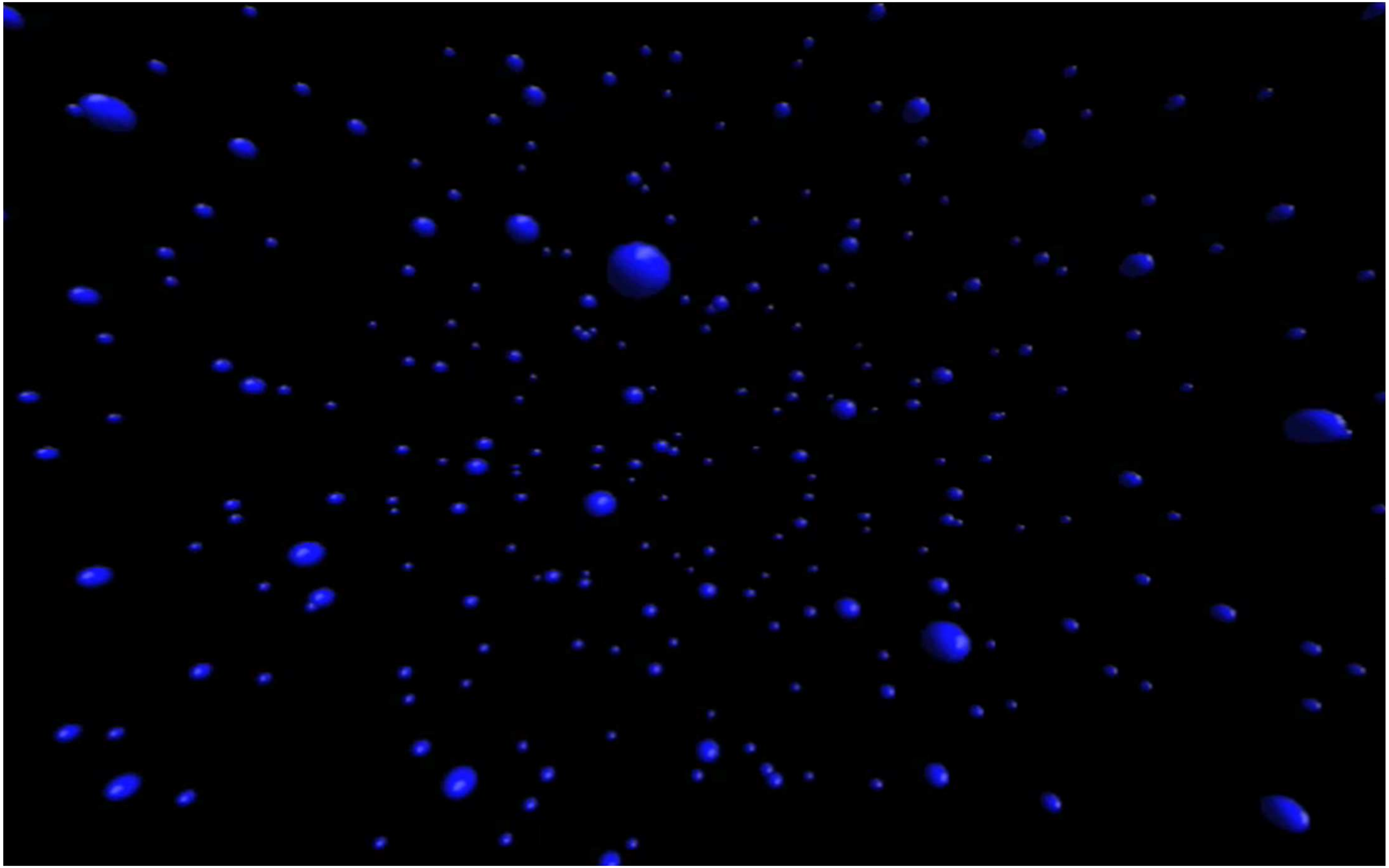
**(e) Connected Regime**
- Single giant component
- No isolated nodes or clusters
- Size of the giant component: $N_G = N$
- Giant component has loops

# Network evolution in graph theory

A graph has a given property $Q$ if the probability of having $Q$ approaches 1 as $N \to \infty$. That is, for a given $z$ either almost every graph has the property $Q$ or almost no graph has it. For example, for $z$ less

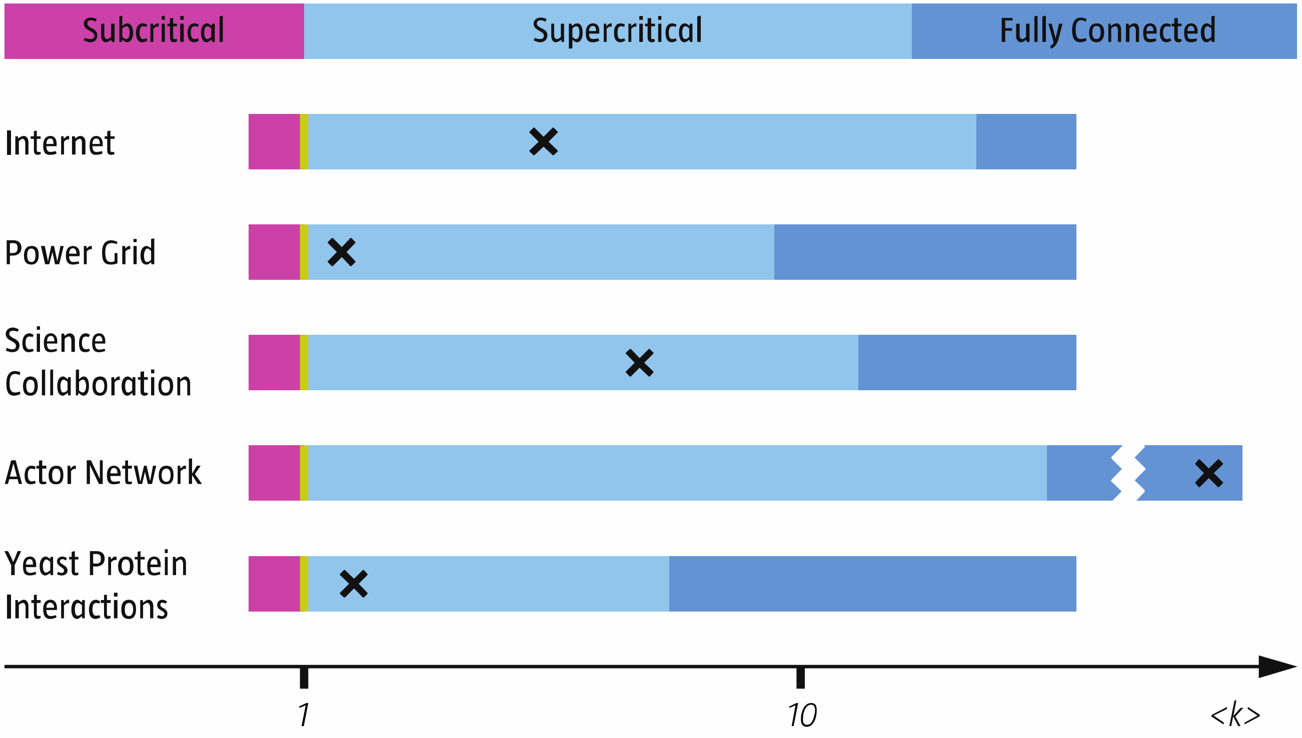



$$p = \langle k \rangle / (N-1)$$



# Real networks are supercritical

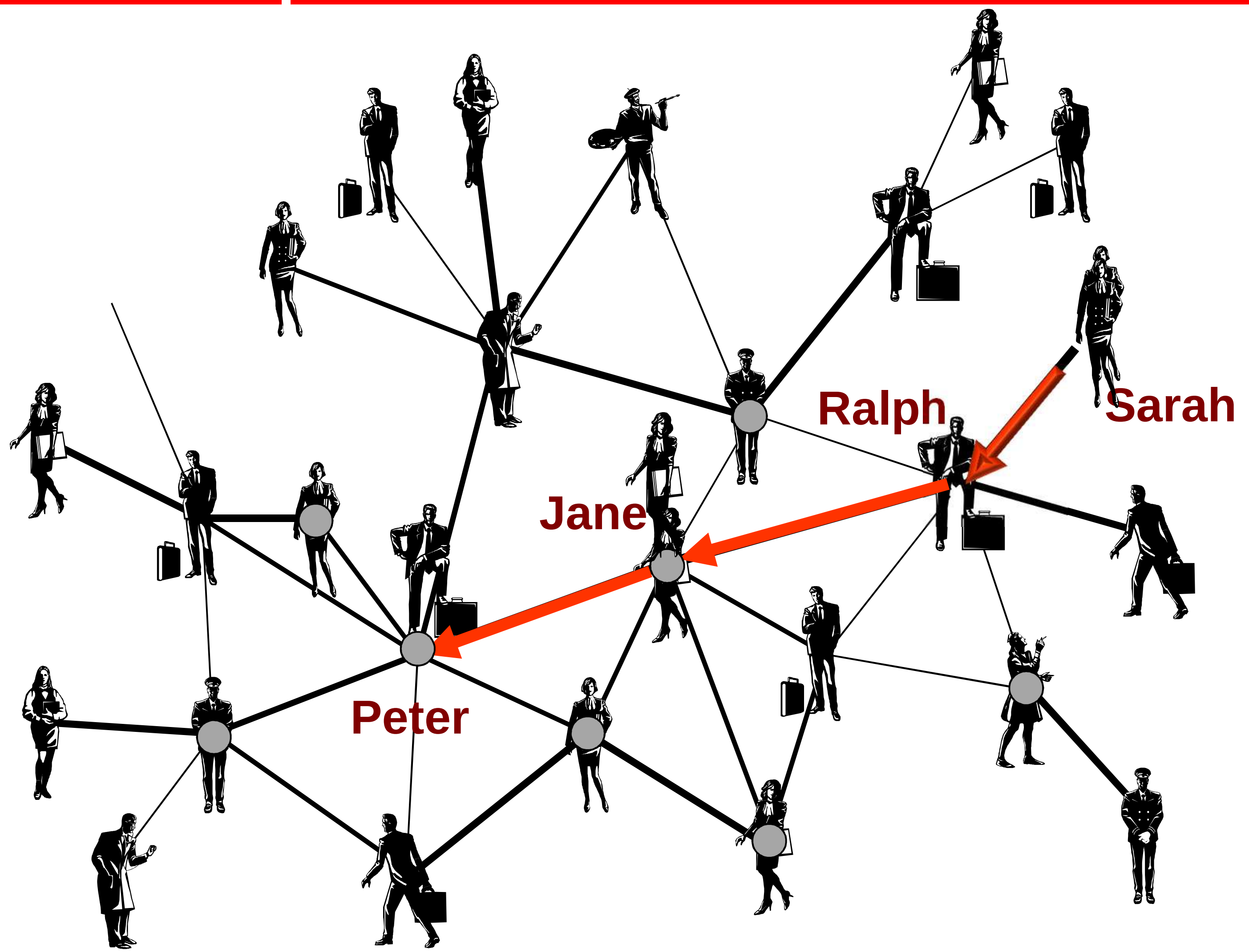
## Section 13.2




| Network | $N$ | $L$ | $\langle k \rangle$ | $\ln N$ |
|---|---|---|---|---|
| Internet | 192,244 | 609,066 | 6.34 | 12.17 |
| Power Grid | 4,941 | 6,594 | 2.67 | 8.51 |
| Science Collaboration | 23,133 | 186,936 | 8.08 | 10.04 |
| Actor Network | 212,250 | 3,054,278 | 28.78 | 12.27 |
| Yeast Protein Interactions | 2,018 | 2,930 | 2.90 | 7.61 |

# Small worlds

## SIX DEGREES | small worlds




*Frigyes Karinthy, 1929*
*Stanley Milgram, 1967*

# SIX DEGREES — 1929: Frigyes Kartinthy

*Frigyes Karinthy (1887-1938)*
*Hungarian Writer*

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.”

"The worker knows the manager in the shop, who knows Ford; Ford is on friendly terms with the general director of Hearst Publications, who last year became good friends with Arpad Pasztor, someone I not only know, but to the best of my knowledge a good friend of mine. So I could easily ask him to send a telegram via the general director telling Ford that he should talk to the manager and have the worker in the shop quickly hammer together a car for me, as I happen to need one."

## SIX DEGREES | 1967: Stanley Milgram

HOW TO TAKE PART IN THIS STUDY

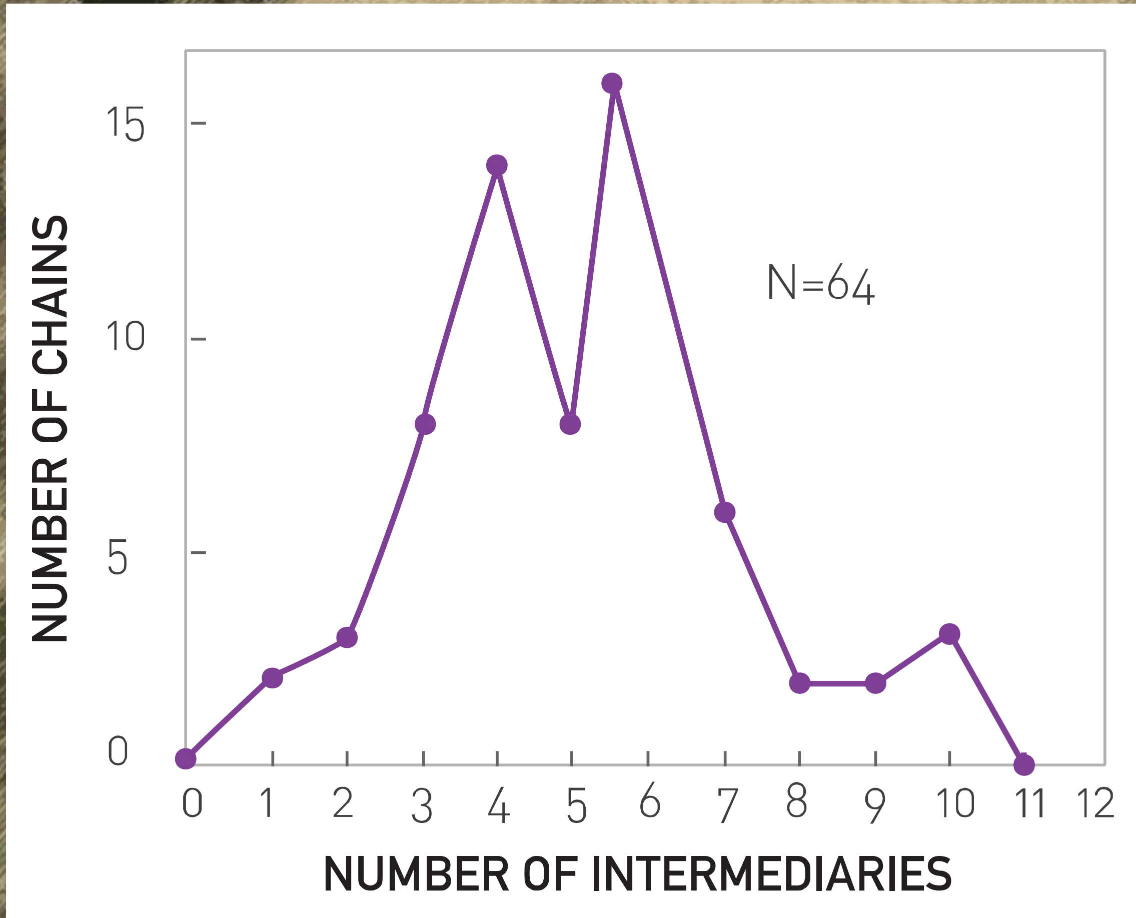
1. ADD YOUR NAME TO THE ROSTER AT THE BOTTOM OF THIS SHEET, so that the next person who receives this letter will know who it came from.

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.

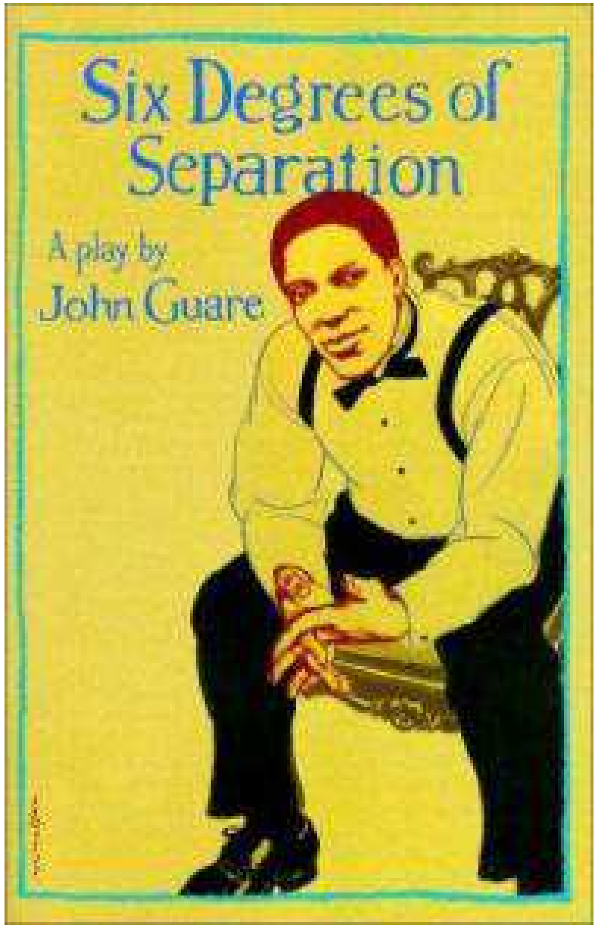
3. IF YOU KNOW THE TARGET PERSON ON A PERSONAL BASIS, MAIL THIS FOLDER DIRECTLY TO HIM (HER). Do this only if you have previously met the target person and know each other on a first name basis.

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.

## SIX DEGREES | 1967: Stanley Milgram



## 1991: John Guare




"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."

# WWW: 19 DEGREES OF SEPARATION

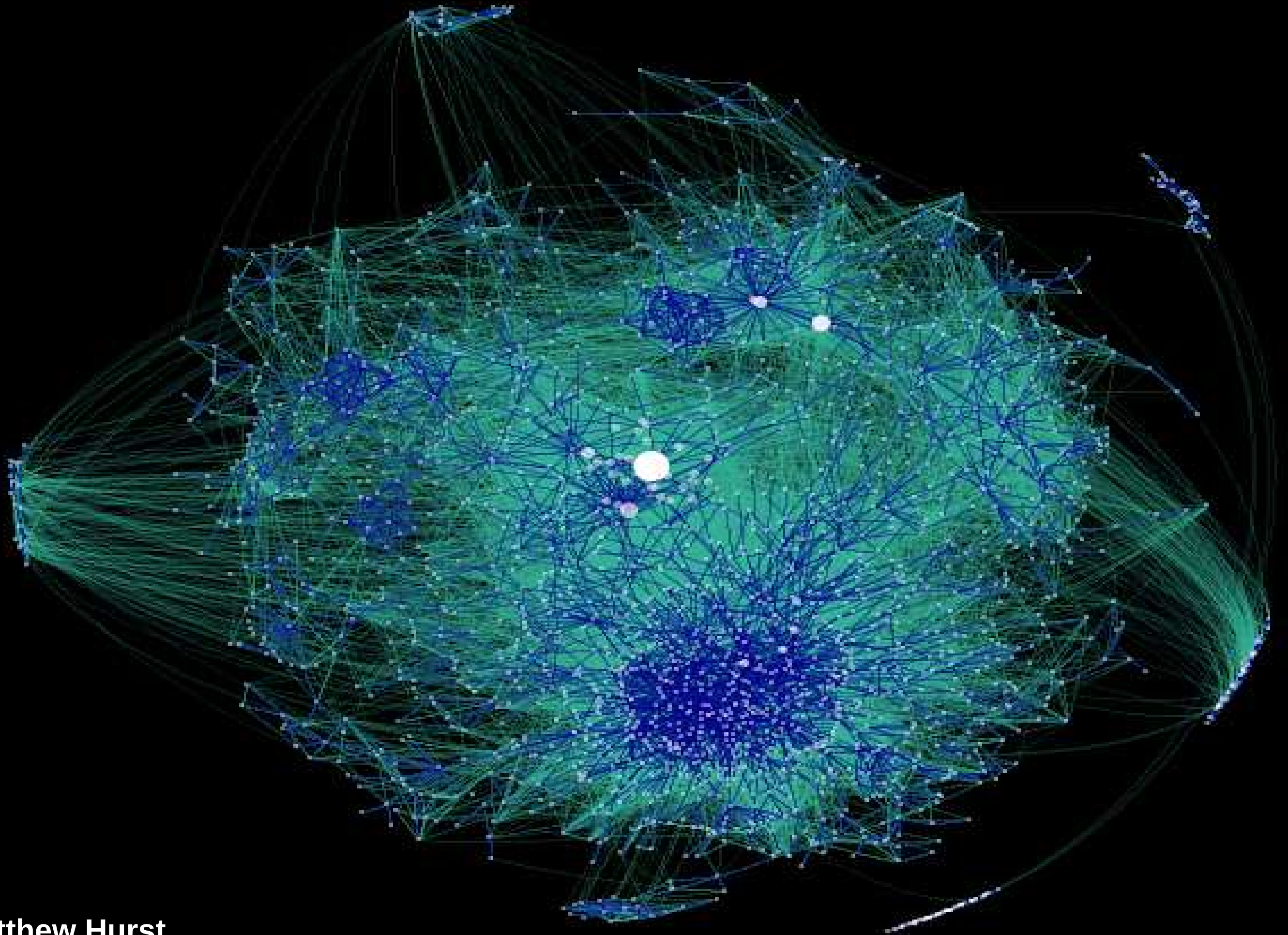
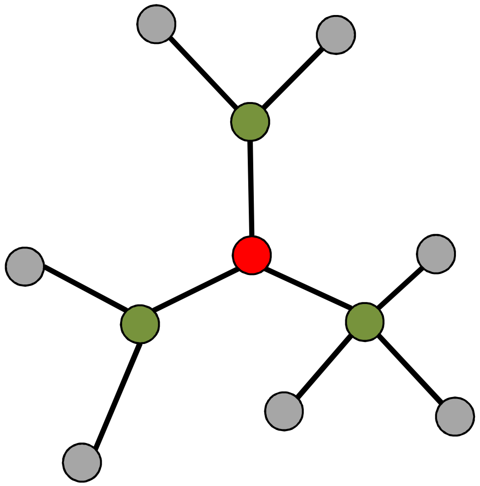


Image by **Matthew Hurst**
*Blogosphere*

# DISTANCES IN RANDOM GRAPHS

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



<$k$> nodes at distance one ($d$=1).

<$k$>$^2$ nodes at distance two ($d$=2).

<$k$>$^3$ nodes at distance three ($d$ =3).

...

<$k$>$^d$ nodes at distance $d$.

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

# DISTANCES IN RANDOM GRAPHS

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

In most networks this offers a better approximation to the average distance between two randomly chosen nodes, ⟨d⟩, than to $d_{max}$ .

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

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

# Average Degree

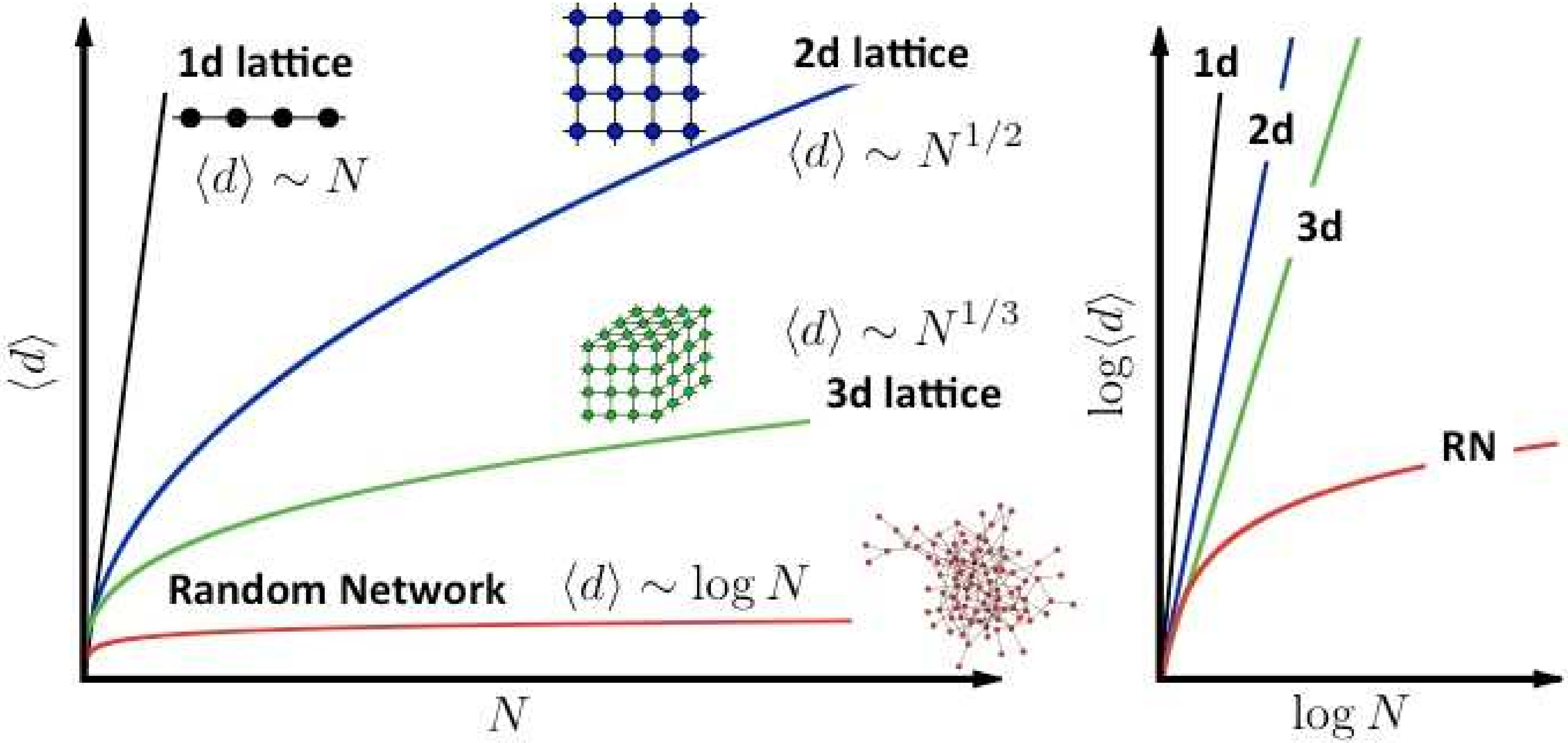
| NETWORK | NODES | LINKS | DIRECTED UNDIRECTED | N | L | ⟨k⟩ |
|---|---|---|---|---|---|---|
| **Internet** | Routers | Internet connections | Undirected | 192,244 | 609,066 | 6.33 |
| **WWW** | Webpages | Links | Directed | 325,729 | 1,497,134 | 4.60 |
| **Power Grid** | Power plants, transformers | Cables | Undirected | 4,941 | 6,594 | 2.67 |
| **Mobile Phone Calls** | Subscribers | Calls | Directed | 36,595 | 91,826 | 2.51 |
| **Email** | Email addresses | Emails | Directed | 57,194 | 103,731 | 1.81 |
| **Science Collaboration** | Scientists | Co-authorship | Undirected | 23,133 | 93,439 | 8.08 |
| **Actor Network** | Actors | Co-acting | Undirected | 702,388 | 29,397,908 | 83.71 |
| **Citation Network** | Paper | Citations | Directed | 449,673 | 4,689,479 | 10.43 |
| **E. Coli Metabolism** | Metabolites | Chemical reactions | Directed | 1,039 | 5,802 | 5.58 |
| **Protein Interactions** | Proteins | Binding interactions | Undirected | 2,018 | 2,930 | 2.90 |

## DISTANCES IN RANDOM GRAPHS compare with real data

| NETWORK | $N$ | $L$ | $\langle k \rangle$ | $\langle d \rangle$ | $d_{max}$ | $\frac{\ln N}{\ln \langle k \rangle}$ |
|---|---|---|---|---|---|---|
| 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 |
| Email | 57,194 | 103,731 | 1.81 | 5.88 | 18 | 18.4 |
| Science Collaboration | 23,133 | 93,439 | 8.08 | 5.35 | 15 | 4.81 |
| Actor Network | 702,388 | 29,397,908 | 83.71 | 3.91 | 14 | 3.04 |
| Citation Network | 449,673 | 4,707,958 | 10.43 | 11.21 | 42 | 5.55 |
| E. Coli Metabolism | 1,039 | 5,802 | 5.58 | 2.98 | 8 | 4.04 |
| Protein Interactions | 2,018 | 2,930 | 2.90 | 5.61 | 14 | 7.14 |

Given the huge differences in scope, size, and average degree, the agreement is excellent.

## Why are small worlds surprising? Suprising compared to what?



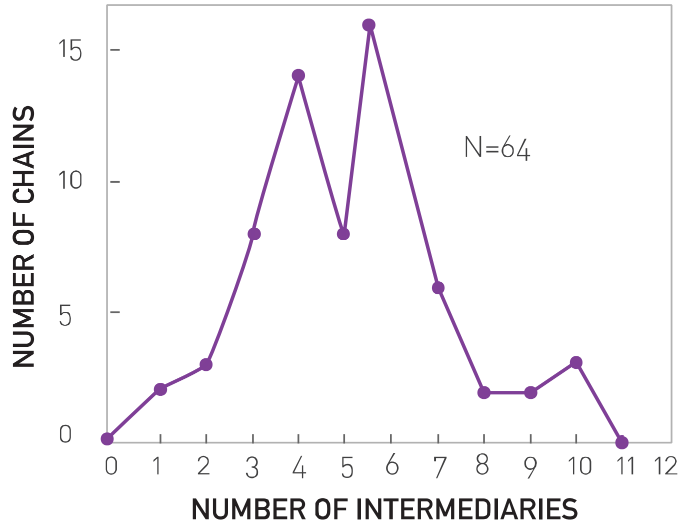
## Three, Four or Six Degrees?

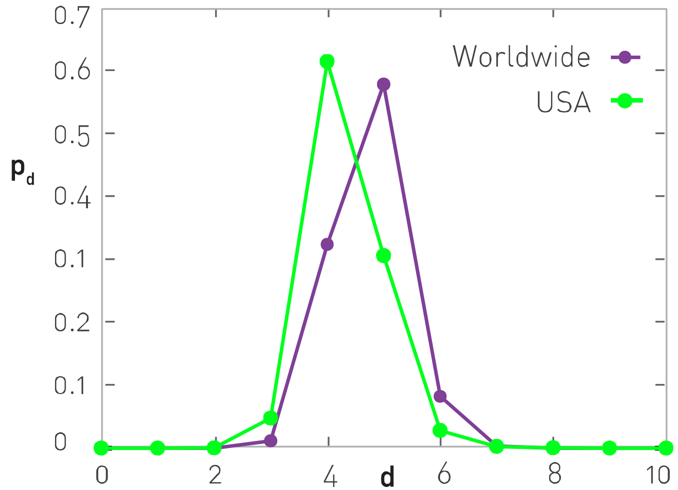
For the globe's social networks:

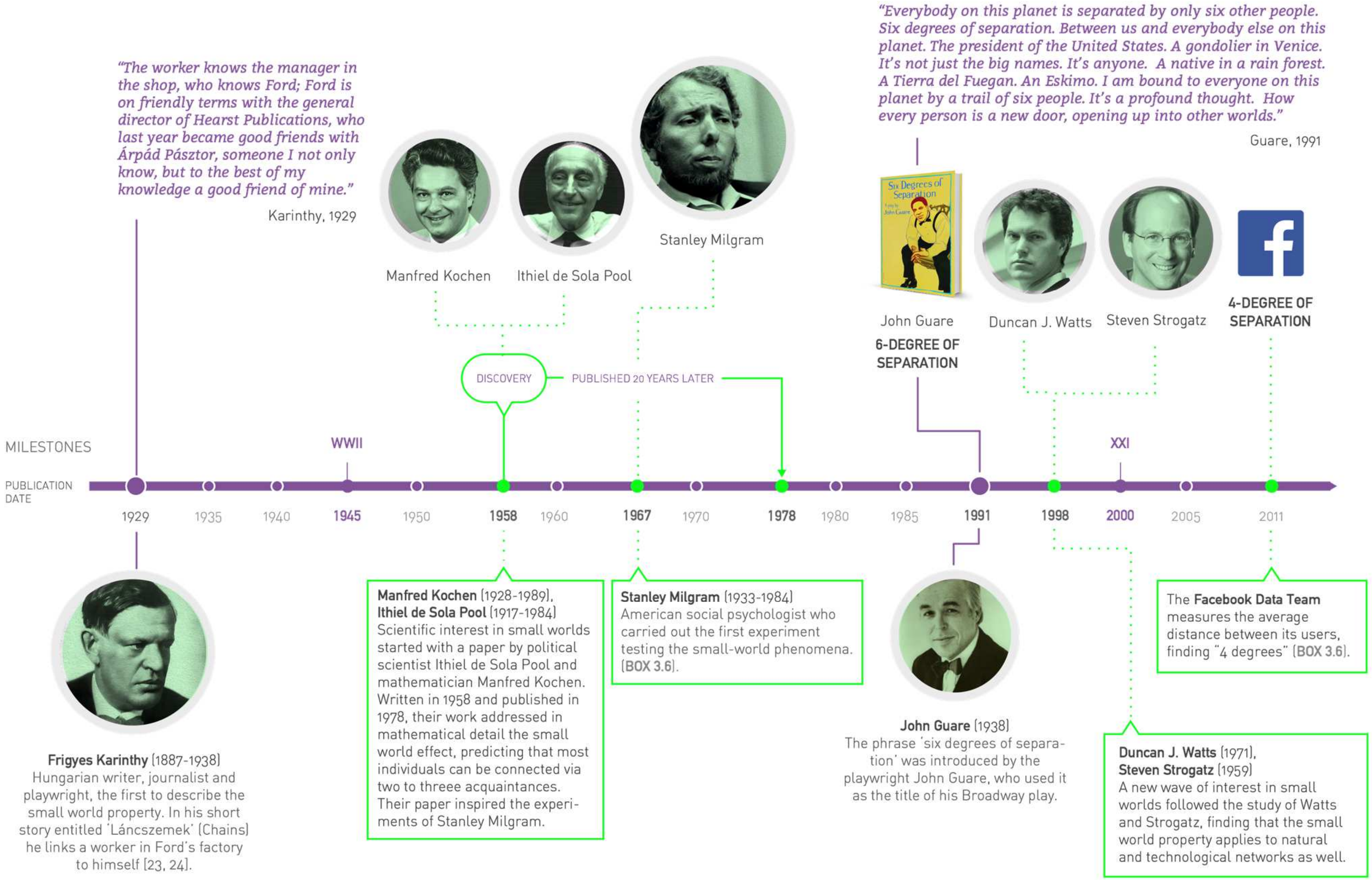
$\langle k \rangle \simeq 10^3$

$N \simeq 7 \times 10^9$ for the world's population.

$$< d >= \frac{\ln(N)}{\ln\langle k \rangle} = 3.28$$



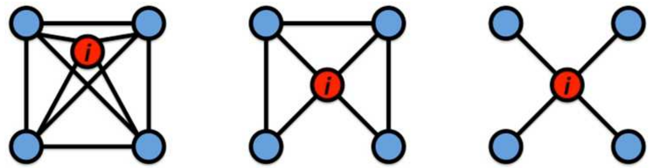




"The worker knows the manager in the shop, who knows Ford; Ford is on friendly terms with the general director of Hearst Publications, who last year became good friends with Árpád Pásztor, someone I not only know, but to the best of my knowledge a good friend of mine."
Karinthy, 1929
Manfred Kochen
Ithiel de Sola Pool
Stanley Milgram
"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."
Guare, 1991
Six Degrees of Separation
John Guare
6-DEGREE OF SEPARATION
Duncan J. Watts
Steven Strogatz
4-DEGREE OF SEPARATION
DISCOVERY
PUBLISHED 20 YEARS LATER
MILESTONES
WWII
XXI
PUBLICATION DATE
1929
1935
1940
1945
1950
1958
1960
1967
1970
1978
1980
1985
1991
1998
2000
2005
2011
**Frigyes Karinthy** (1887-1938)
Hungarian writer, journalist and playwright, the first to describe the small world property. In his short story entitled 'Láncszemek' (Chains) he links a worker in Ford's factory to himself [23, 24].
**Manfred Kochen** (1928-1989), **Ithiel de Sola Pool** (1917-1984)
Scientific interest in small worlds started with a paper by political scientist Ithiel de Sola Pool and mathematician Manfred Kochen. Written in 1958 and published in 1978, their work addressed in mathematical detail the small world effect, predicting that most individuals can be connected via two to threee acquaintances. Their paper inspired the experiments of Stanley Milgram.
**Stanley Milgram** (1933-1984)
American social psychologist who carried out the first experiment testing the small-world phenomena. (BOX 3.6).
**John Guare** (1938)
The phrase 'six degrees of separation' was introduced by the playwright John Guare, who used it as the title of his Broadway play.
The **Facebook Data Team** measures the average distance between its users, finding "4 degrees" (BOX 3.6).
**Duncan J. Watts** (1971), **Steven Strogatz** (1959)
A new wave of interest in small worlds followed the study of Watts and Strogatz, finding that the small world property applies to natural and technological networks as well.

# Clustering coefficient

# CLUSTERING COEFFICIENT

$$C_i \equiv \frac{2 < L_i >}{k_i(k_i - 1)}$$



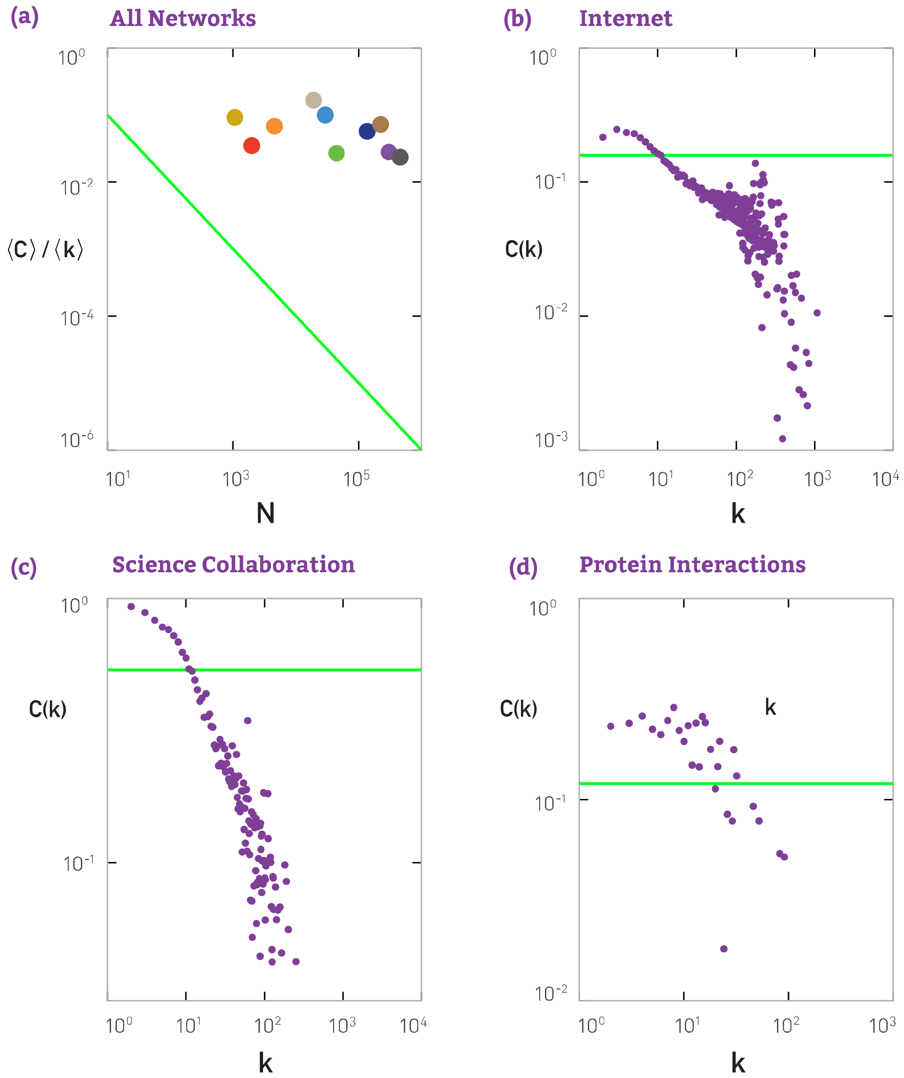

$C_i = 1$ $C_i = 1/2$ $C_i = 0$

Since edges are independent and have the same probability *p*,

$$< L_i > \cong p\frac{k_i(k_i - 1)}{2} \quad \Rightarrow \quad C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}.$$

- The clustering coefficient of random graphs is small.
- For fixed degree C decreases with the system size N.
- C is independent of a node's degree k.

# CLUSTERING COEFFICIENT

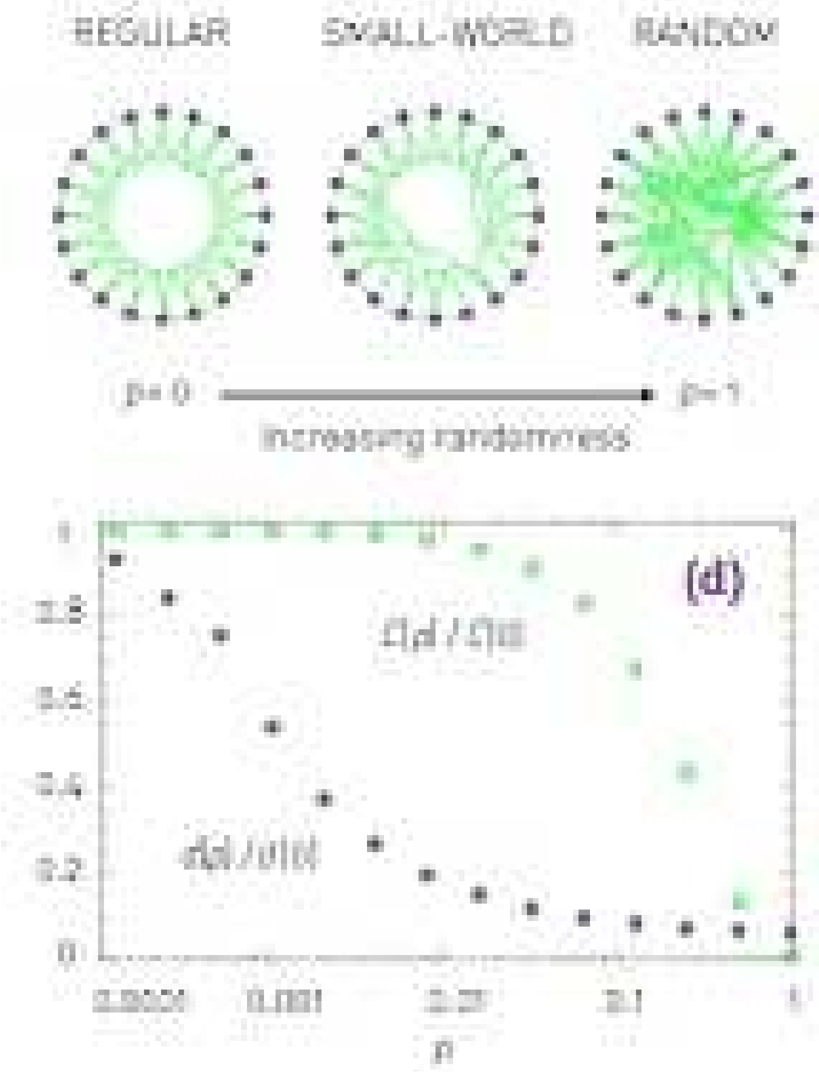



$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}.$$

C decreases with the system size *N*.

C is independent of a node's degree *k*.

## Watts-Strogatz Model



# Real networks are not random

# ARE REAL NETWORKS LIKE RANDOM GRAPHS?

As quantitative data about real networks became available, we can compare their topology with the predictions of random graph theory.

Note that once we have N and <k> for a random network, from it we can derive every measurable property. Indeed, we have:

<u>Average path length:</u>

$$< l_{rand} > \approx \frac{\log N}{\log \langle k \rangle}$$

<u>Clustering Coefficient:</u>

$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}.$$

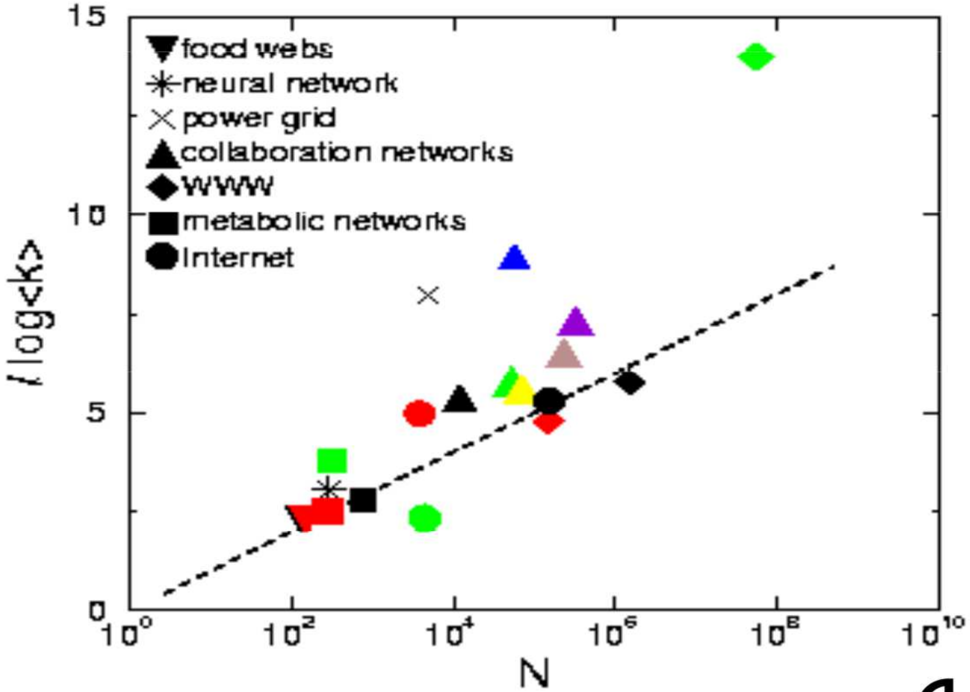
<u>Degree Distribution:</u>

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

## PATH LENGTHS IN REAL NETWORKS

**Prediction:**

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

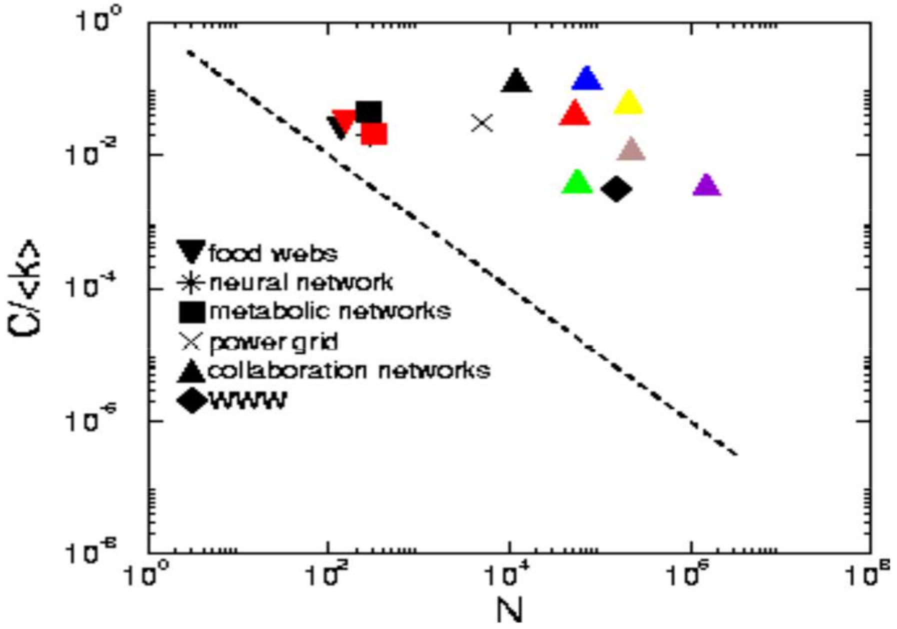



Real networks have short distances
like random graphs.

## CLUSTERING COEFFICIENT

**Prediction:**

$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}.$$




$C_{rand}$ underestimates with orders of magnitudes the clustering coefficient of real networks.

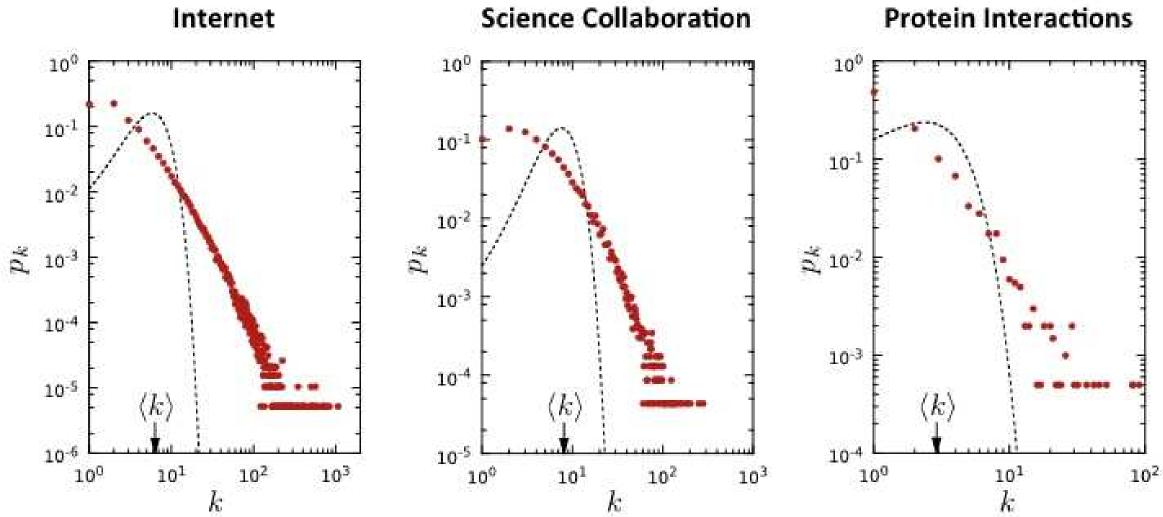
# THE DEGREE DISTRIBUTION

## Prediction:

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

## Data:

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




# ARE REAL NETWORKS LIKE RANDOM GRAPHS?

As quantitative data about real networks became available, we can compare their topology with the predictions of random graph theory.

Note that once we have N and <k> for a random network, from it we can derive every measurable property. Indeed, we have:

<u>Average path length:</u>

$$< l_{rand} > \approx \frac{\log N}{\log\langle k \rangle}$$

<u>Clustering Coefficient:</u>

$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}.$$

<u>Degree Distribution:</u>

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

# IS THE RANDOM GRAPH MODEL RELEVANT TO REAL SYSTEMS?

(B) Most important: we need to ask ourselves, are real networks random?

The answer is simply: NO

**There is no network in nature that we know of that would be described by the random network model.**

## IF IT IS WRONG AND IRRELEVANT, WHY DID WE DEVOT TO IT A FULL CLASS?

It is the reference model for the rest of the class.

It will help us calculate many quantities, that can then be compared to the real data, understanding to what degree is a particular property the result of some random process.

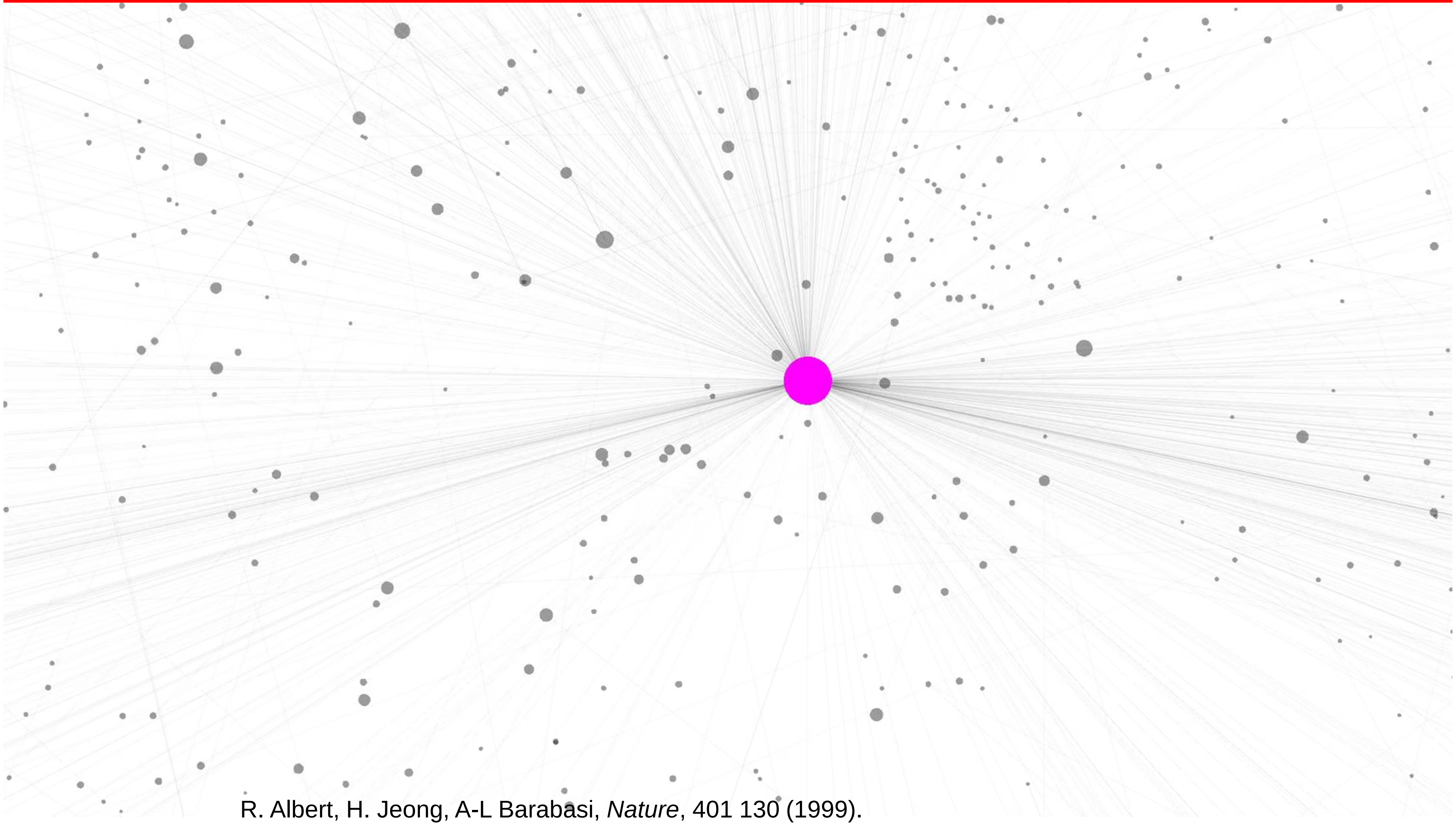
Patterns in real networks that are shared by a large number of real networks, yet which deviate from the predictions of the random network model.

In order to identify these, we need to understand how would a particular property look like if it is driven entirely by random processes.

**While WRONG and IRRELEVANT, it will turn out to be extremly USEFUL!**

# Scale-free property

# WORLD WIDE WEB



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

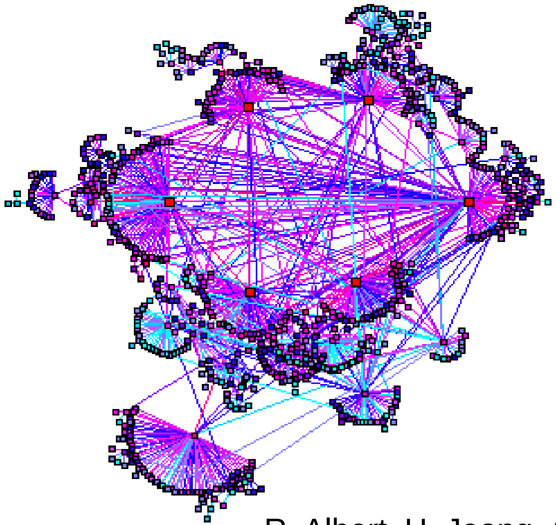
# Power laws and scale-free networks

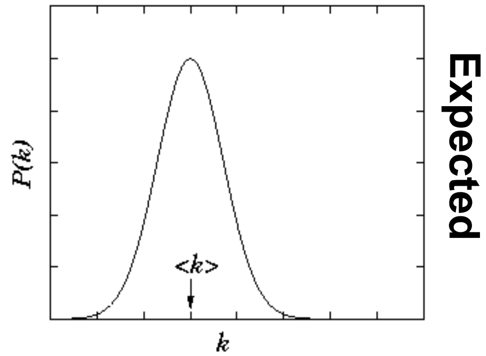
# WORLD WIDE WEB

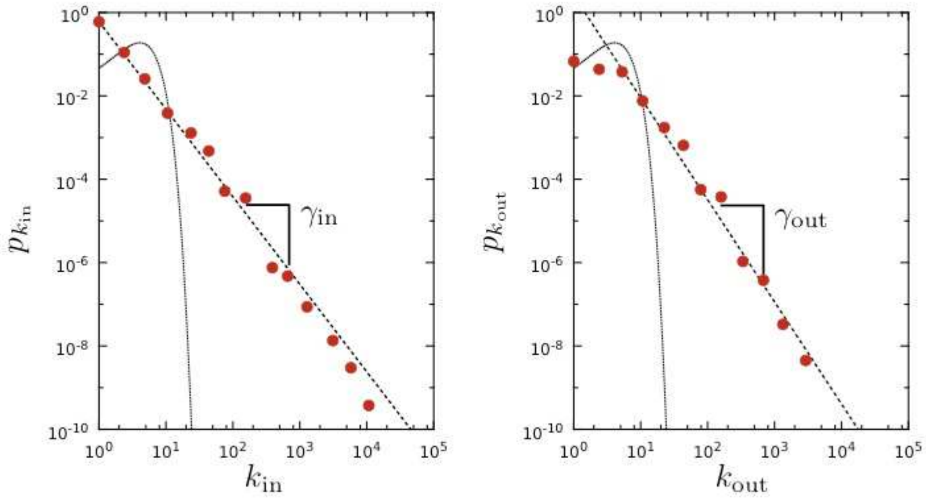
Nodes: **WWW documents**
Links: **URL links**

Over 3 billion documents

ROBOT: collects all URL's found in a document and follows them recursively






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

# Discrete vs. Continuum formalism

**Discrete Formalism**

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

$$p_k = Ck^{-\gamma} .$$

$$\sum_{k=1}^{\infty} p_k = 1 .$$

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

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

**Continuum Formalism**

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

$$p(k) = Ck^{-\gamma} .$$

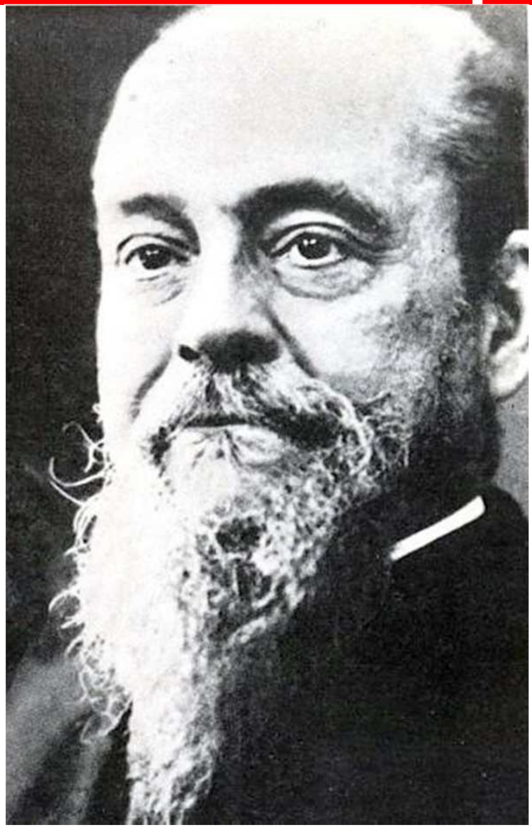
$$\int_{k_{\min}}^{\infty} p(k)dk = 1$$

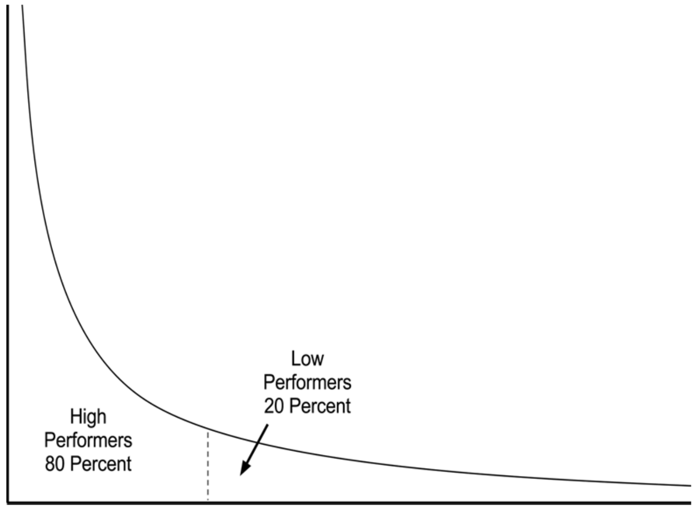
$$C = \frac{1}{\int_{k_{\min}}^{\infty} k^{-\gamma}dk} = (\gamma - 1)k_{\min}^{\gamma-1}$$

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

**INTERPRETATION:** $p_k$ $\qquad \int_{k_1}^{k_2} p(k)dk$

# 80/20 RULE

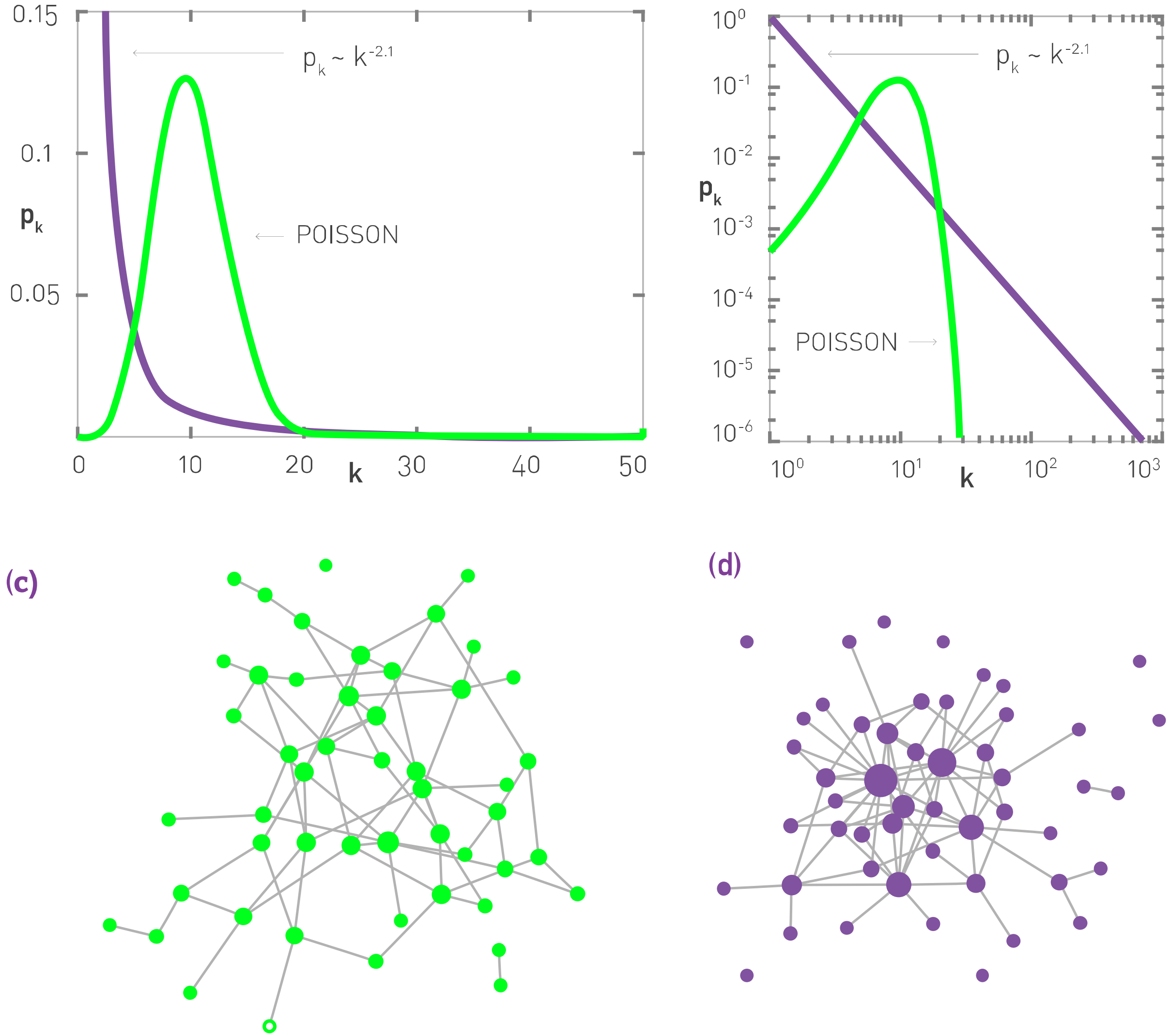





**Vilfredo Federico Damaso Pareto (1848 – 1923),** Italian economist, political scientist and philosopher, who had important contributions to our understanding of income distribution and to the analysis of individuals choices. A number of fundamental principles are named after him, like Pareto efficiency, Pareto distribution (another name for a power-law distribution), the Pareto principle (or 80/20 law).

# Hubs

## The difference between a power law and an exponential distribution

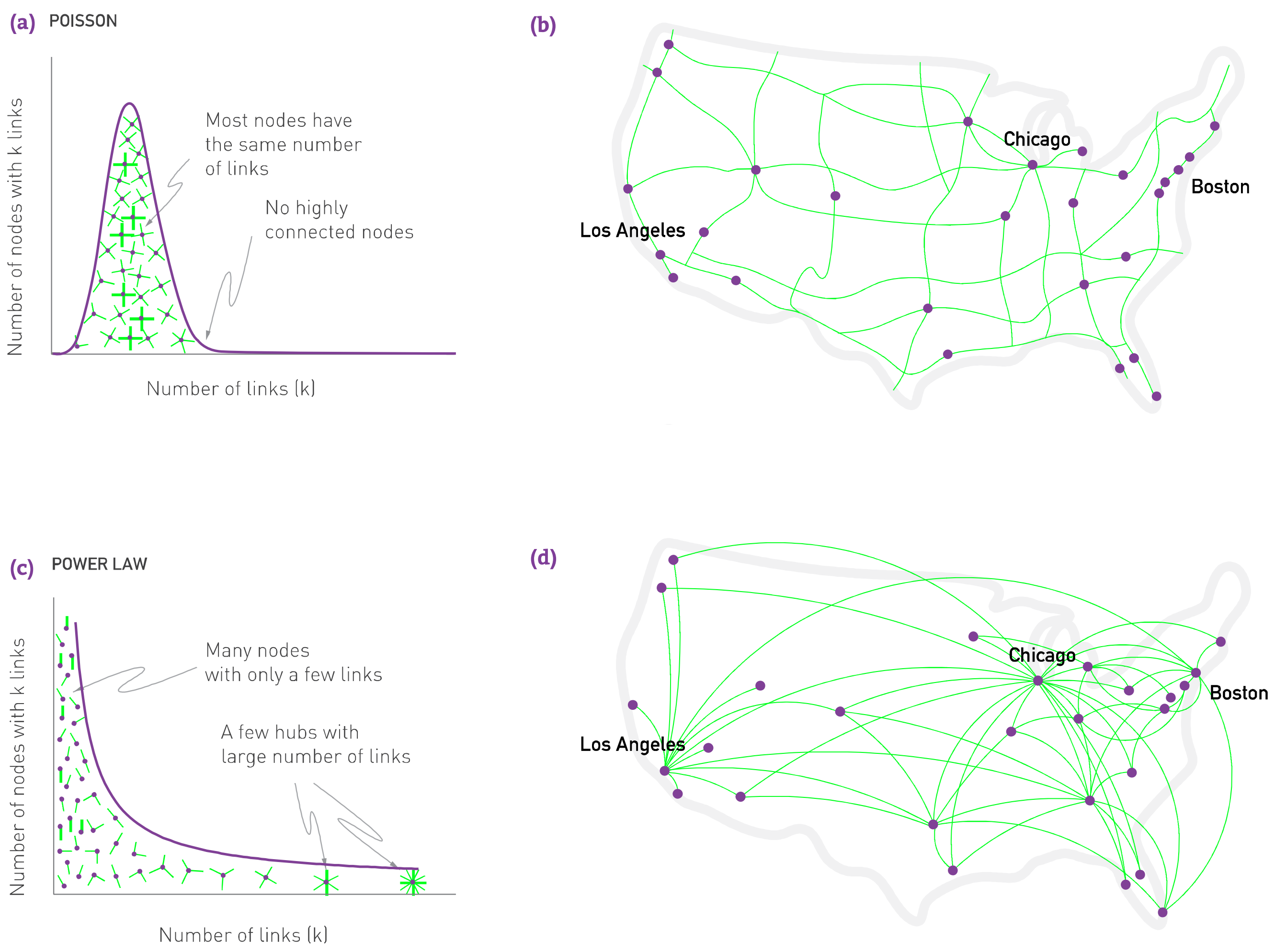


## The difference between a power law and an exponential distribution

**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* $N_{k>100} = 10^9$ *k>100 degree nodes*


(a) POISSON
Most nodes have the same number of links
No highly connected nodes
Number of nodes with k links
Number of links (k)
(b)
Chicago
Boston
Los Angeles
(c) POWER LAW
Many nodes with only a few links
A few hubs with large number of links
Number of nodes with k links
Number of links (k)
(d)
Chicago
Boston
Los Angeles

## The size of the biggest hub

All real networks are finite → let us explore its consequences.
→ We have an expected maximum degree, $k_{max}$

<u>Estimating $k_{max}$</u>

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

Why: 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

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

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

## The size of the biggest hub

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

To illustrate the difference in the maximum degree of an exponential and a scale-free network let us return to the WWW sample of Figure 4.1, consisting of $N \simeq 3 \times 10^5$ nodes. As $k_{min} = 1$, if the degree distribution were to follow an exponential, (4.17) predicts that the maximum degree should be $k_{max} \simeq 13$. In a scale-free network of similar size and $\gamma = 2.1$, (4.18) predicts $k_{max} \simeq 85{,}000$, a remarkable difference. Note that the largest in-degree of the WWW map of Figure 4.1 is 10,721, which is comparable to $k_{max}$ predicted by a scale-free network. This reinforces our conclusion that *in a random network hubs are effectivelly forbidden, while in scale-free networks they are naturally present.*

# Finite scale-free networks

Expected maximum degree, $k_{max}$

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

•$k_{max}$, increases with the size of the network

→the larger a system is, the larger its biggest hub

•For γ>2 $k_{max}$ increases slower than N

→the largest hub will contain a decreasing fraction of links as N increases.

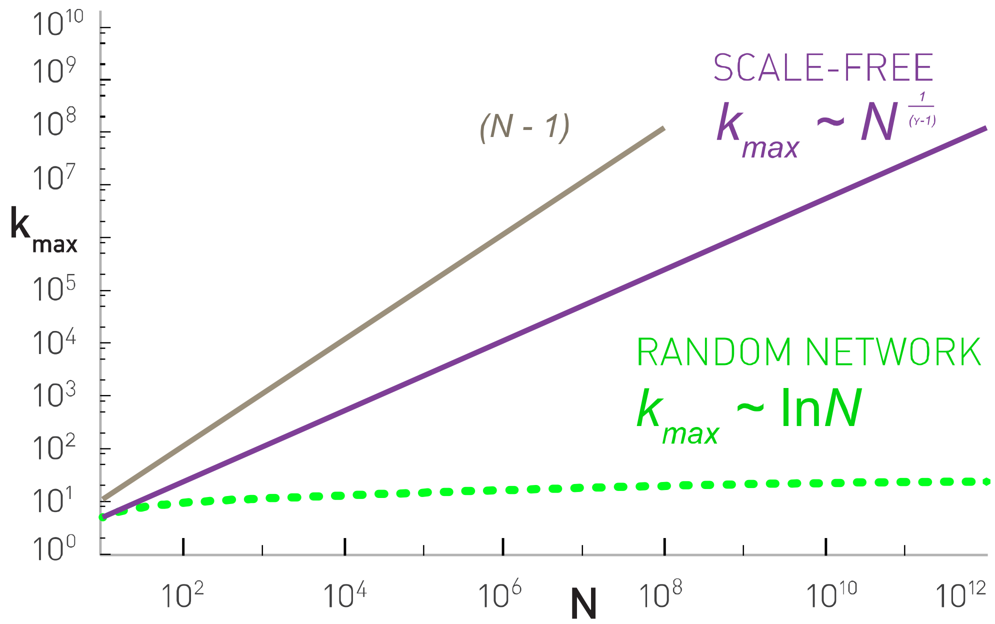
•For γ=2 $k_{max}$~N.

→ The size of the biggest hub is O(N)

•For γ<2 $k_{max}$ increases faster than N: condensation phenomena

→ the largest hub will grab an increasing fraction of links. Anomaly!

# The size of the largest hub




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

# The meaning of scale-free

# Scale-free networks: Definition

**Definition:**

**Networks with a power law tail in their degree distribution are called 'scale-free networks'**

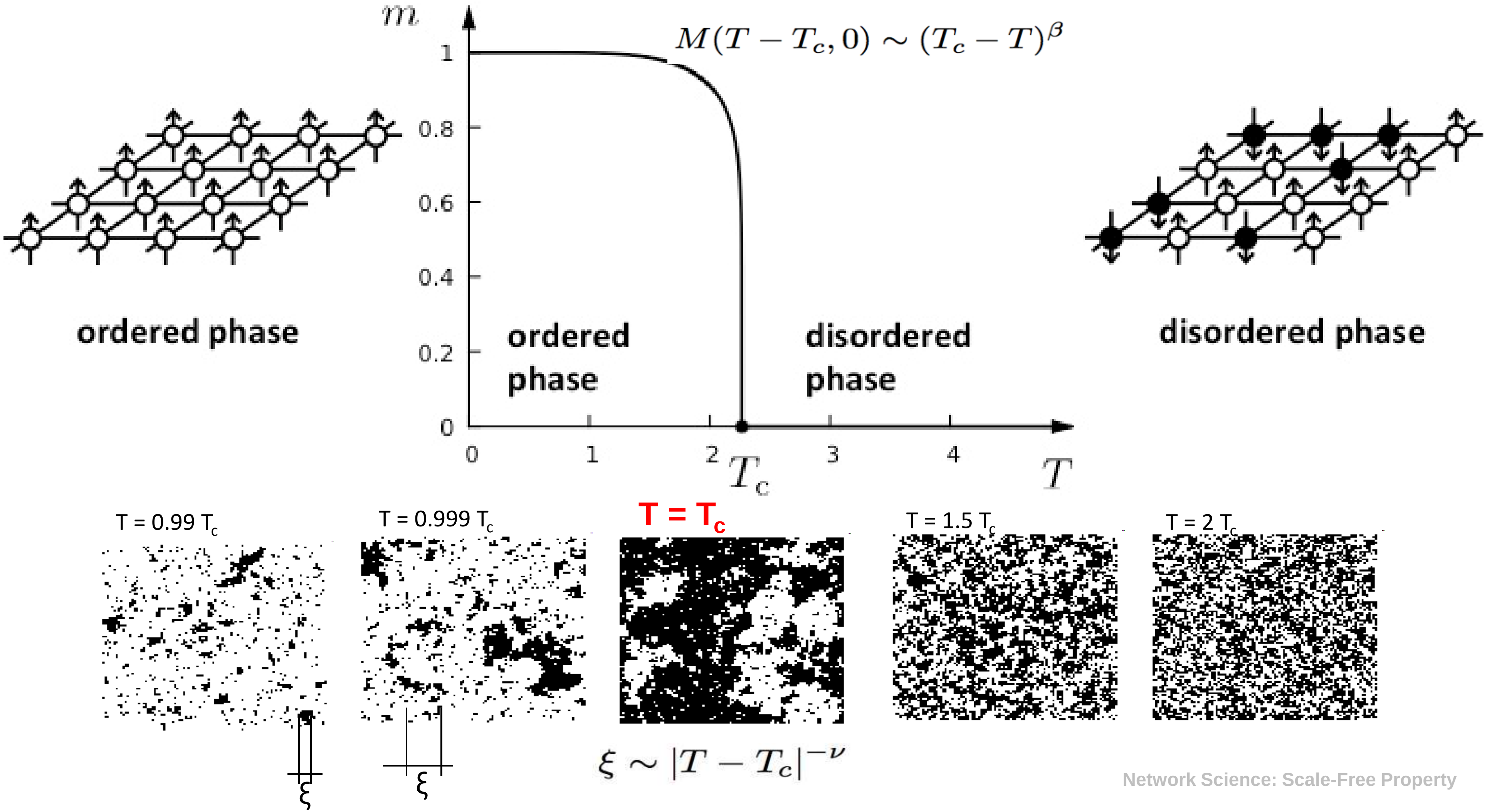
Where does the name come from?

## Critical Phenomena and scale-invariance

(a detour)

Slides after **Dante R. Chialvo**

# Phase transitions in complex systems I: Magnetism



## Scale-free behavior in space

$$\xi \sim |T - T_c|^{-\nu}$$


**At T = Tc:**

correlation length diverges

Fluctuations emerge at all scales:

*scale-free behavior*

## CRITICAL PHENOMENA

- Correlation length diverges at the critical point: the whole system is correlated!
- **Scale invariance**: there is no characteristic scale for the fluctuation **(scale-free behavior).**
- **Universality**: exponents are independent of the system’s details.

# Divergences in scale-free distributions

$$P(k) = Ck^{-\gamma} \qquad k = [k_{\min}, \infty) \qquad \int_{k_{\min}}^{\infty} P(k)dk = 1 \qquad C = \frac{1}{\int_{k_{\min}}^{\infty} k^{-\gamma}\, dk} = (\gamma - 1)k_{\min}^{\gamma-1}$$

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

$$< k^m >= \int_{k_{\min}}^{\infty} k^m P(k)\, dk \qquad < k^m >= (\gamma - 1)k_{\min}^{\gamma-1} \int_{k_{\min}}^{\infty} k^{m-\gamma}\, dk = \frac{(\gamma - 1)}{(m - \gamma + 1)} k_{\min}^{\gamma-1} \left[ k^{m-\gamma+1} \right]_{k_{\min}}^{\infty}$$

If m-γ+1<0:

$$< k^m >= -\frac{(\gamma - 1)}{(m - \gamma + 1)} k_{\min}^{m}$$

If m-γ+1>0, the integral diverges.

For a fixed γ this means that all moments with m>γ-1 diverge.

# DIVERGENCE OF THE HIGHER MOMENTS

$$< k^m >= (\gamma - 1) k_{\min}^{\gamma-1} \int_{k_{\min}}^{\infty} k^{m-\lambda}\, dk = \frac{(\gamma-1)}{(m-\gamma+1)} k_{\min}^{\gamma-1} \left[ k^{m-\gamma+1} \right]_{k_{\min}}^{\infty}$$

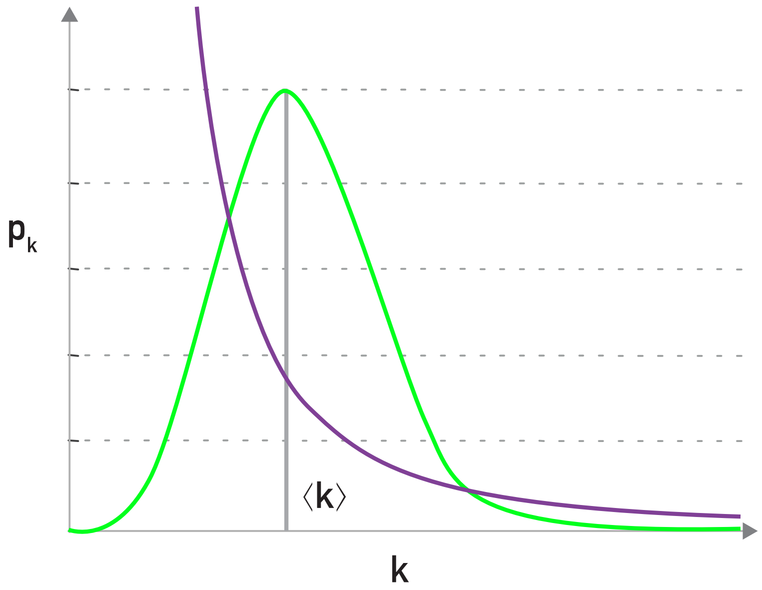
For a fixed λ this means all moments m>γ-1 diverge.

| Network | Size | $\langle k \rangle$ | $\kappa$ | $\gamma_{out}$ | $\gamma_{in}$ |
|---|---|---|---|---|---|
| WWW | 325 729 | 4.51 | 900 | 2.45 | 2.1 |
| WWW | $4\times10^7$ | 7 | | 2.38 | 2.1 |
| WWW | $2\times10^8$ | 7.5 | 4000 | 2.72 | 2.1 |
| WWW, site | 260 000 | | | | 1.94 |
| Internet, domain* | 3015–4389 | 3.42–3.76 | 30–40 | 2.1–2.2 | 2.1–2.2 |
| Internet, router* | 3888 | 2.57 | 30 | 2.48 | 2.48 |
| Internet, router* | 150 000 | 2.66 | 60 | 2.4 | 2.4 |
| Movie actors* | 212 250 | 28.78 | 900 | 2.3 | 2.3 |
| Co-authors, SPIRES* | 56 627 | 173 | 1100 | 1.2 | 1.2 |
| Co-authors, neuro.* | 209 293 | 11.54 | 400 | 2.1 | 2.1 |
| Co-authors, math.* | 70 975 | 3.9 | 120 | 2.5 | 2.5 |
| Sexual contacts* | 2810 | | | 3.4 | 3.4 |
| Metabolic, *E. coli* | 778 | 7.4 | 110 | 2.2 | 2.2 |
| Protein, *S. cerev.** | 1870 | 2.39 | | 2.4 | 2.4 |
| Ythan estuary* | 134 | 8.7 | 35 | 1.05 | 1.05 |
| Silwood Park* | 154 | 4.75 | 27 | 1.13 | 1.13 |
| Citation | 783 339 | 8.57 | | | 3 |
| Phone call | $53\times10^6$ | 3.16 | | 2.1 | 2.1 |
| Words, co-occurrence* | 460 902 | 70.13 | | 2.7 | 2.7 |
| Words, synonyms* | 22 311 | 13.48 | | 2.8 | 2.8 |

Many degree exponents are smaller than 3

→ $<k^2>$ diverges in the N→∞ limit!!!

# The meaning of scale-free




Random Network

Randomly chosen node: $k = \langle k \rangle \pm \langle k \rangle^{1/2}$

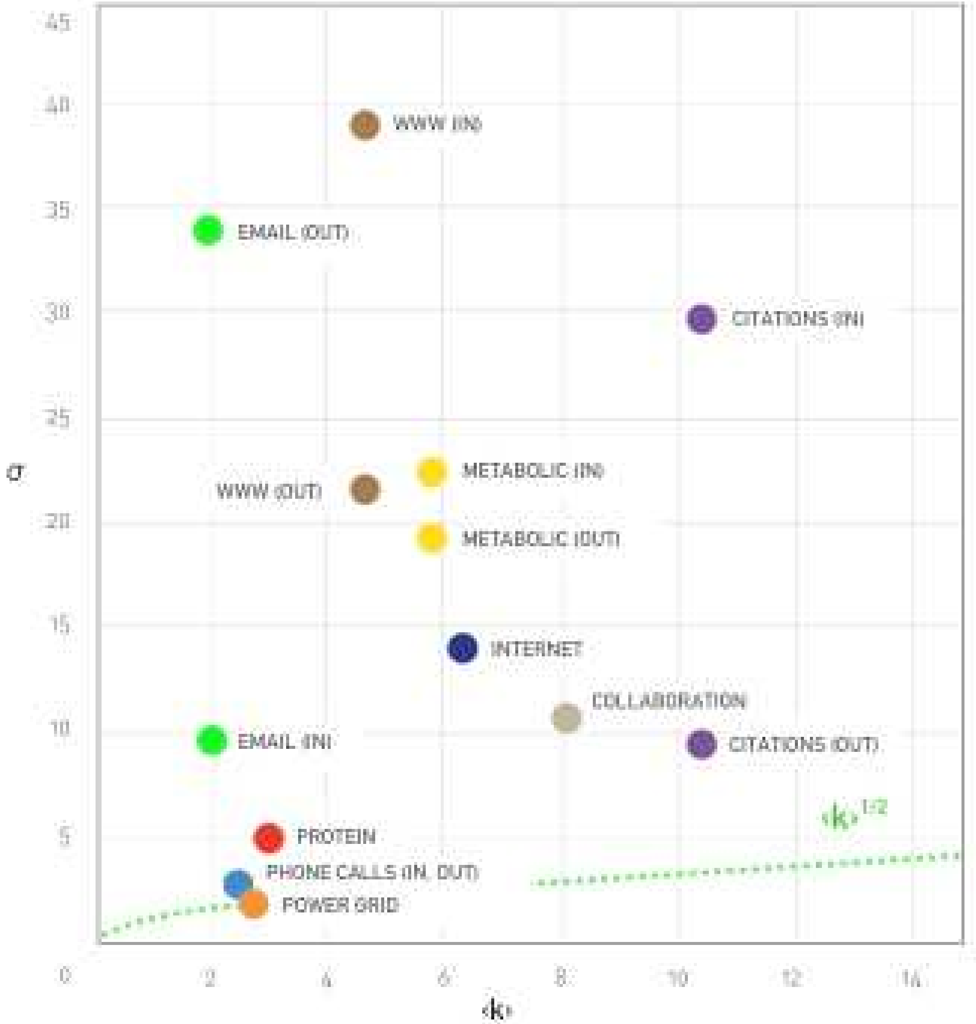
Scale: $\langle k \rangle$

Scale-Free Network

Randomly chosen node: $k = \langle k \rangle \pm \infty$

Scale: none

# The meaning of scale-free

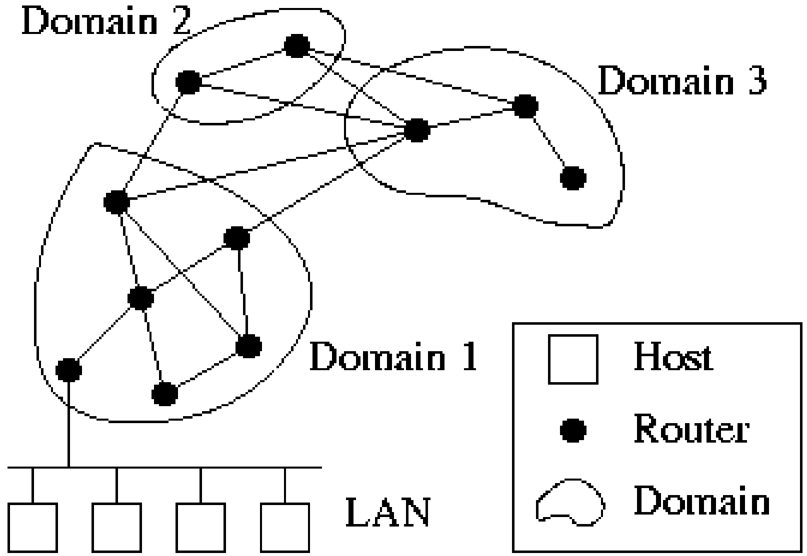


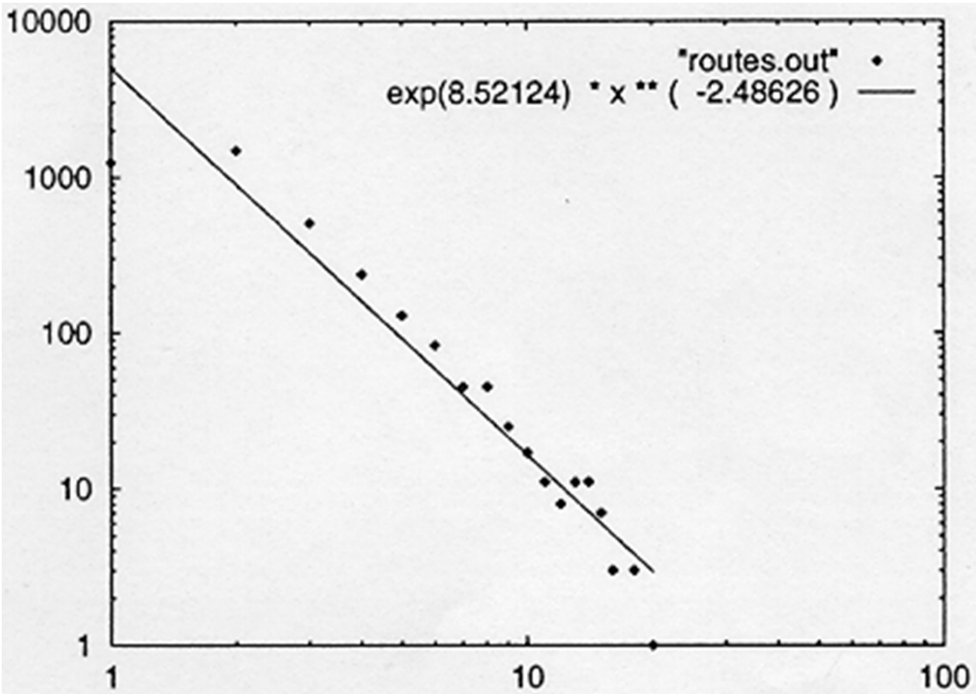

$$k = \langle k \rangle \pm \sigma_k$$

# universality

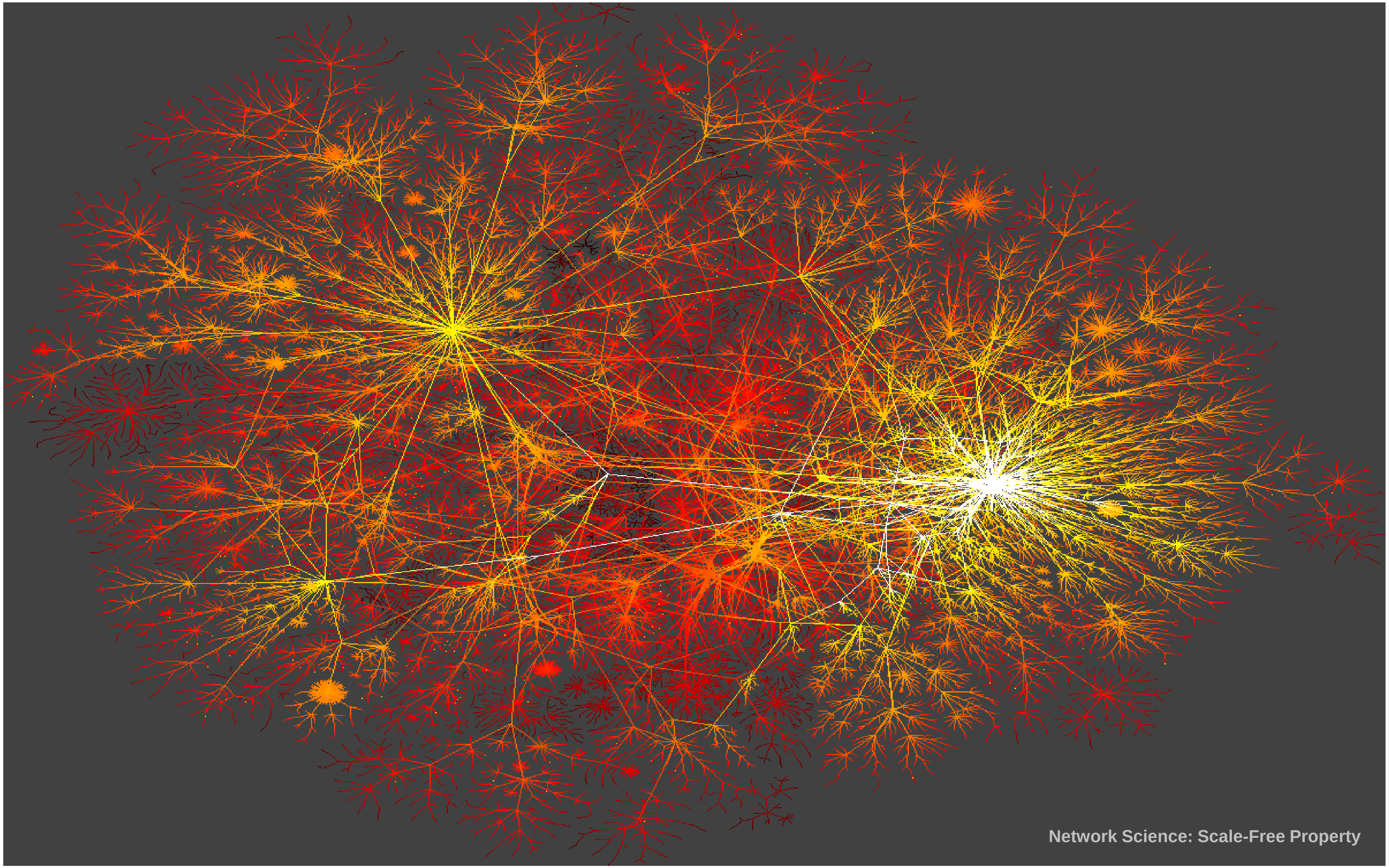
# INTERNET BACKBONE

**Nodes**: computers, routers
**Links**: physical lines



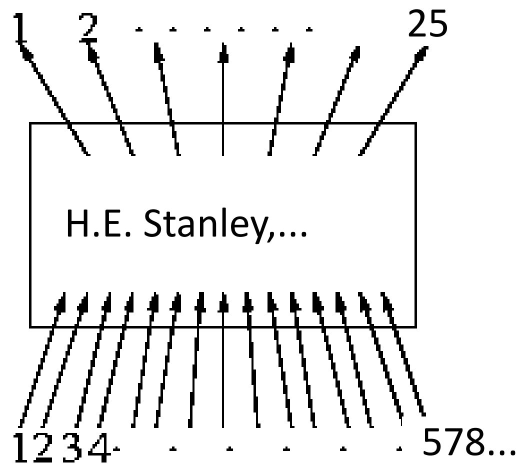

(Faloutsos, Faloutsos and Faloutsos, 1999)


Network Science: Scale-Free Property

# SCIENCE CITATION INDEX

Out of over 500,000 Examined
(see http://www.sst.nrel.gov)

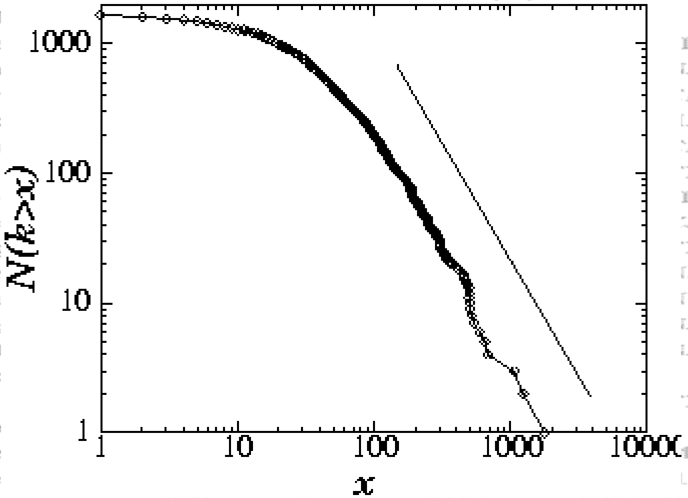
**Nodes**: papers
**Links**: citations




$P(k) \sim k^{-\gamma}$

$(\gamma = 3)$

1736 PRL papers (1988)

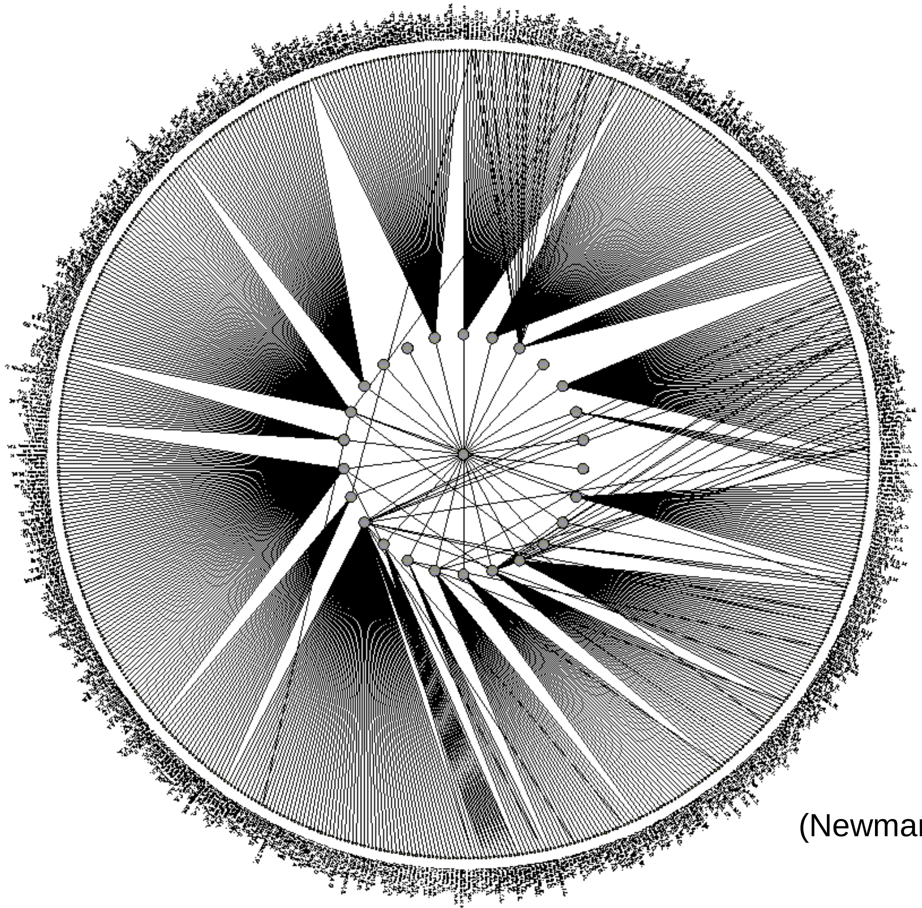


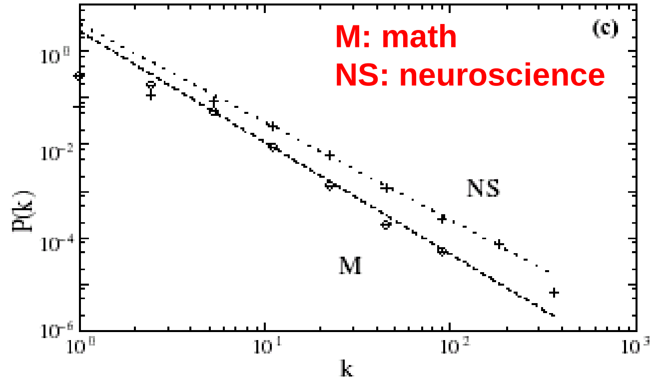

(S. Redner, 1998)

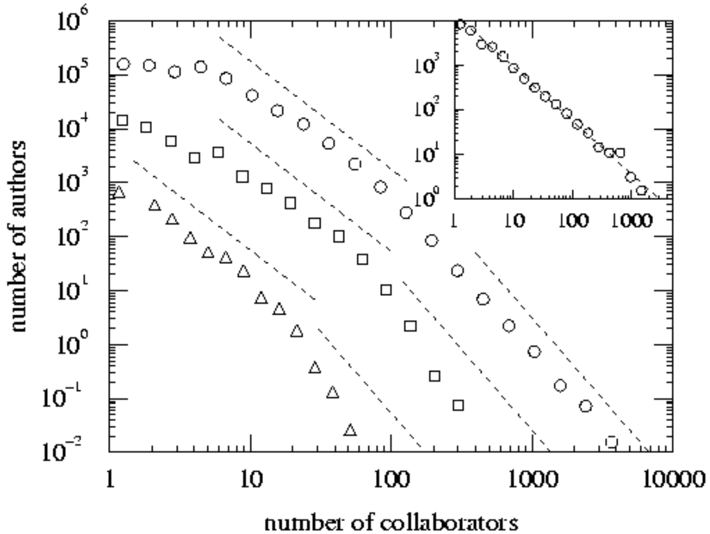
* citation total may be skewed because of multiple authors with the same name

## SCIENCE COAUTHORSHIP

**Nodes**: scientist (authors)
**Links**: joint publication





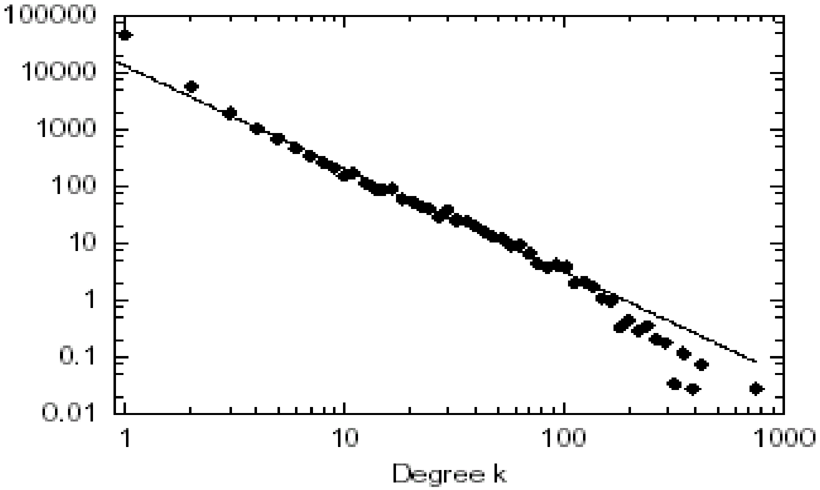




(Newman, 2000, Barabasi et al 2001)

# ONLINE COMMUNITIES

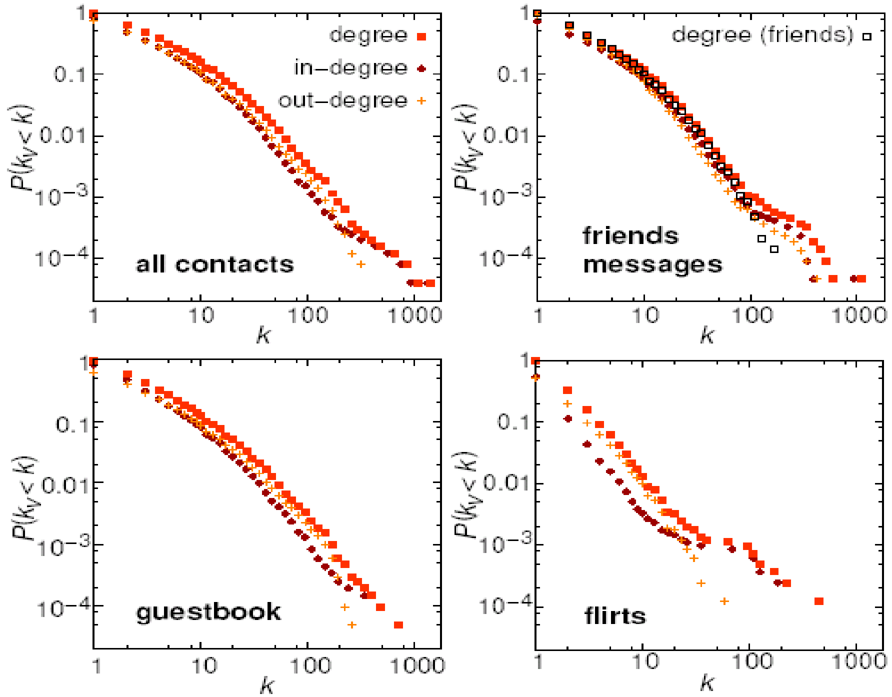
**Nodes**: online user
**Links**: email contact

Kiel University log files
112 days, N=59,912 nodes




Ebel, Mielsch, Bornholdtz, PRE 2002.

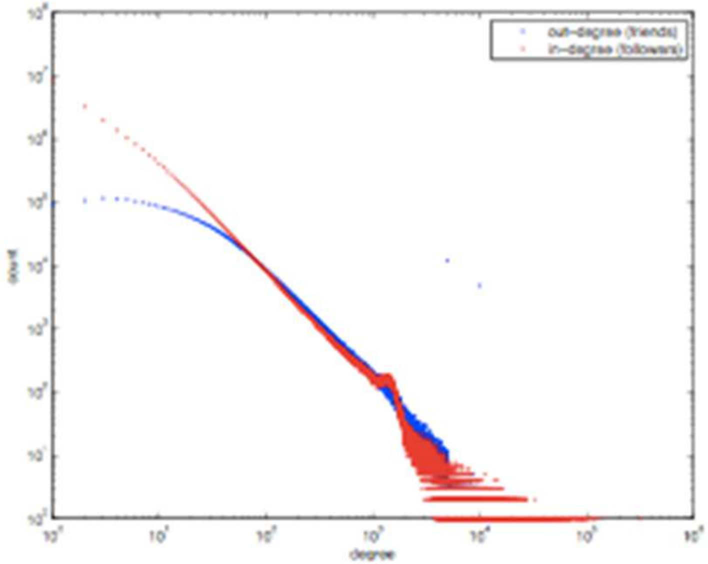
Pussokram.com online community;
512 days, 25,000 users.



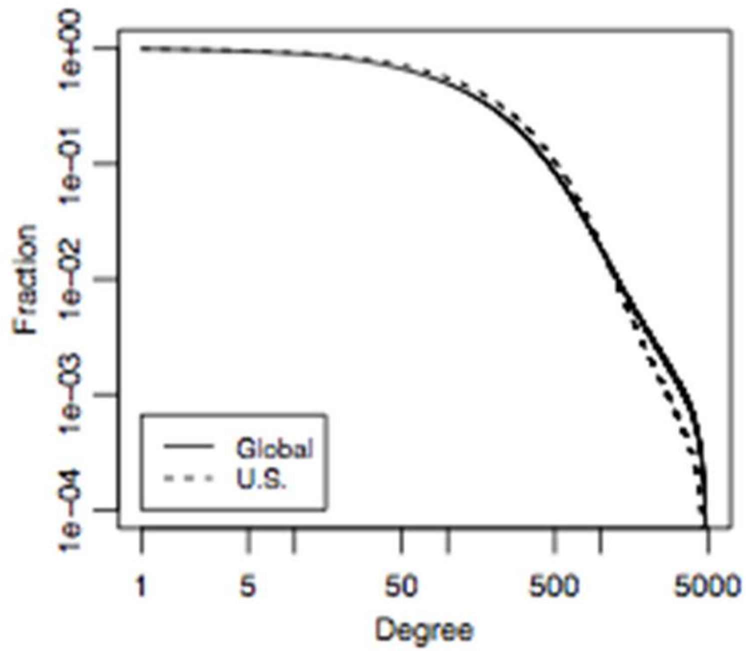

Holme, Edling, Liljeros, 2002.

# ONLINE COMMUNITIES

Twitter:

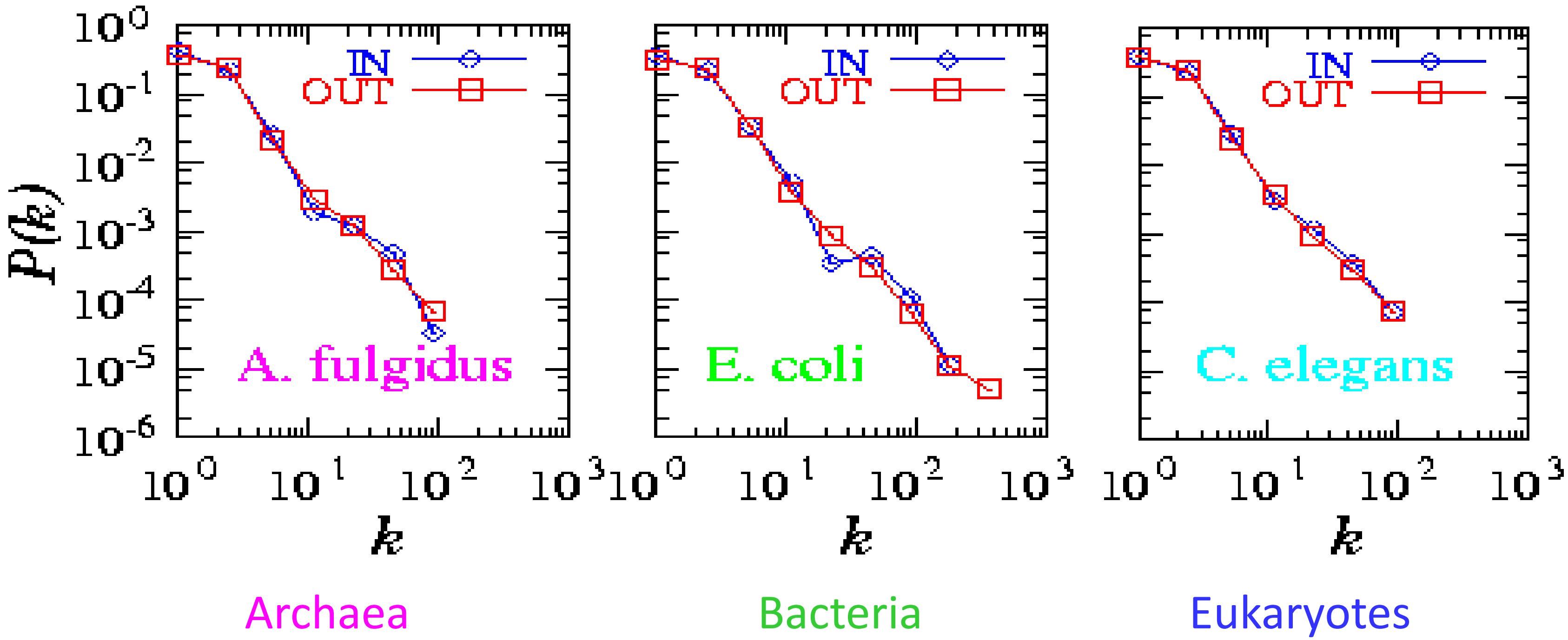


Facebook




Brian Karrer, Lars Backstrom, Cameron Marlowm 2011

# METABOLIC NETWORK




Organisms from all three domains of life are scale-free!

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

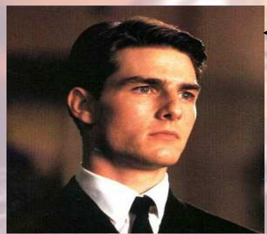
$$P_{out}(k) \approx k^{-2.2}$$

H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabasi, *Nature*, 407 651 (2000)

# ACTOR NETWORK

**Nodes**: actors
**Links**: cast jointly

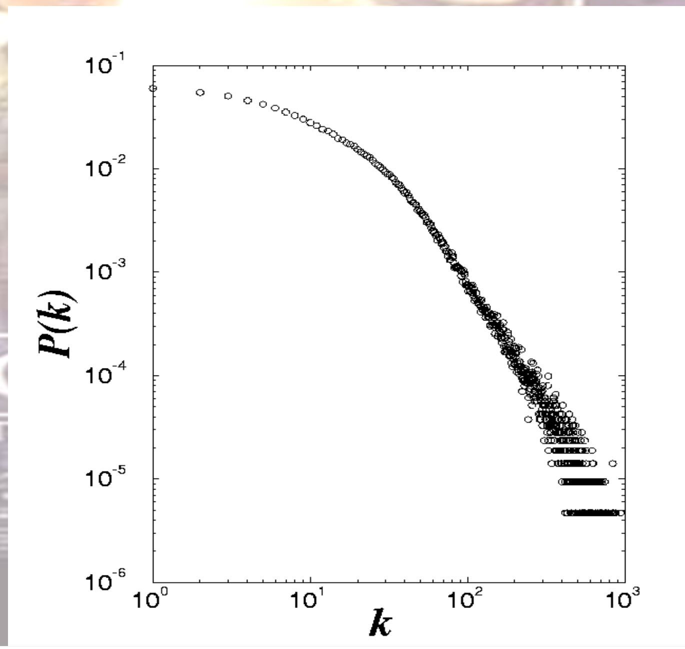
IMDb Internet Movie Database



***Days of Thunder*** *(1990)*
***Far and Away*** *(1992)*
***Eyes Wide Shut*** *(1999)*

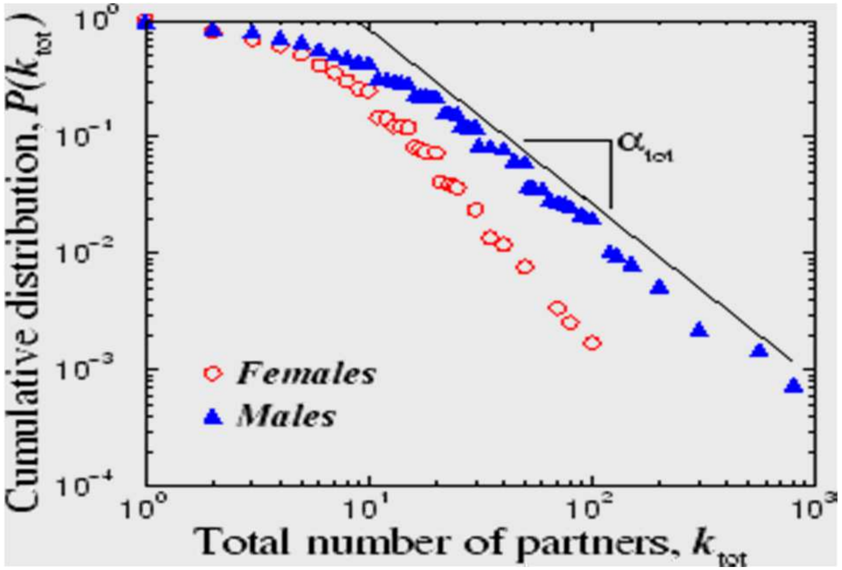
$N = 212{,}250$ actors
$\langle k \rangle = 28.78$

$P(k) \sim k-\gamma$
$\gamma=2.3$



## SWEDISH SE-WEB

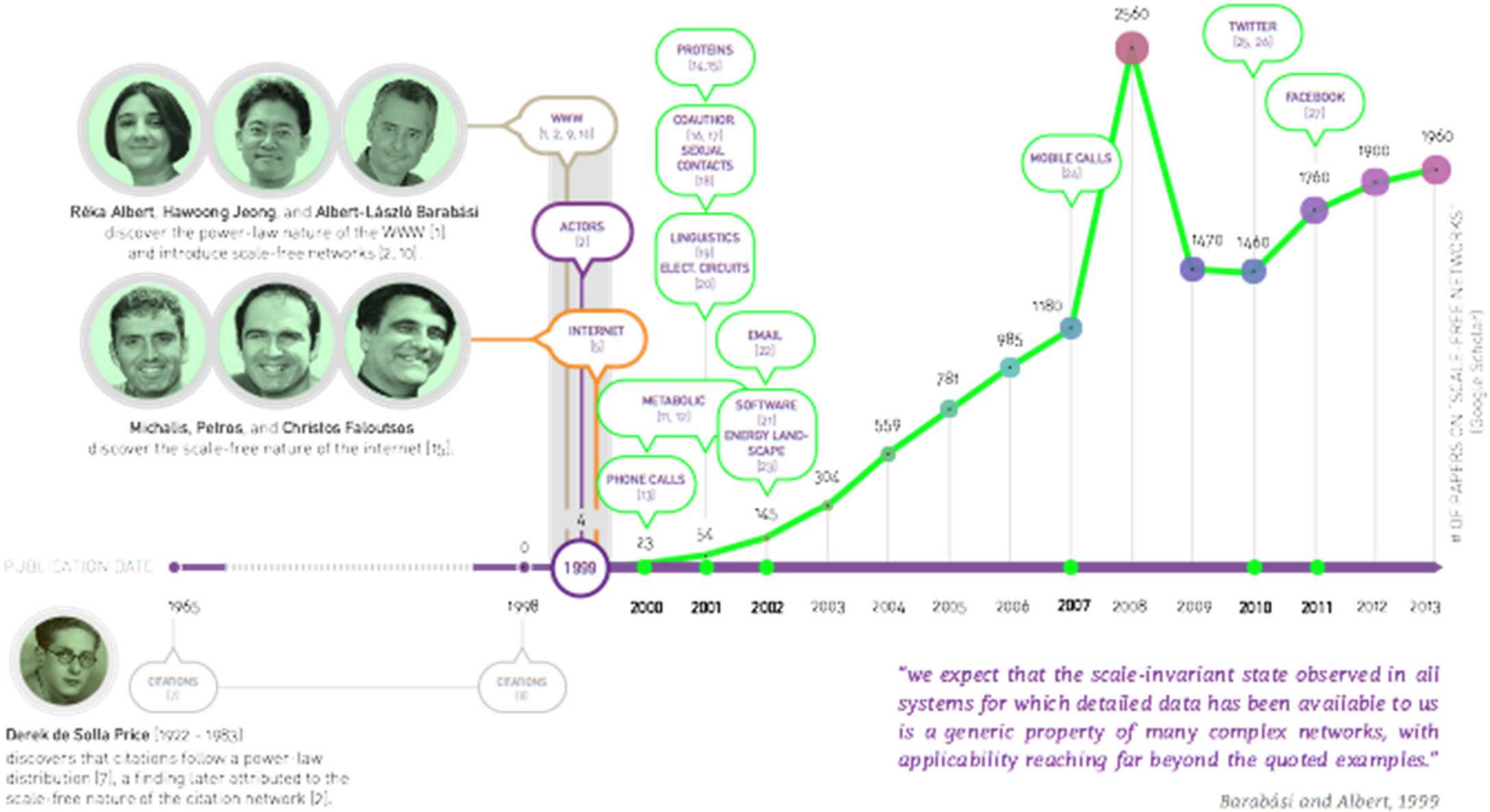
**Nodes:** people (Females; Males)
**Links:** sexual relationships




4781 Swedes; 18-74;
59% response rate.

Liljeros et al. Nature 2001

## TIMELINE: SCALE-FREE NETWORKS




*"we expect that the scale-invariant state observed in all systems for which detailed data has been available to us is a generic property of many complex networks, with applicability reaching far beyond the quoted examples."*

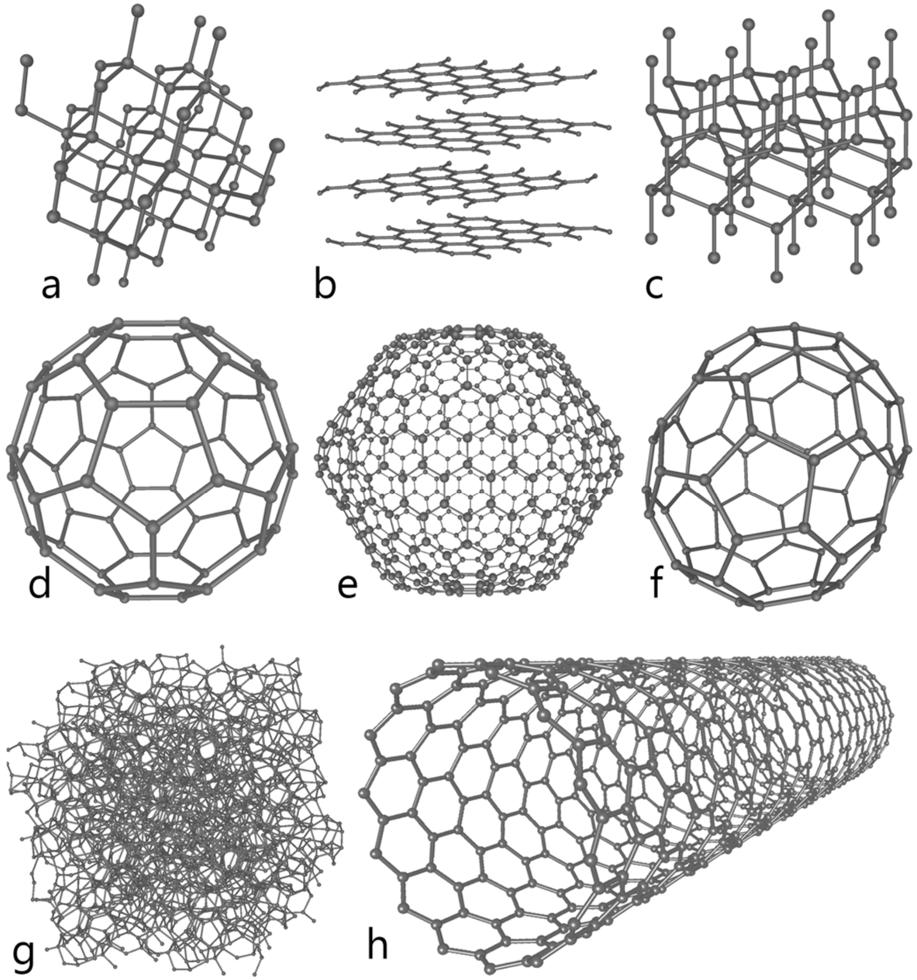
*Barabási and Albert, 1999*

# Not all networks are scale-free

•*Networks appearing in material science, like the network describing the bonds between the atoms in crystalline or amorphous materials, where each node has exactly the same degree.*

•*The neural network of the C.elegans worm.*

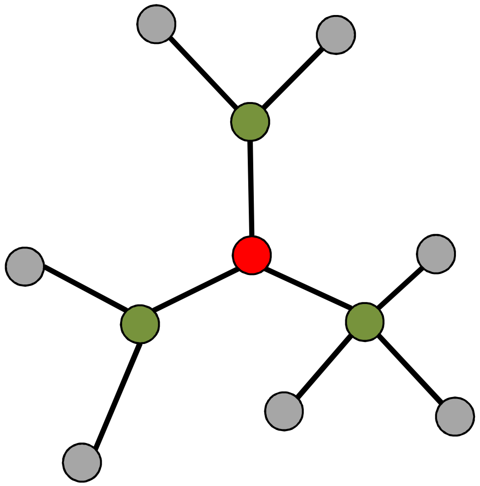
•*The power grid, consisting of generators and switches connected by transmission lines*



# Ultra-small property

# DISTANCES IN RANDOM GRAPHS

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



• nr. of first neighbors: $N_1 \cong \langle k \rangle$

• nr. of second neighbors: $N_2 \cong \langle k \rangle^2$

•nr. of neighbours at distance d: $N_d \cong \langle k \rangle^d$

• estimate maximum distance:

$$1 + \sum_{l=1}^{l_{max}} \langle k \rangle^i = N \Rightarrow l_{max} = \frac{\log N}{\log \langle k \rangle}$$

# SMALL WORLD BEHAVIOR IN SCALE-FREE NETWORKS

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

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

$\gamma = 2$: 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.

**Ultra Small World**

$2 < \gamma < 3$: 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.

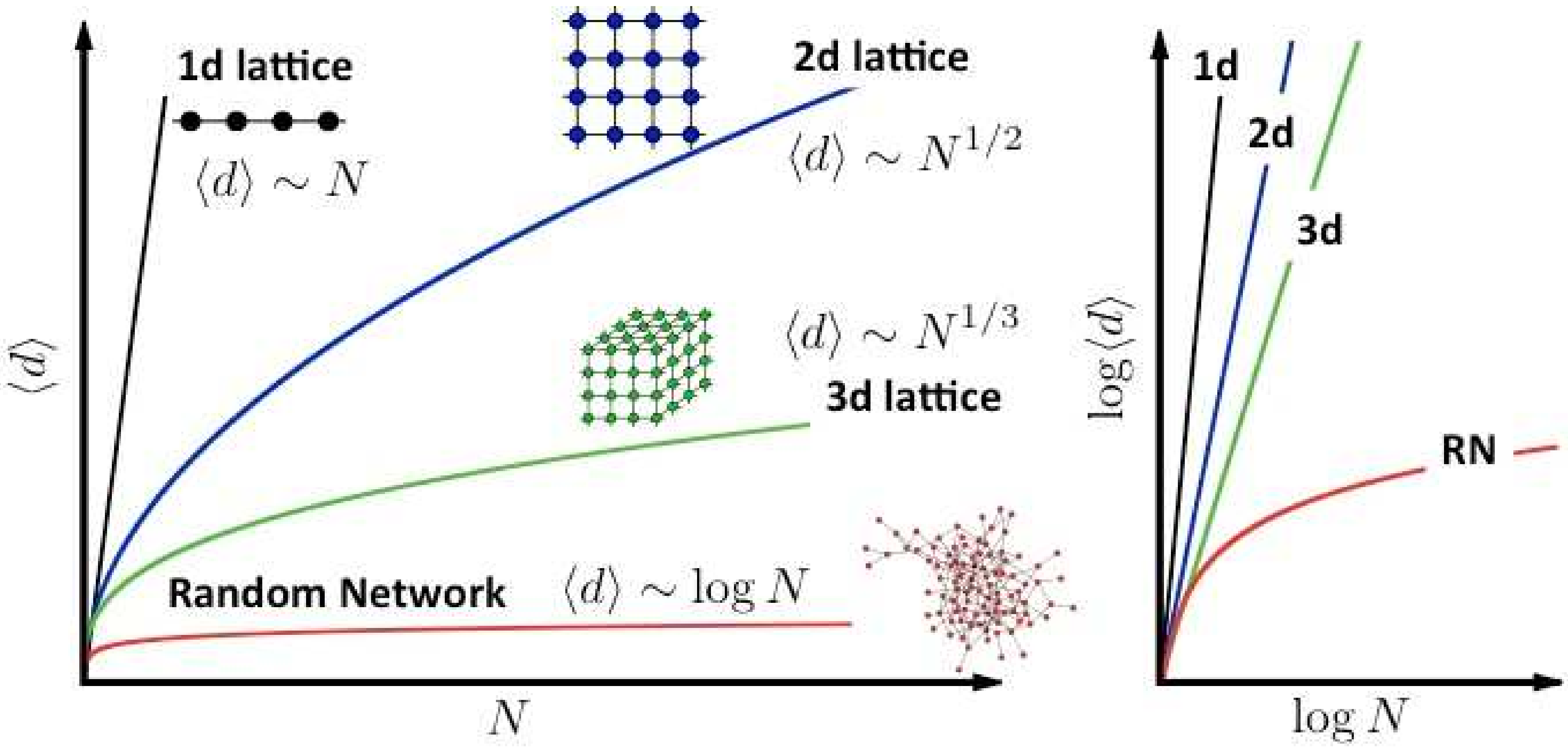
$\gamma = 3$: Some key models produce γ=3, so the result is of particular importance for them. This was first derived by Bollobas and collaborators for the network diameter in the context of a dynamical model, but it holds for the average path length as well.

**Small World**

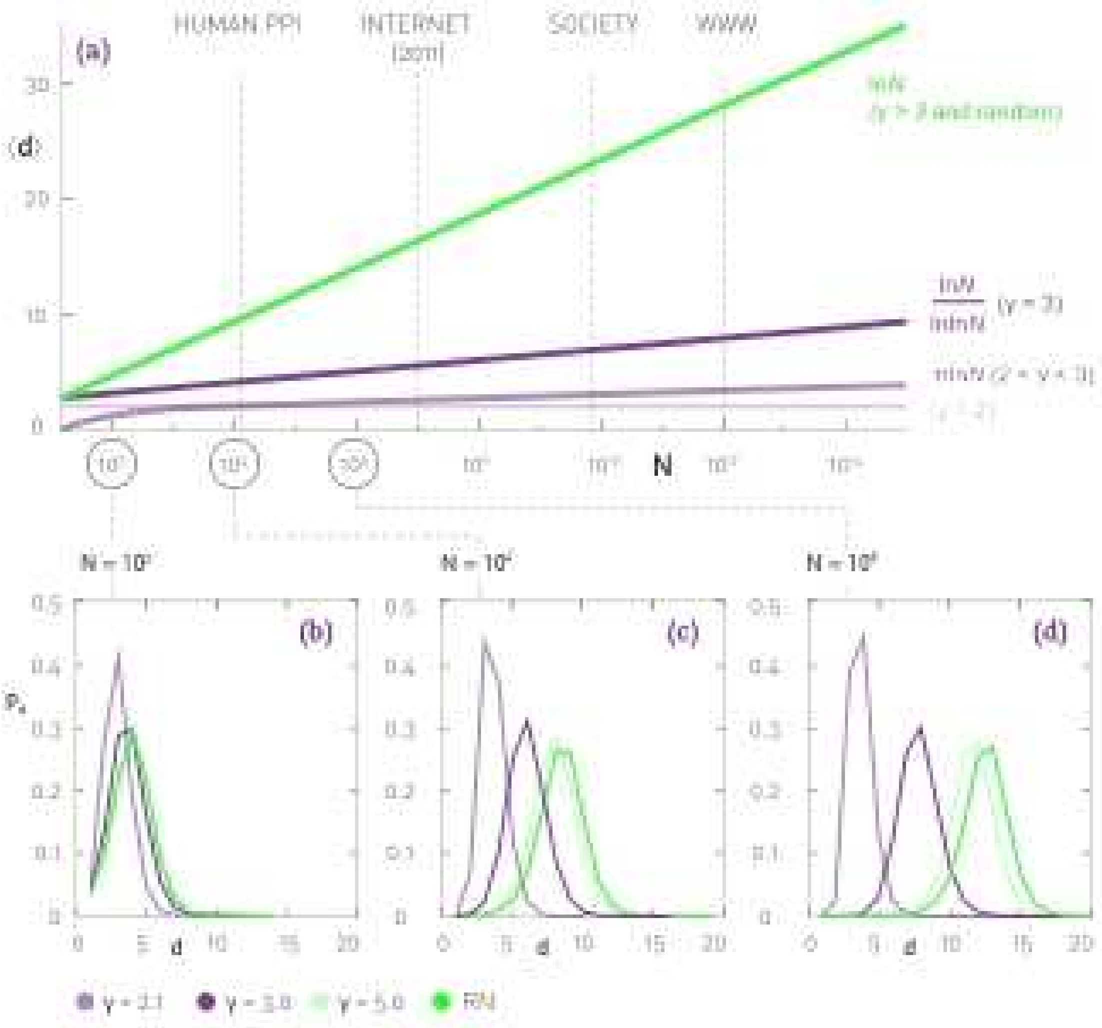
$\gamma > 3$: **T**he second moment of the distribution is finite, thus in many ways the network behaves as a random network. Hence the average path length follows the result that we derived for the random network model earlier.

Cohen, Havlin Phys. Rev. Lett. 90, 58701(2003); Cohen, Havlin and ben-Avraham, in Handbook of Graphs and Networks, Eds. Bornholdt and Shuster (Willy-VCH, NY, 2002) Chap. 4; Confirmed also by: Dorogovtsev et al (2002), Chung and Lu (2002); (Bollobas, Riordan, 2002; Bollobas, 1985; Newman, 2001

## Why are small worlds surprising? Suprising compared to what?



## SMALL WORLD BEHAVIOR IN SCALE-FREE NETWORKS

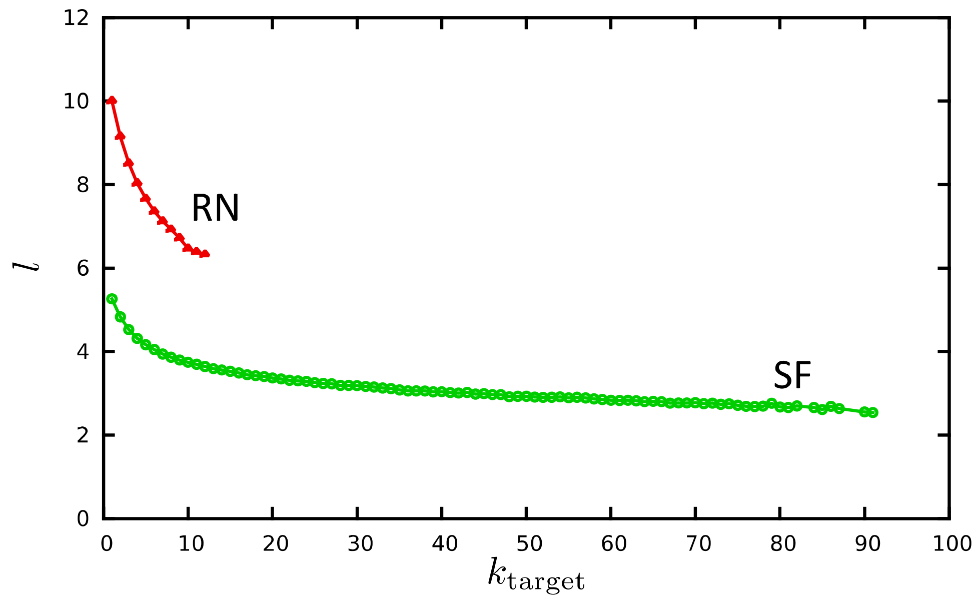



$$\langle d \rangle \sim \begin{cases} \text{const.} & \gamma = 2, \\ \dfrac{\ln \ln N}{\ln(\gamma - 1)} & 2 < \gamma < 3, \\ \dfrac{\ln N}{\ln \ln N} & \gamma = 3, \\ \ln N & \gamma > 3. \end{cases}$$

## We are always close to the hubs

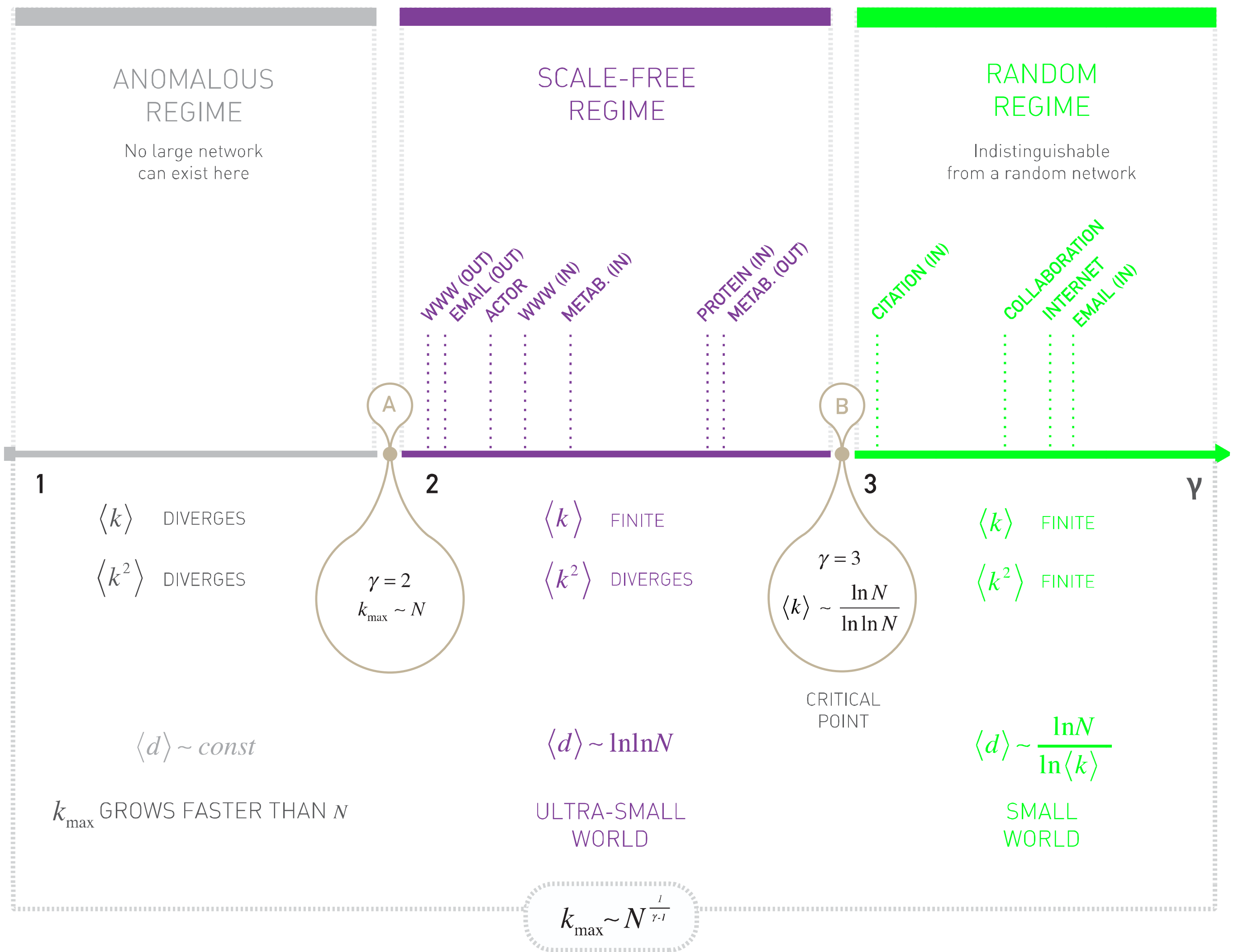
" it's always easier to find someone who knows a famous or popular figure than some run-the-mill, insignificant person."

(Frigyes Karinthy, 1929)

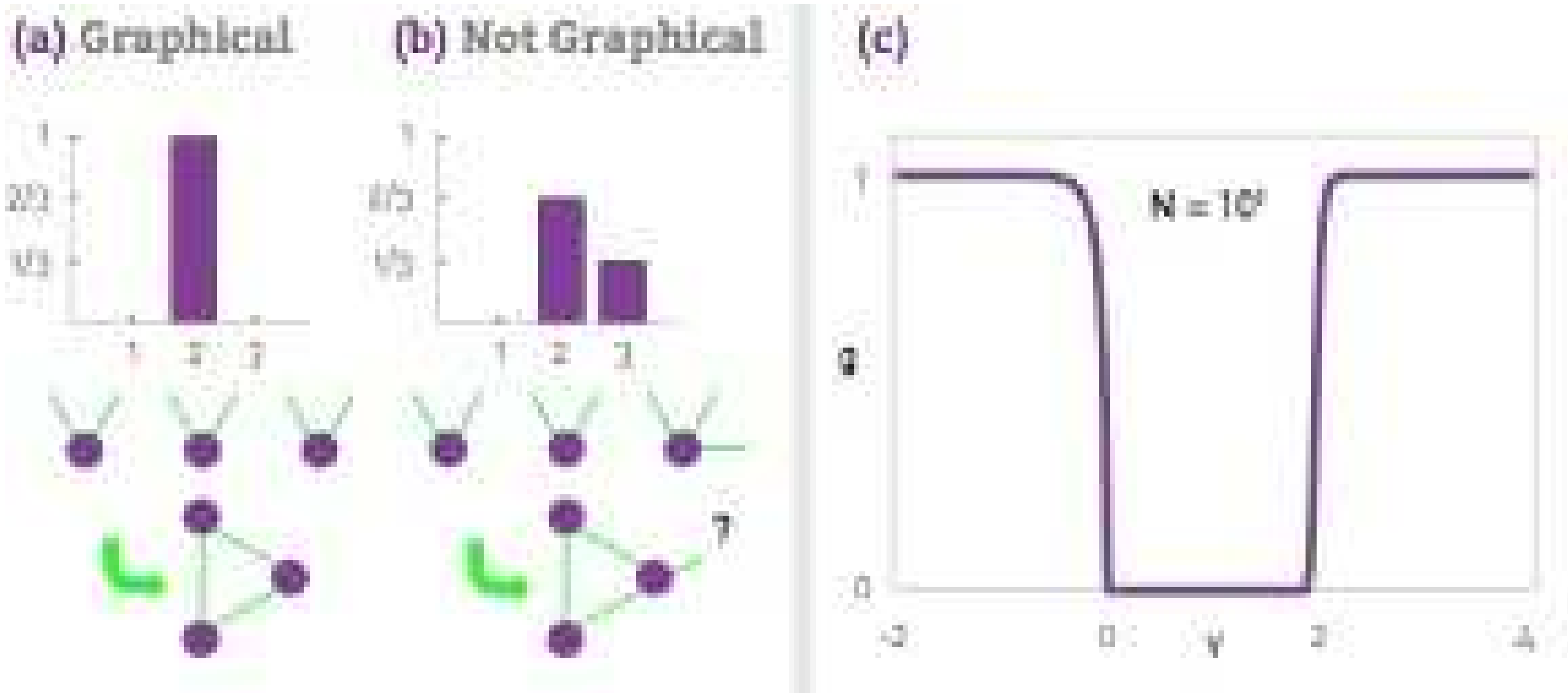


# The role of the degree exponent

# SUMMARY OF THE BEHAVIOR OF SCALE-FREE NETWORKS



## Graphicality: No large networks for γ<2




In scale-free networks: $$k_{\max} = k_{\min} N^{\frac{1}{\gamma-1}}$$ For γ<2: 1/(γ-2)>1

# Why don't we see networks with exponents in the range of γ=4,5,6,  etc?

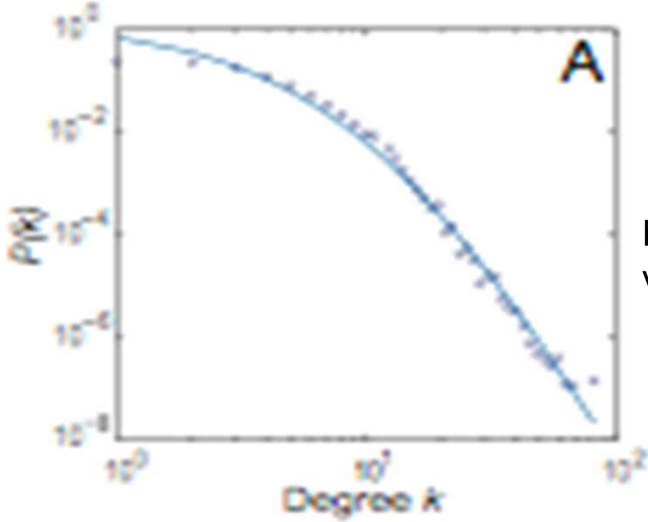
In order to document a scale-free networks, we need 2-3 orders of magnitude scaling.
That is, $K_{max} \sim 10^3$

However, that constrains on the system size we require to document it.
For example, to measure an exponent $\gamma=5$,we need to maximum degree a system size of the order of

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

$$N = \left( \frac{K_{\max}}{K_{\min}} \right)^{\gamma-1} \approx 10^8$$

Mobile Call Network



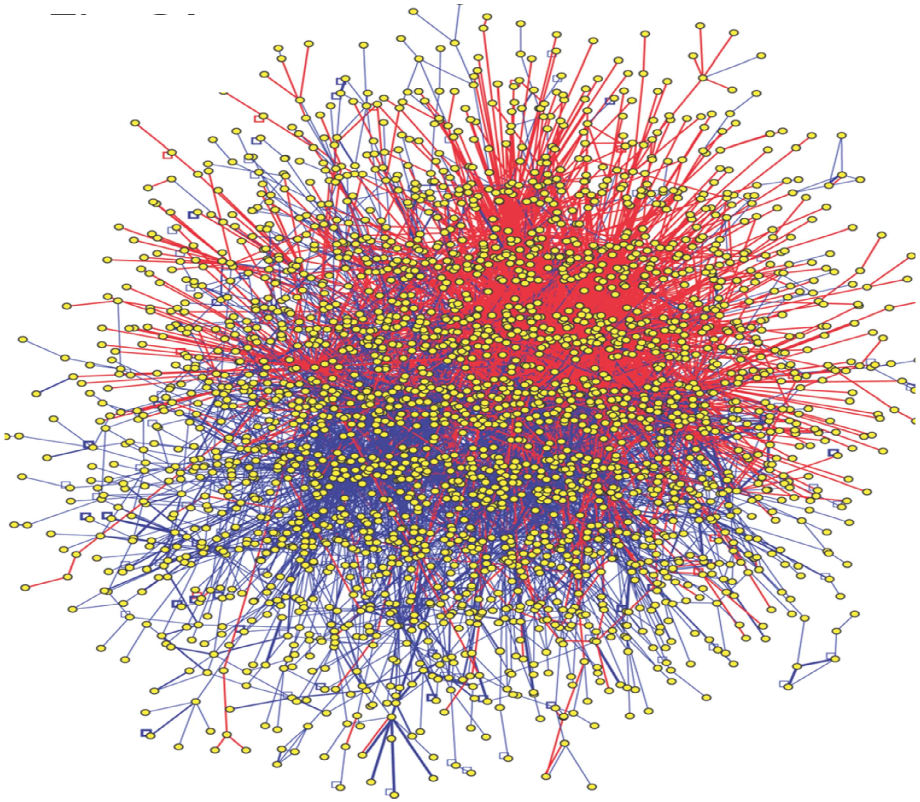

N=4.6x10$^6$
γ=8.4

Fig. 1. Characterizing the large-scale structure and the tie strengths of the mobile call graph. (A and B) Vertex degree (A) and tie strength distribution (B). Each distribution was fitted with $P(x) = a(x + x_0)^{-\gamma} \exp(-x/x_c)$, shown as a blue curve, where $x$ corresponds to either $k$ or $w$. The parameter values for the fits are $k_0 = 10.9$, $\gamma_k = 8.4$, $k_c = \infty$ (A, degree), and $w_0 = 280$, $\gamma_w = 1.9$, $w_c = 3.45 \times$

Onella et al. PNAS 2007

# PLOTTING POWER LAWS

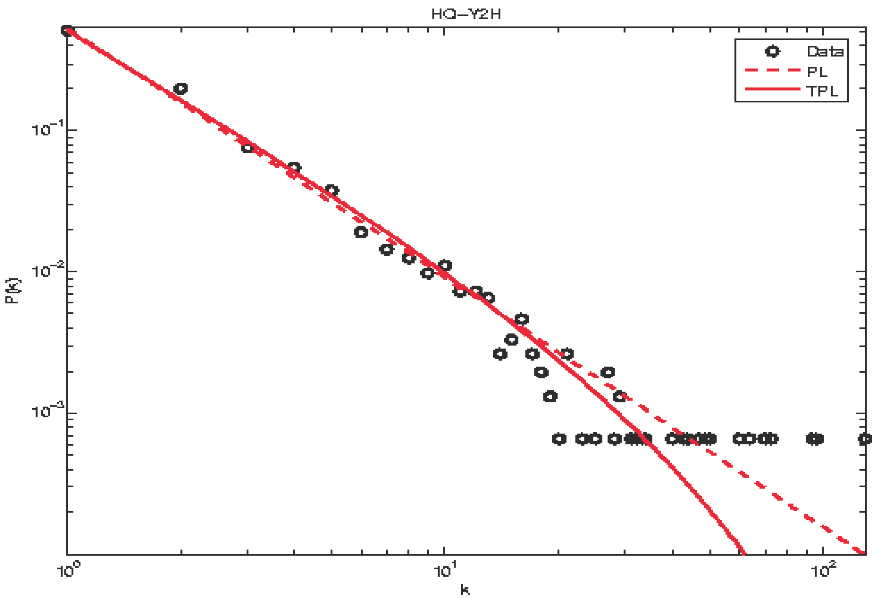
# HUMAN INTERACTION NETWORK



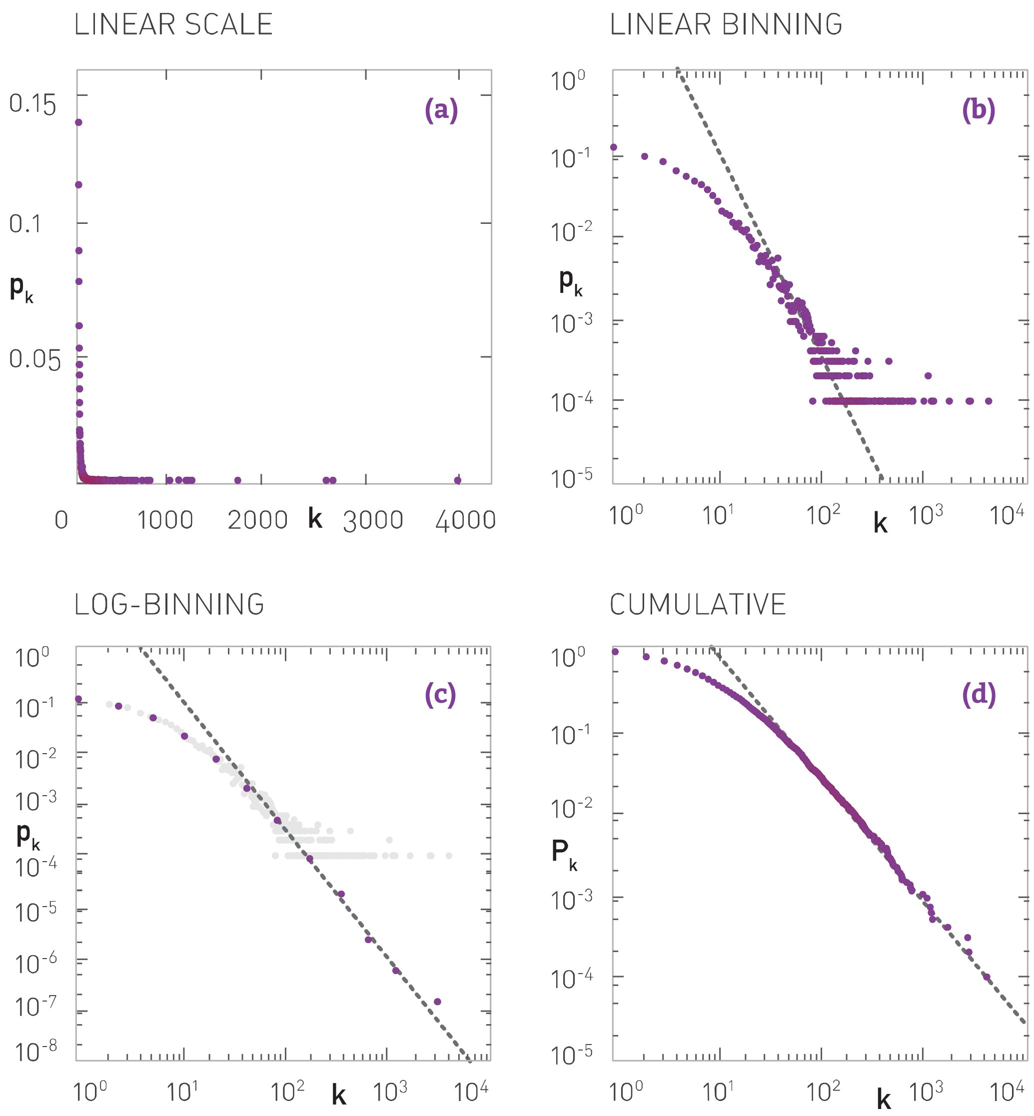
2,800 Y2H interactions

4,100 binary LC interactions

(HPRD, MINT, BIND, DIP, MIPS)




Rual et al. Nature 2005; Stelze et al. Cell 2005



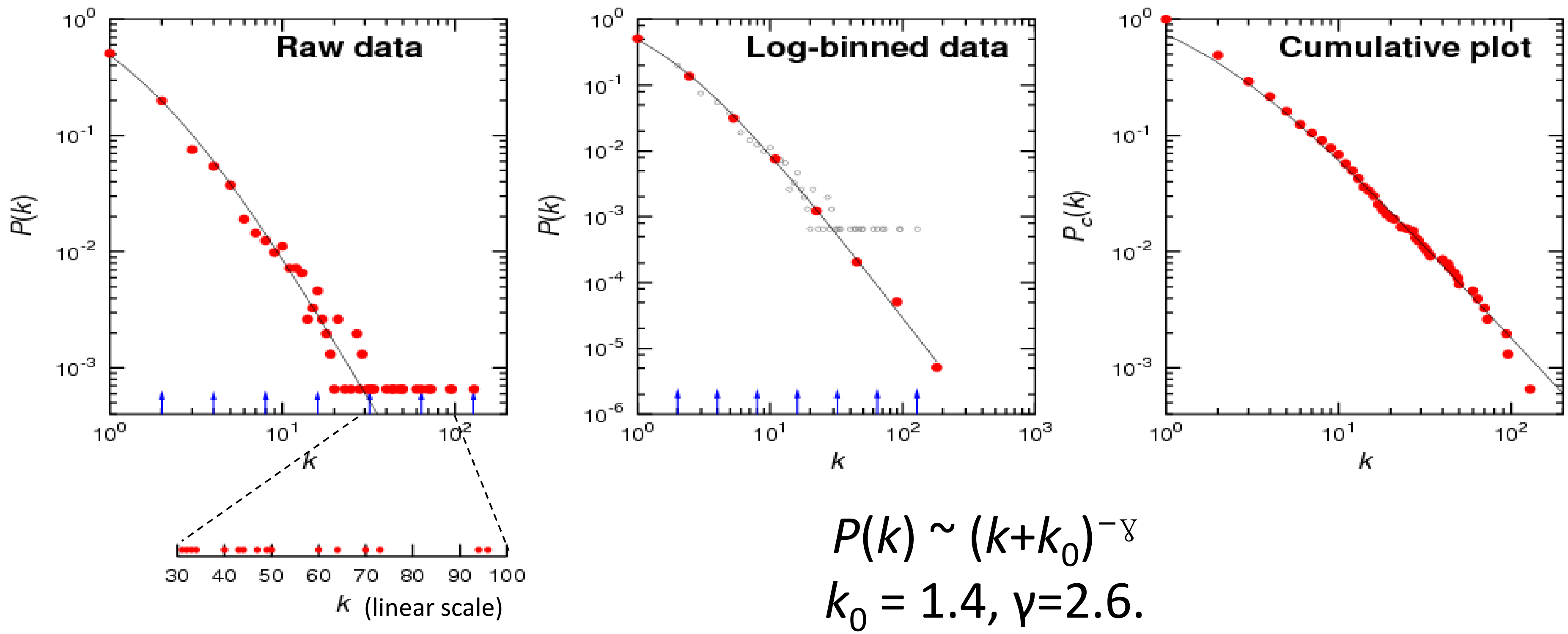

# Use a Log-Log Plot

# Avoid Linear Binning

## Use Logarithmic Binning

## Use Cumulative Distribution

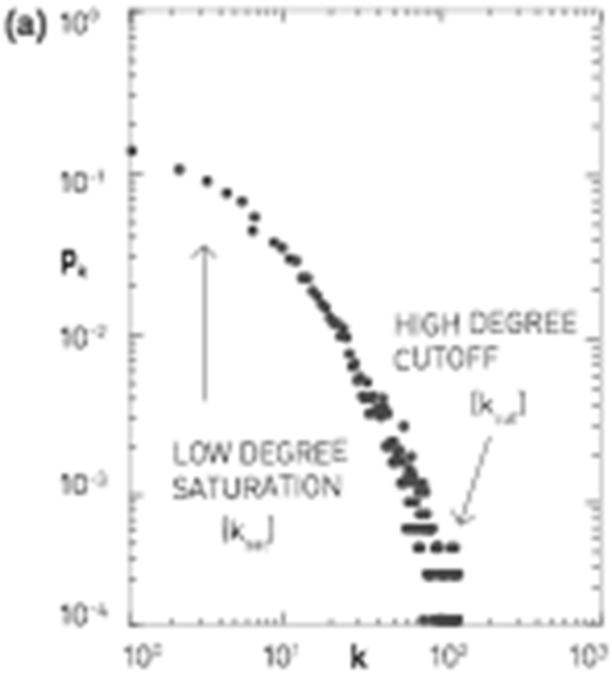
# HUMAN INTERACTION DATA BY RUAL ET AL.




$$P(k) \sim (k+k_0)^{-\gamma}$$

$$k_0 = 1.4,\ \gamma=2.6.$$

# COMMON MISCONCEPTIONS

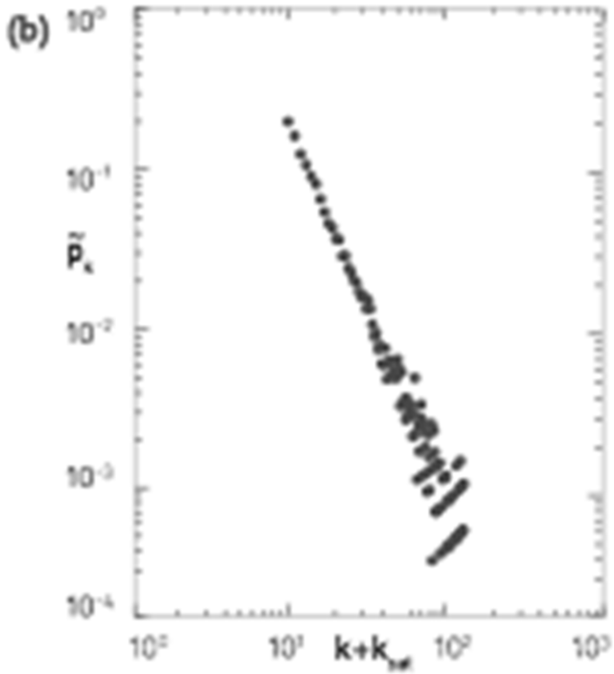



$$p_k = a(k + k_{sat})^{-\gamma} \exp\left(-\frac{k}{k_{cut}}\right),$$

$$\tilde{p}_k = p_k \exp\left(\frac{k}{k_{cut}}\right)$$

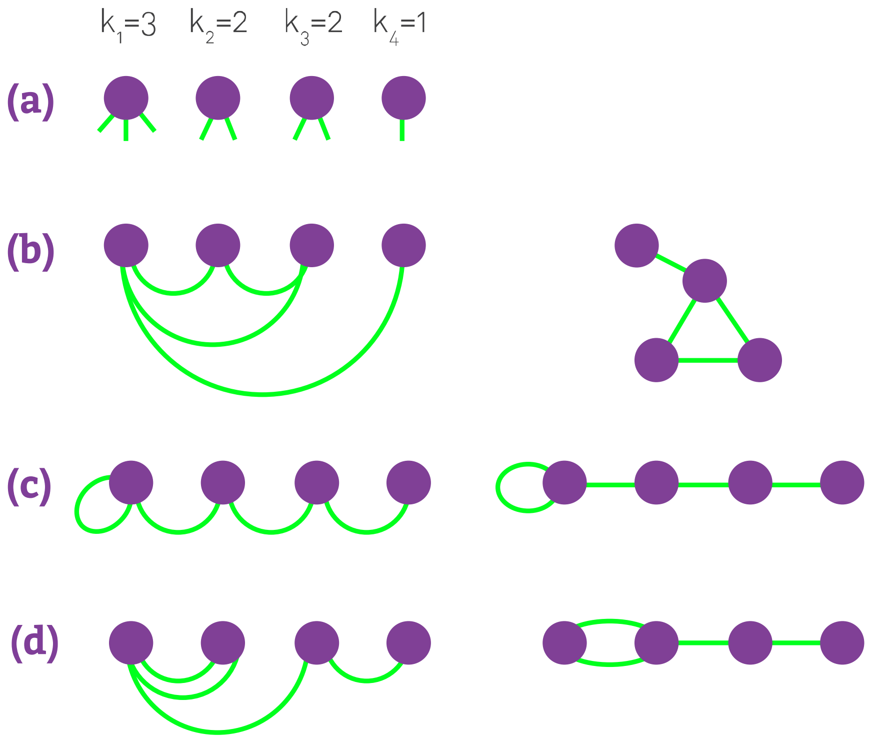
$$\tilde{k} = k + k_{sat},$$

$$\tilde{p} \sim \tilde{k}^{-\gamma},$$



# Generating networks with a pre-defined $p_k$

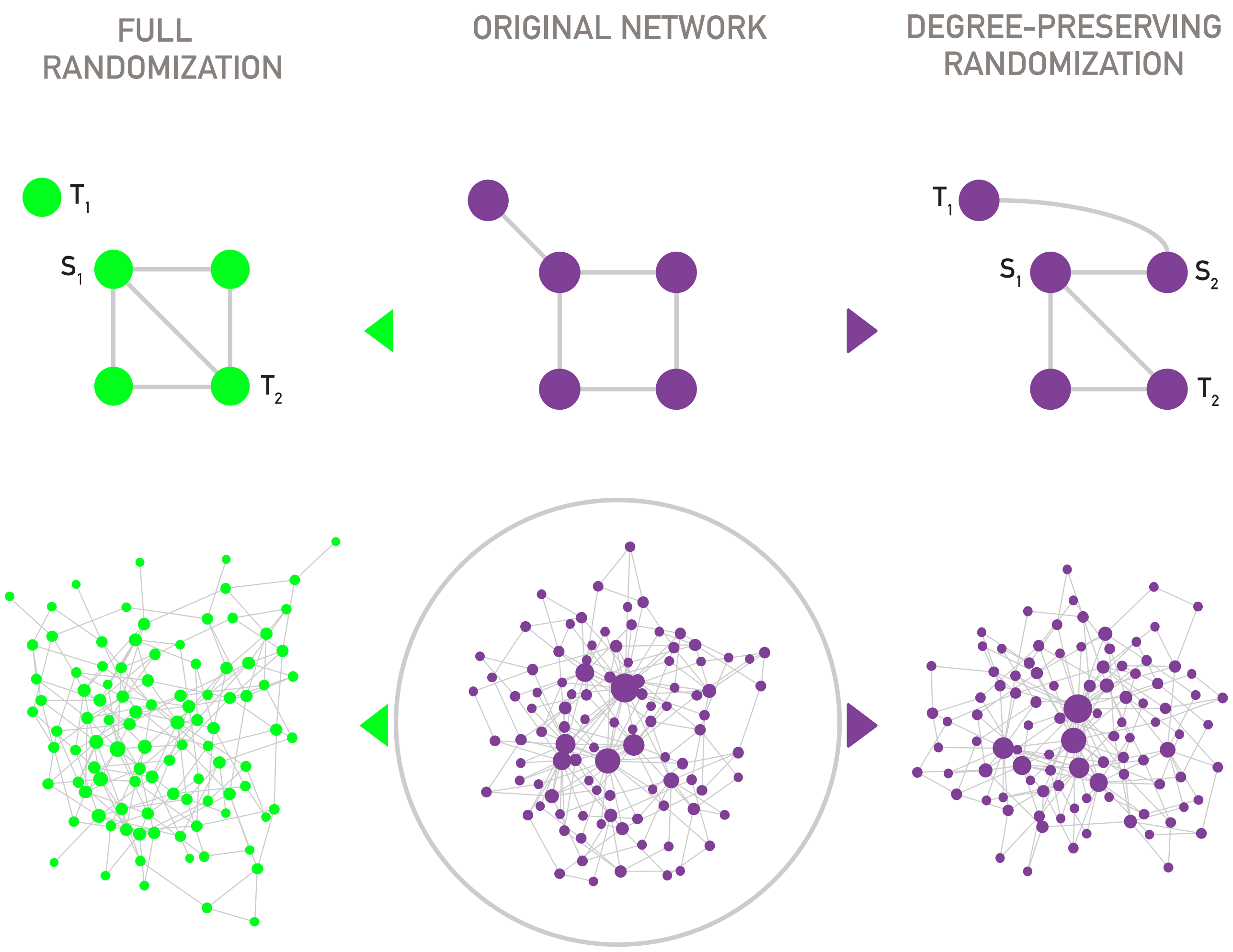
# Configuration model



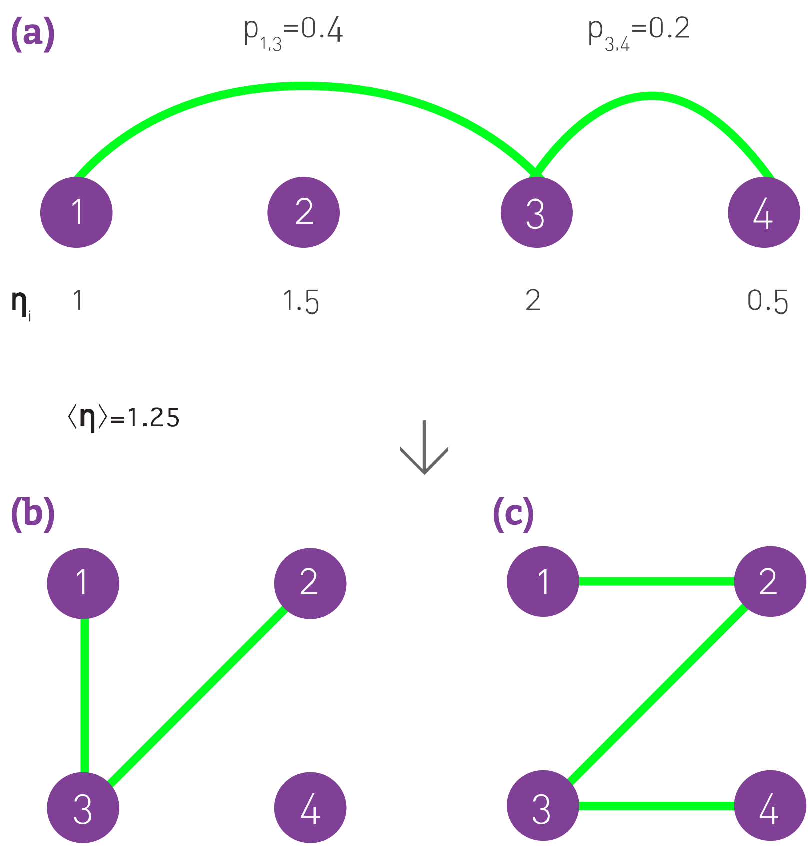

$$p_{ij} = \frac{k_i k_j}{2L - 1}$$

(1) ***Degree sequence****: Assign a degree to each node, represented as stubs or half-links. The degree sequence is either generated analytically from a preselected distribution (Box 4.5), or it is extracted from the adjacency matrix of a real network. We must start from an even number of stubs, otherwise we will be left with unpaired stubs.* *(2)* ***Network assembly****: Randomly select a stub pair and connect them. Then randomly choose another pair from the remaining stubs and connect them. This procedure is repeated until all stubs are paired up. Depending on the order in which the stubs were chosen, we obtain different networks. Some networks include cycles (2a), others self-edges (2b) or multi-edges (2c). Yet, the expected number of self- and multi-edges goes to zero in the limit.*

# Degree Preserving randomization



# Hidden parameter model




$$p(\eta_i, \eta_j) = \frac{\eta_i \eta_j}{\langle \eta \rangle N}$$

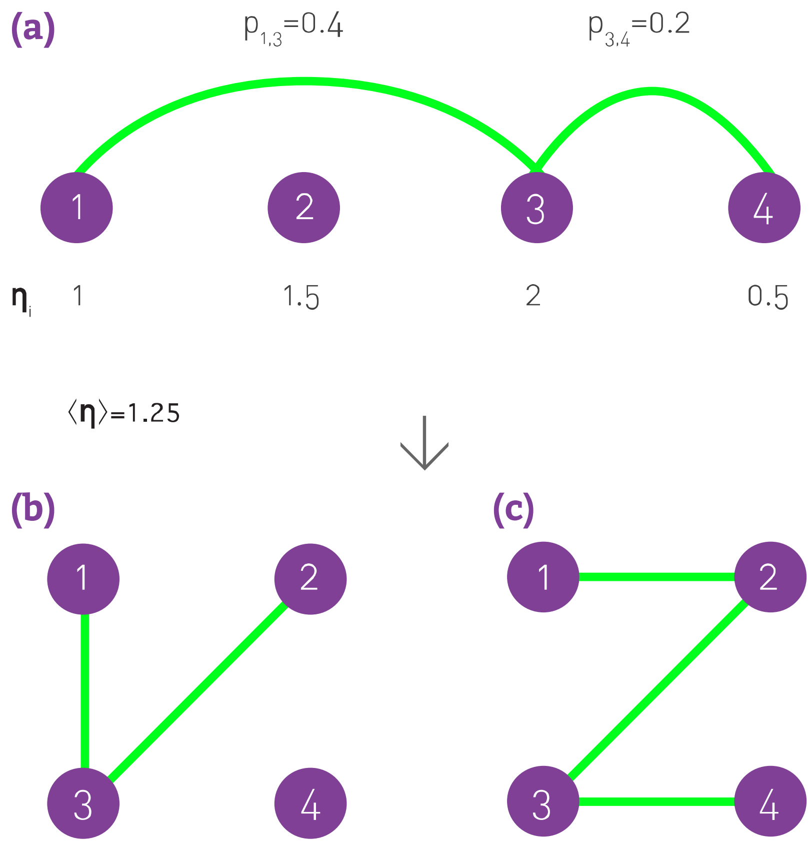
$$p_k = \int \frac{e^{-\eta}\eta^k}{k!} p(\eta) d\eta.$$

$$\{\eta_1, \eta_2, \ldots, \eta_N\}$$

$$p_k = \frac{1}{N} \sum_j \frac{e^{-\eta_j} \eta_j^{\,k}}{k!}.$$

$$\eta_j = \frac{c}{i^\alpha}, i = 1, \ldots, N \qquad p_k \sim k^{-(1+\frac{1}{\alpha})}$$

# Hidden parameter model




Start with N isolated nodes and assign to each node a "hidden parameter" η , which can be randomly selected from a ρ(η) distribution. We next connect each node pair with probability

$$p(\eta_i, \eta_j) = \frac{\eta_i \eta_j}{\langle \eta \rangle N}$$

For example, the figure shows the probability to connect nodes (1,3) and (3,4). After connecting the nodes, we end up with

the networks shown in (b) or (c), representing two independent realizations generated by the same hidden parameter sequence (a). The expected number of links in the obtained network is

$$L = \frac{1}{2} \sum_{N}^{i,j} \frac{\eta_i \eta_j}{\langle \eta \rangle N} = \frac{1}{2} \langle \eta \rangle N$$

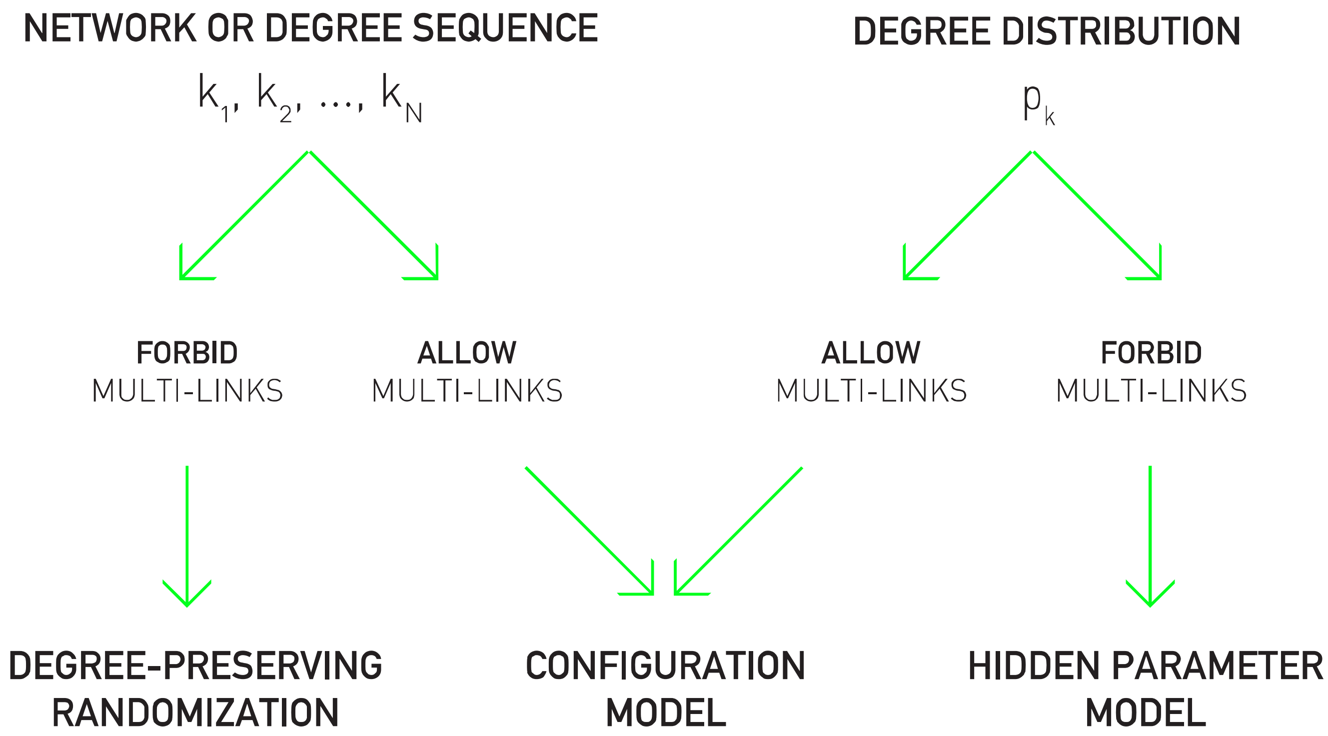
$$\eta_j = \frac{c}{i^{\alpha}}, i = 1, \ldots, N$$

$$p_k = \int \frac{e^{-\eta} \eta^k}{k!} p(\eta) d\eta.$$

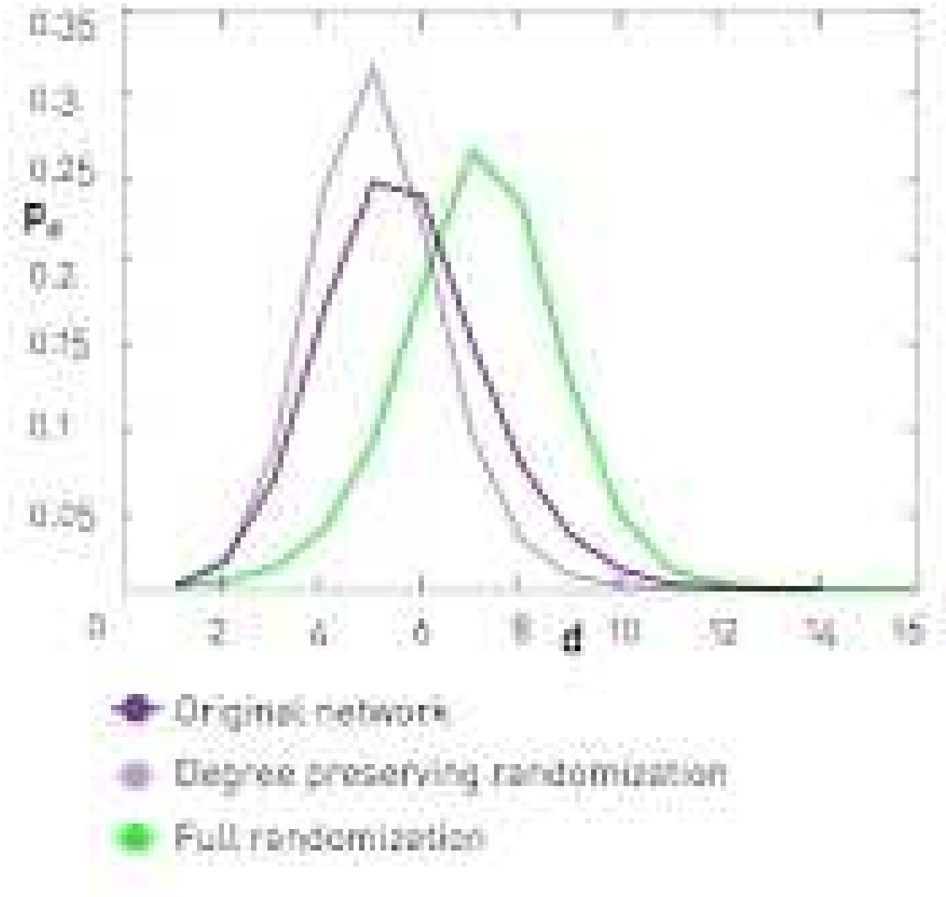
$$\{\eta_1, \eta_2, \ldots, \eta_N\} \quad p_k = \frac{1}{N} \sum_j \frac{e^{-\eta_j} \eta_j^{\,k}}{k!}.$$

$$p_k \sim k^{-(1+\frac{1}{\alpha})}$$

# Decision tree

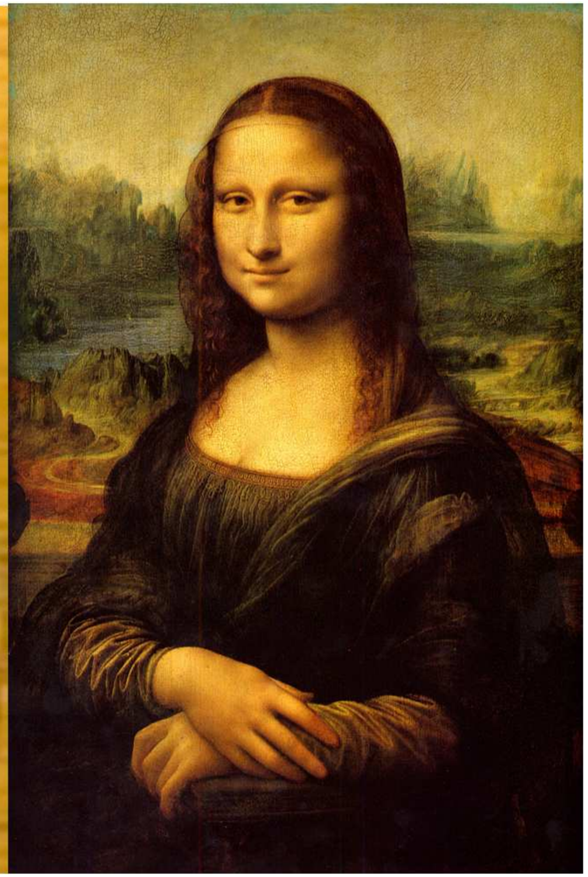


# Case Study: PPI Network




We have: $\langle d \rangle$=5.61±1.64 (original), $\langle d \rangle$=7.13 ± 1.62 (full randomization), $\langle d \rangle$=5.08 ± 1.34 (degree-preserving randomization).

## Something to keep in mind



# summary

# Section 9

DEGREE DISTRIBUTION

Discrete form:

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

Continuous form:

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

SIZE OF THE LARGEST HUB

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

MOMENTS OF $p_k$ *for* $N \to \infty$

$2 < \gamma < 3$: $\langle k \rangle$ finite, $\langle k^2 \rangle$ diverges.

$\gamma > 3$: $\langle k \rangle$ and $\langle k^2 \rangle$ finite.

DISTANCES

$$\langle d \rangle \sim \begin{cases} \text{const.} & \gamma = 2, \\ \frac{\ln \ln N}{\ln(\gamma - 1)} & 2 < \gamma < 3, \\ \frac{\ln N}{\ln \ln N} & \gamma = 3, \\ \ln N & \gamma > 3. \end{cases}$$

## Bounded Networks

We call a network *bounded* if its degree distribution decrease exponentially or faster for high $k$. As a consequence $<k^2>$ is smaller than $<k>$, implying that we lack significant degree variations. Examples of $p_k$ in this class include the Poisson, Gaussian, or the simple exponential distribution (Table 4.2). The Erdős-Rényi and the Watts-Strogatz networks are the best known network models belonging to this class. Bounded networks lack outliers, consequently most nodes have comparable degrees. Real networks in this class include highway networks and the power grid.

## Unbounded Networks

We call a network *unbounded* if its degree distribution has a fat tail in the high-$k$ region. As a consequence $<k^2>$ is much larger than $<k>$, resulting in considerable degree variations. Scale-free networks with a power-law degree distribution (4.1) offer the best known example of networks belonging to this class. Outliers, or exceptionally high-degree nodes, are not only allowed but are expected in these networks. Networks in this class include the WWW, the Internet, the protein interaction networks, and most social and online networks.

# Barabàsi-Albert (BA) model

## Section 1

Hubs represent the most striking difference between a random and a scale-free network. Their emergence in many real systems raises several fundamental questions:

•Why does the random network model of Erdős and Rényi fail to reproduce the hubs and the power laws observed in many real networks?

• Why do so different systems as the WWW or the cell converge to a similar scale-free architecture?

Section 2

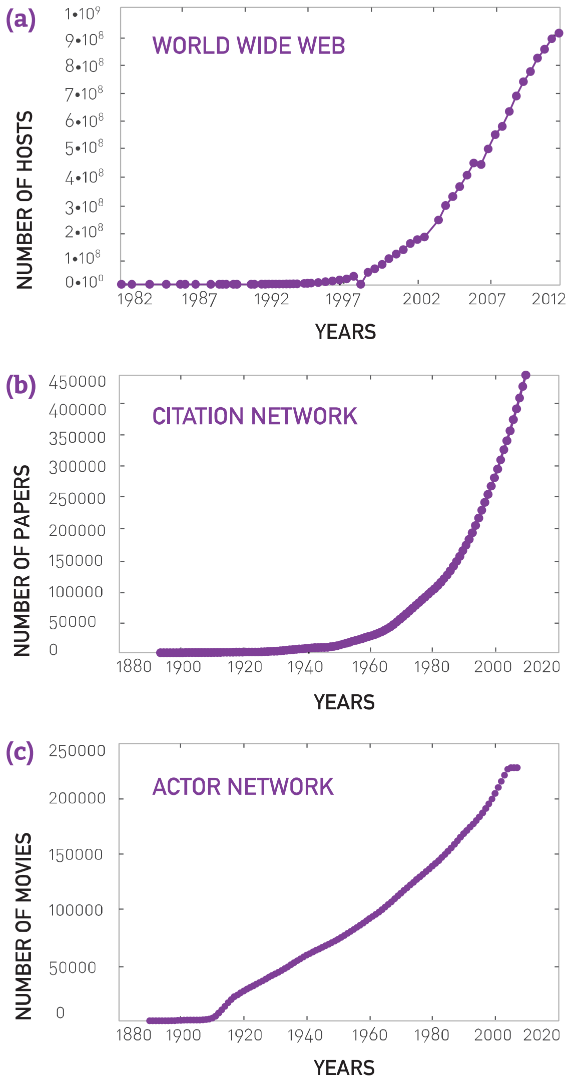
# Growth and preferential attachment

# BA MODEL: Growth

**ER model**:
the number of nodes, N, is fixed (static models)

**networks expand through the addition of new nodes**

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



# BA MODEL: Preferential attachment

ER model: links are added randomly to the network

**New nodes prefer to connect to  the more connected nodes**

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

# Section 2: Growth and Preferential Sttachment

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.

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

# The Barabási-Albert model

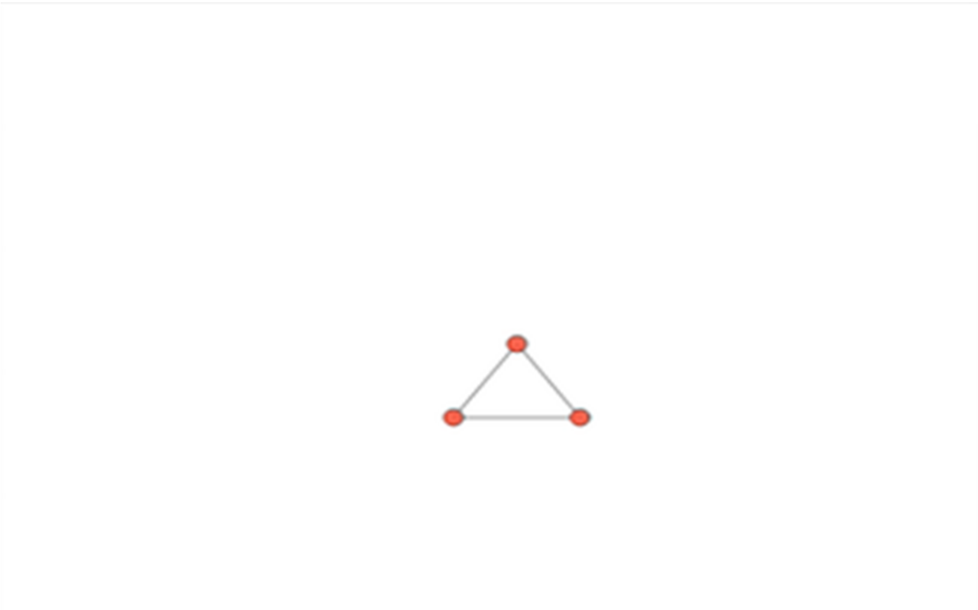
# Origin of SF networks: Growth and preferential attachment

(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)

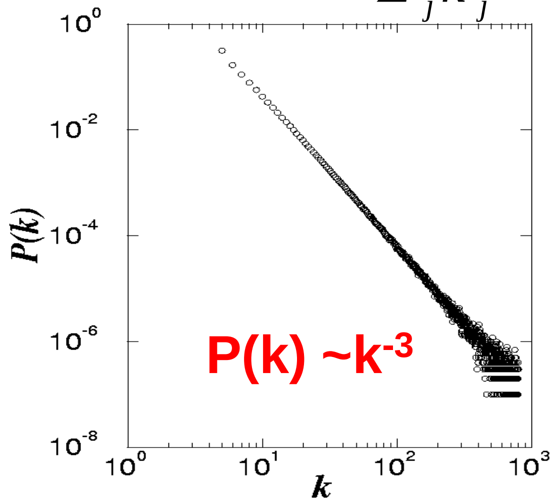
**GROWTH:**

add a new node with m links

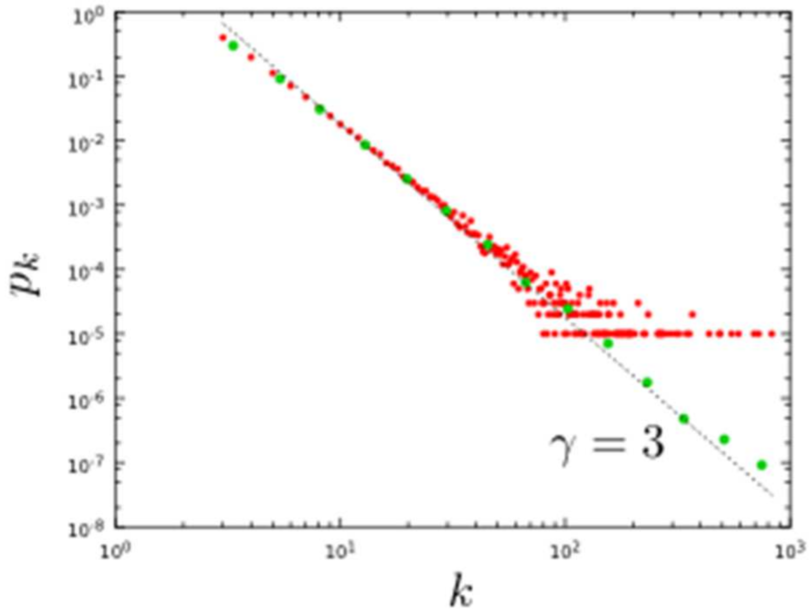
**PREFERENTIAL ATTACHMENT:**

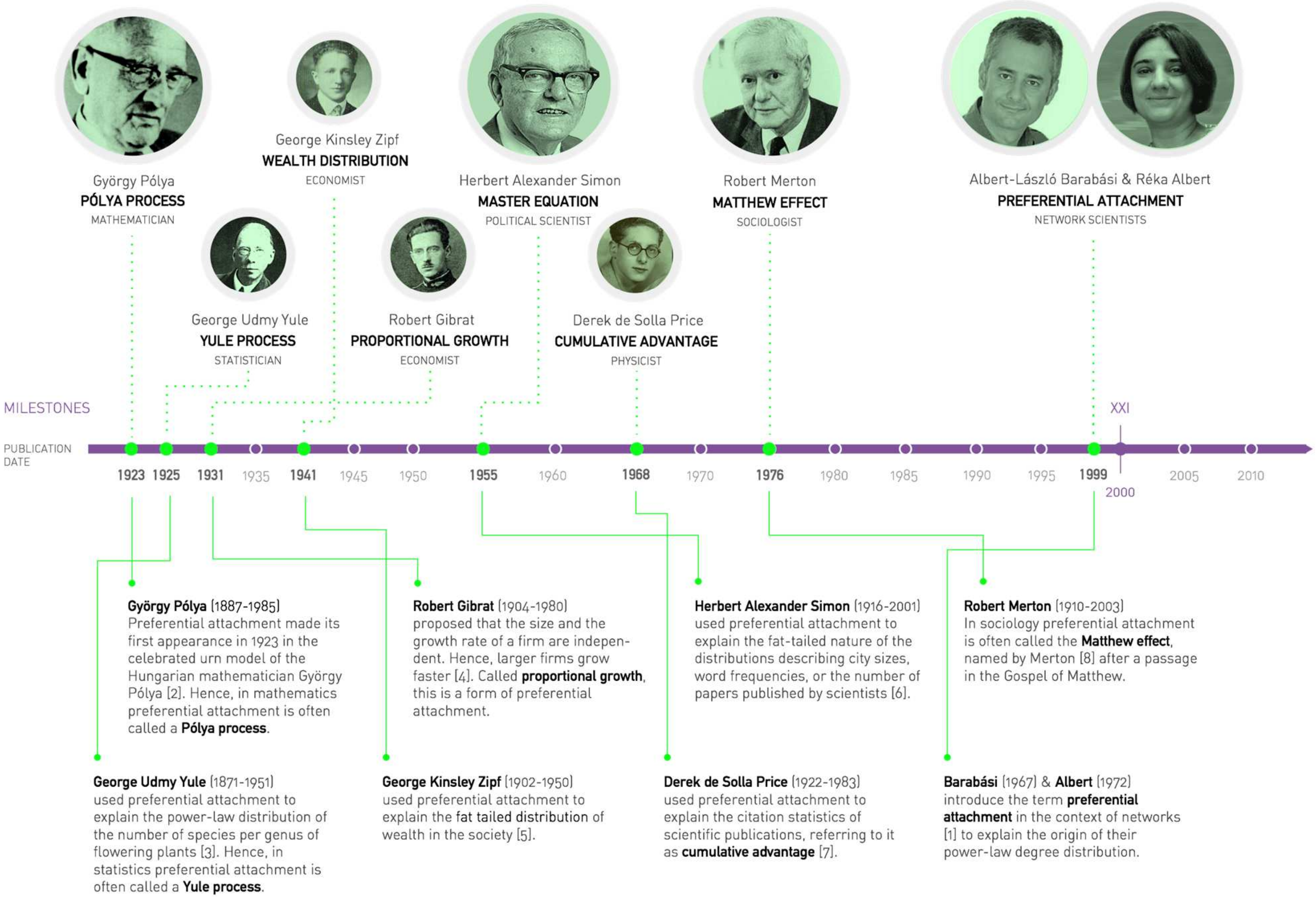
the probability that a node connects to a node with *k* links is proportional to *k*.

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



# Section 4



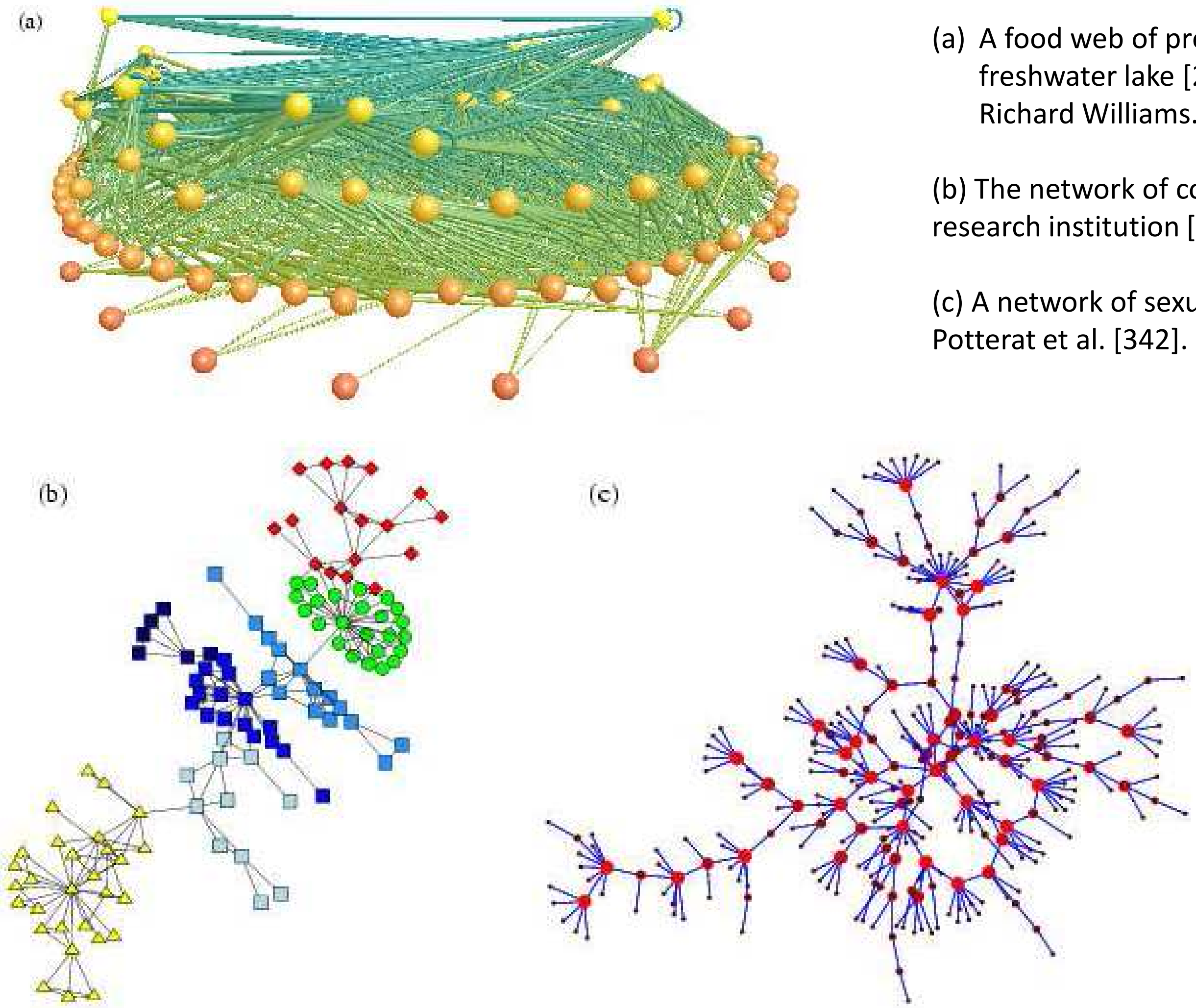


György Pólya
PÓLYA PROCESS
MATHEMATICIAN
George Kinsley Zipf
WEALTH DISTRIBUTION
ECONOMIST
Herbert Alexander Simon
MASTER EQUATION
POLITICAL SCIENTIST
Robert Merton
MATTHEW EFFECT
SOCIOLOGIST
Albert-László Barabási & Réka Albert
PREFERENTIAL ATTACHMENT
NETWORK SCIENTISTS
George Udmy Yule
YULE PROCESS
STATISTICIAN
Robert Gibrat
PROPORTIONAL GROWTH
ECONOMIST
Derek de Solla Price
CUMULATIVE ADVANTAGE
PHYSICIST
MILESTONES
XXI
PUBLICATION DATE
1923 1925 1931 1935 1941 1945 1950 1955 1960 1968 1970 1976 1980 1985 1990 1995 1999 2000 2005 2010
**György Pólya** (1887-1985)
Preferential attachment made its first appearance in 1923 in the celebrated urn model of the Hungarian mathematician György Pólya [2]. Hence, in mathematics preferential attachment is often called a **Pólya process**.
**Robert Gibrat** (1904-1980)
proposed that the size and the growth rate of a firm are independent. Hence, larger firms grow faster [4]. Called **proportional growth**, this is a form of preferential attachment.
**Herbert Alexander Simon** (1916-2001)
used preferential attachment to explain the fat-tailed nature of the distributions describing city sizes, word frequencies, or the number of papers published by scientists [6].
**Robert Merton** (1910-2003)
In sociology preferential attachment is often called the **Matthew effect**, named by Merton [8] after a passage in the Gospel of Matthew.
**George Udmy Yule** (1871-1951)
used preferential attachment to explain the power-law distribution of the number of species per genus of flowering plants [3]. Hence, in statistics preferential attachment is often called a **Yule process**.
**George Kinsley Zipf** (1902-1950)
used preferential attachment to explain the fat tailed distribution of wealth in the society [5].
**Derek de Solla Price** (1922-1983)
used preferential attachment to explain the citation statistics of scientific publications, referring to it as **cumulative advantage** [7].
**Barabási** (1967) & **Albert** (1972)
introduce the term **preferential attachment** in the context of networks [1] to explain the origin of their power-law degree distribution.

**Section 3**

# The structure and function of complex networks

# Section 4




(a) A food web of predator-prey interactions between species in a freshwater lake [272]. Picture courtesy of Neo Martinez and Richard Williams.

(b) The network of collaborations between scientists at a private research institution [171].

(c) A network of sexual contacts between individuals in the study by Potterat et al. [342].

# Degree Correlation

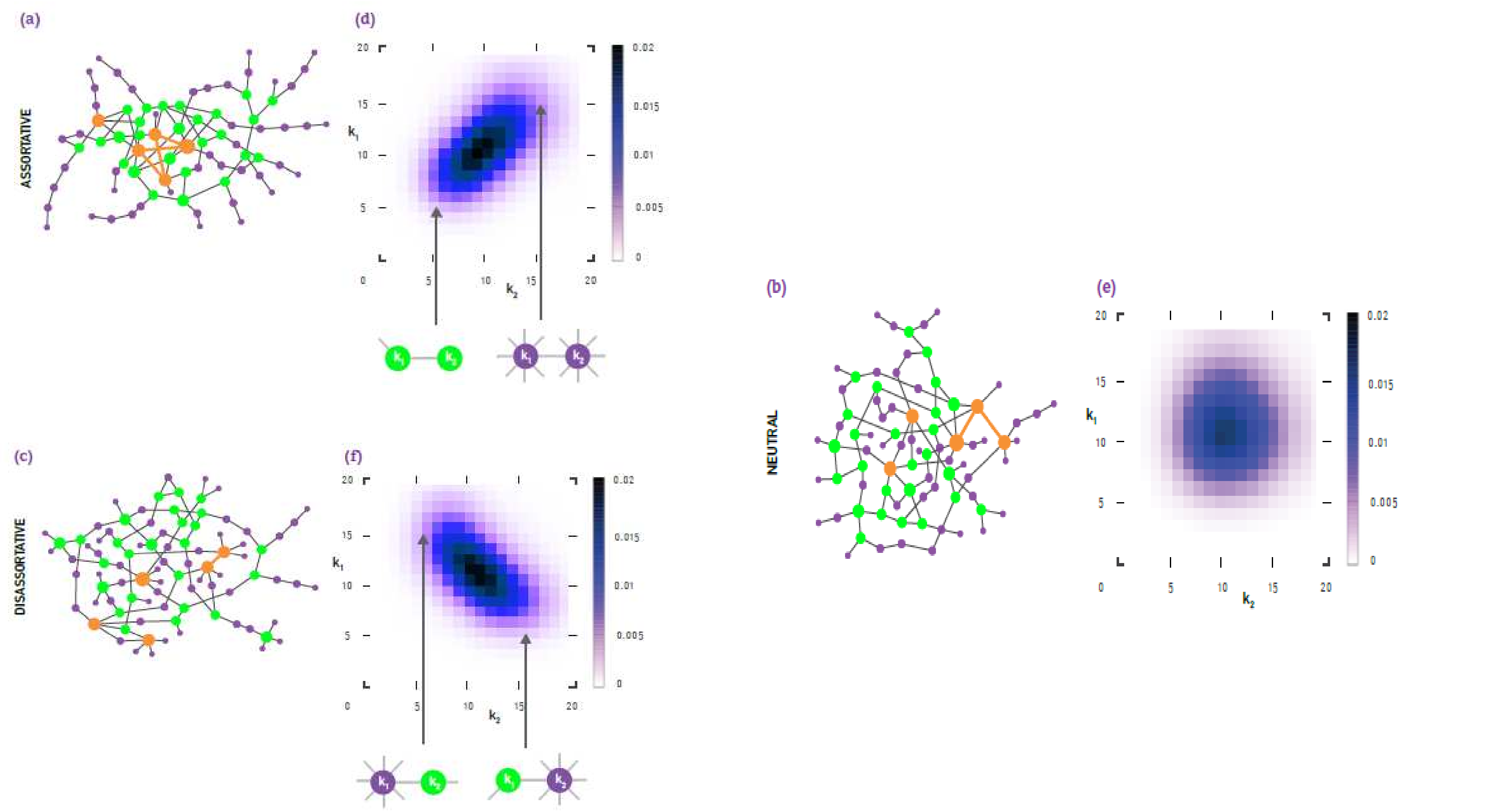
The probability that nodes with degrees *k* and *k*′ link to each other:

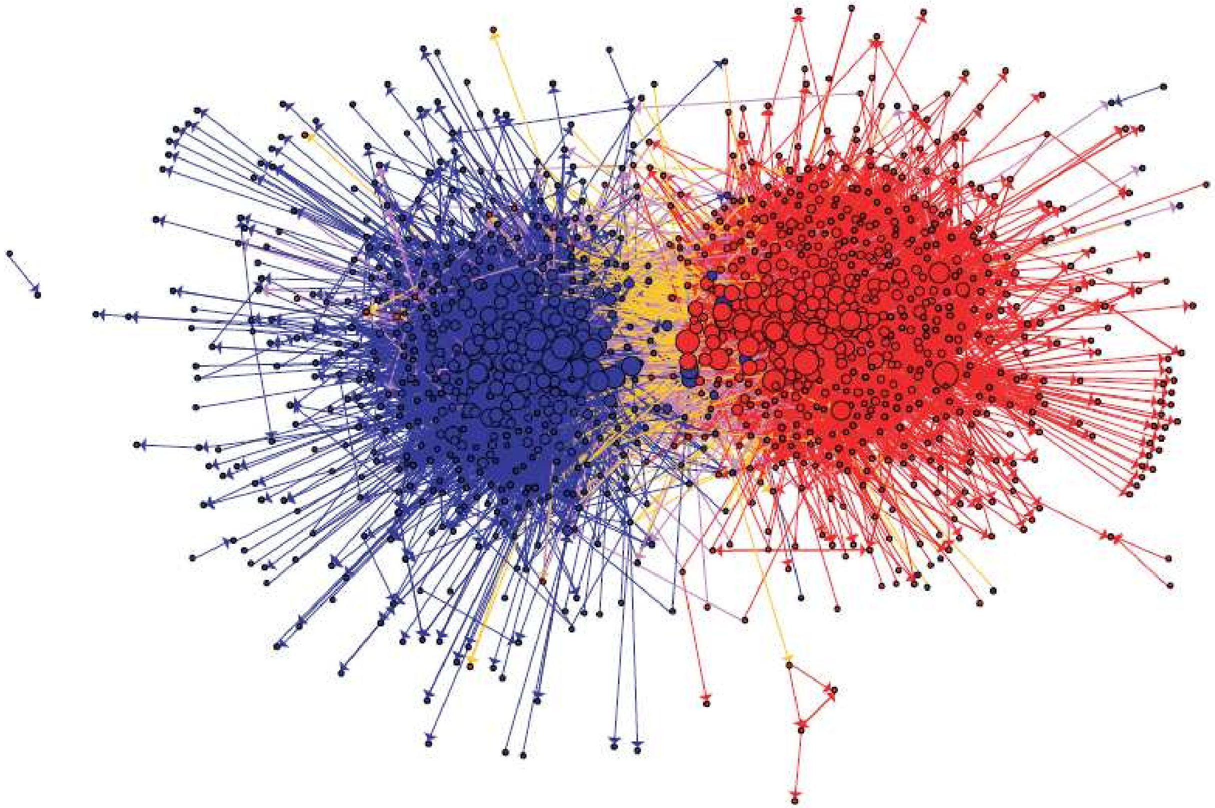
$$p_{k,k'} = \frac{kk'}{2L}.$$

Celebrity Couples
**Hubs Dating Hubs**

Protein interaction map of yeast
**Hubs Avoiding hubs**

# Section 4 Associativity and Dissociativity





**Politics is Never Neutral**

The network behind the US political blogosphere illustrates the presence of associative mixing, as used in sociology, nodes of similar caracteristics tend to link to each other.

# Network Robustness

# Section 4

“Robust” comes from the latin Quercus Robur, meaning oak, the symbol of strength and longevity in the ancient world.

The tree in the figure stands near the Hungarian village Diosviszlo and is documented at www.dendromania.
hu, a site that catalogs Hungary's oldest and largest trees.

Image courtesy of Gyorgy Posfai.

## Section 4

| NETWORK | RANDOM FAILURES (REAL NETWORK) | RANDOM FAILURES (RANDOMIZED NETWORK) | ATTACK (REAL NETWORK) |
|---|---|---|---|
| Internet | 0.92 | 0.84 | 0.16 |
| WWW | 0.88 | 0.85 | 0.12 |
| Power Grid | 0.61 | 0.63 | 0.20 |
| Mobile-Phone Call | 0.78 | 0.68 | 0.20 |
| Email | 0.92 | 0.69 | 0.04 |
| Science Collaboration | 0.92 | 0.88 | 0.27 |
| Actor Network | 0.98 | 0.99 | 0.55 |
| Citation Network | 0.96 | 0.95 | 0.76 |
| E. Coli Metabolism | 0.96 | 0.90 | 0.49 |
| Yeast Protein Interactions | 0.88 | 0.66 | 0.06 |

**Breakdown Thresholds Under Random Failures and Attacks**

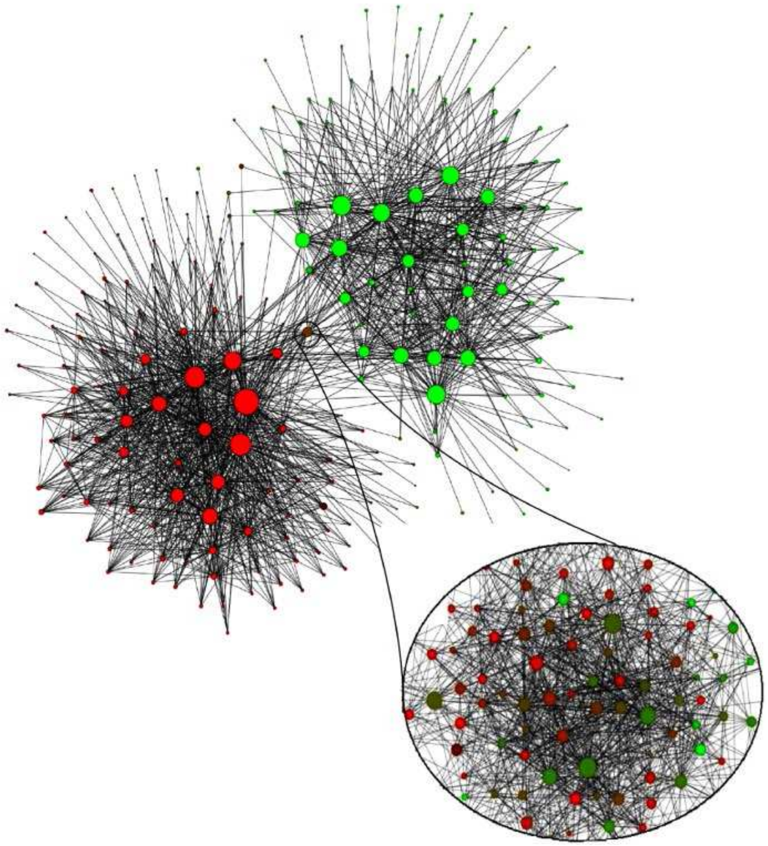
The table shows the estimated *fc* for random node failures (second column) and attacks (fourth column) for ten reference networks. The third column (randomized network) offers *fc* for a network whose *N* and *L* coincides with the original network, but whose nodes are connected randomly to each other (randomized network.

For most networks *fc* for random failures exceeds *fc* for the corresponding randomized network, indicating that these networks display enhanced robustness. Three networks lack

this property: the power grid, a consequence of the fact that its degree distribution is exponential and the actor and the citation networks, which have a very high $\langle k \rangle$, diminishing the role of the high $\langle k2 \rangle$.

# Communities

# Section 4



**Communities in Belgium**
Communities extracted from the call pattern of the consumers of the largest Belgian mobile phone company. The network has about two million mobile phone users. The nodes correspond to communities, the size of each node being proportional to the number of individuals in the corresponding community.
The color of each community on a red–green scale represents the language spoken in the particular community, red for French and green for Dutch. Only communities of more than 100 individuals are shown. The community that connects the two main clusters consists of several smaller communities with less obvious language separation, capturing the culturally mixed Brussels, the country’s capital.

# Section 4

## HIERARCHICAL CLUSTERING

**AGGLOMERATIVE PROCEDURES: THE RAVASZ ALGORITHM**

Step 1: Define the Similarity Matrix
Step 2: Decide Group Similarity
Step 3: Apply Hierarchical Clustering
Step 4: Dendrogram

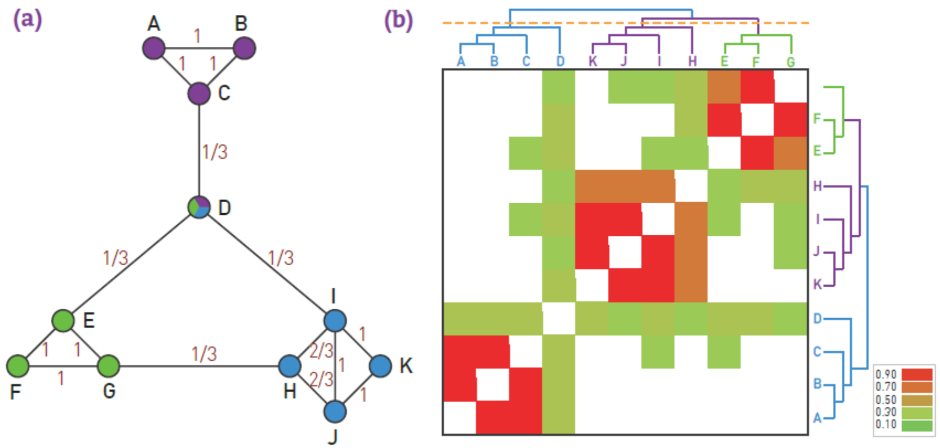
**DIVISIVE PROCEDURES: THE GIRVAN-NEWMAN ALGORITHM**

Step 1: Define Centrality
Step 2: Hierarchical Clustering

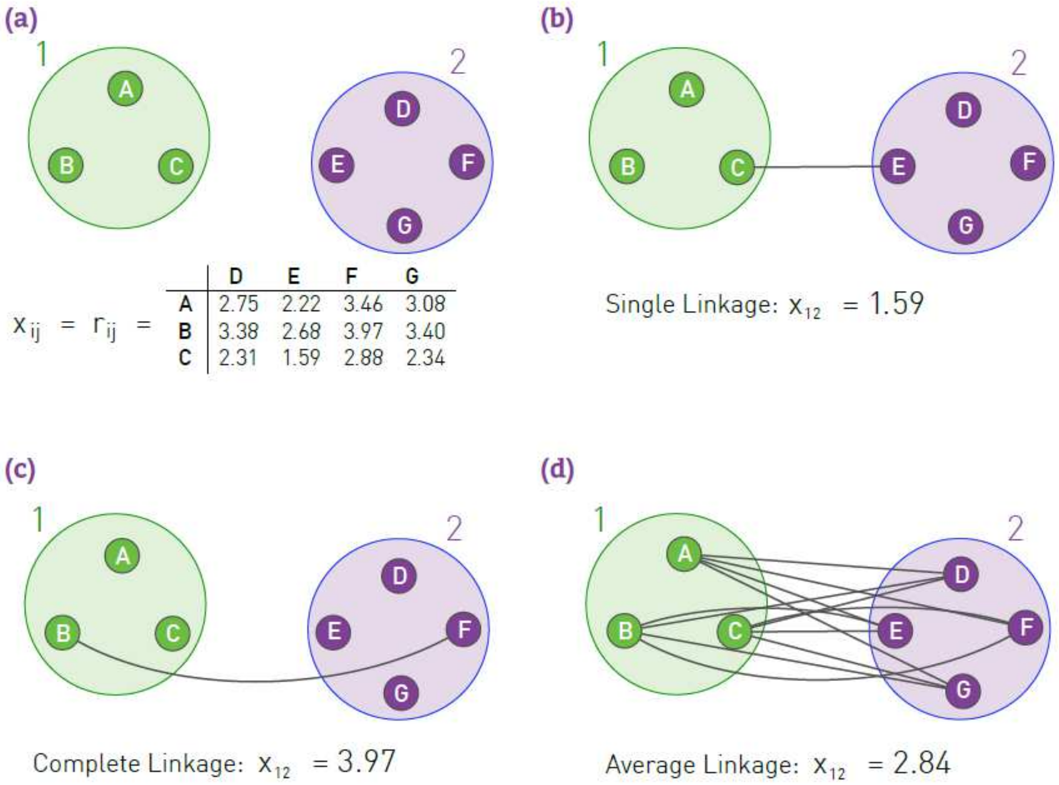
## Topological Overlap Matrix

$$x_{ij}^{o} = \frac{J(i,j)}{\min(k_i, k_j) + 1 - \Theta(A_{ij})}$$

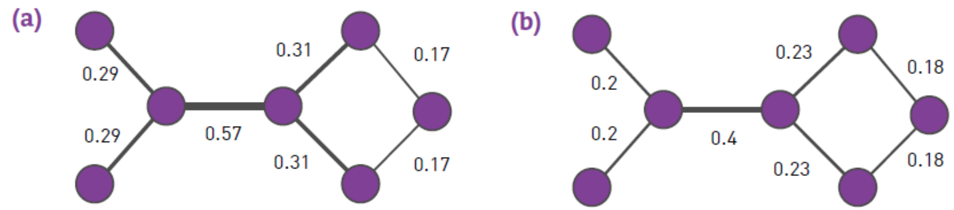
Here $\Theta$(x) is the Heaviside step function, which is zero for x≤0 and one for x>0; *J*(*i, j*) is the number of common neighbors of node *i* and *j*, to which we add one (+1) if there is a direct link between *i* and *j*; min(*ki,kj*) is the smaller of the degrees *ki* and *kj*.



## Cluster Similarity



## Centrality Measure




**Centrality Measures**

Divisive algorithms require a centrality measure that is high for nodes that belong to different communities and is low for node pairs in the same community. Two frequently used measures can achieve this:

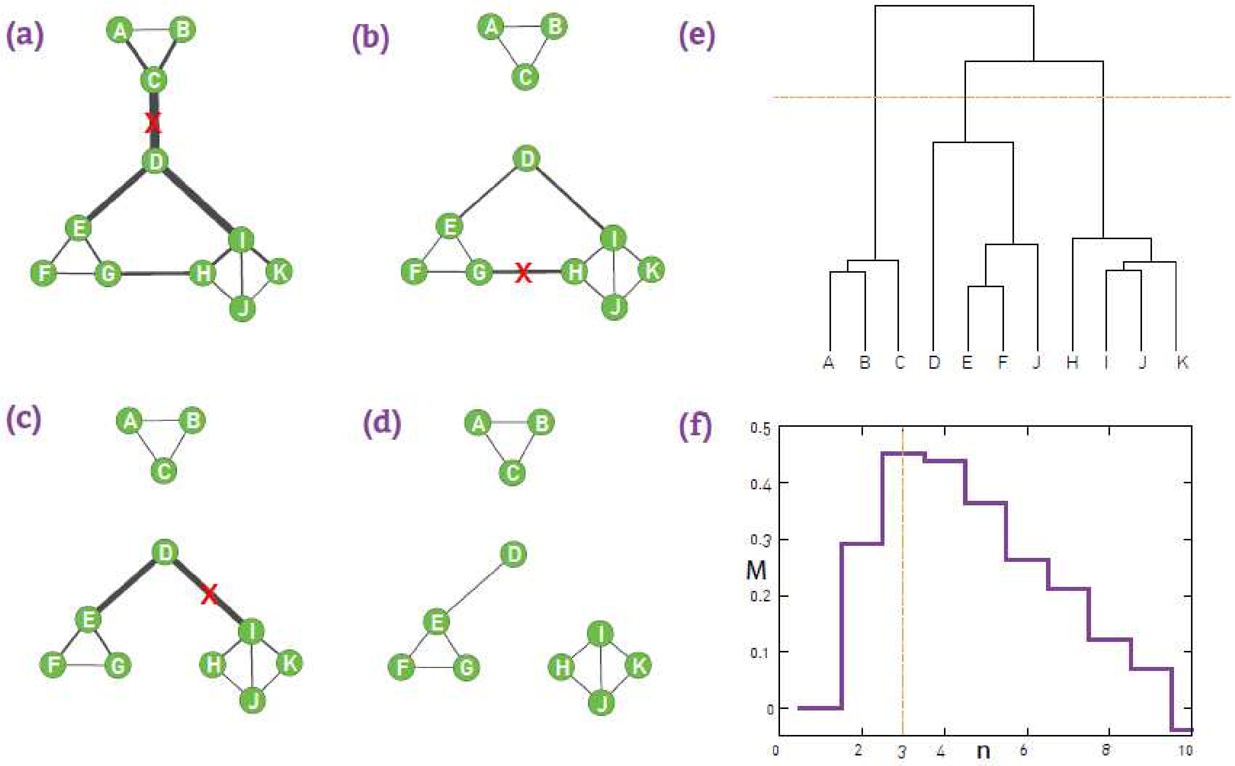
**(a) Link Betweenness**

Link betweenness captures the role of each link in information transfer. Hence x*ij* is proportional to the number of shortest paths between all node pairs that run along the link (*i,j*). Consequently, inter-community links, like the central link in the figure with *xij* =0.57, have large betweenness.

**(b) Random-Walk Betweenness**

A pair of nodes *m* and *n* are chosen at random. A walker starts at *m*, following each adjacent link with equal probability until it reaches *n*. *Random walk betweenness xij* is the probability that the link *i*→*j* was crossed by the walker after averaging over all possible choices for the starting nodes *m* and *n*.

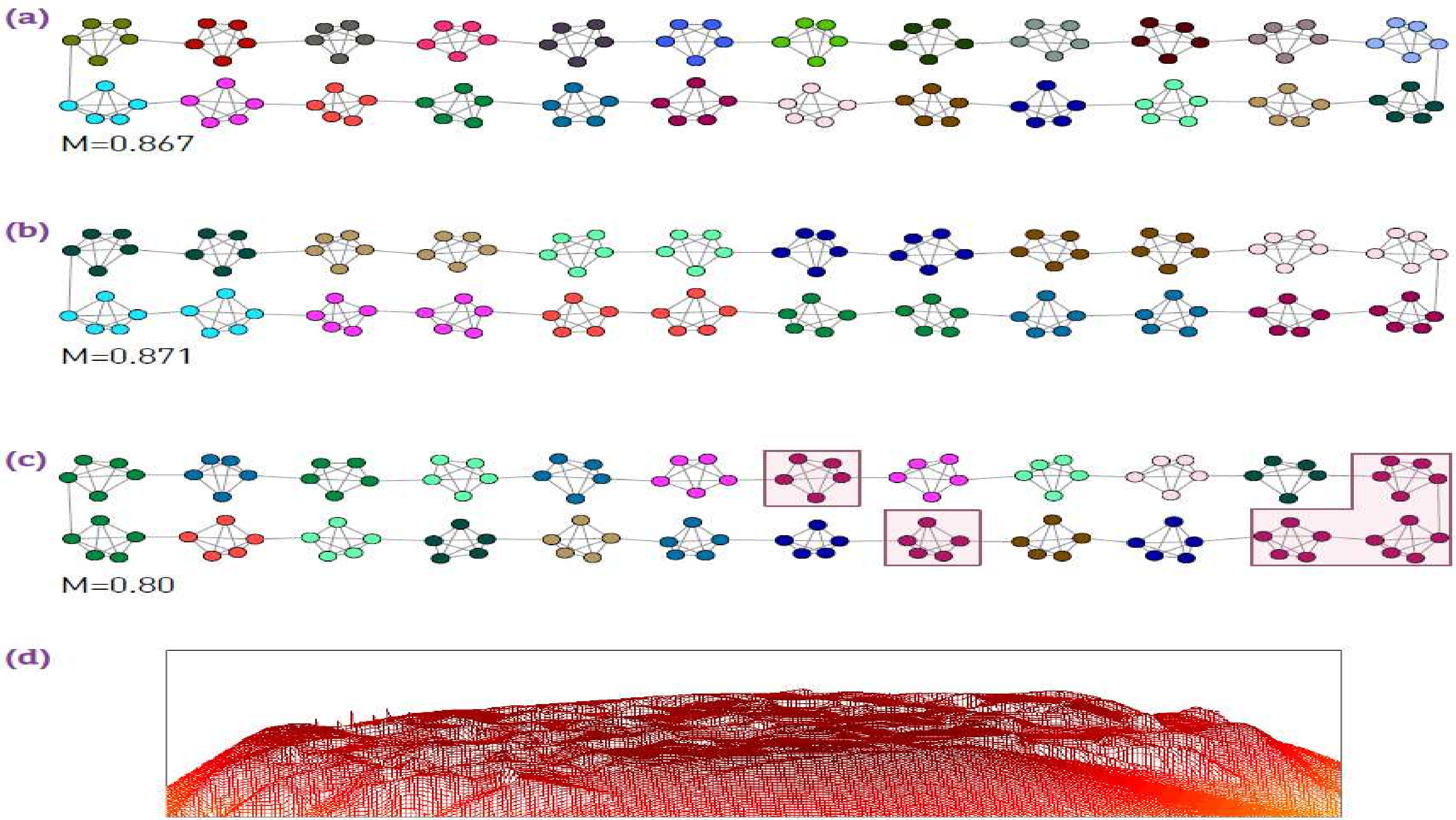
## The Girvan-Newman Algorithm

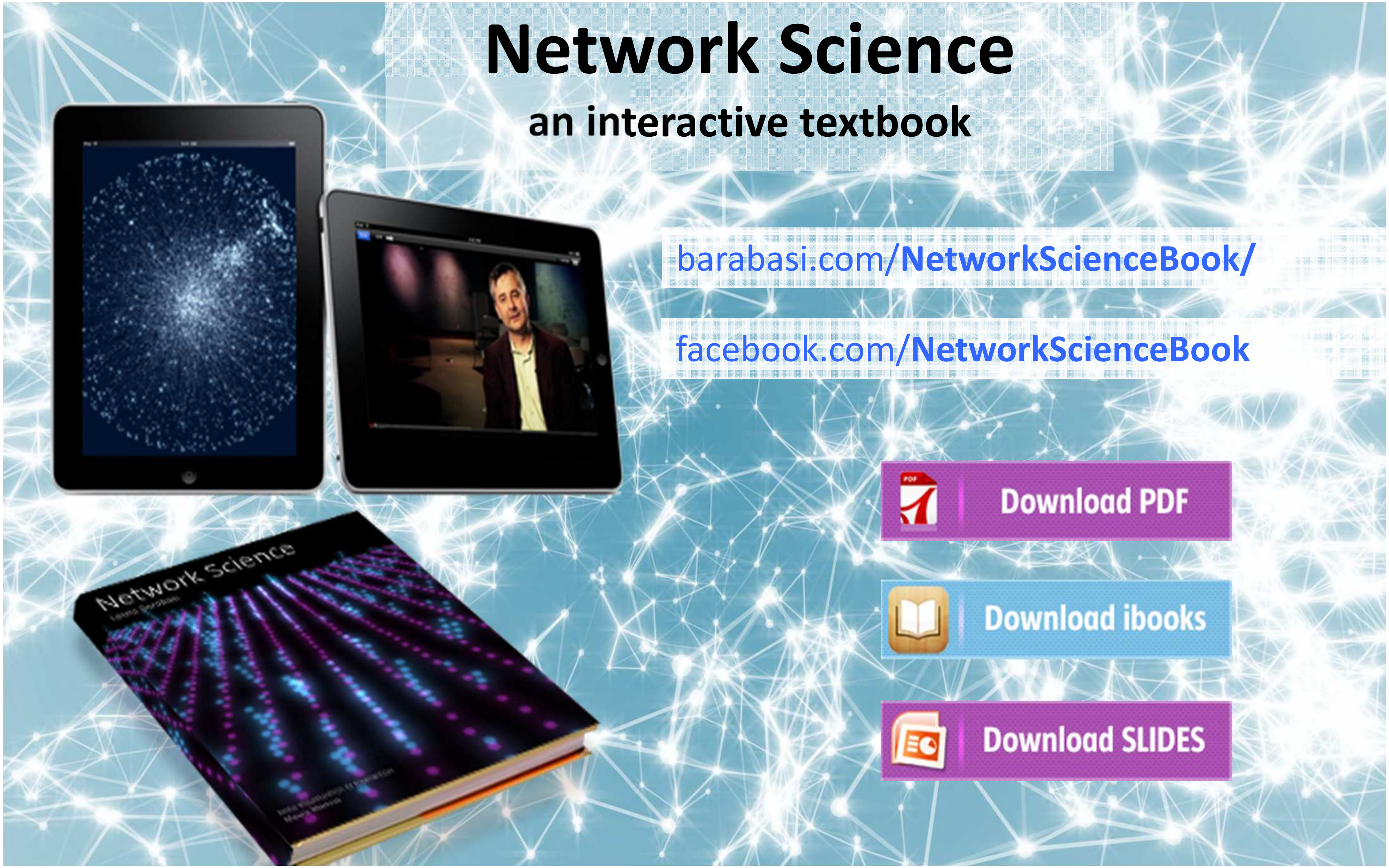



The **partition's *modularity*** *is obtained* by summing over all *nc* communities

$$M = \sum_{c=1}^{n_c} \left[ \frac{L_c}{L} - \left( \frac{k_c}{2L} \right)^2 \right].$$

# Section 4




Network Science
an interactive textbook
barabasi.com/NetworkScienceBook/
facebook.com/NetworkScienceBook
Network Science
Download PDF
Download ibooks
Download SLIDES