Network Science

Class 3: Random Networks

(Chapter 3 in textbook)

Albert-László Barabási

with

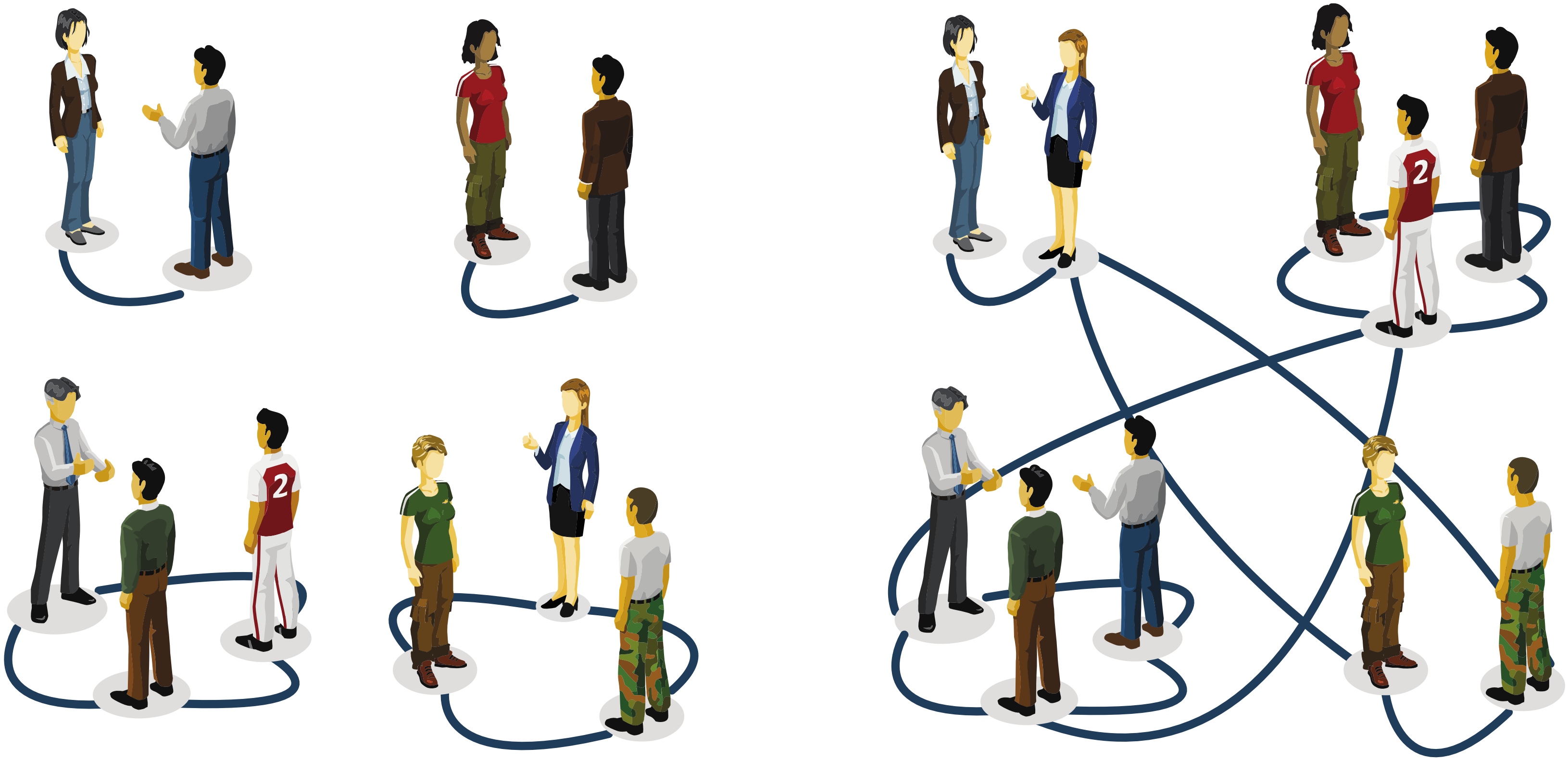
Emma K. Towlson, Michael Danziger,
Sebastian Ruf, Louis Shekhtman

www.BarabasiLab.com

Section 1

Introduction

RANDOM NETWORK MODEL

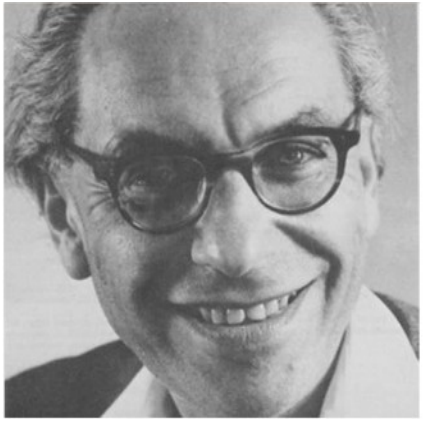


Section 3.2

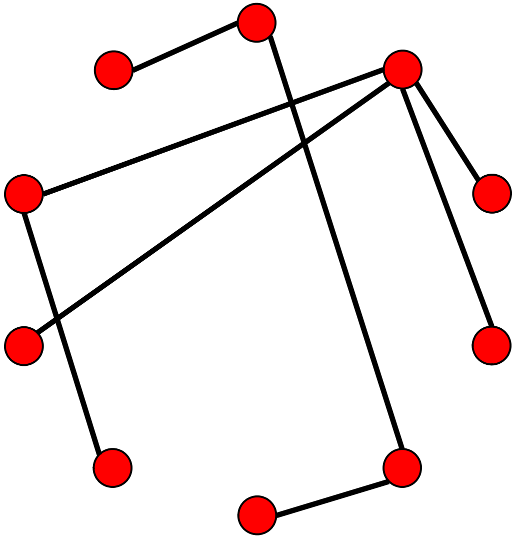
The random network model

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**

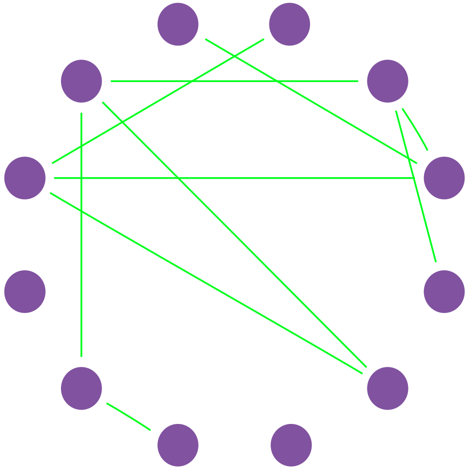
N labeled nodes are connected with *L* randomly placed links. Erdős and Rényi used this definition in their string of papers on random networks [2-9].

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

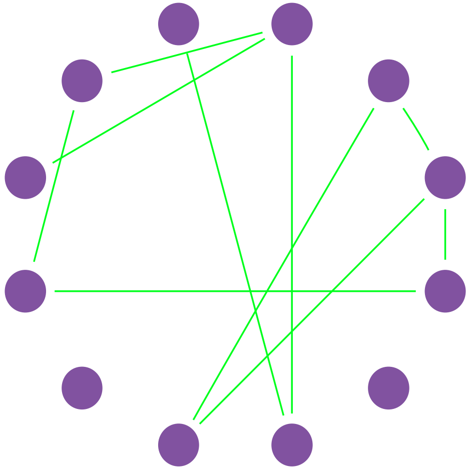
Each pair of N labeled nodes is connected with probability *p*, a model introduced by Gilbert [10].

RANDOM NETWORK MODEL

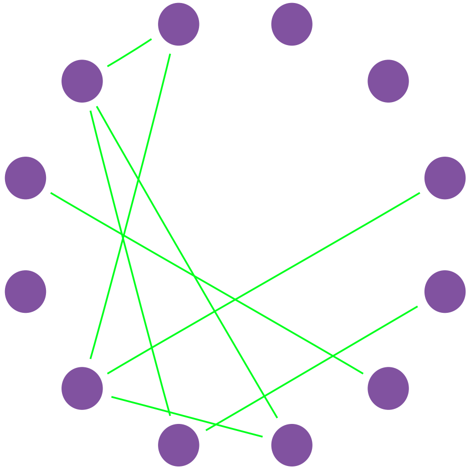
p=1/6
N=12



L=8
Prob=?



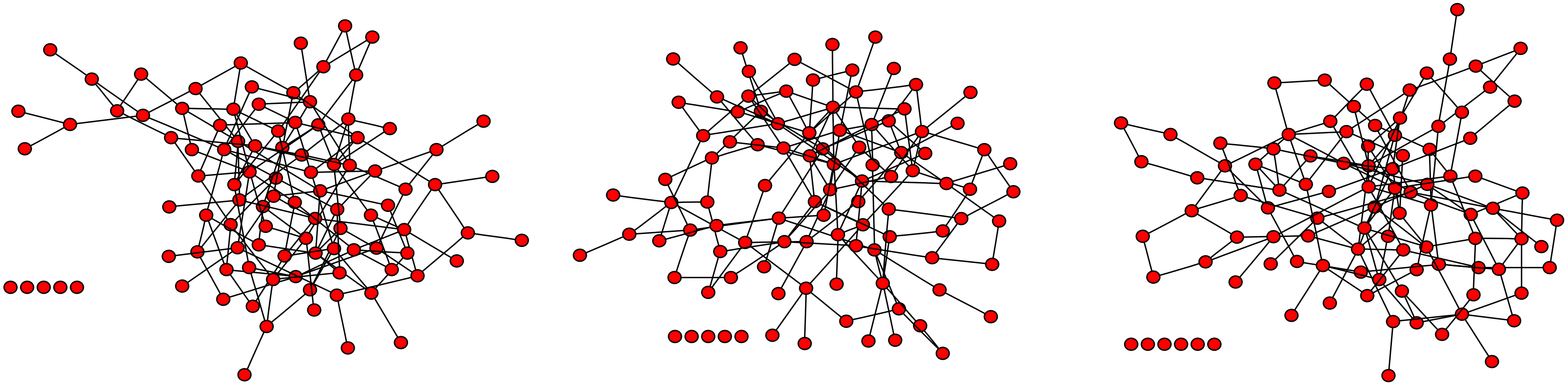
L=10
Prob=?



L=7
Prob=?

RANDOM NETWORK MODEL

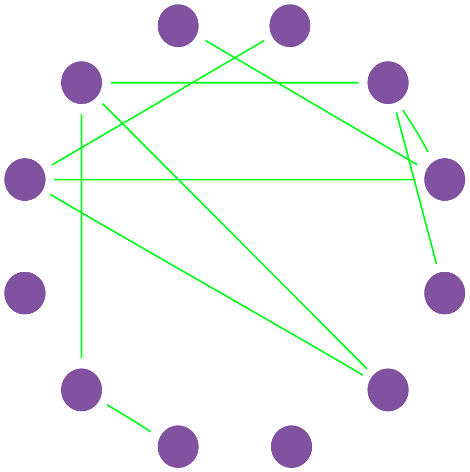
p=0.03
N=100



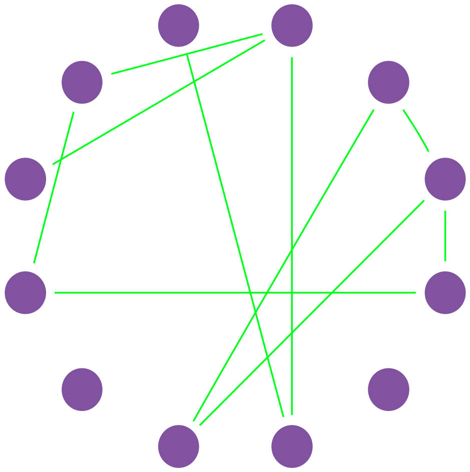
The number of links is variable

RANDOM NETWORK MODEL

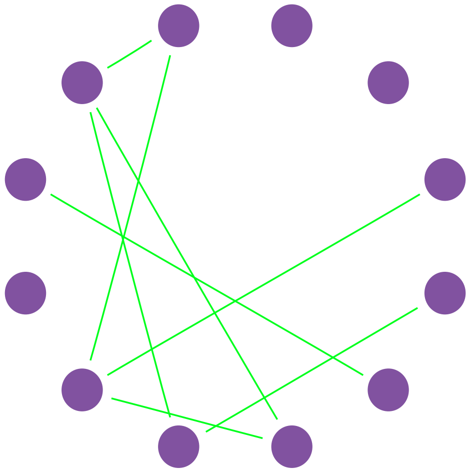
p=1/6
N=12



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.

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

Binomial distribution...

MATH TUTORIAL | Binomial Distribution: The bottom line

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

$$<x> = pN$$

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

$$s_x=(<x^2>-<x>^2)^{1/2}=[p(1-p)N]^{1/2}$$

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

$$P(L)=\left(\begin{pmatrix} N \\ 2 \end{pmatrix} \atop L\right) p^{L}(1-p)^{\frac{N(N-1)}{2}-L}$$

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

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

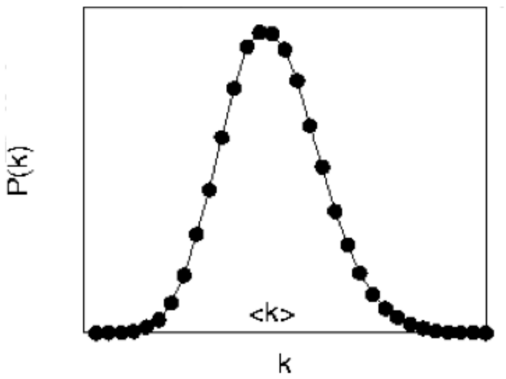
$$<k>=2L/N=p(N-1)$$

- The standard deviation

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

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

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

$$s_k^2 = p(1-p)(N-1)$$

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

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

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

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

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

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

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

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

POISSON DEGREE DISTRIBUTION

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

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

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

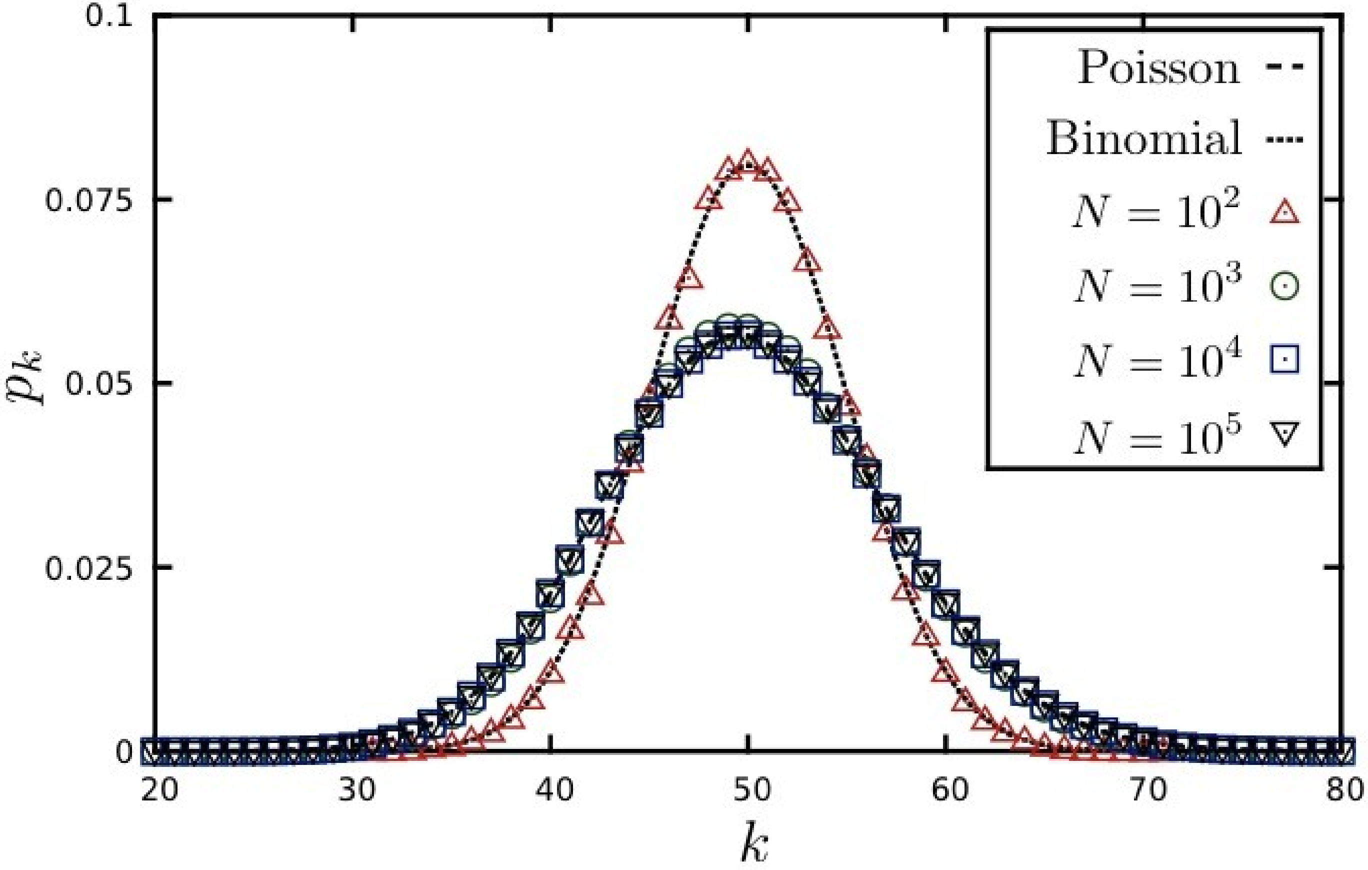
For large *N* and small *k*, we arrive at the Poisson distribution:

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

DEGREE DISTRIBUTION OF A RANDOM GRAPH

<k>=50

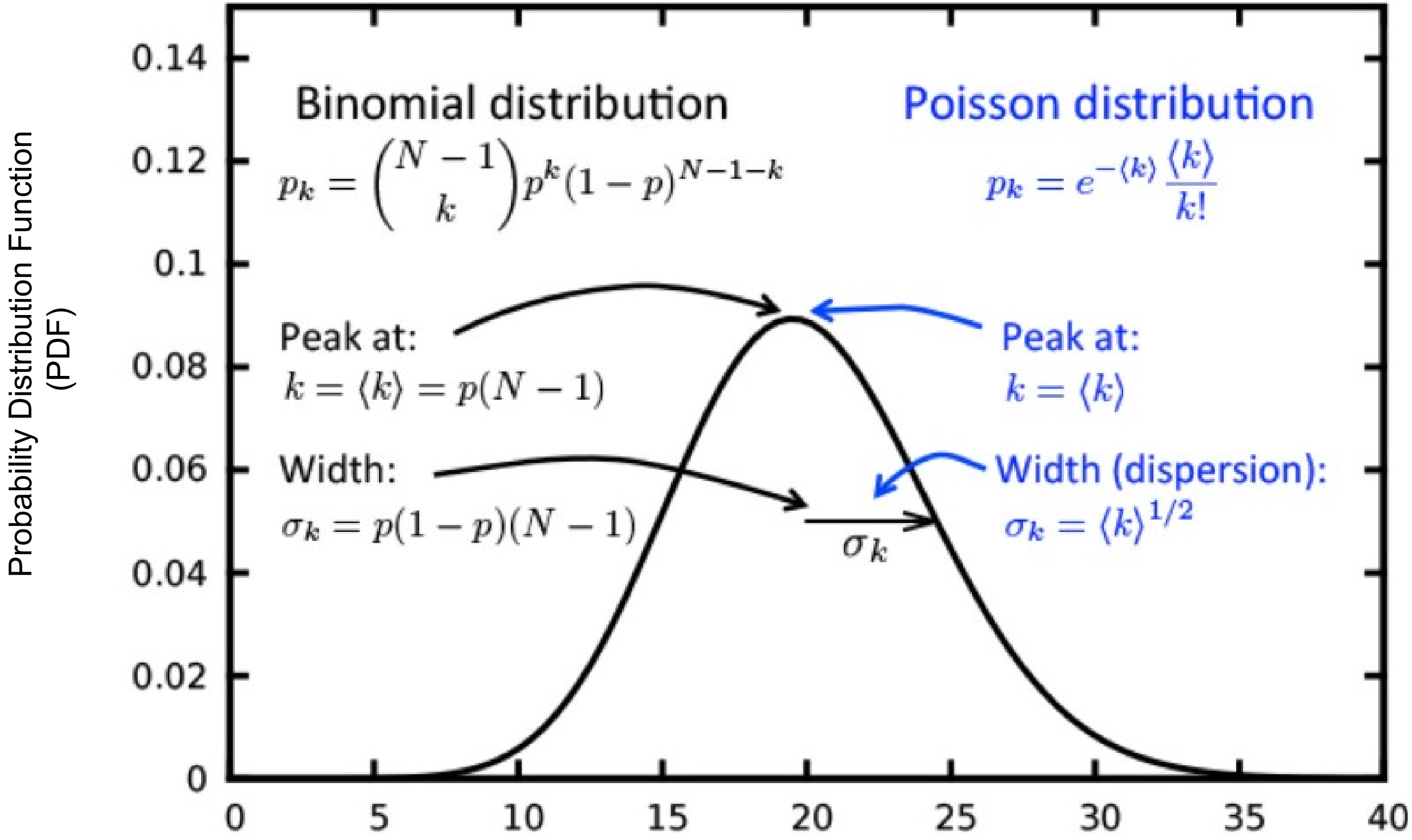
$$P(k)=e^{-<k>}\frac{<k>^{k}}{k!}$$



DEGREE DISTRIBUTION OF A RANDOM NETWORK

Exact Result
-binomial distribution-

Large N limit
-Poisson distribution-

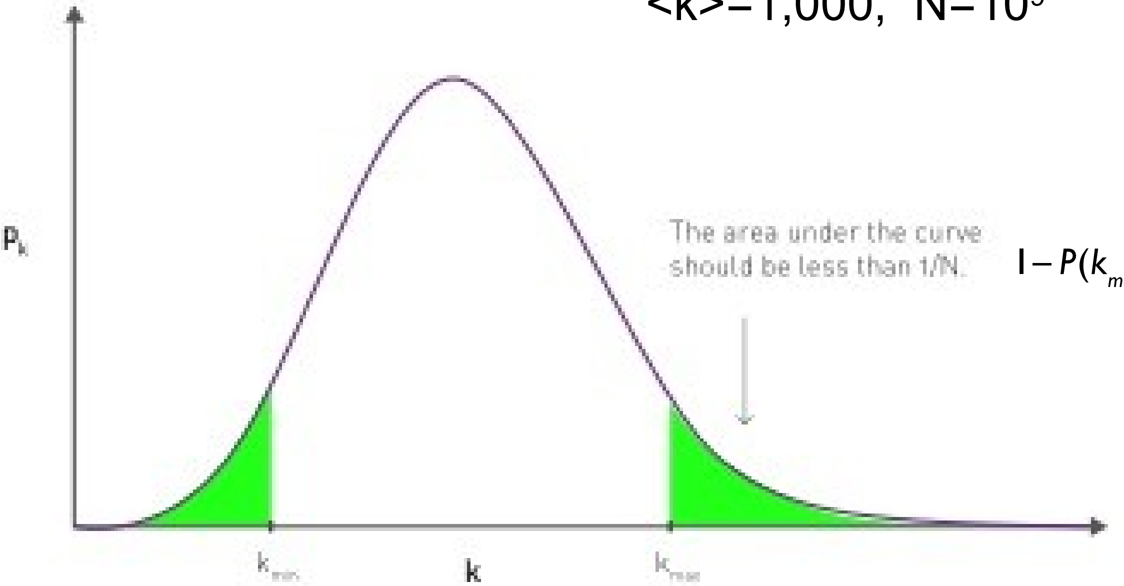


Section 3.4

Real Networks are not Poisson

Section 3.5 Maximum and minimum degree

$<k>=1{,}000, \quad N=10^9$



The area under the curve should be less than 1/N.

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

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

$<k>=1{,}000, \quad N=10^9$

$k_{max}=1{,}185$

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

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

$k_{min}=816$

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

$$\sigma_k = 31.62.$$

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 nonexistent in a random society.

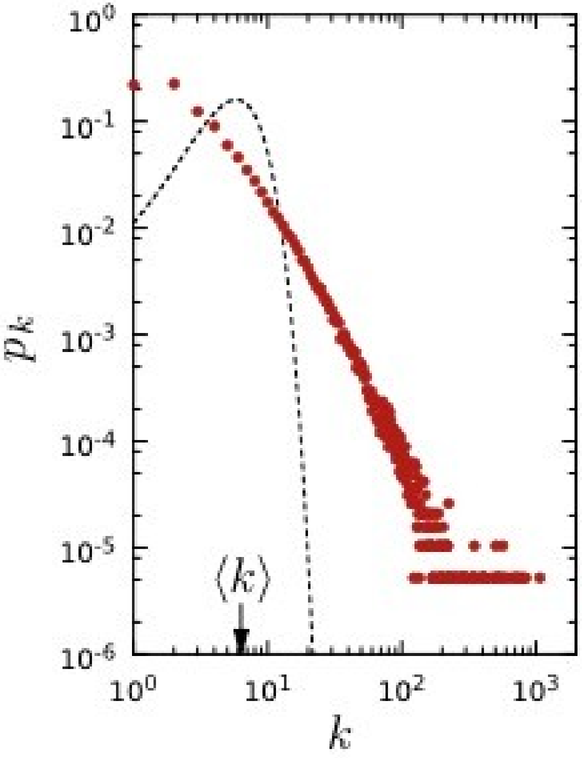
A random society would consist of mainly average individuals, with everyone with roughly the same number of friends.

It would lack outliers, individuals that are either highly popular or recluse.

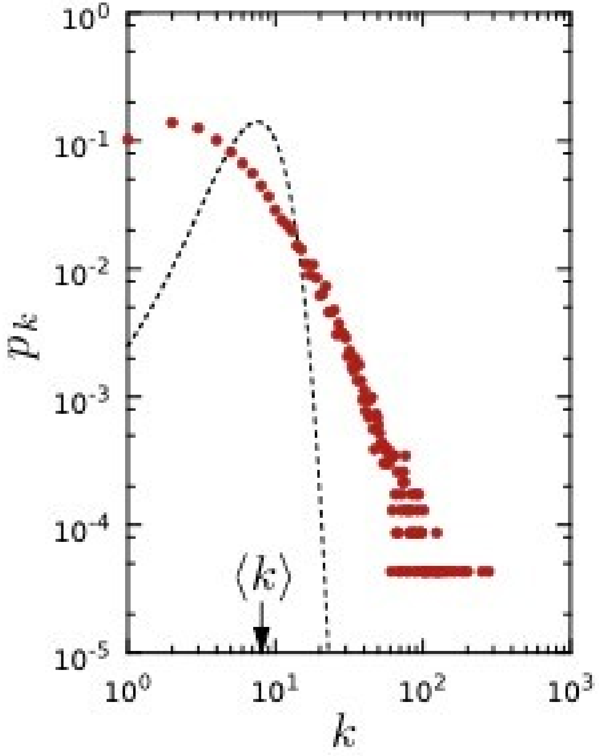
FACING REALITY: Degree distribution of real networks

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

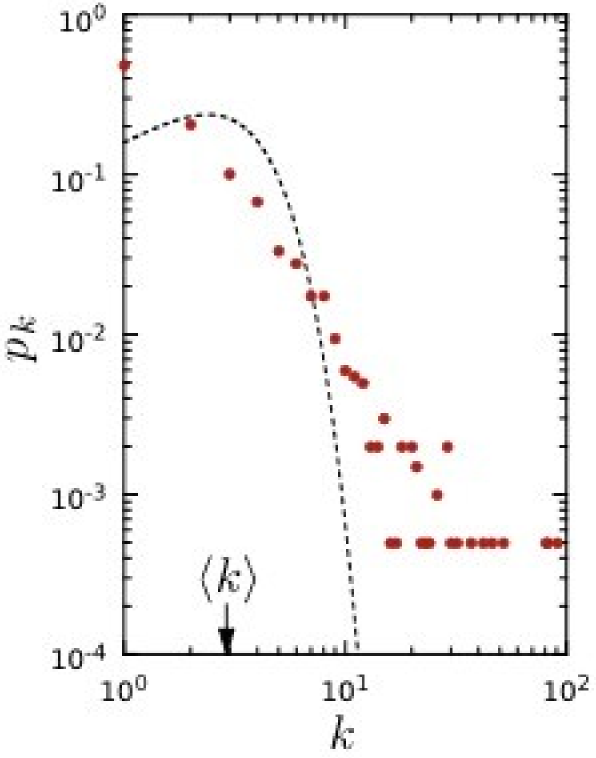
Internet



Science Collaboration

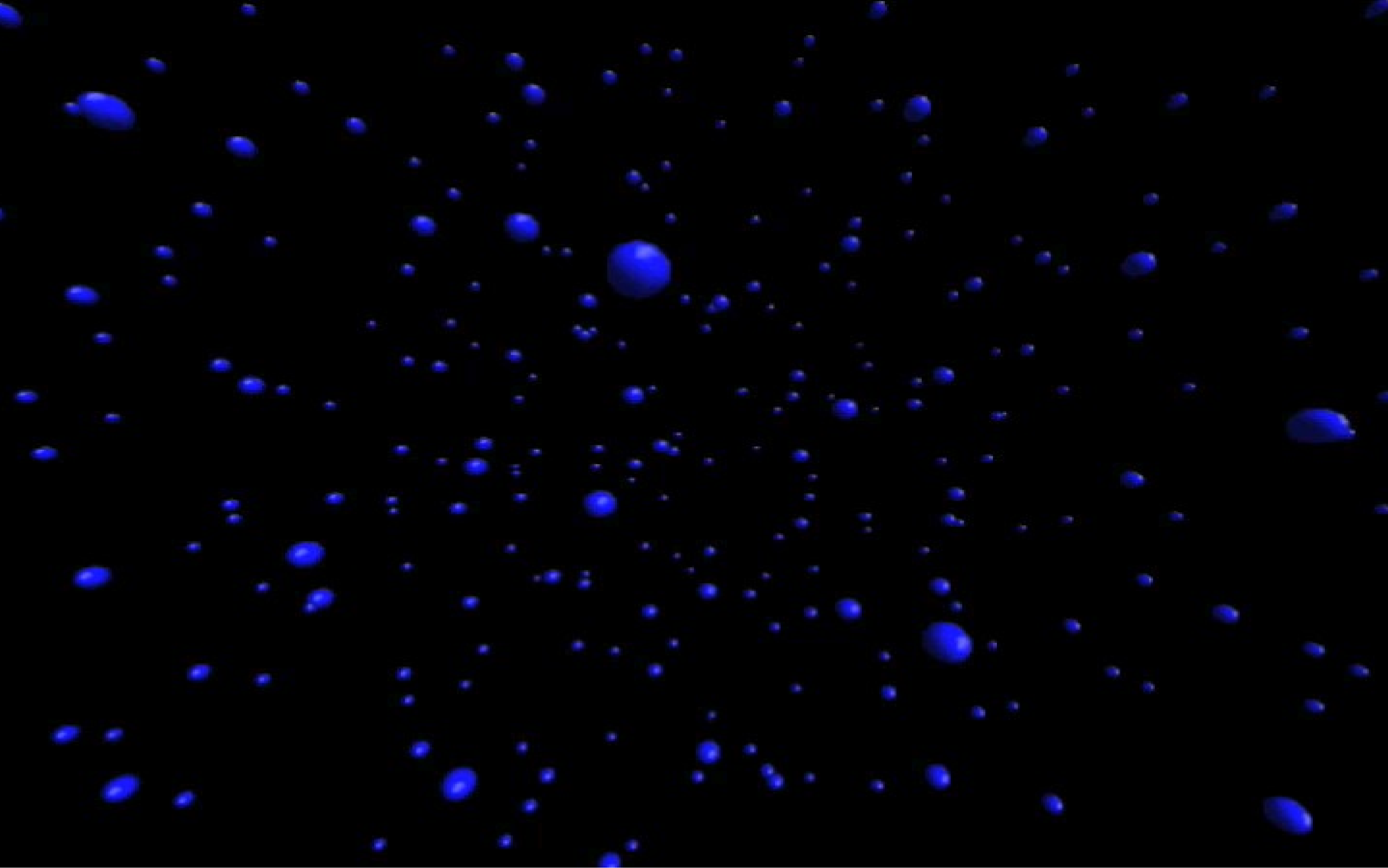


Protein Interactions



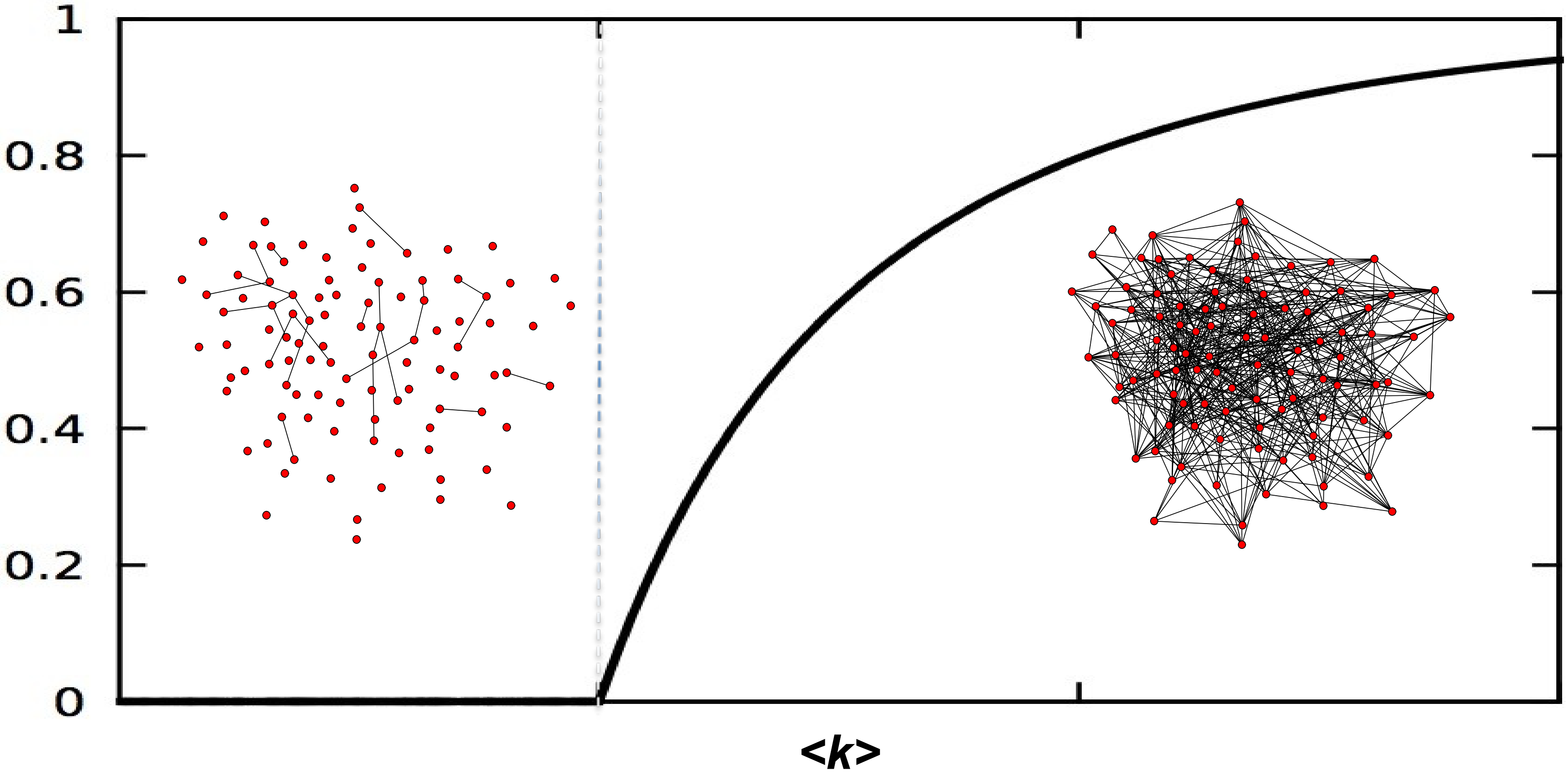
Section 6

The evolution of a random network



EVOLUTION OF A RANDOM NETWORK

disconnected nodes ➔ **NETWORK**.



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

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

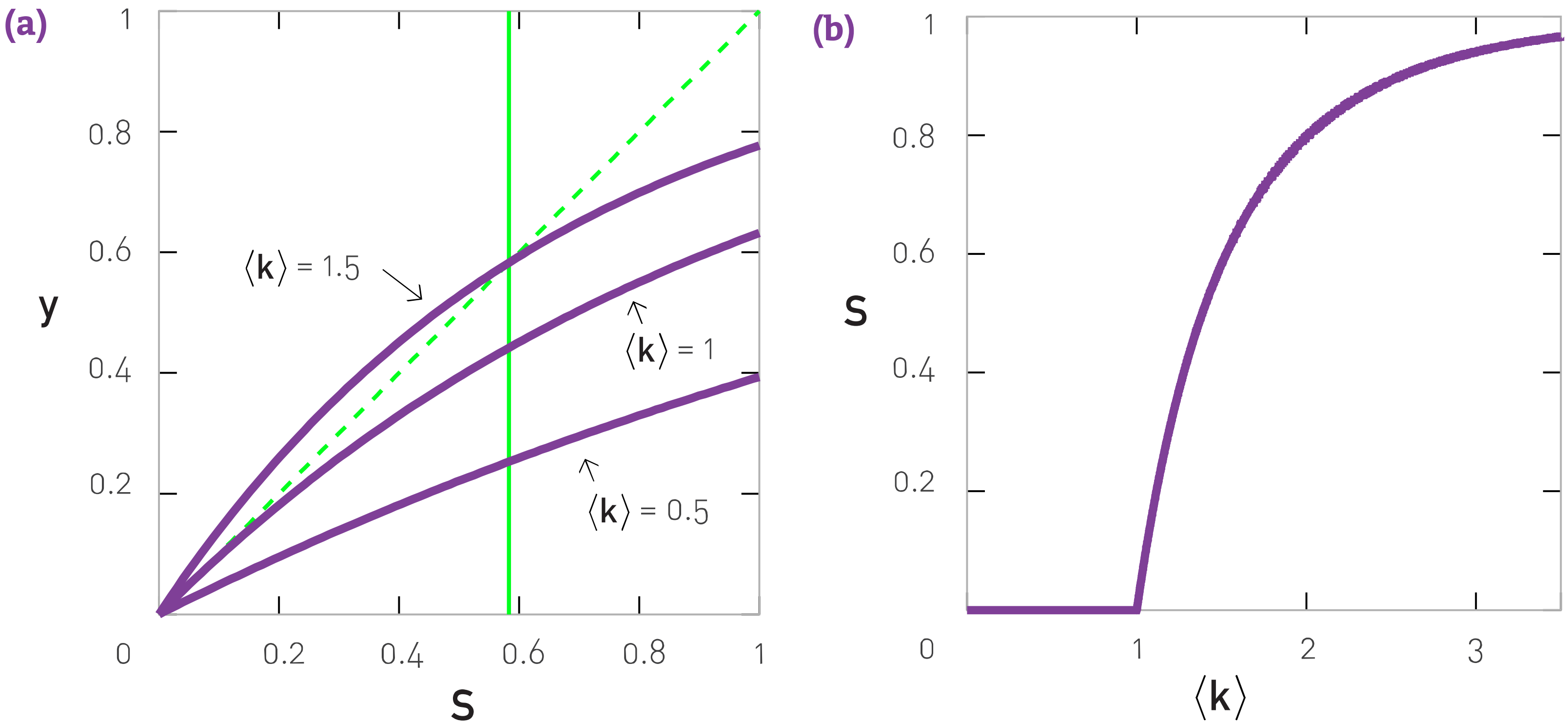
provides the size of the giant component via $N_G = N(1 - u)$. Using $p = \langle k \rangle / (N - 1)$ and taking the log of both sides, for $\langle k \rangle \ll N$ we obtain

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

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

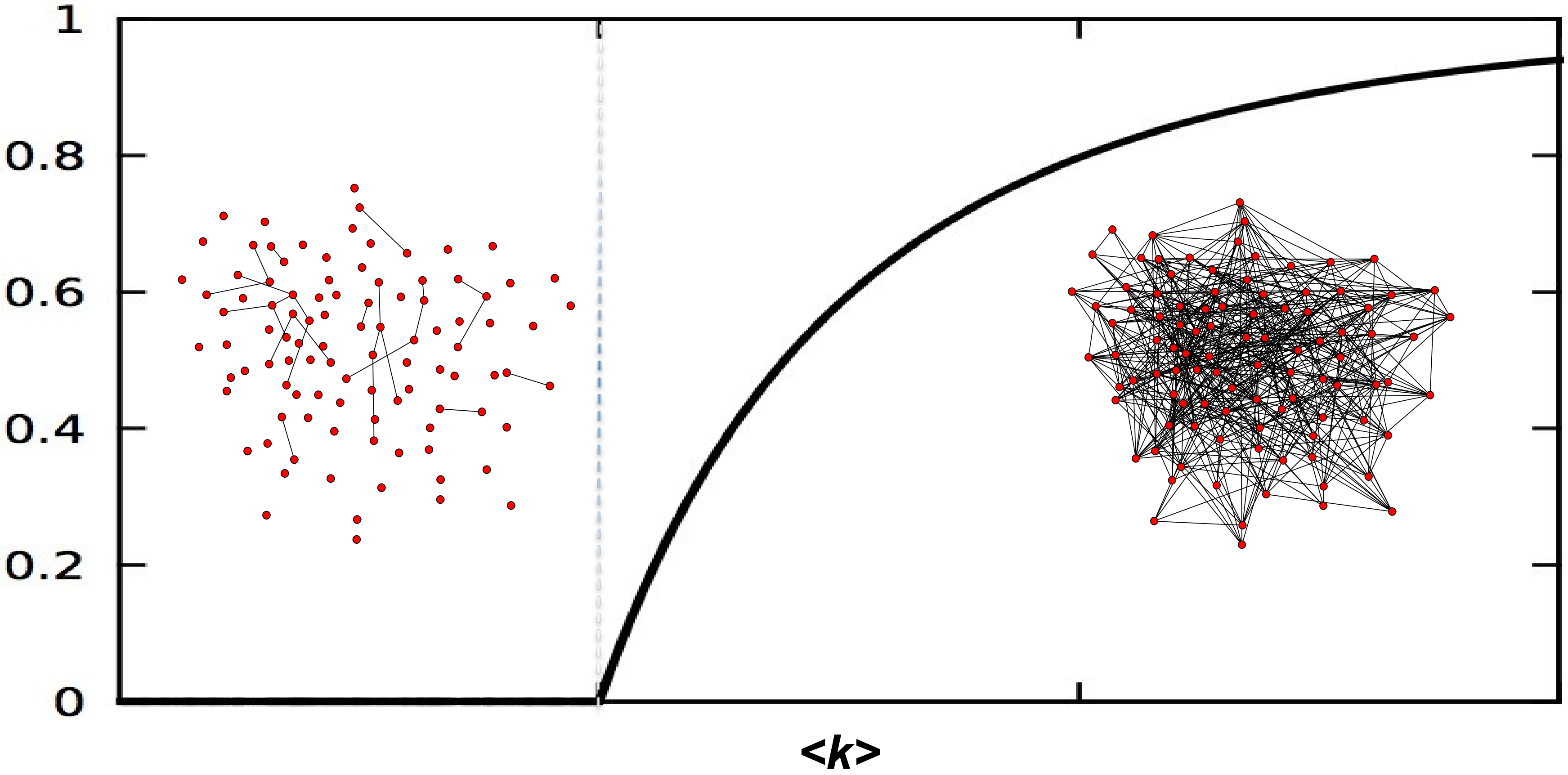
$$S = 1 - e^{-\langle k \rangle S}.$$

$$S = 1 - e^{-\langle k \rangle S}. \qquad (3.32)$$



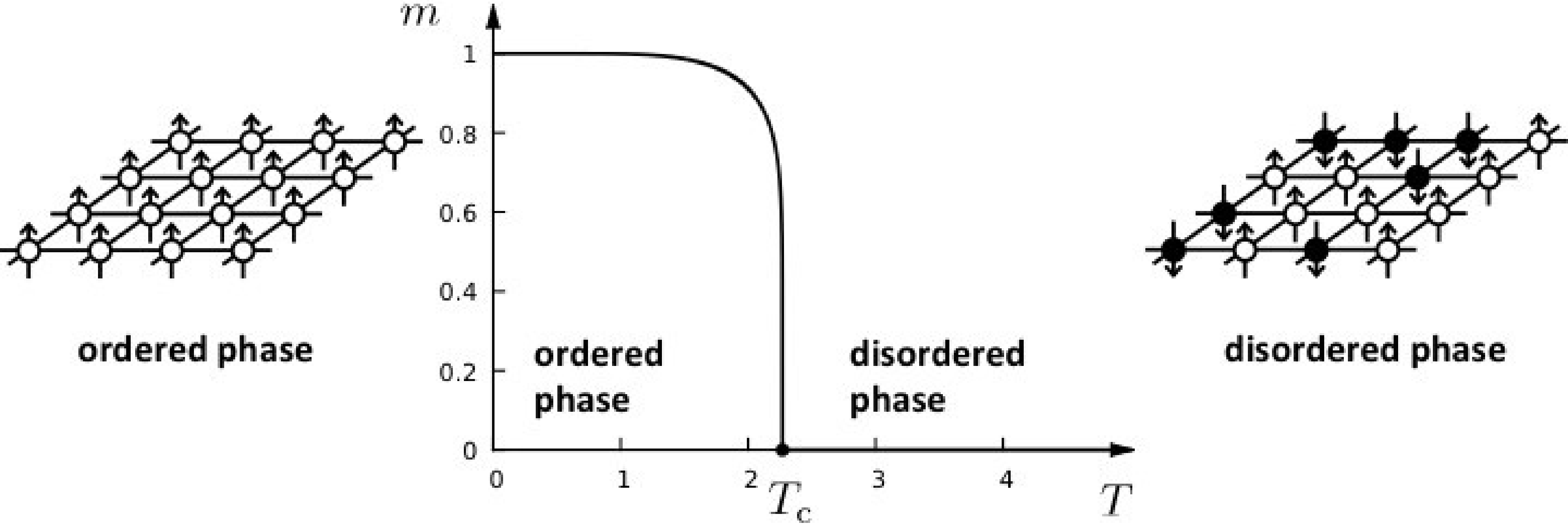
EVOLUTION OF A RANDOM NETWORK

disconnected nodes ➔ **NETWORK**.

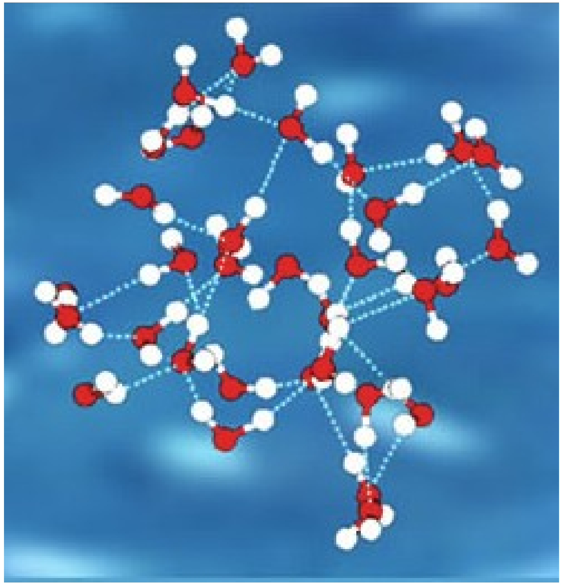


How does this transition happen?

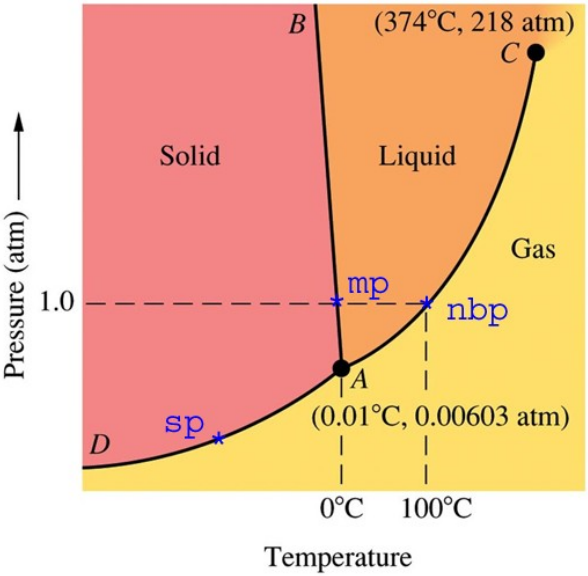
Phase transitions in complex systems I: Magnetism

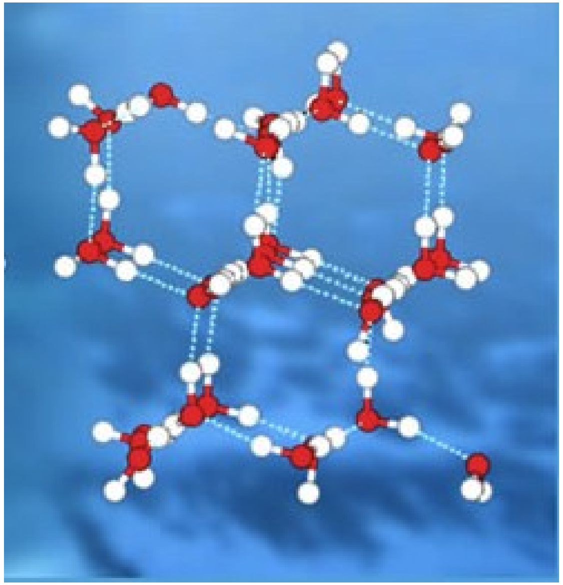


Phase transitions in complex systems I: liquids



Water





Ice

CLUSTER SIZE DISTRIBUTION

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

$$p(s) = e^{-<k> s}$$

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

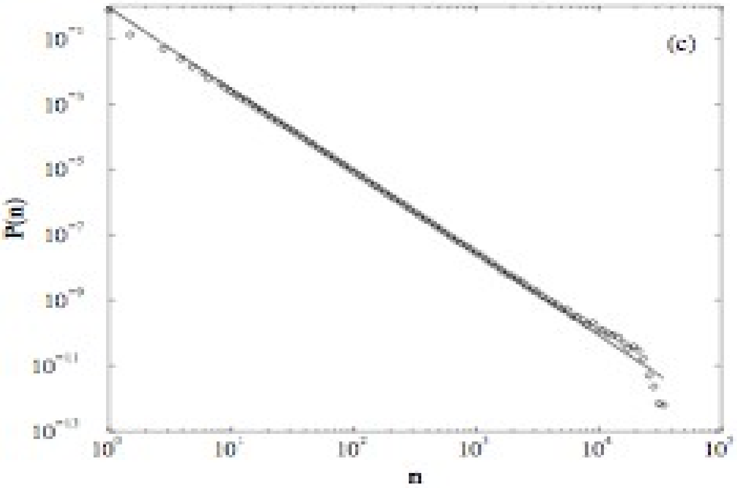
$$p(s) = \frac{s^{s-1}}{s!} e^{-\langle k\rangle s + (s-1)\ln\langle k\rangle}$$

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

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

At the critical point <k>=1

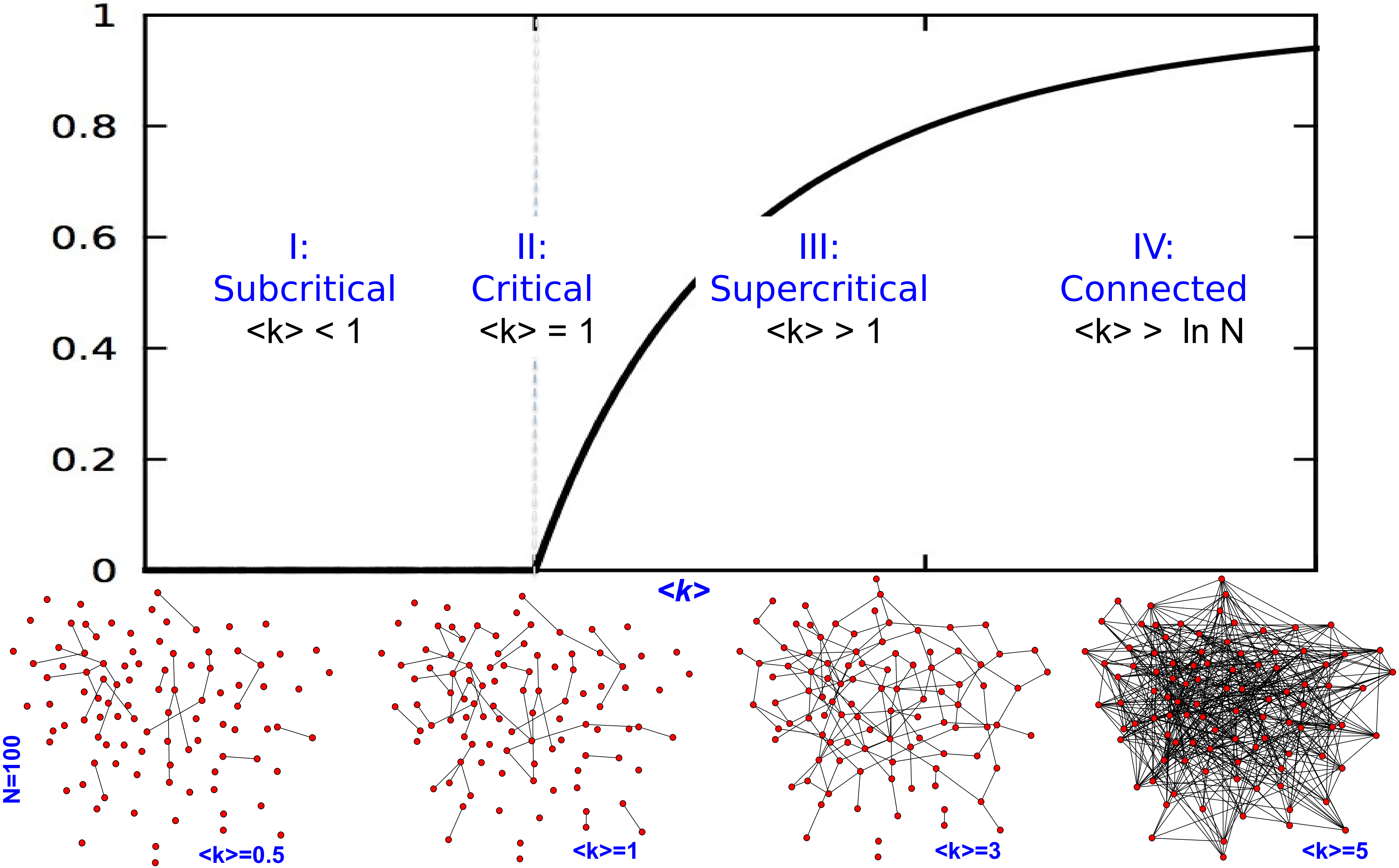
$$p(s) \sim s^{-3/2}$$

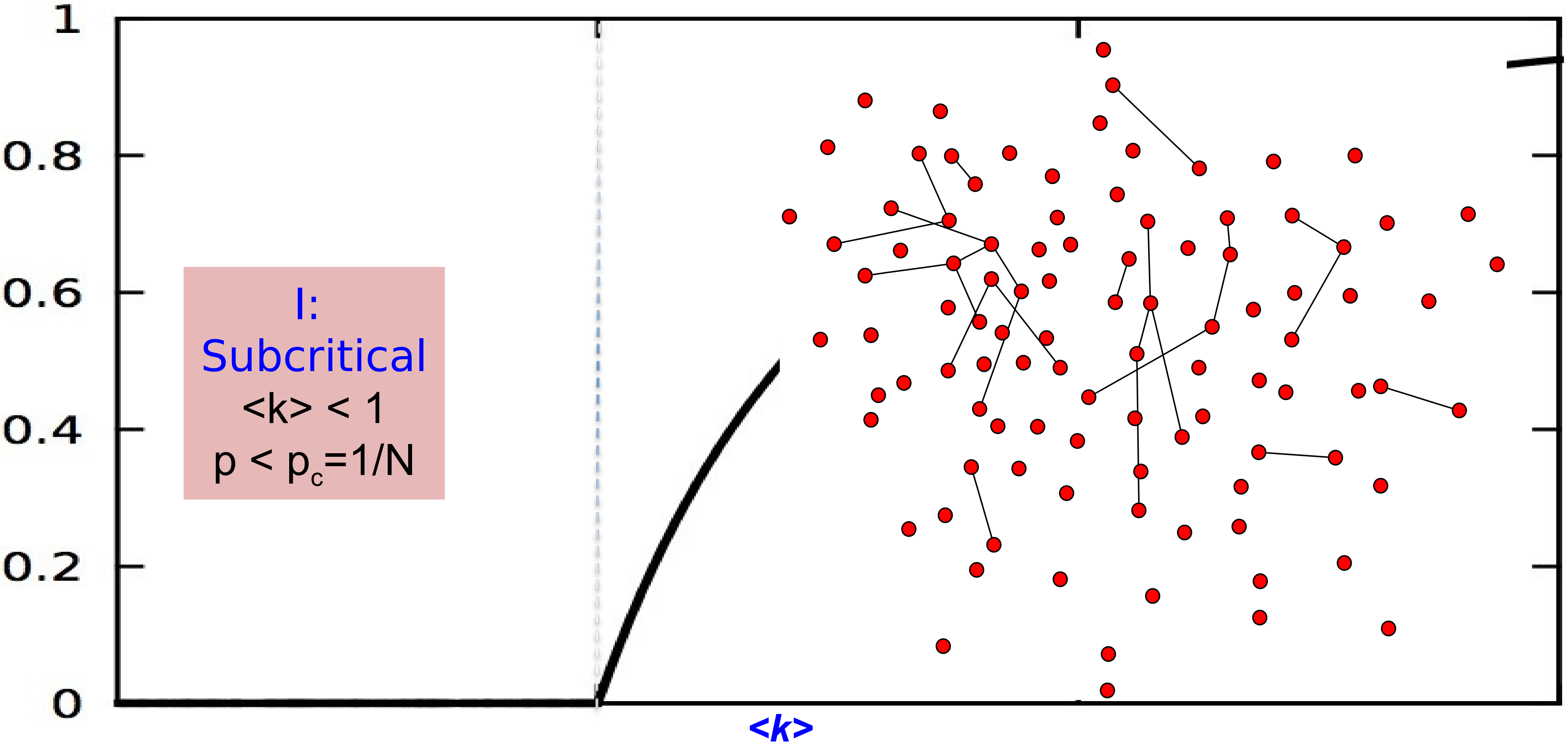


The distribution of cluster sizes at the critical point, displayed in a log-log plot. The data represent an average over 1000 systems of sizes The dashed line has a slope of

$$-t_n = -2.5$$

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



I:

Subcritical

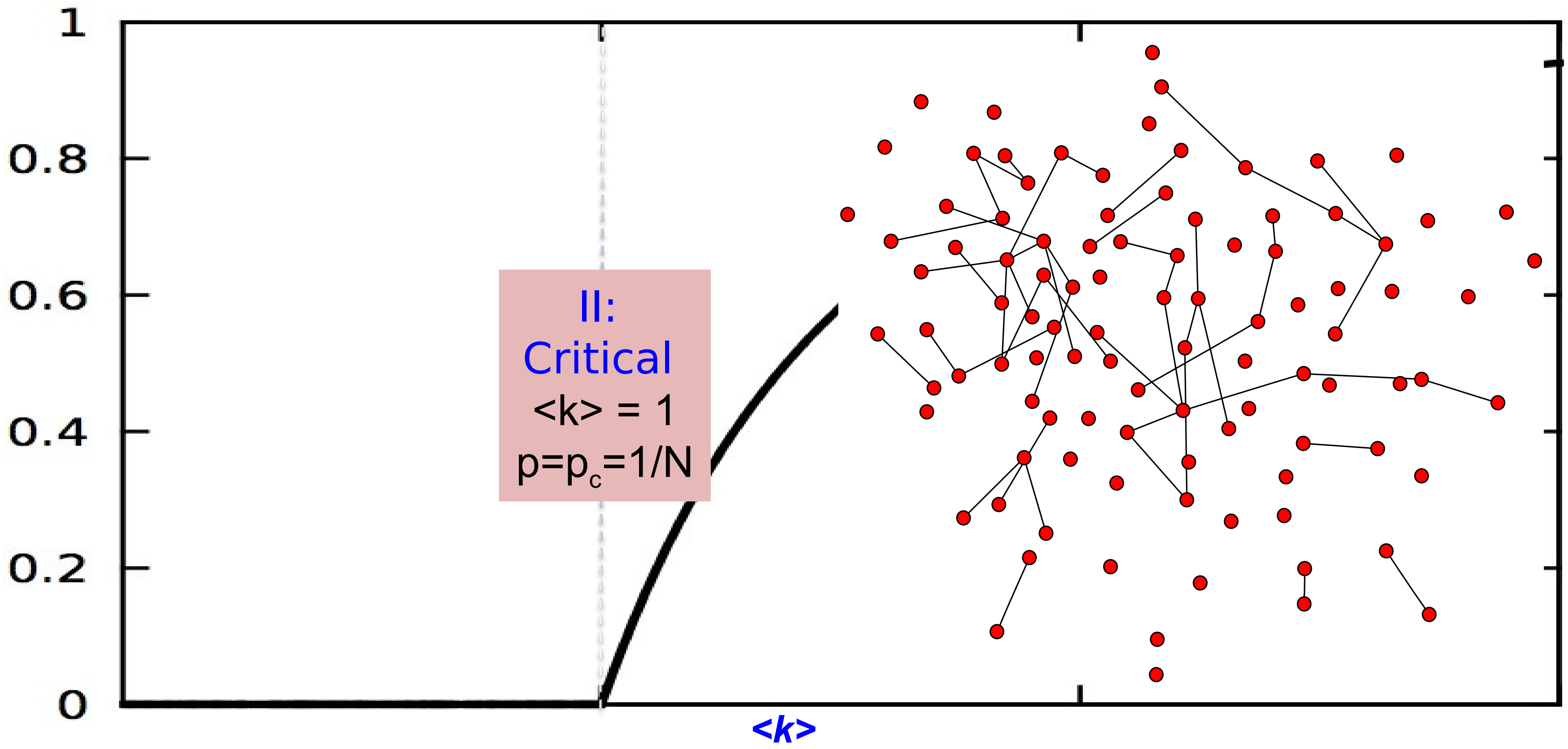
<k> < 1

$p < p_c = 1/N$

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*



II:
Critical
<k> = 1
$p=p_c=1/N$

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

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

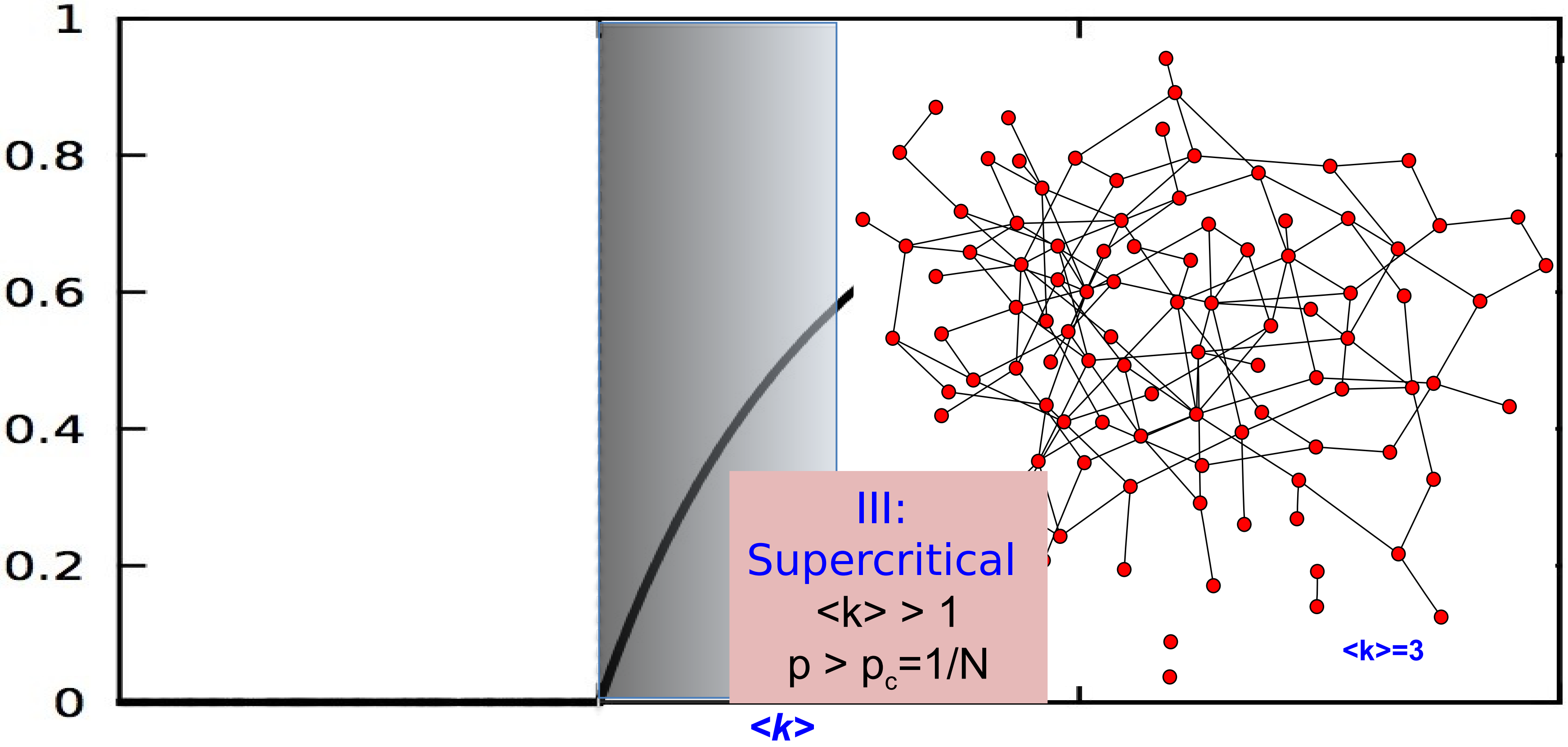
→ Small components are trees, GC has loops.

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



III: Supercritical

<k> > 1

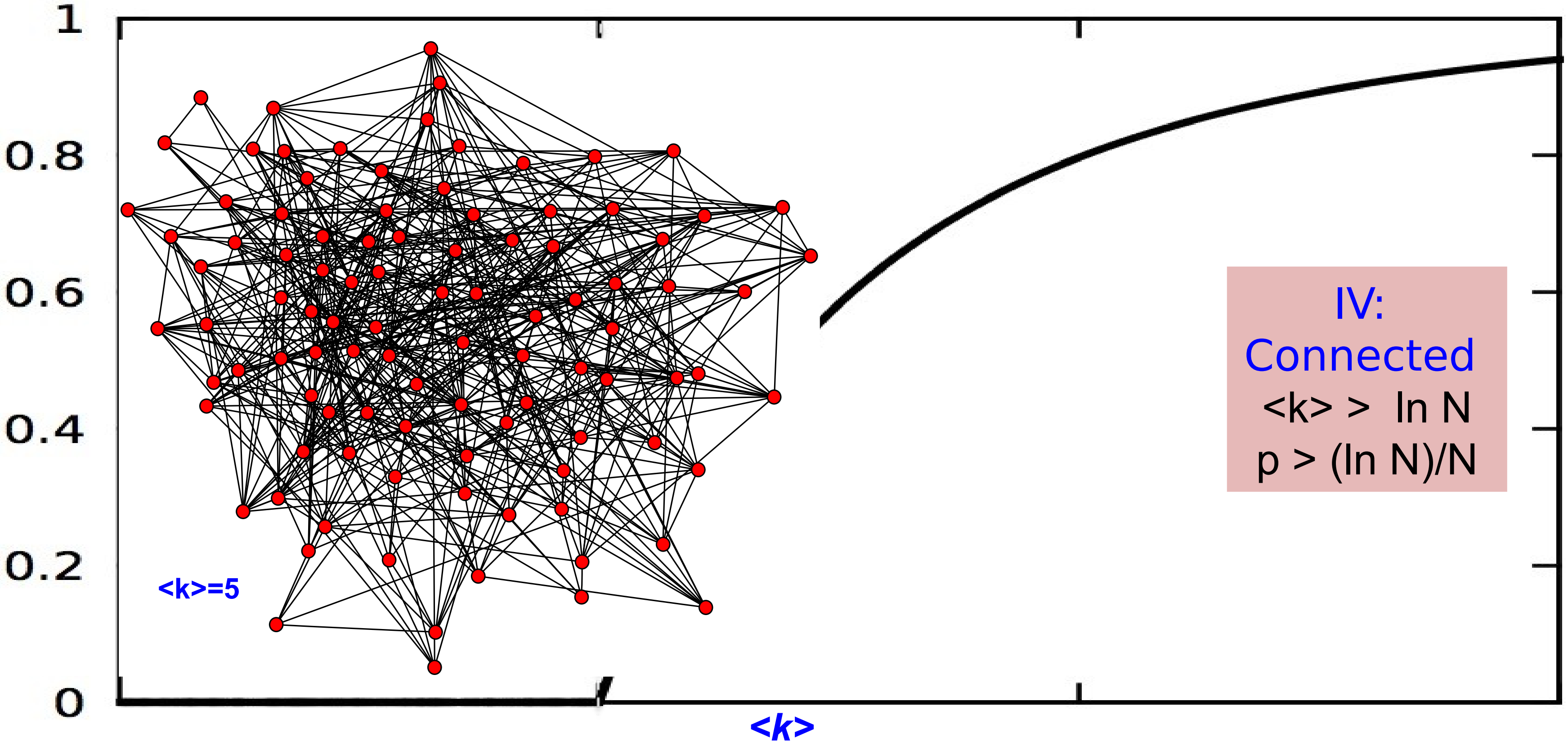
$p > p_c = 1/N$

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

→GC has loops.

Cluster size distribution: exponential

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

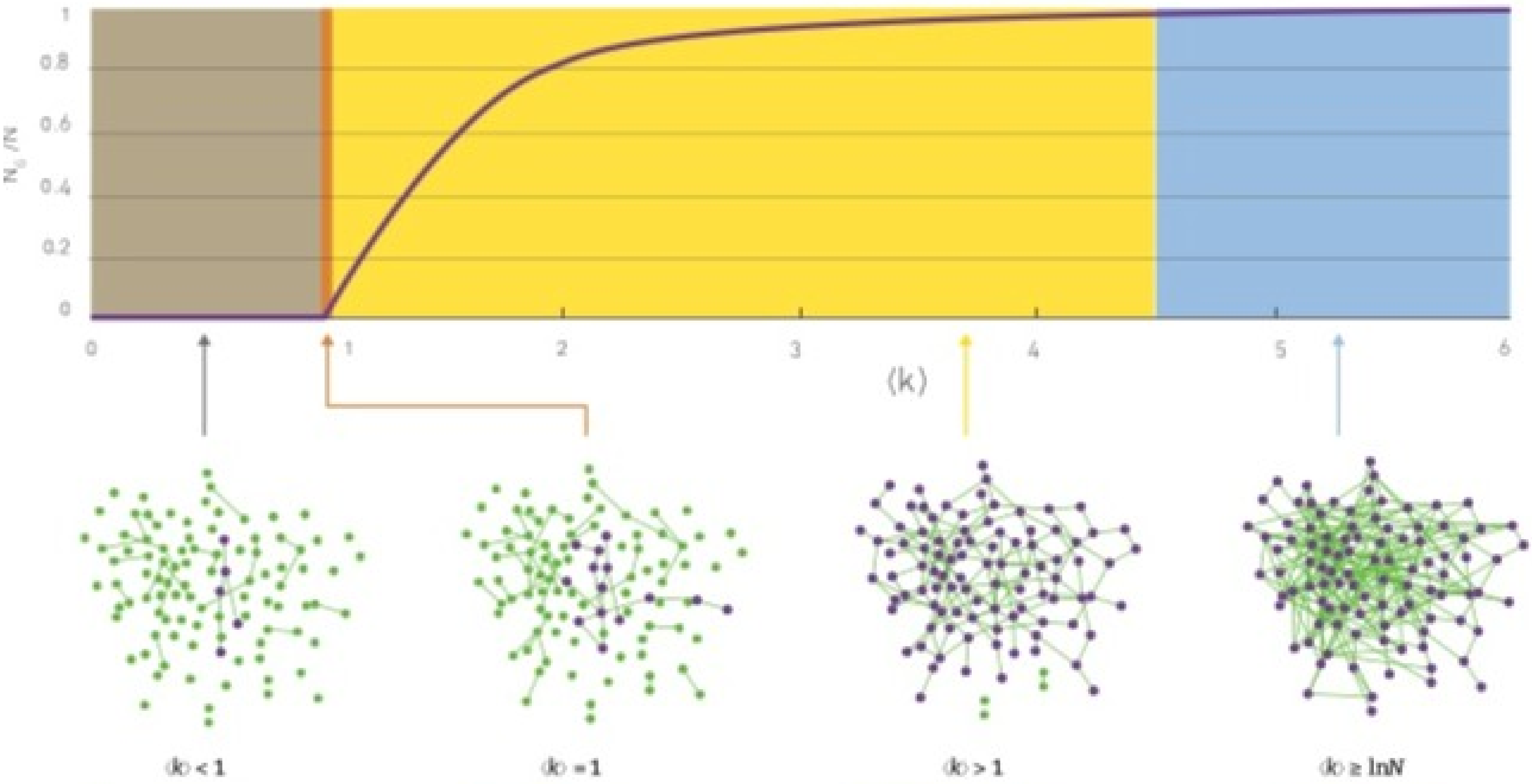


IV:
Connected
<k> > ln N
p > (ln N)/N

Only one cluster: N_G=N

→GC is dense.

Cluster size distribution: None



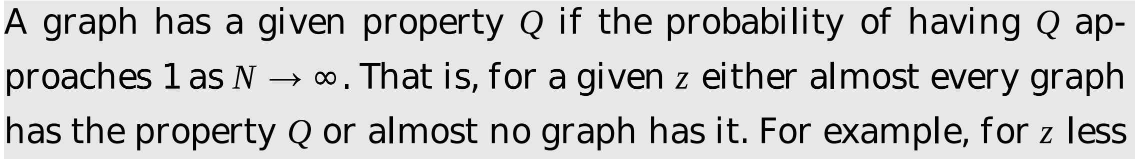
(b) Subcritical Regime
- No giant component
- Cluster size distribution: $p_s \sim s^{-3/2} e^{-\alpha s}$
- Size of the largest cluster: $N_G \sim \ln N$
- The clusters are trees

(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

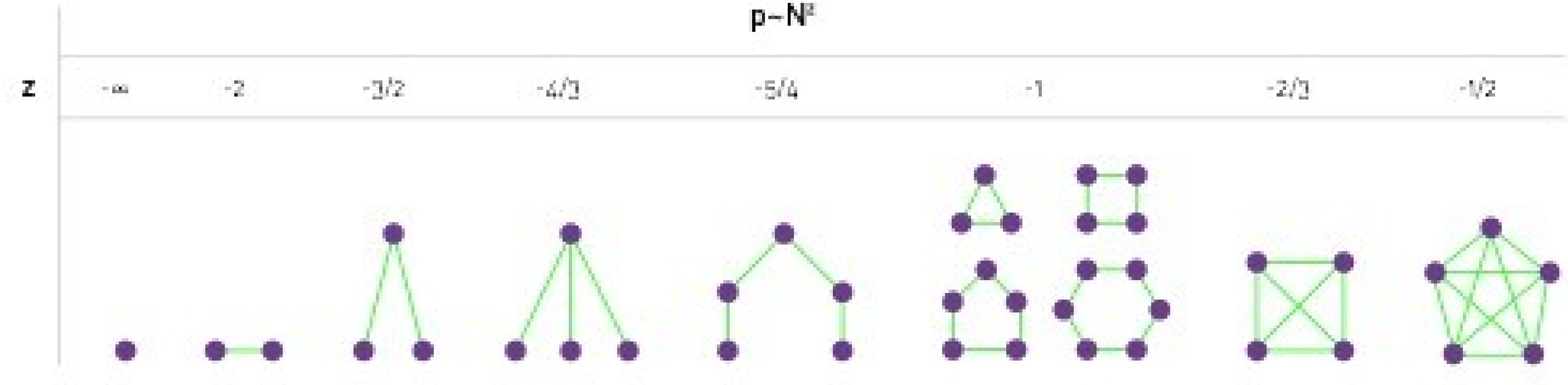
(d) Supercritical Regime
- Single giant component
- Cluster size distribution: $p_s \sim s^{-3/2} e^{-\alpha s}$
- Size of the giant component: $N_G \sim (p - p_c)N$
- The small clusters are trees
- Giant component has loops

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

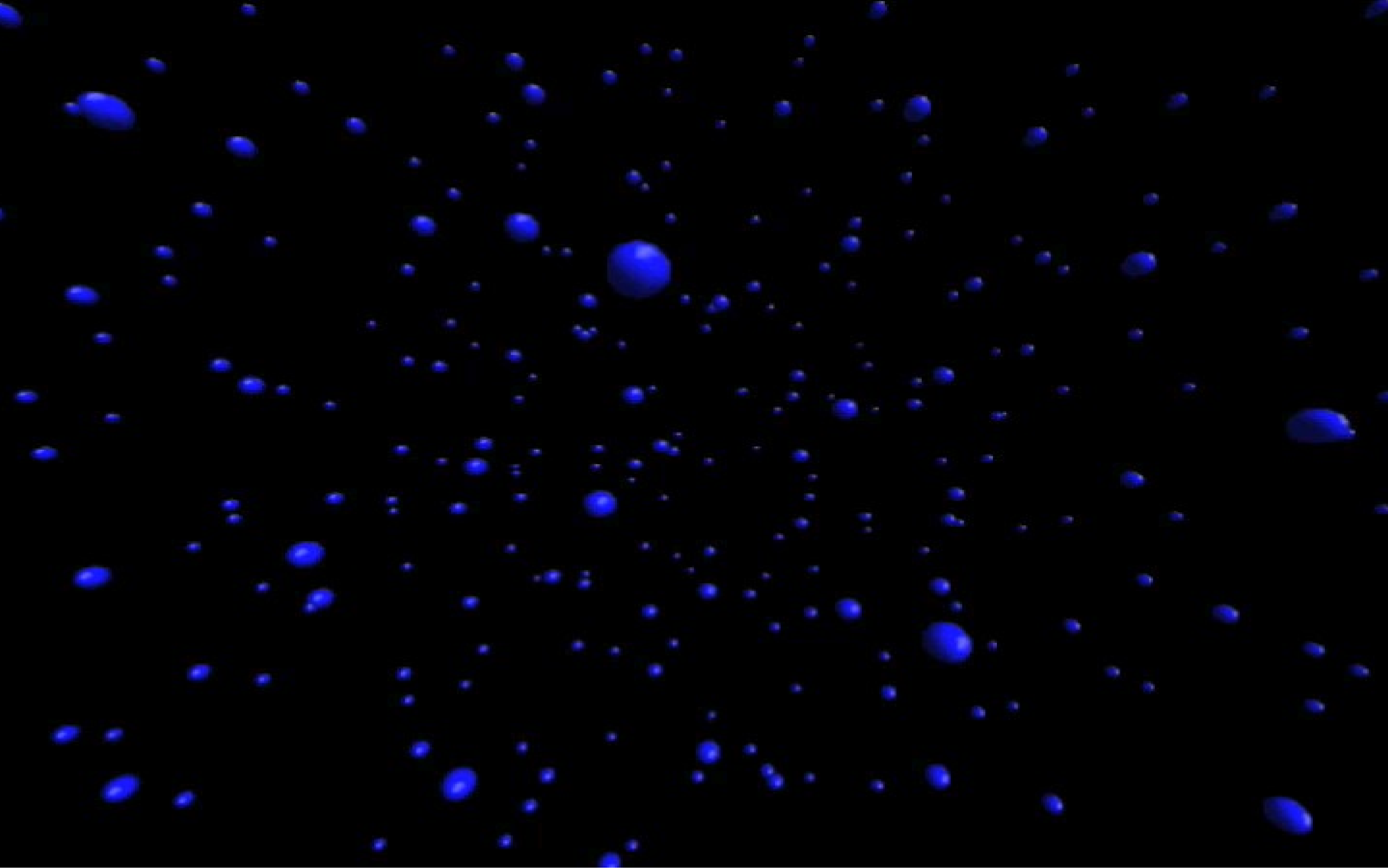
Network evolution in graph theory



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



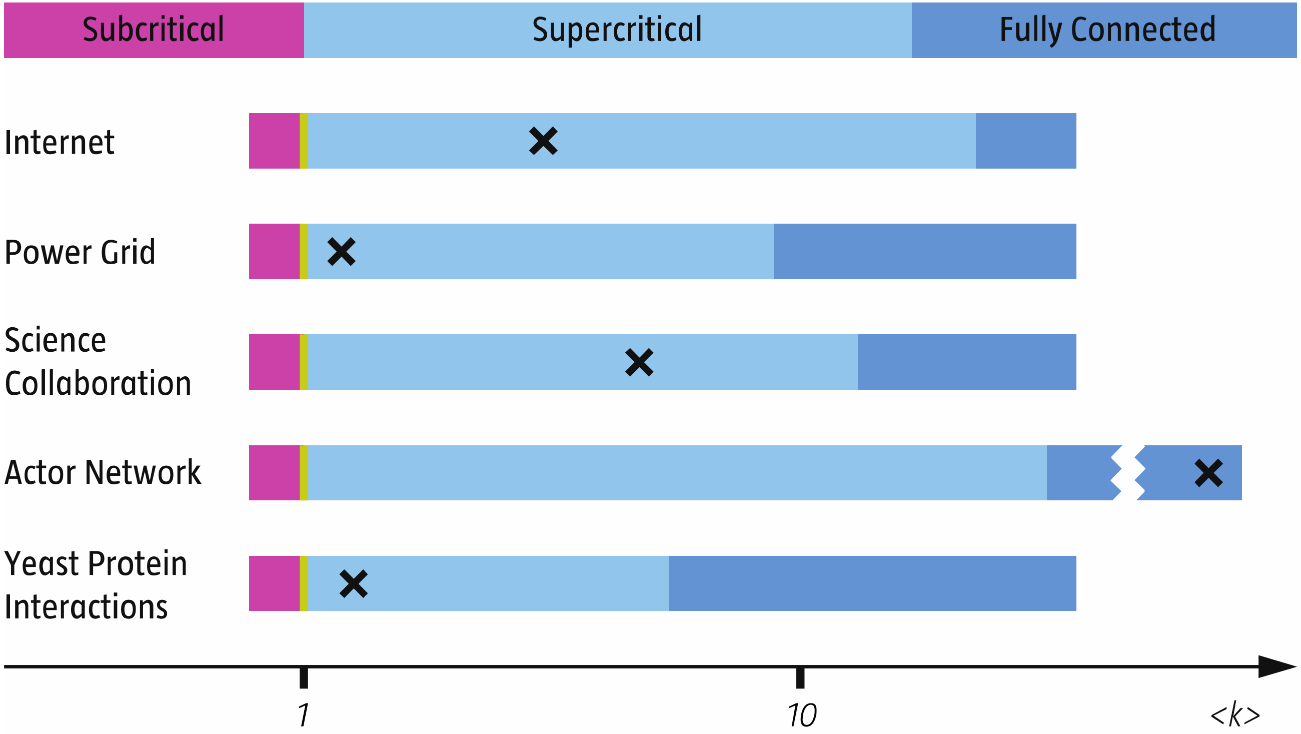
$$p = < k > /(N - 1)$$

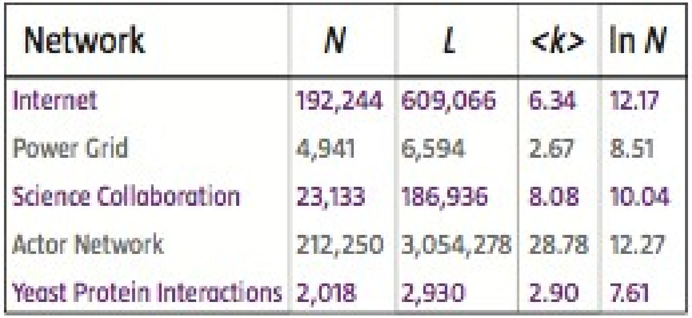


Section 7

Real networks are supercritical

Section 7



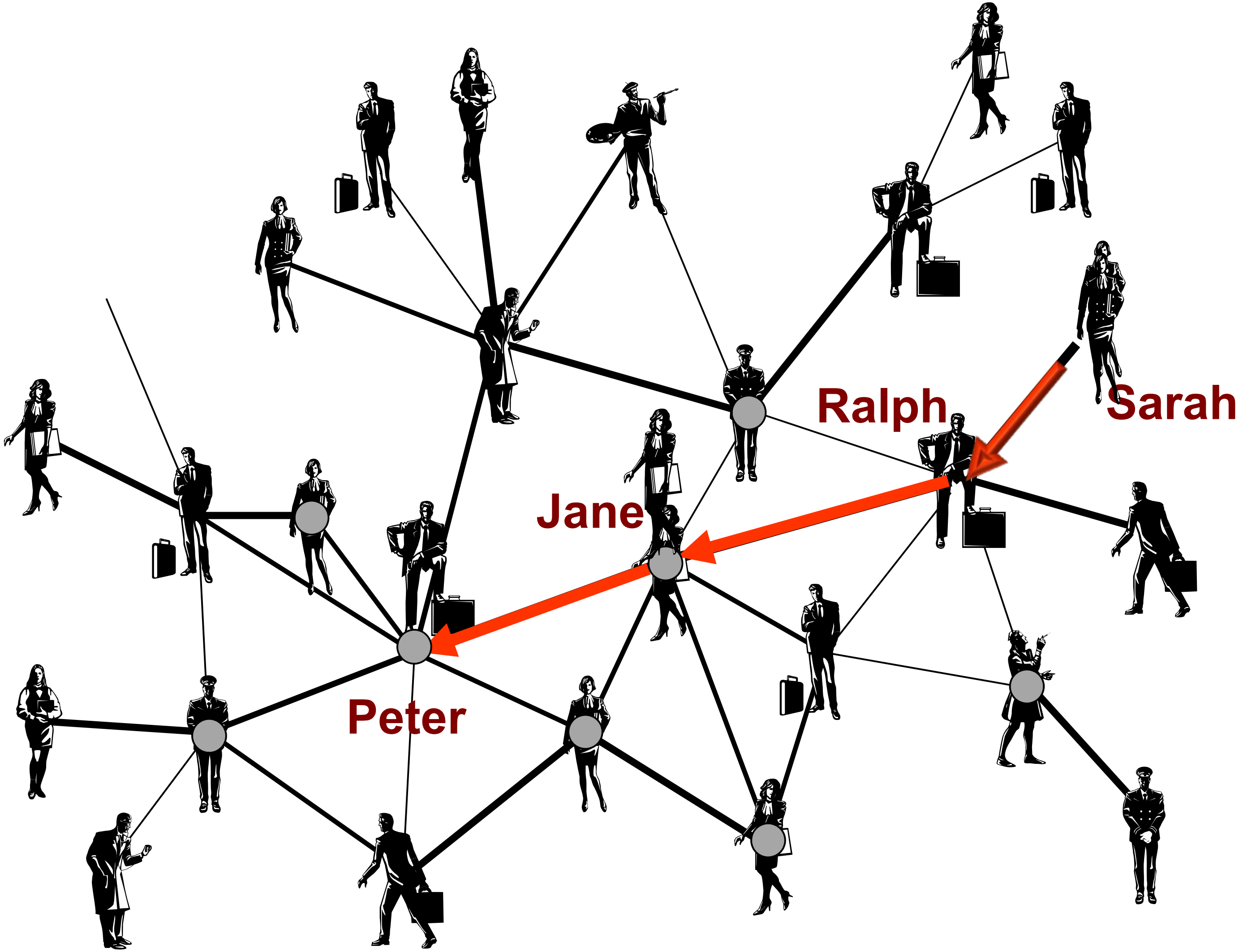


Network	N	L	$<k>$	ln N
Internet	192,244	609,066	6.34	12.17
Power Grid	4,941	6,594	2.67	8.51
Science Collaboration	23,133	186,936	8.08	10.04
Actor Network	212,250	3,054,278	28.78	12.27
Yeast Protein Interactions	2,018	2,930	2.90	7.61

Section 3.8

Small worlds

SIX DEGREES small worlds



Frigyes Karinthy, 1929
Stanley Milgram, 1967

1929: Frigyes Kartinthy

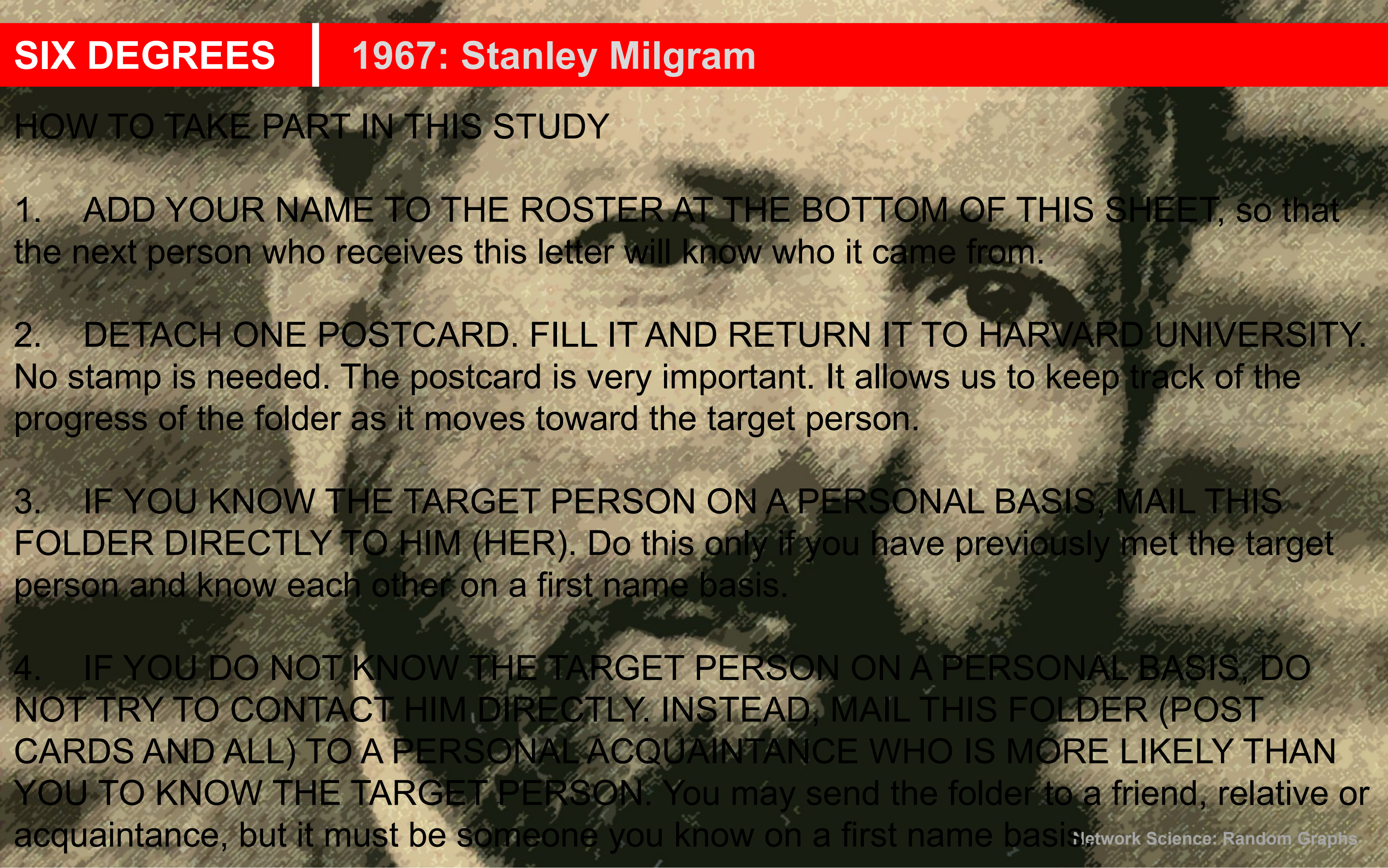
Frigyes Karinthy (1887-1938)
Hungarian Writer

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

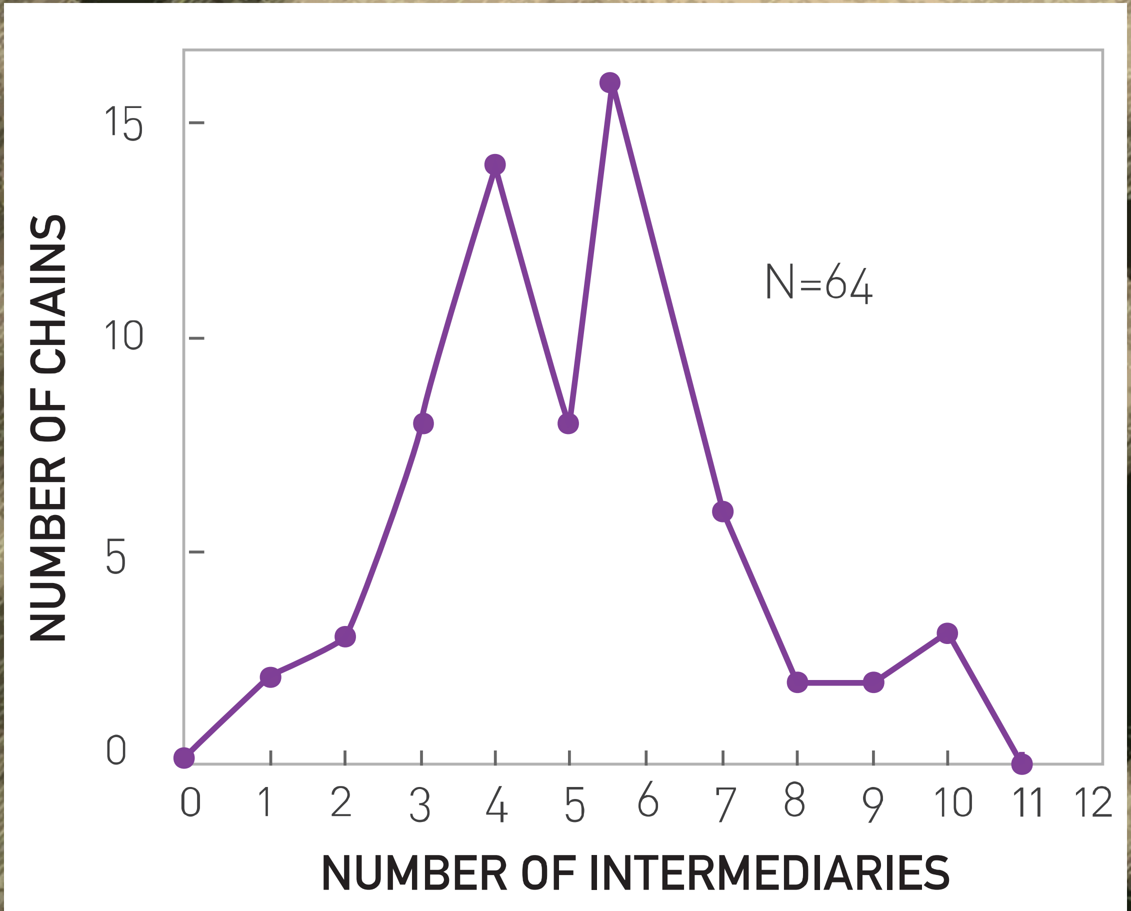
1. ADD YOUR NAME TO THE ROSTER AT THE BOTTOM OF THIS SHEET, so that the next person who receives this letter will know who it came from.

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.

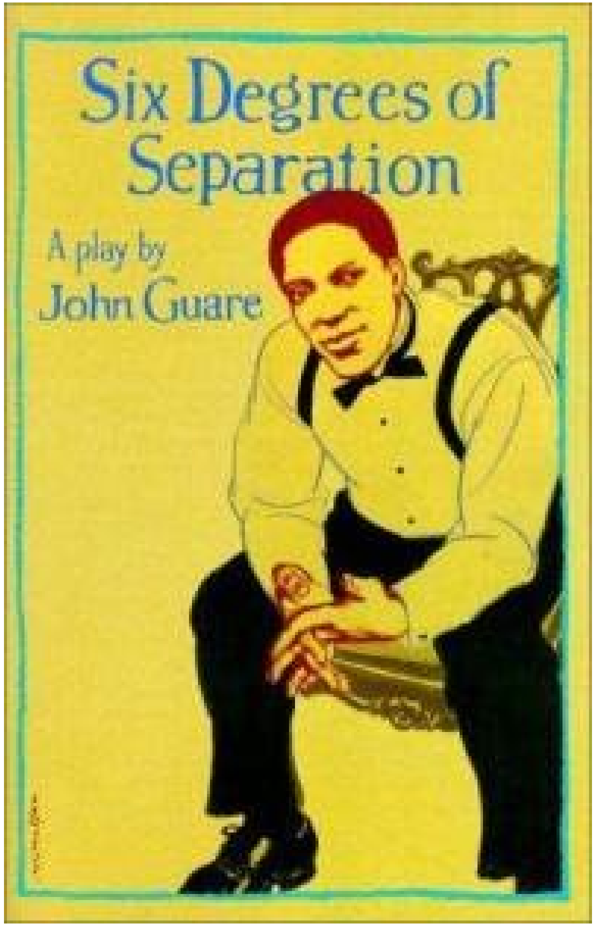
4. IF YOU DO NOT KNOW THE TARGET PERSON ON A PERSONAL BASIS, DO NOT TRY TO CONTACT HIM DIRECTLY. INSTEAD, MAIL THIS FOLDER (POST CARDS AND ALL) TO A PERSONAL ACQUAINTANCE WHO IS MORE LIKELY THAN YOU TO KNOW THE TARGET PERSON. You may send the folder to a friend, relative or acquaintance, but it must be someone you know on a first name basis.

SIX DEGREES | 1967: Stanley Milgram



SIX DEGREES

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

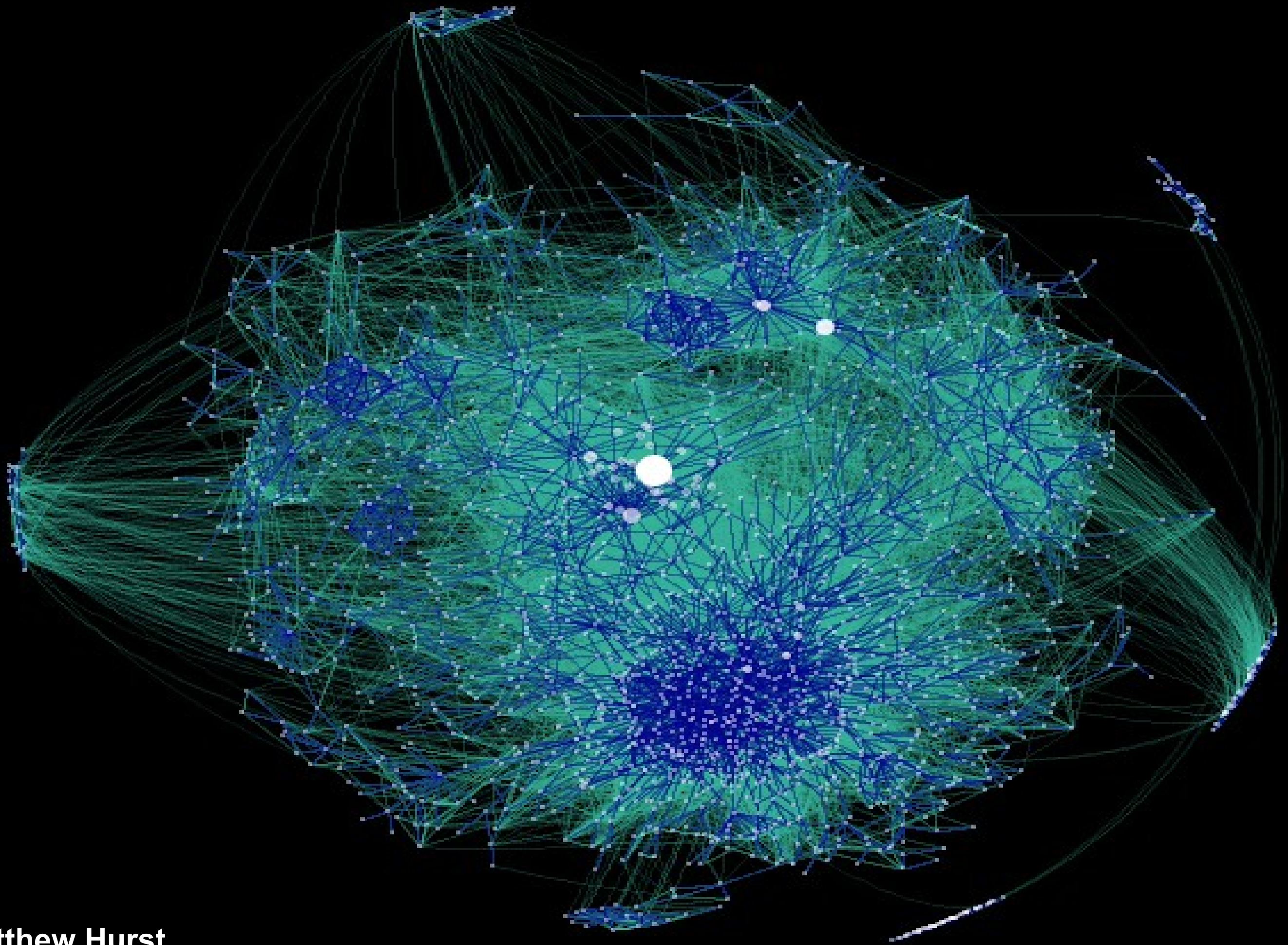
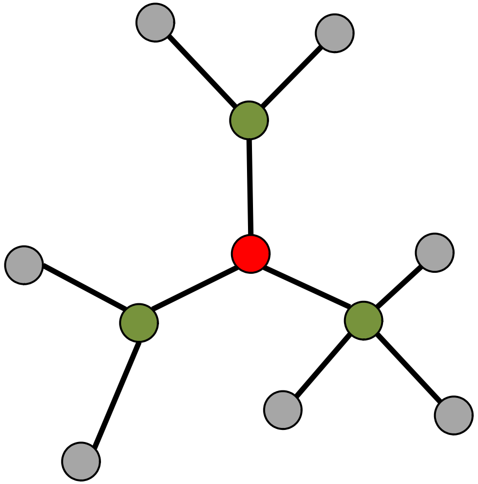


Image by **Matthew Hurst**
Blogosphere

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.

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

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 $d_{\max}$.

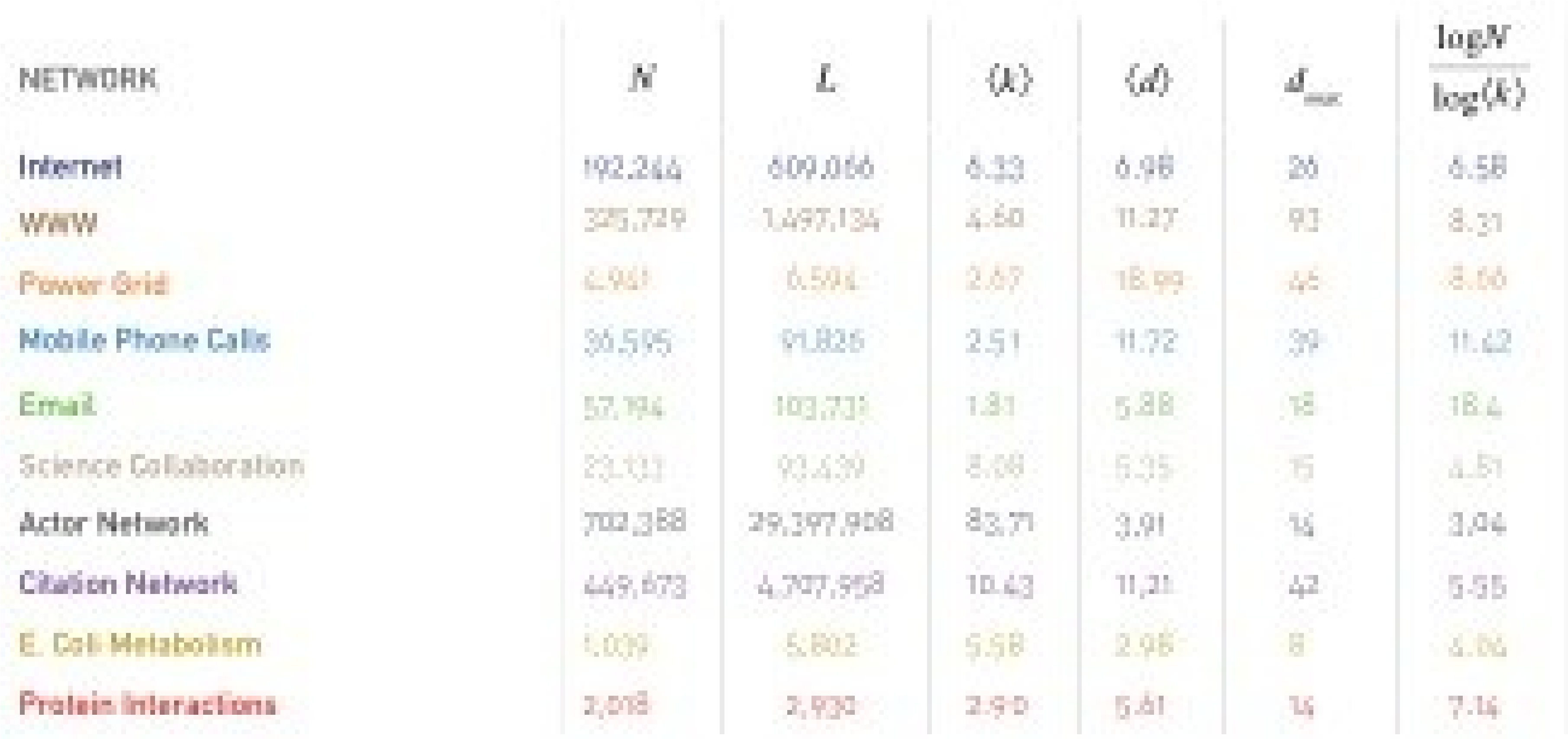
$$d >= \frac{\log N}{\log\langle k\rangle}$$

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.

DISTANCES IN RANDOM GRAPHS

compare with real data

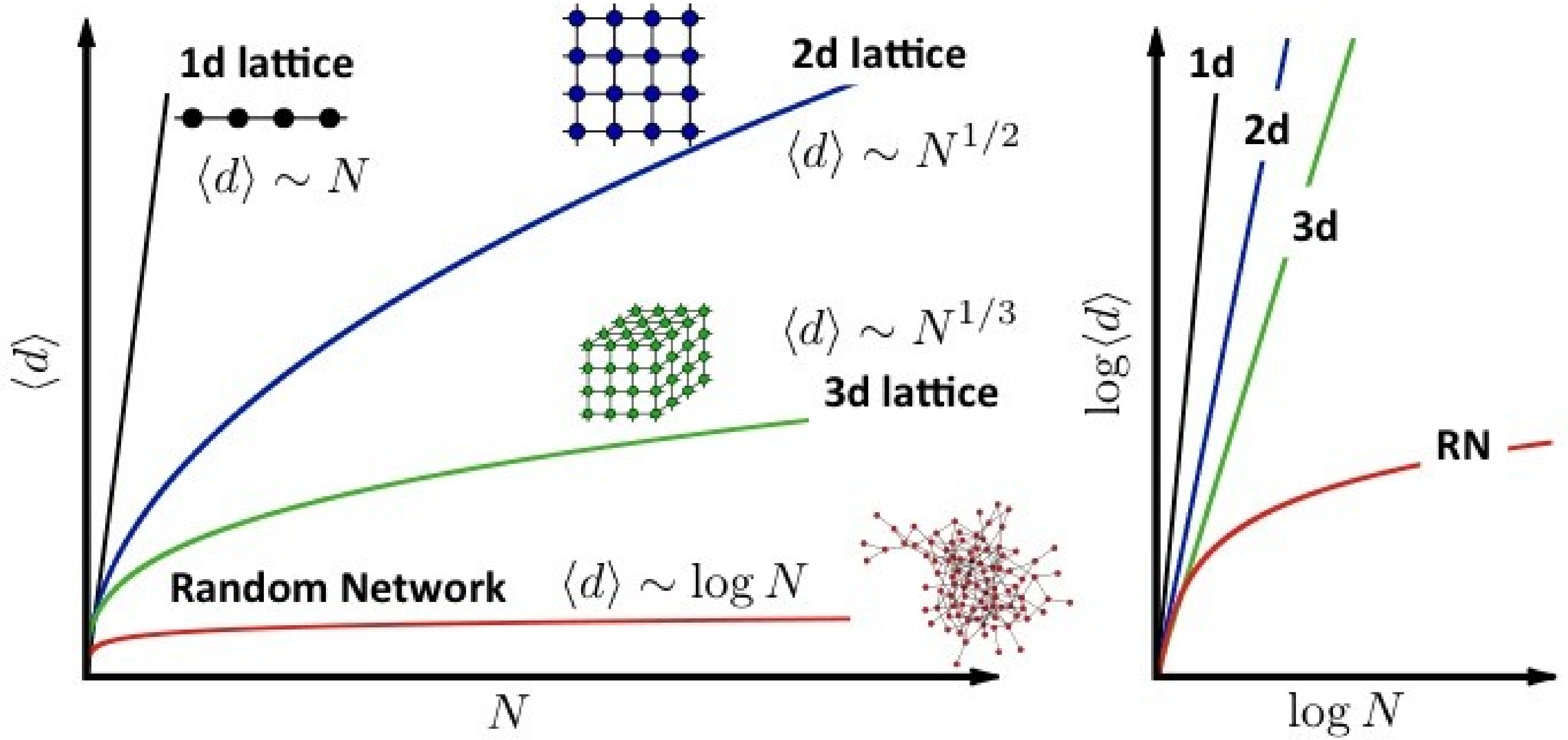


NETWORK	N	L	$\langle k \rangle$	$\langle d \rangle$	d_{max}	$\frac{\log N}{\log \langle k \rangle}$
Internet	192,244	609,066	6.33	6.98	26	6.58
WWW	325,729	1,497,134	4.60	11.27	93	8.31
Power Grid	4,941	6,594	2.67	18.99	46	8.66
Mobile Phone Calls	36,595	91,826	2.51	11.72	39	11.42
Email	57,194	103,731	1.81	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?



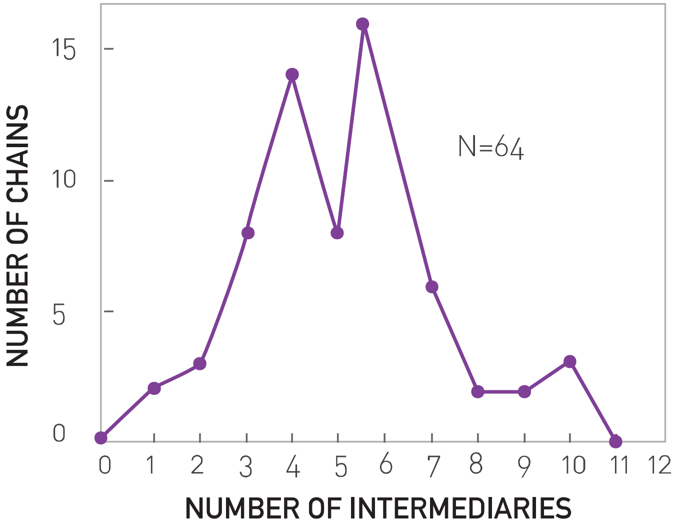
Three, Four or Six Degrees?

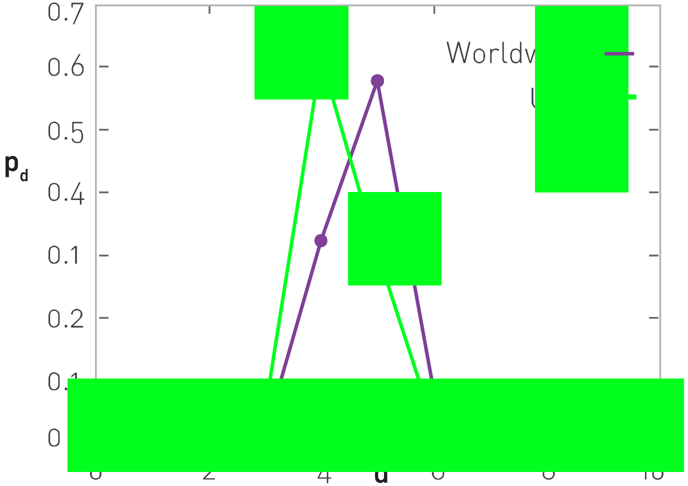
For the globe's social networks:

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

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

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

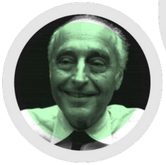




"The worker knows the manager in the shop, who knows Ford; Ford is on friendly terms with the general director of Hearst Publications, who last year became good friends with Árpád Pásztor, someone I not only know, but to the best of my knowledge a good friend of mine."

Karinthy, 1929

Manfred Kochen

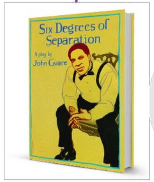


Ithiel de Sola Pool

Stanley Milgram

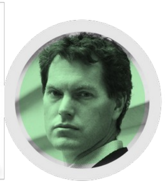
"Everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. The president of the United States. A gondolier in Venice. It's not just the big names. It's anyone. A native in a rain forest. A Tierra del Fuegan. An Eskimo. I am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



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.

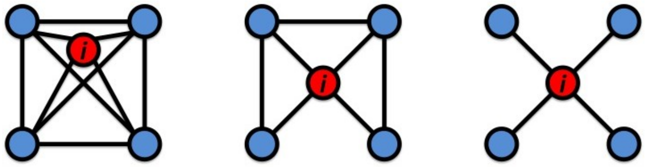
The **Facebook Data Team** measures the average distance between its users, finding "4 degrees" (**BOX 3.6**).

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.

Section 9

Clustering coefficient

CLUSTERING COEFFICIENT



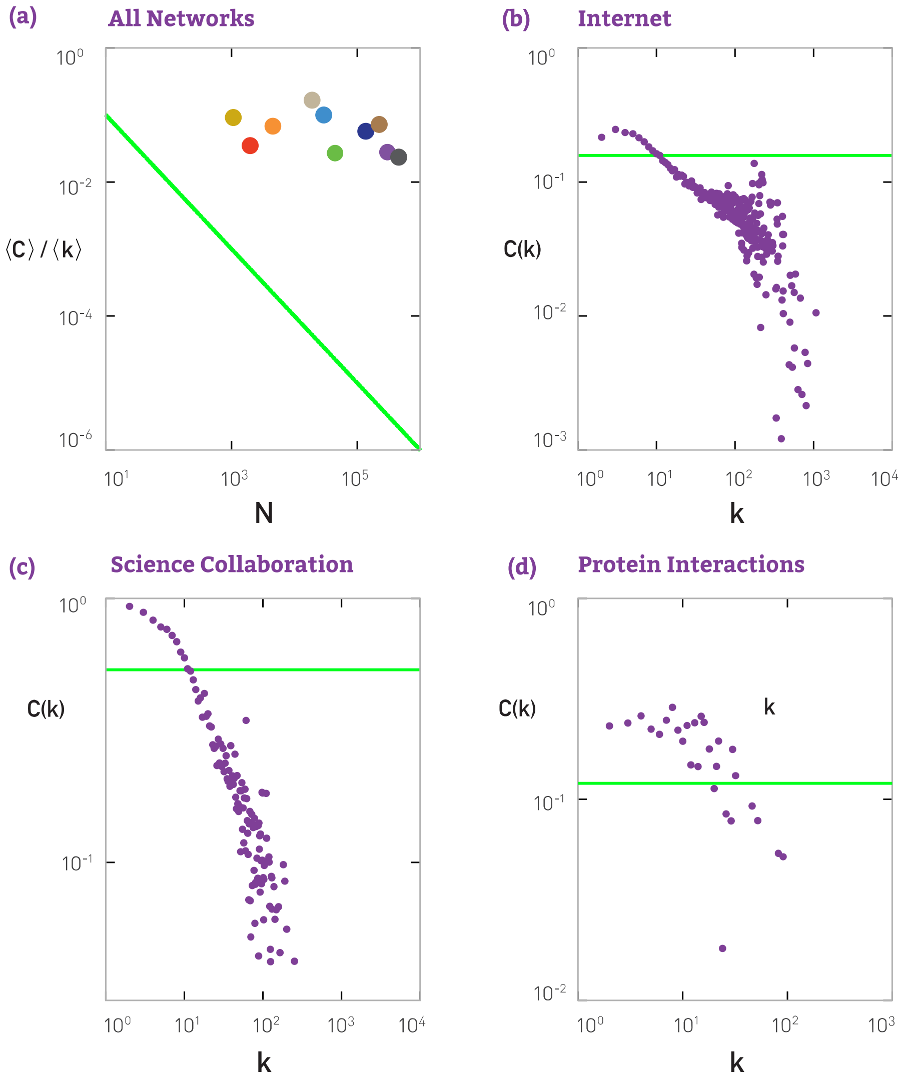
$C_i = 1$ $C_i = 1/2$ $C_i = 0$

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

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

- 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

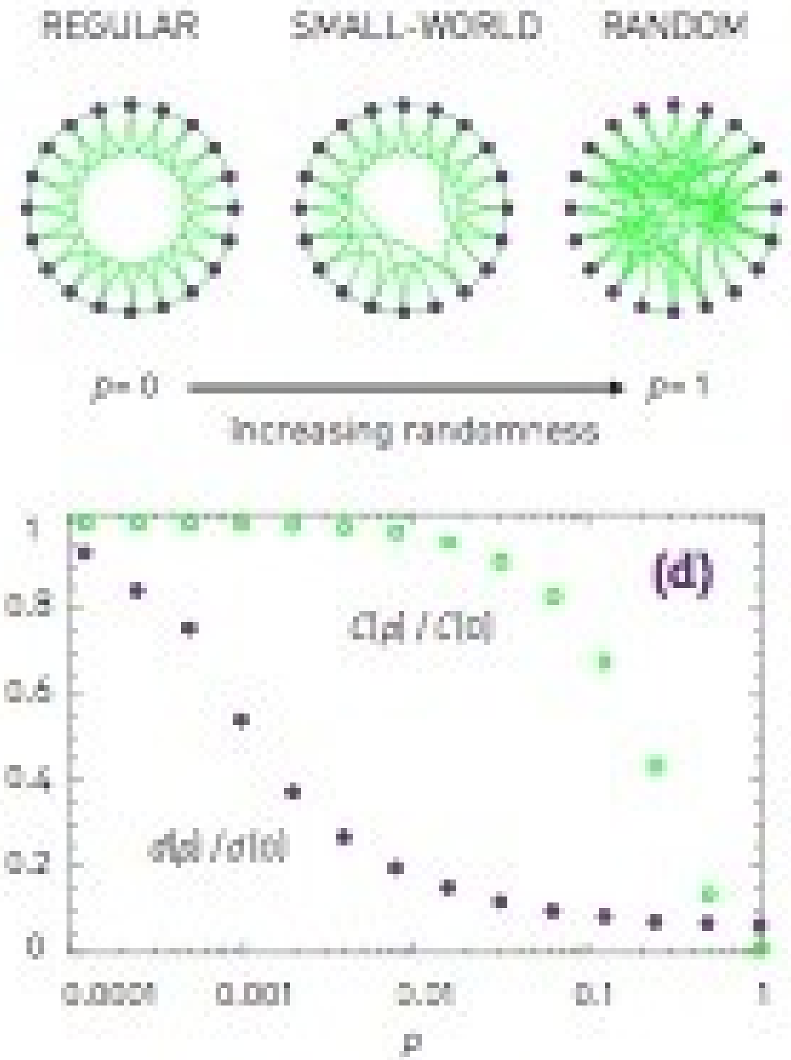


$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}.$$

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:

Average path length:

$$l_{\text{rand}} >\gg \frac{\log N}{\log\langle k\rangle}$$

Clustering Coefficient:

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

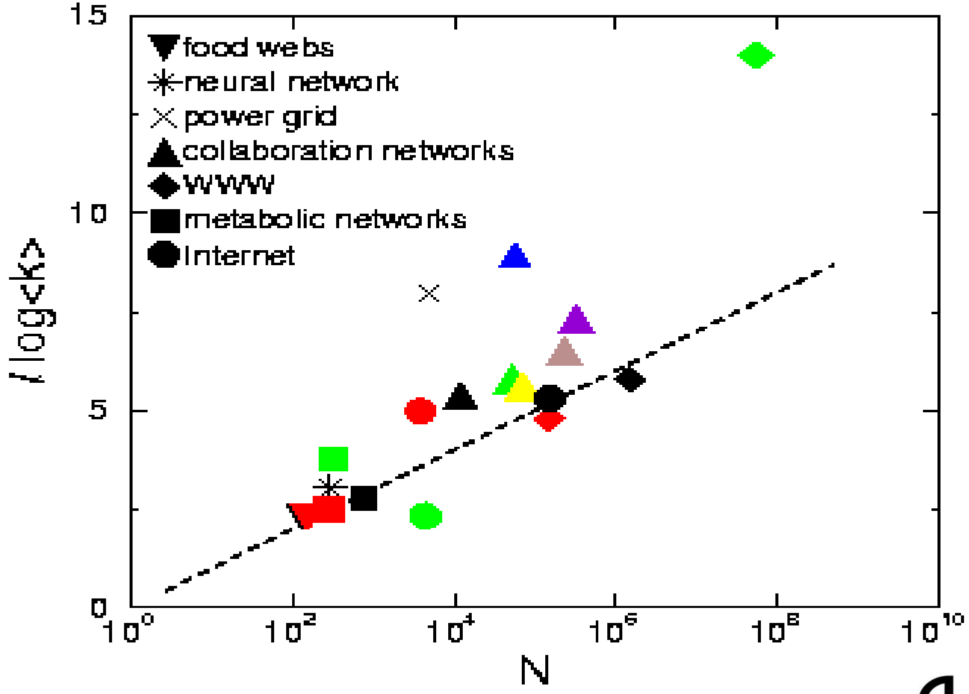
Degree Distribution:

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

PATH LENGTHS IN REAL NETWORKS

Prediction:

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

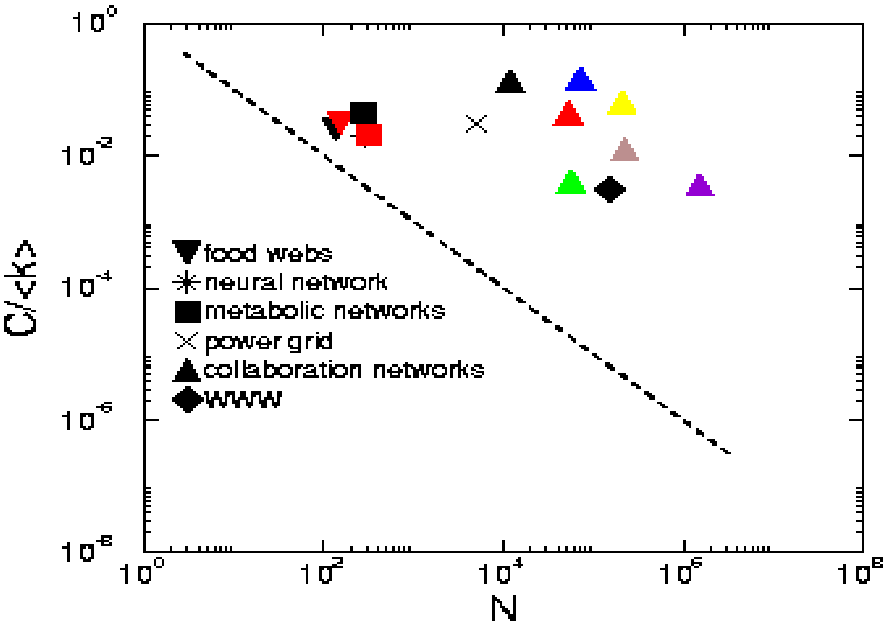


Real networks have short distances like random graphs.

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.

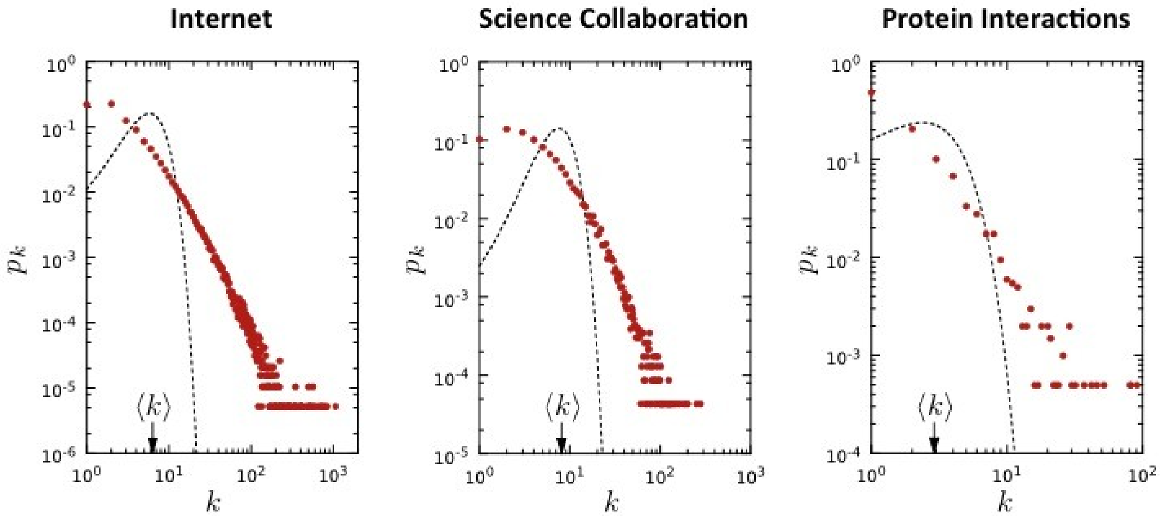
THE DEGREE DISTRIBUTION

Prediction:

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

Data:

$$P(k) \text{»} k^{-g}$$



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>

$$l_{\mathrm{rand}} > \gg \frac{\log N}{\log \langle k \rangle}$$

👍

<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>

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

👎

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.

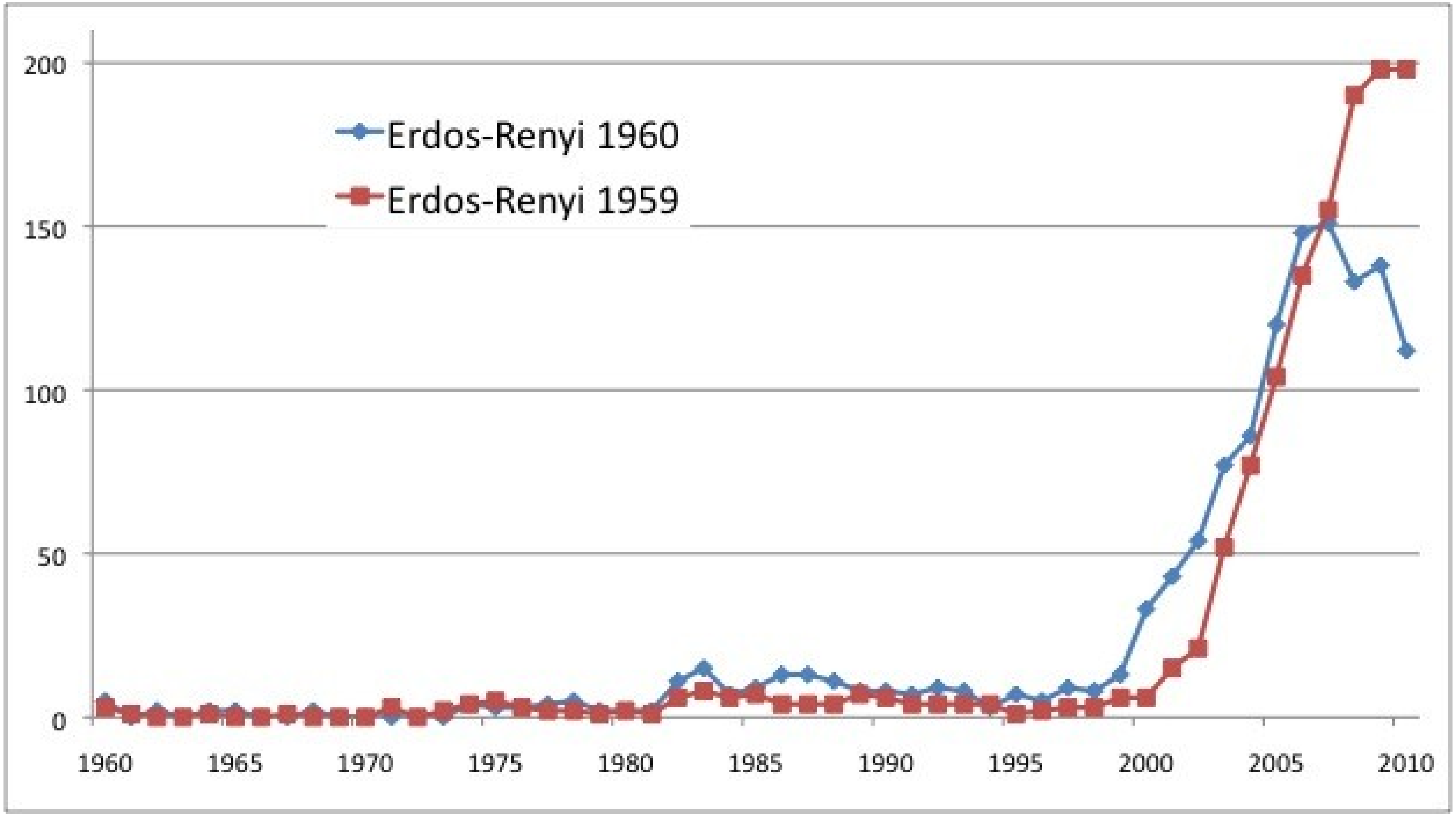
In order to identify these, we need to understand how would a particular property look like if it is driven entirely by random processes.

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

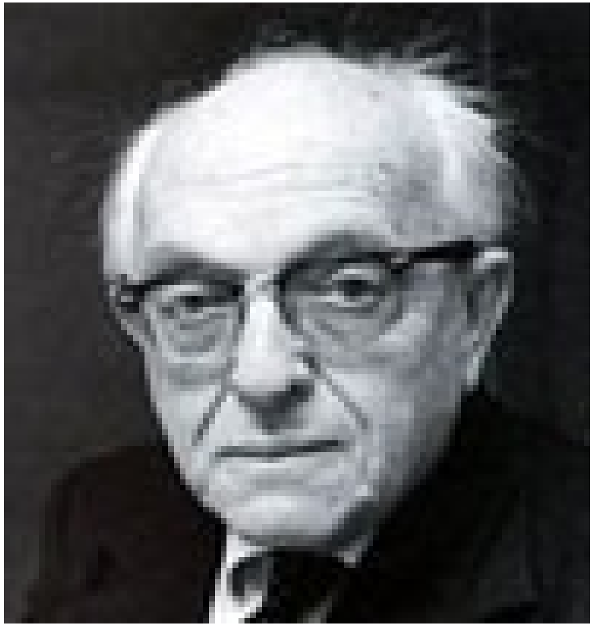
Section 11

Summary

Erdös-Rényi MODEL (1960)



HISTORICAL NOTE



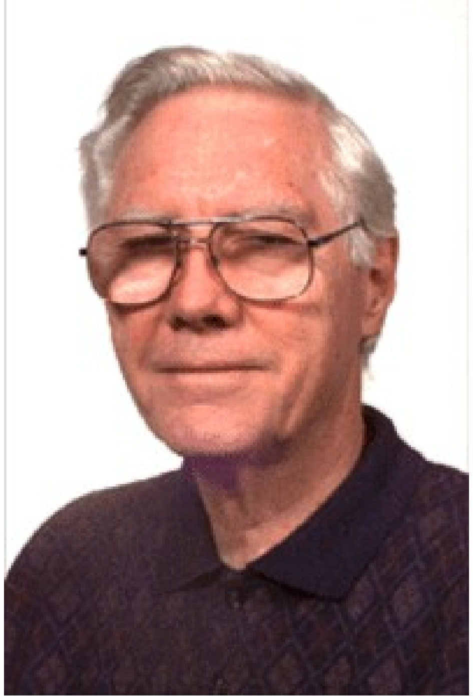
Anatol Rapoport
1911- 2007

1951, Rapoport and Solomonoff:

→ first systematic study of a random graph.
→demonstrates the phase transition.

→natural systems: neural networks; the social networks of physical contacts (epidemics); genetics.

1959: *G(N,p)*

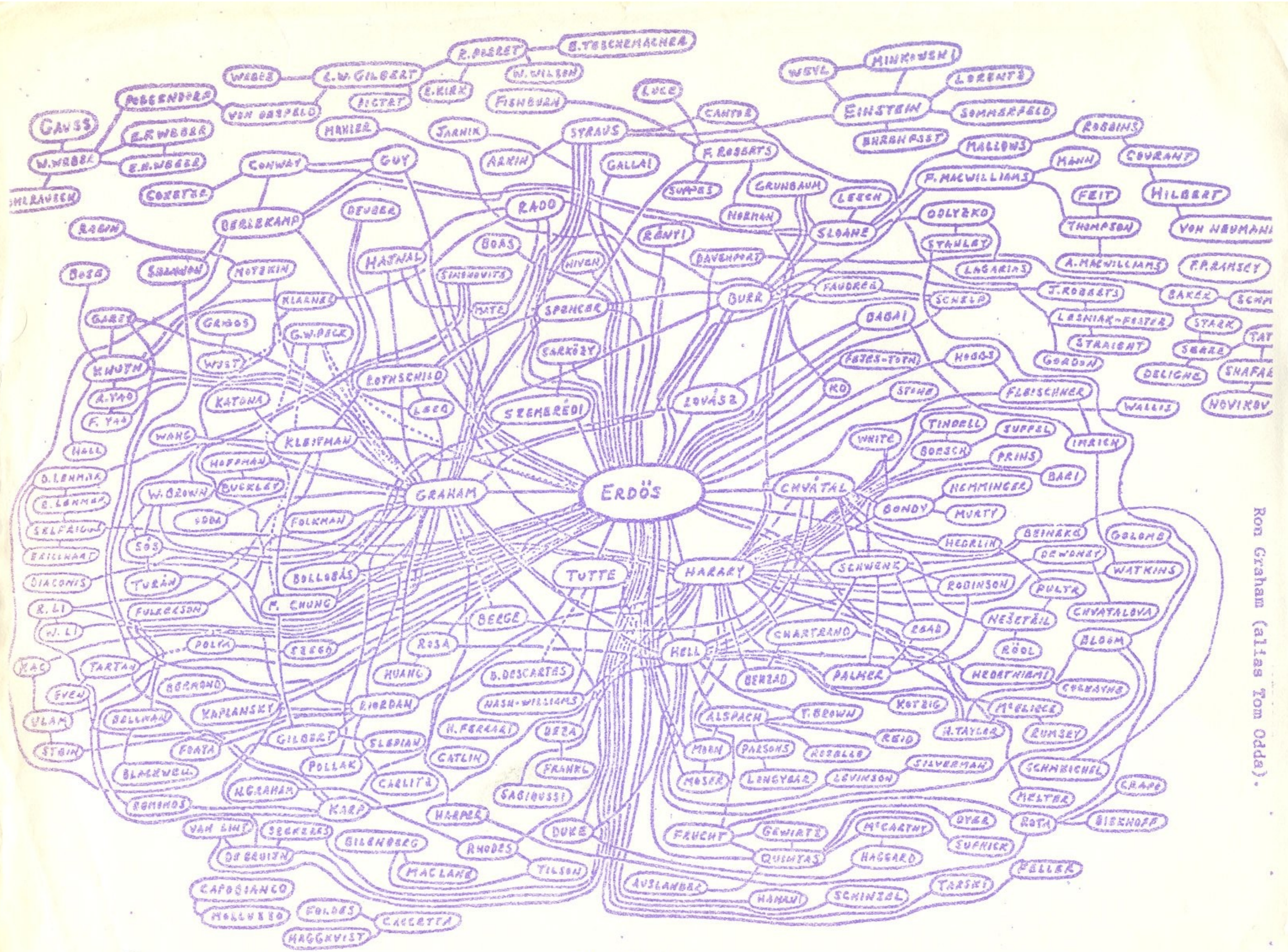


Edgar N. Gilbert
(**b**.1923)

Why do we call it the Erdos-Renyi random model?

HISTORICAL NOTE

NETWORK DATA: SCIENCE COLLABORATION NETWORKS



Ron Graham (alias Tom Odda).

Figure 1

To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

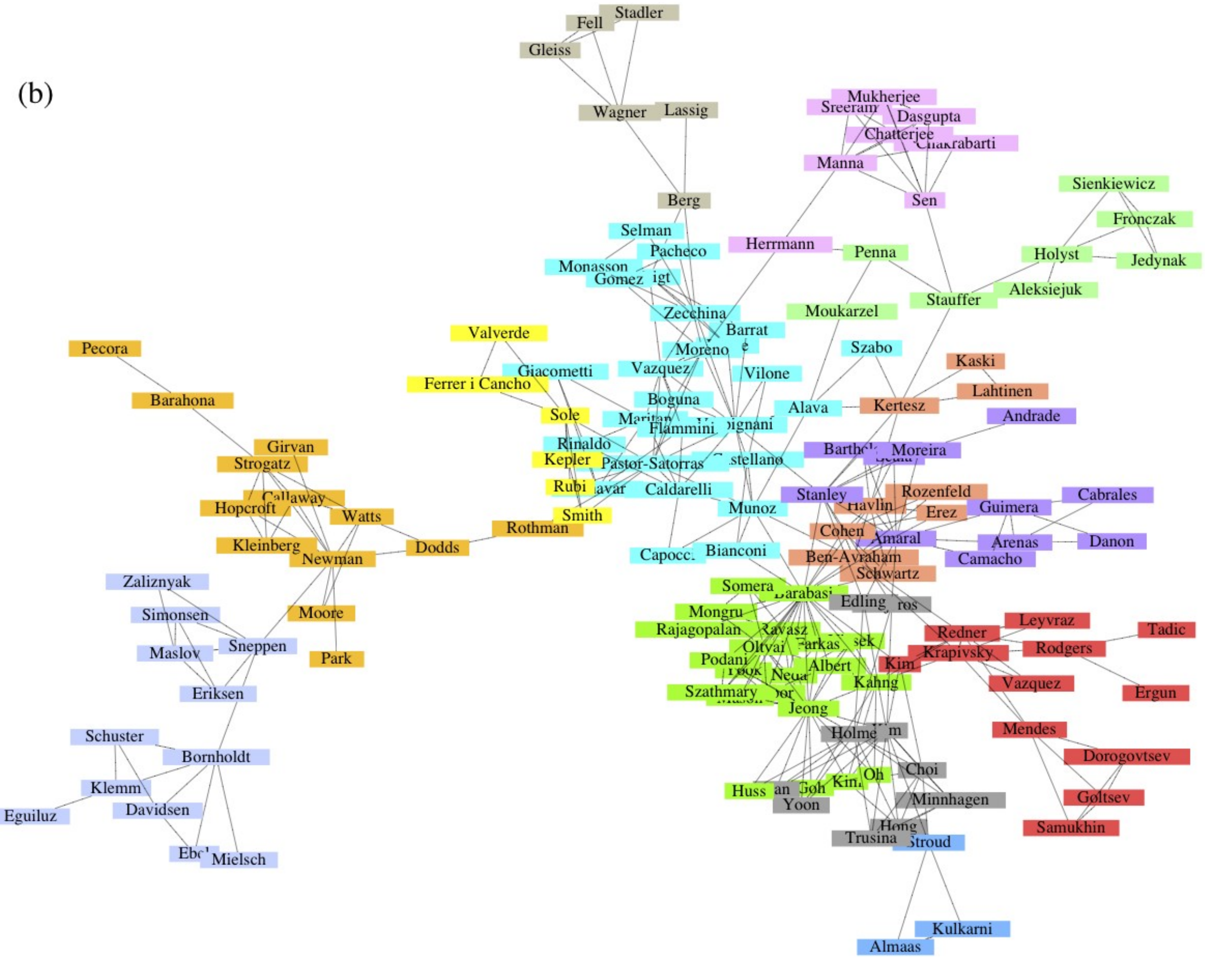
Erdos:
1,400 papers
507 coauthors

Einstein: EN=2
Paul Samuelson EN=5

….

ALB: EN: 3

NETWORK DATA: SCIENCE COLLABORATION NETWORKS

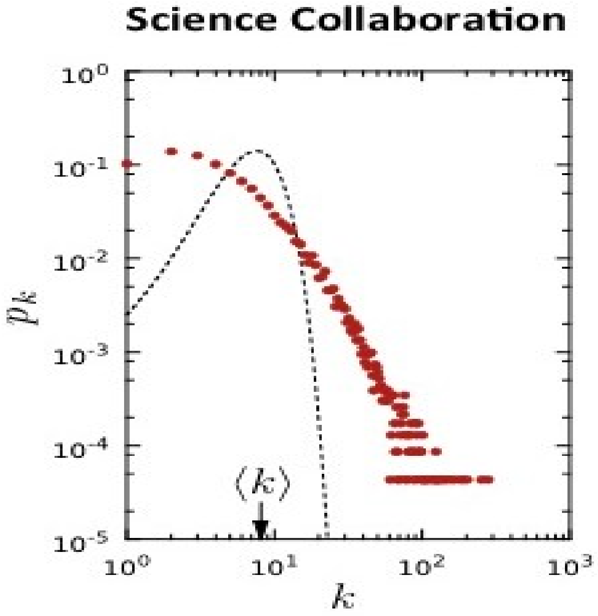


Collaboration Network:
Nodes: Scientists
Links: Joint publications

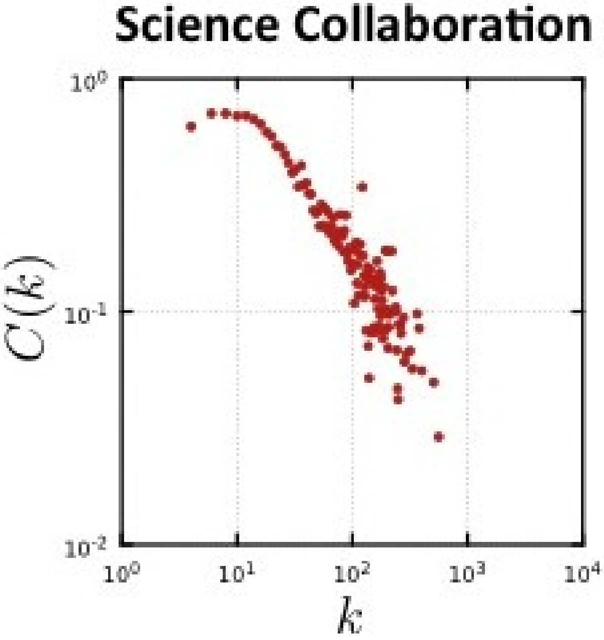
Physical Review:
1893 – 2009.

N=449,673
L=4,707,958

See also Stanford Large Network database
http://snap.stanford.edu/data/#canets



Scale-free



Hierarchical

THE END

FINAL PROJECTS

PROJECT PAIRS

1. **NETSI PHD STUDENTS**
   You will complete your projects individually.

2. **EVERYONE ELSE**
   Work in pairs; we are sharing a spreadsheet to help identify mutual interests.
   Find someone who shares a DIFFERENT academic background to you!

COMPONENTS OF THE PROJECT

1. **DATA ACQUISITION**
   Downloading the data and putting it in a usable format

2. **NETWORK RESPRESENTATION**
   What are the nodes and links

3. **NETWORK ANALYSIS**
   What questions do you want to answer with this network, and which tools/measurements will you use?

DATA ACQUISITION

- Many online data sources will have an **API** (application programming interface) that allows querying and downloading the data in a targeted way
  - Example: What are all movies from 1984-1995 starring Kevin Bacon and distributed by Paramount Pictures?
  - This is done either through a web interface or through a library within a programming language

- Other sources will provide raw bulk data (e.g., Excel spreadsheets) that require processing, either manually or through a program you will write

$\text{“GRAPH”} \neq \text{“NETWORK”}$

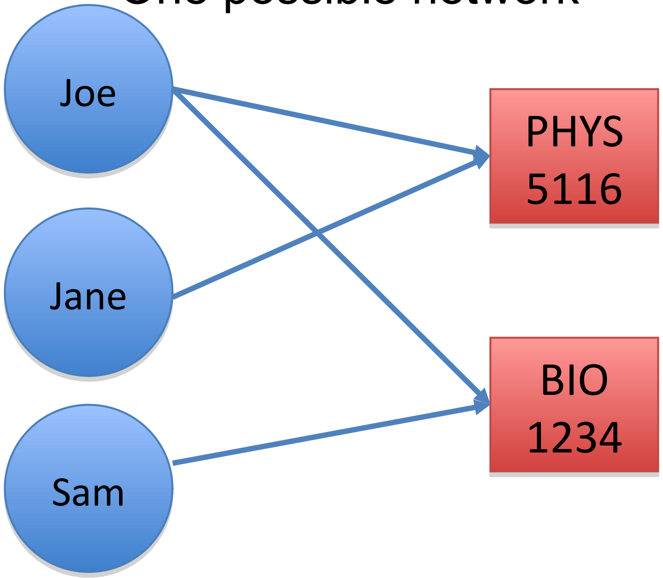
- Most datasets will admit more than one representation as a network
- Some representations will be more or less informative than others
- Figuring out the “network” that’s buried in your data is part of your project!

NETWORK RECONSTRUCTION

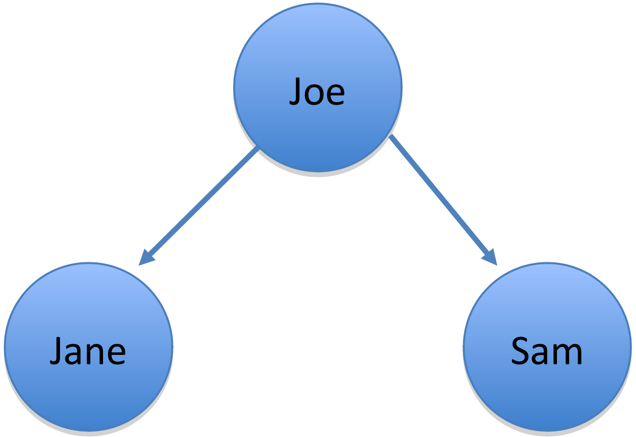
$$\text{"GRAPH"} \neq \text{"NETWORK"}$$

Suppose you have a list of students and the courses they are registered for

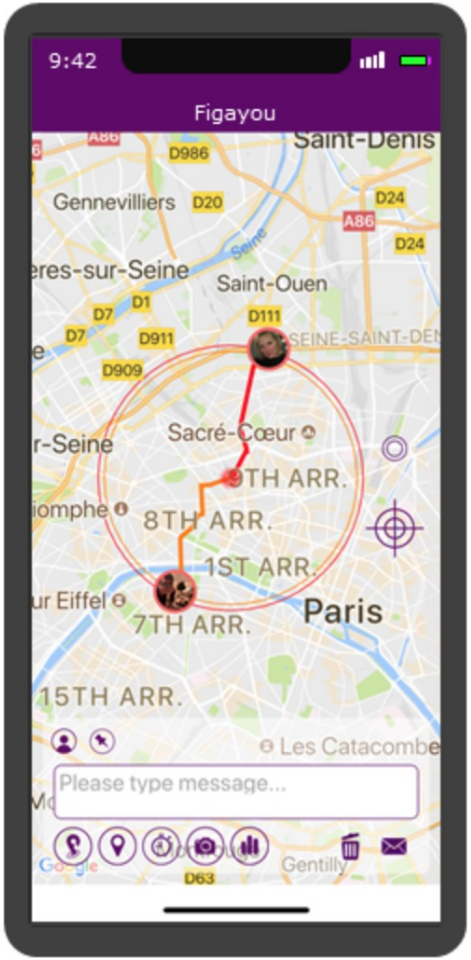
One possible network



Another possibility



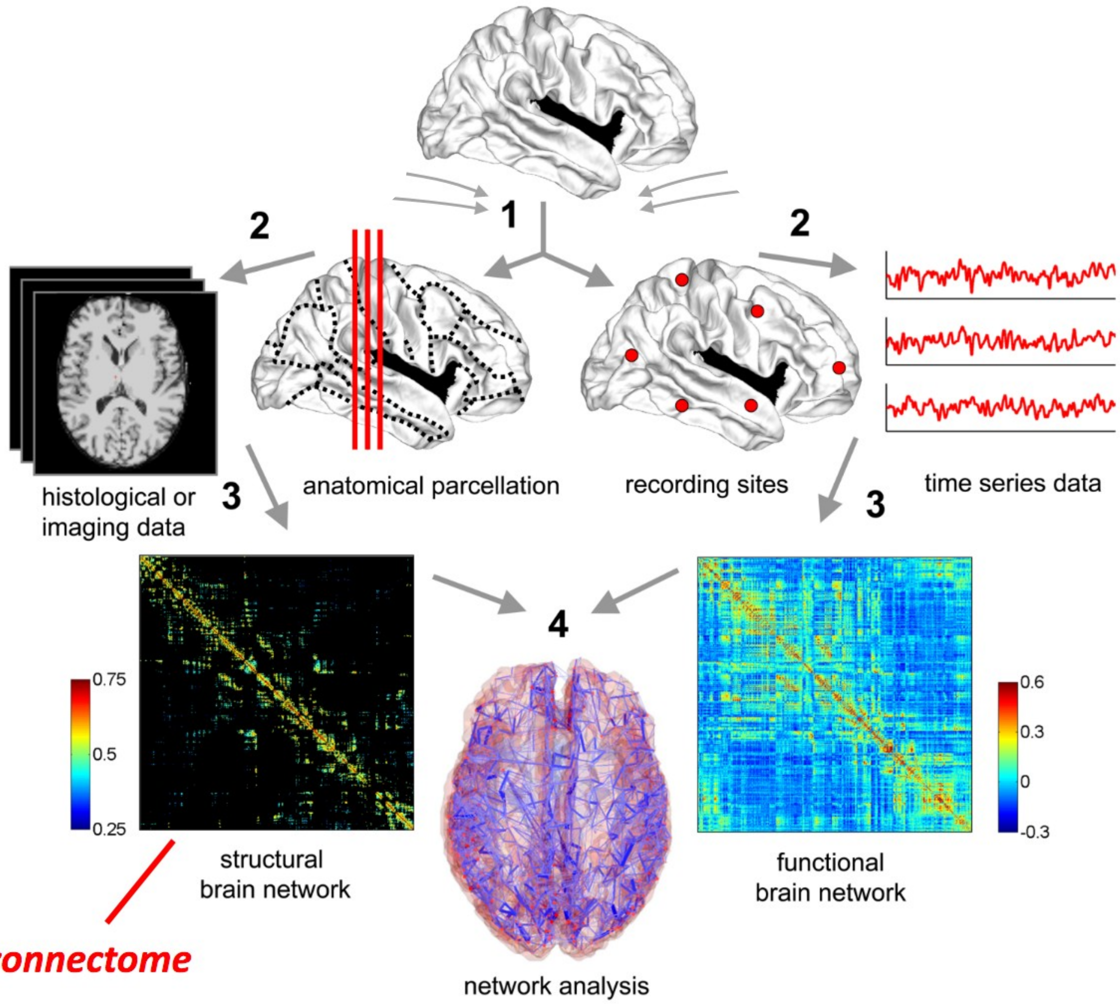
Mobility: Figayou



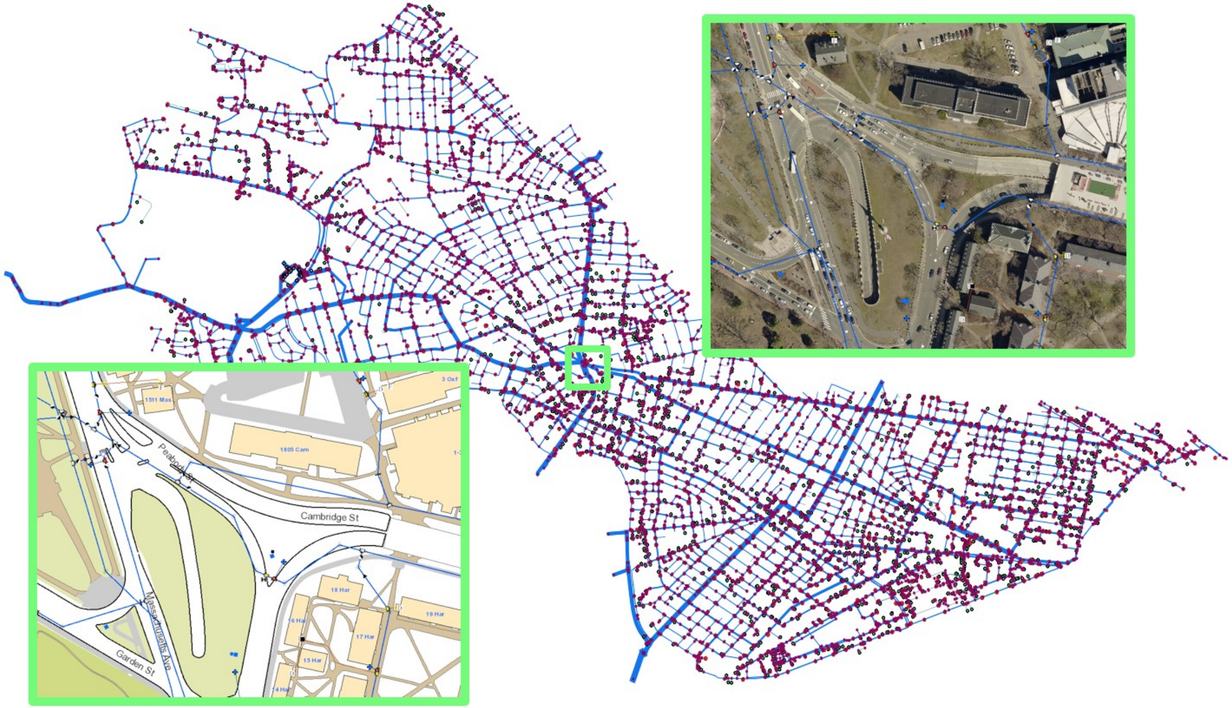
- Mobility data (various settings: social, conferences…)
- Metadata
- Representative (Hamid Benbrahim) in Boston willing to work with you

fMRI

- FMRI timeseries for human brain
- Healthy and patient data
- Collaborators at NEU

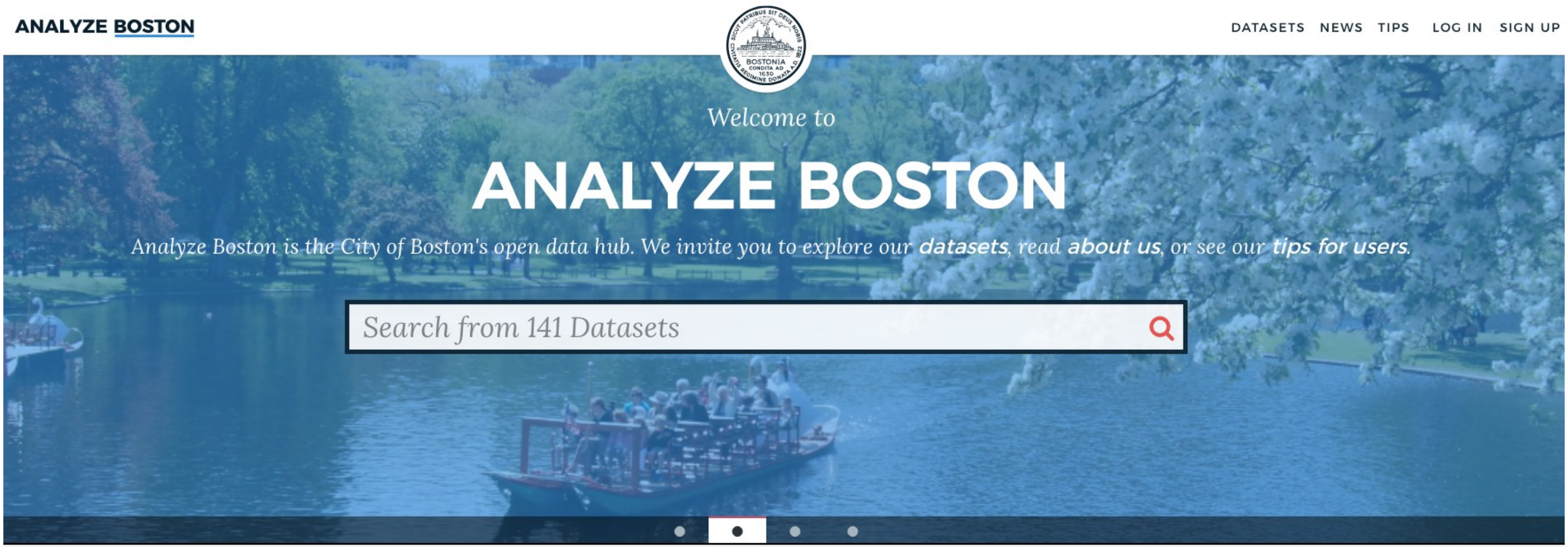


Infrastructure networks



- Eg Cambridge water distribution
- Partially embedded

Boston 311



ANALYZE BOSTON

DATASETS NEWS TIPS LOG IN SIGN UP

Welcome to

ANALYZE BOSTON

Analyze Boston is the City of Boston's open data hub. We invite you to explore our **datasets**, read **about us**, or see our **tips for users**.

Search from 141 Datasets

Final project guidelines

Measure: N(t), L(t) [t- time if you have a time dependent system); P(k) (degree distribution); <l> average path length; C (clustering coefficient), $C_{rand,}$ C(k); Visualization/communities; P(w) if you have a weighted network; network robustness (if appropriate); spreading (if appropriate).

It is not sufficient to measure things– you need to discuss the insights they offer:
What did you learn from each quantity you measured?
What was your expectation?
How do the results compare to your expectations?

Time frame will be strictly enforced. Approx 12min + 3 min questions;
No need to write a report—you will hand in the presentation.
Send us an email with names/titles/program.
Come earlier and try out your slides with the projector. Show an entry of the data source—just to have a sense of how the source looks like. On the slide, give your program/name.

Grading criteria:
Use of network tools (completeness/correctness);
Ability to extract information/insights from your data using the network tools;
Overall quality of the project/presentation.