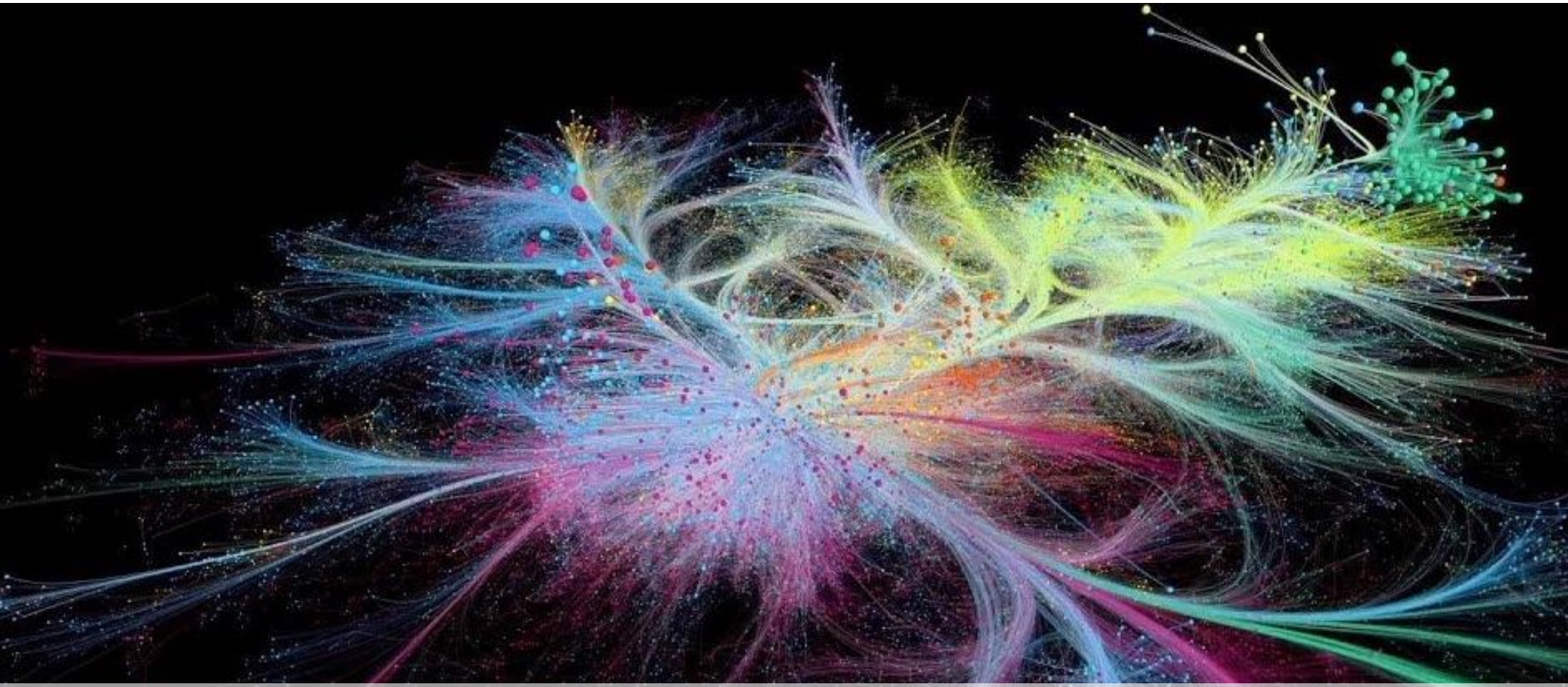


CPSC 572/672: Fundamentals of Social Network Analysis and Data Mining

Admin

Project Matchmaking: Week 1

CPSC 572/672 Project Matchmaking

File Edit View Insert Format Data Tools Add-ons Help Last edit was 2 minutes ago

Share

D8

	A	B	C	D	E	F	G	H	I
1	**Name**	**572/672**	**Skills I have**	**Skills I want to develop**	**Datasets or questions I am interested in**	**Partner**			
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

https://docs.google.com/spreadsheets/d/1vAOhbuaQooo_O5cRZru4QlqgwzTncoXp9URFmTqvfkM/edit?usp=sharing

Project Datasets

Data Sources

You may use any dataset that you have the permissions and expertise to access, and that is NOT already in a network format. You can use APIs (e.g. https://www.flickr.com/services/api/), scrape data yourself if you have the technical expertise, or refer to a data repository. Here are some helpful resources to help you with your search. The links below mostly point to Open Data sources:

- https://github.com/awesomedata/awesome-public-datasets
- Open Calgary (City of Calgary open data portal)
- Kaggle (searchable)
- Github public datasets (categorized by subject)
- Google Dataset Search (currently in beta but quite useful)
- Data Sources for Data Science (UCalgary Library Guide for the Data Science program

Project Datasets

You cannot use these (they are in network format already), but you may be inspired:

- http://snap.stanford.edu/data/index.html
- https://icon.colorado.edu/#!/

List of previous projects - see D2L later today.

Next week installations

Contact

Mailing list
Issue tracker
Source

Releases

Stable (notes)

2.8.6 — August 2022
download | doc | pdf

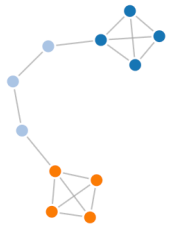
Latest (notes)

3.0 development
github | doc | pdf

Archive

NetworkX
Network Analysis in Python

NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.



Software for complex networks

- Data structures for graphs, digraphs, and multigraphs
- Many standard graph algorithms
- Network structure and analysis measures
- Generators for classic graphs, random graphs, and synthetic networks
- Nodes can be "anything" (e.g., text, images, XML records)
- Edges can hold arbitrary data (e.g., weights, time-series)
- Open source 3-clause BSD license
- Well tested with over 90% code coverage
- Additional benefits from Python include fast prototyping, easy to teach, and multi-platform

Recommend Anaconda

We will use Jupyter notebook

Python 3

Gephi

Gephi
makes graphs handy

Download | Blog | Wiki | Forum | Support | Bug tracker

Home | Features | Learn | Develop | Plugins | Services | Consortium

The Open Graph Viz Platform

Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free.

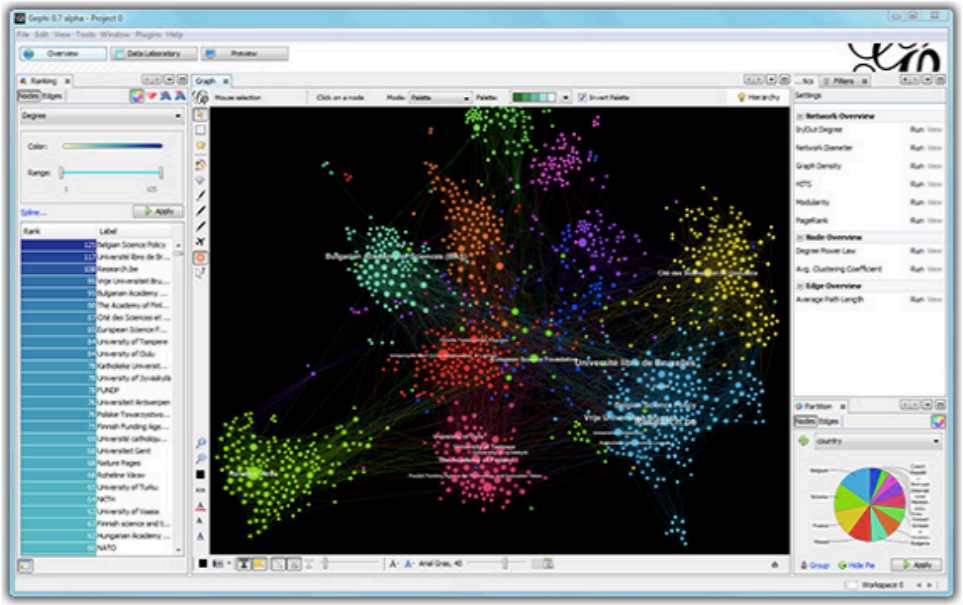
Runs on Windows, Mac OS X and Linux.

Learn More on Gephi Platform »

Download FREE
Gephi 0.9.1

Release Notes | System Requirements

- Features
- Quick start
- Screenshots
- Videos

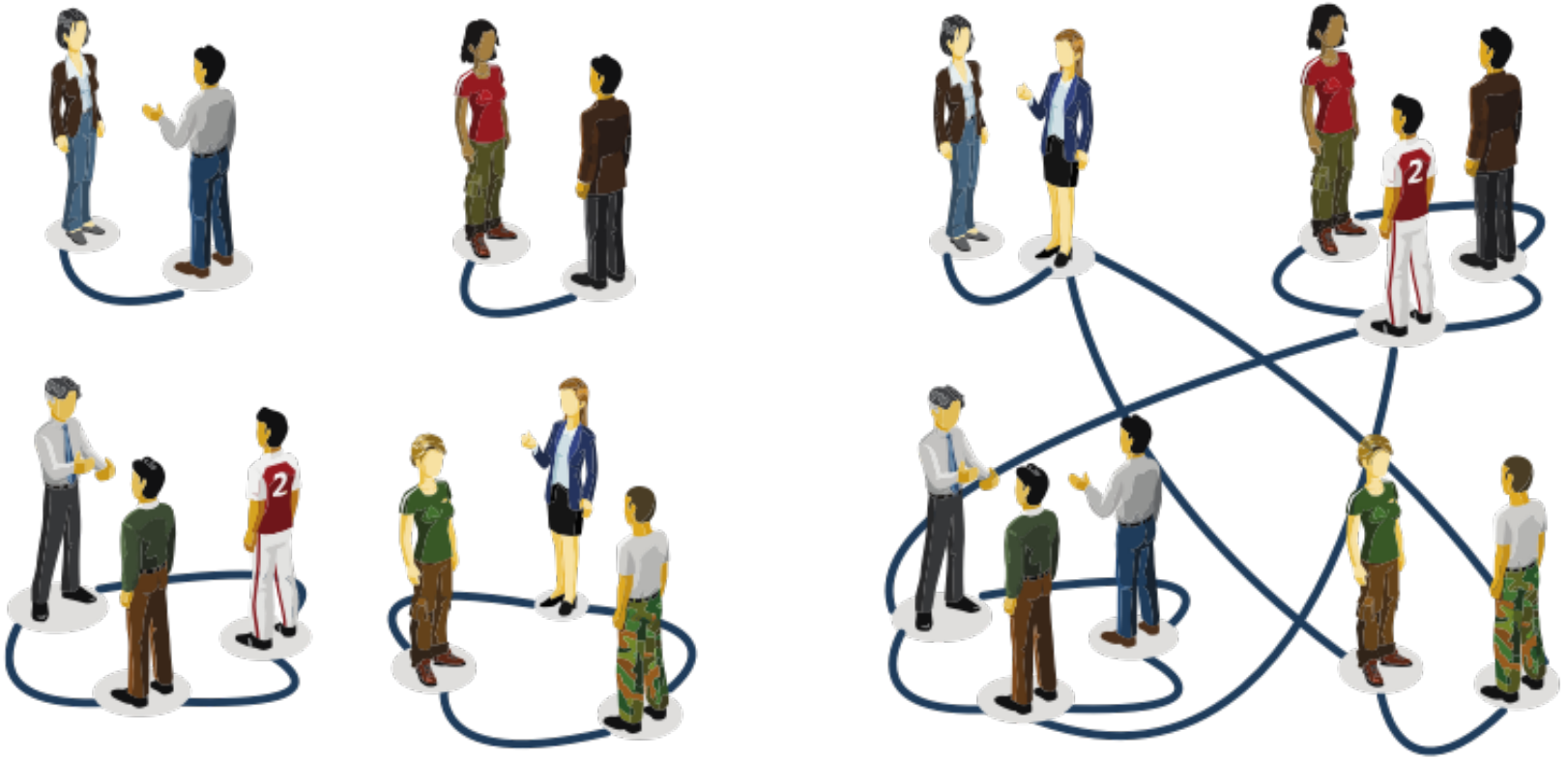


Questions

Section 1

Introduction

RANDOM NETWORK MODEL

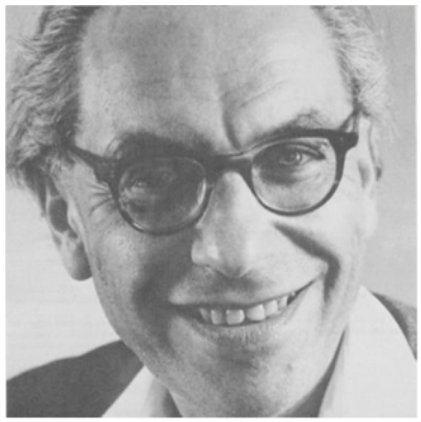


Section 3.2

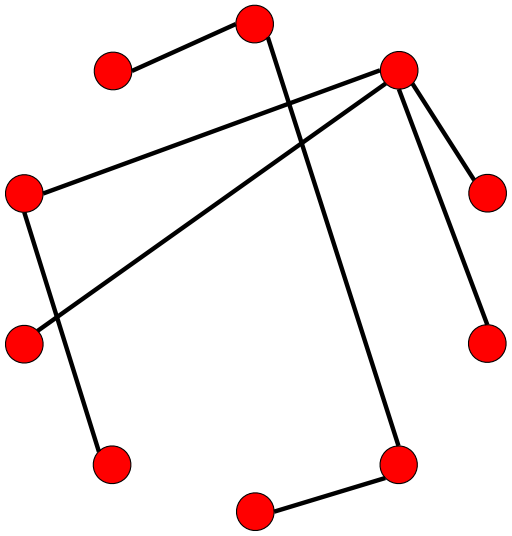
The random network model

RANDOM NETWORK MODEL

Pául 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**.

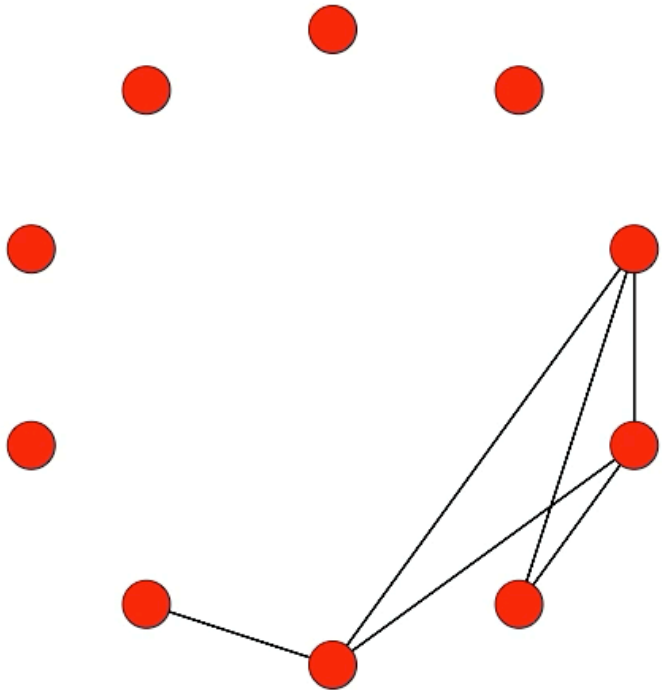
G(*N*, *L*) **Model**

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**

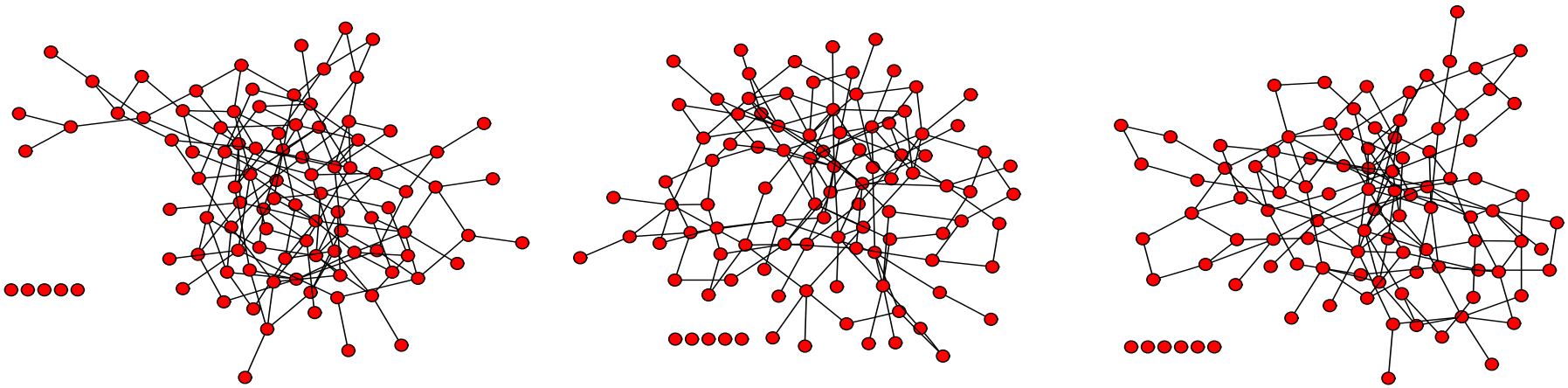
Each pair of N labeled nodes is connected with probability *p*, a model introduced by Gilbert [10].

RANDOM NETWORK MODEL



RANDOM NETWORK MODEL

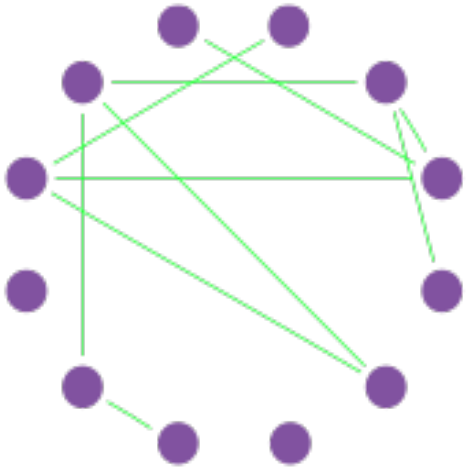
p=0.03
N=100



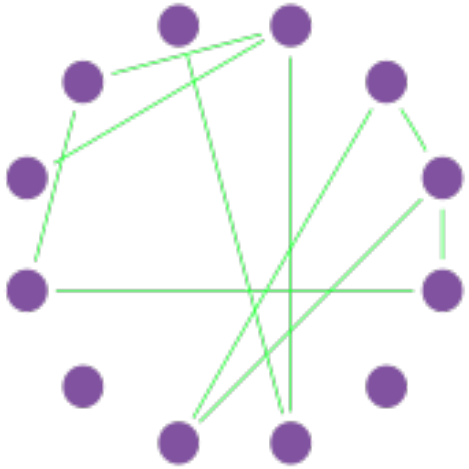
The number of links is variable

RANDOM NETWORK MODEL

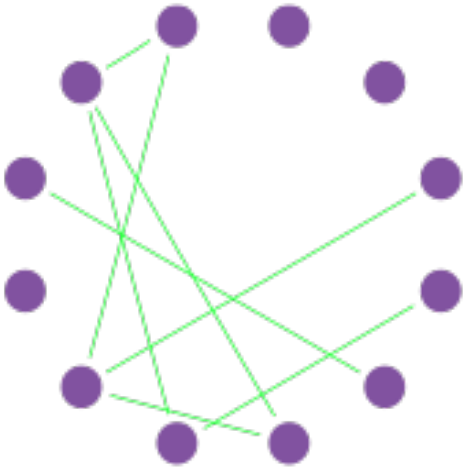
p=1/6
N=12



L=8



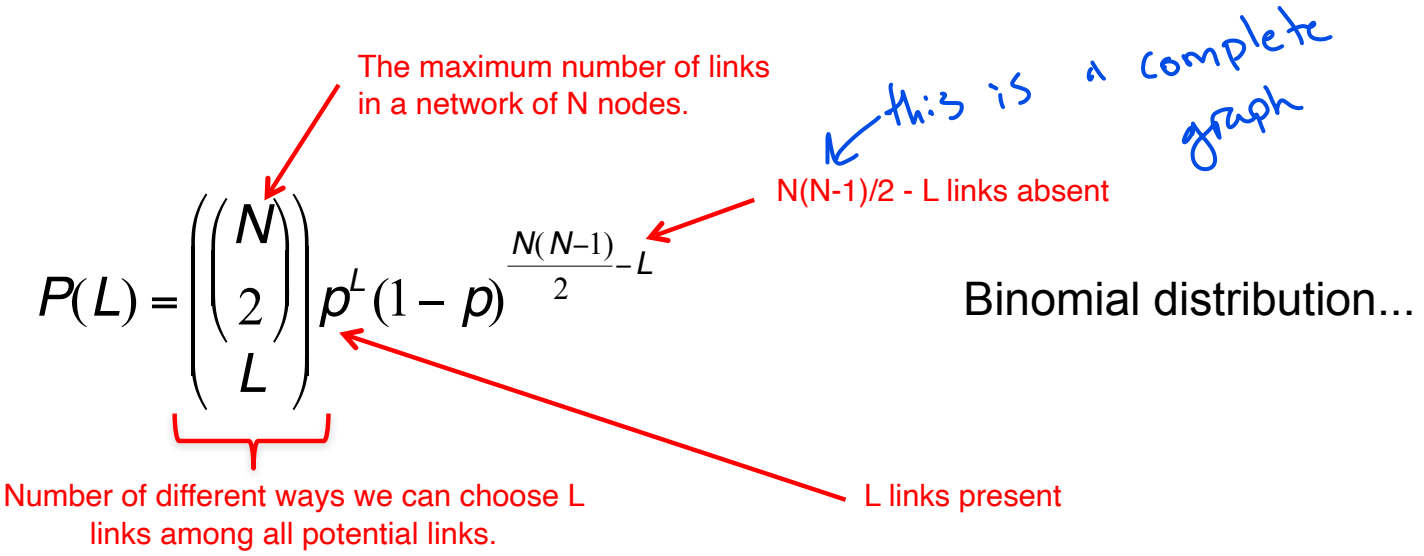
L=10



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*:



MATH TUTORIAL | Binomial Distribution: The bottom line

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

$$< x >= Np$$

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

MATH TUTORIAL | Binomial Distribution: The bottom line

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

← not gonna be assessed

$$< x >= Np$$

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

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

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

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

•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

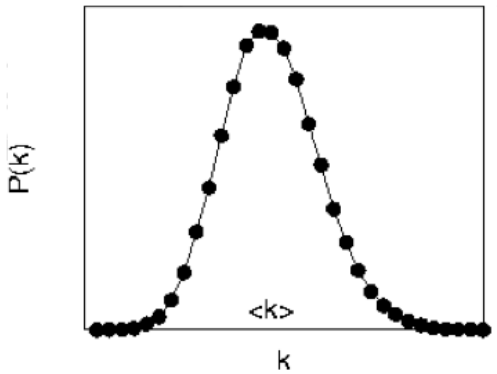
$$\sigma^2 = p(1-p)\frac{N(N-1)}{2}$$

if we know the # of links we know the Avg degree.

Section 3.4

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

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

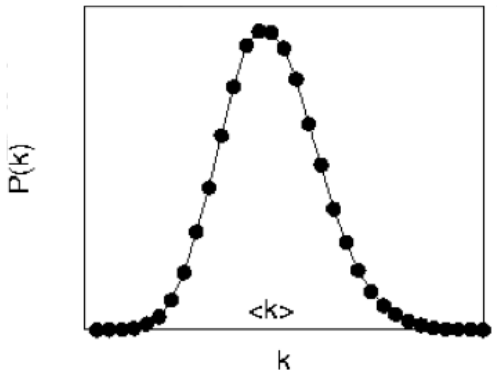
$$\frac{\sigma_k}{< k >} = \left[\frac{1-p}{p}\frac{1}{(N-1)}\right]^{1/2} \approx \frac{1}{(N-1)^{1/2}}$$

Characteristics

more narrow means more nodes are closer to the Avg degree

↳ so if pick a node on Avg that ~~means~~ will have Avg degree to the Avg

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>.

POISSON DEGREE DISTRIBUTION

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

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

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

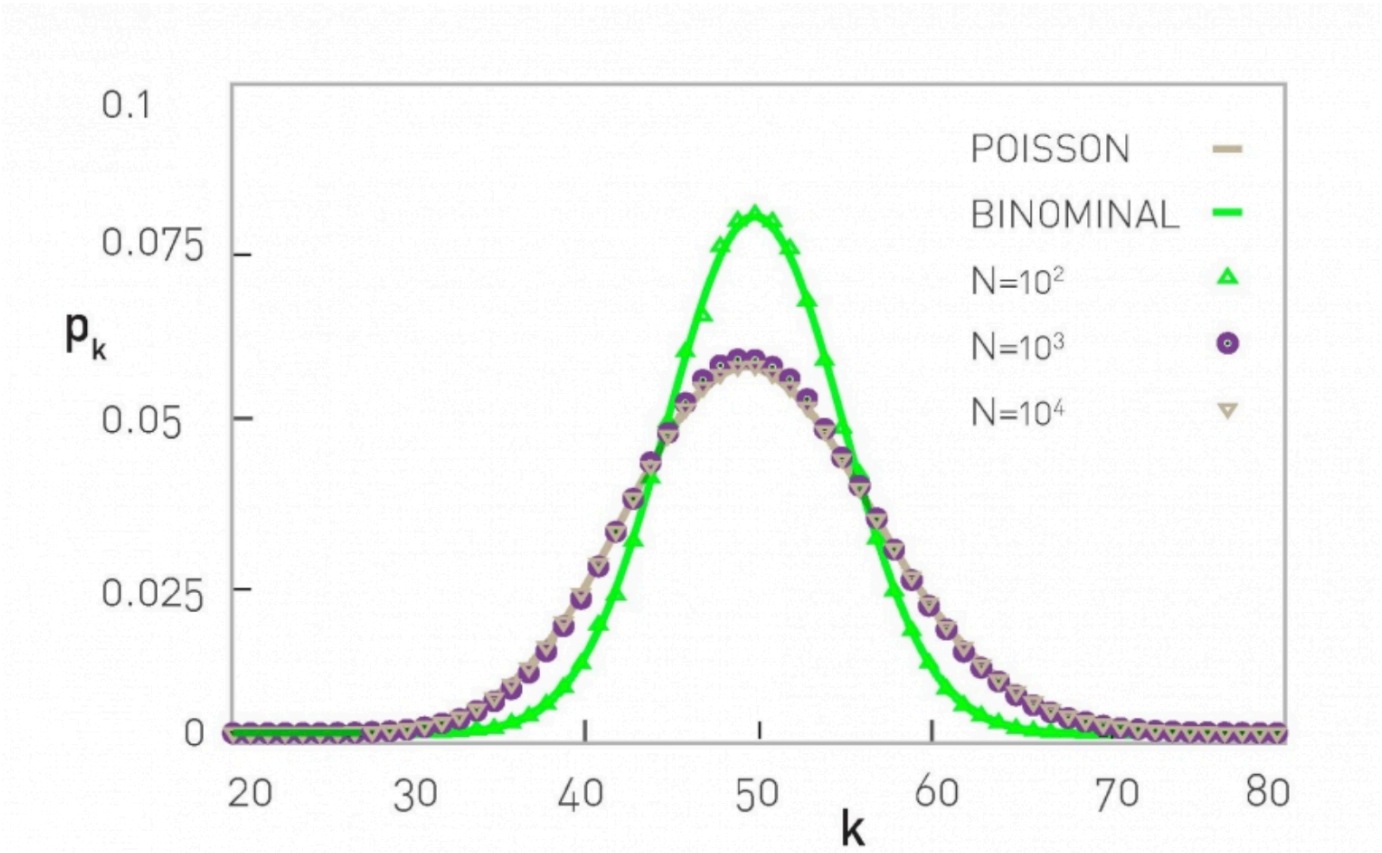
For large N and small k, we arrive to the Poisson distribution:

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

DEGREE DISTRIBUTION OF A RANDOM GRAPH

<k>=50

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

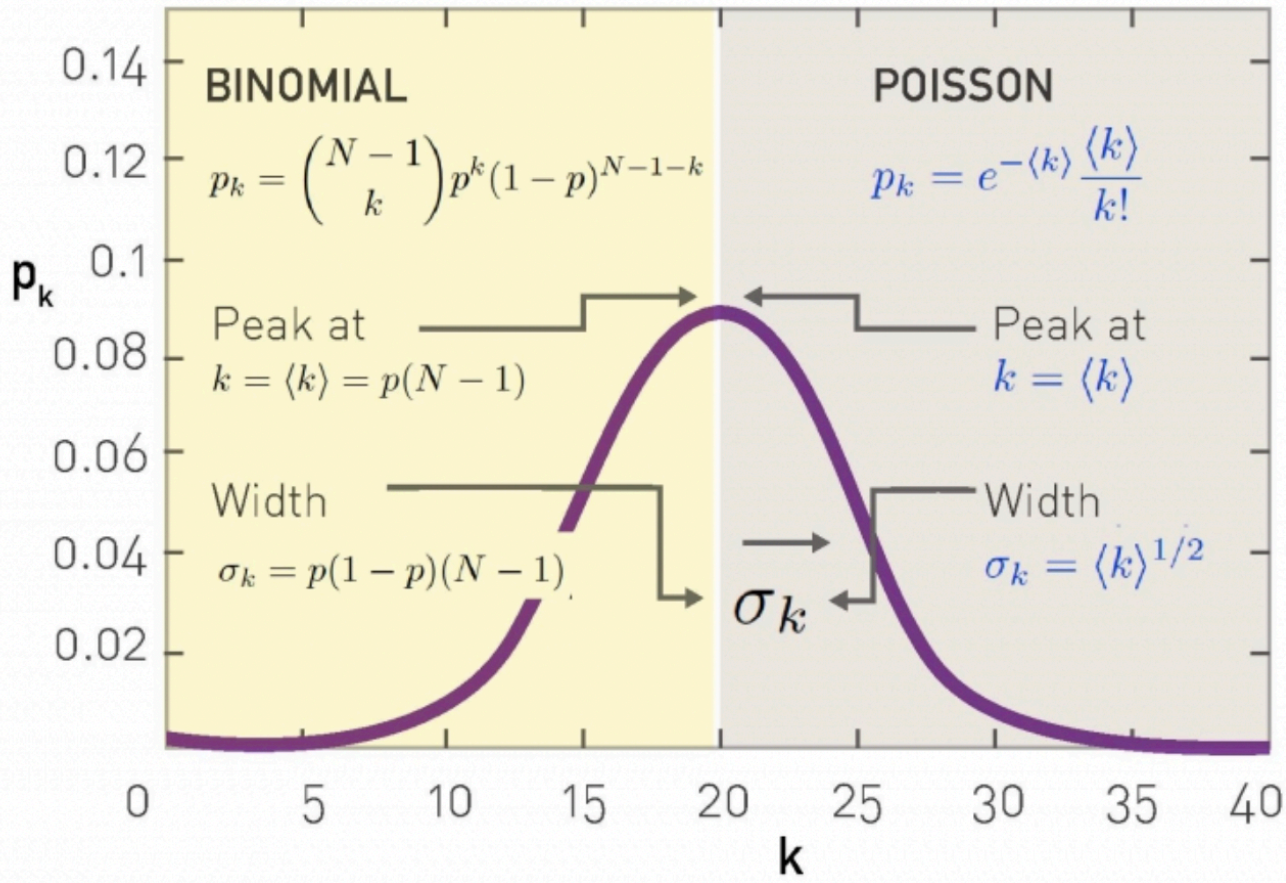


The Poisson distribution is better for models ↓ huge

DEGREE DISTRIBUTION OF A RANDOM NETWORK

Exact Result
-binomial distribution-

Large N limit
-Poisson distribution-



Section 3.4

Real Networks are not Poisson

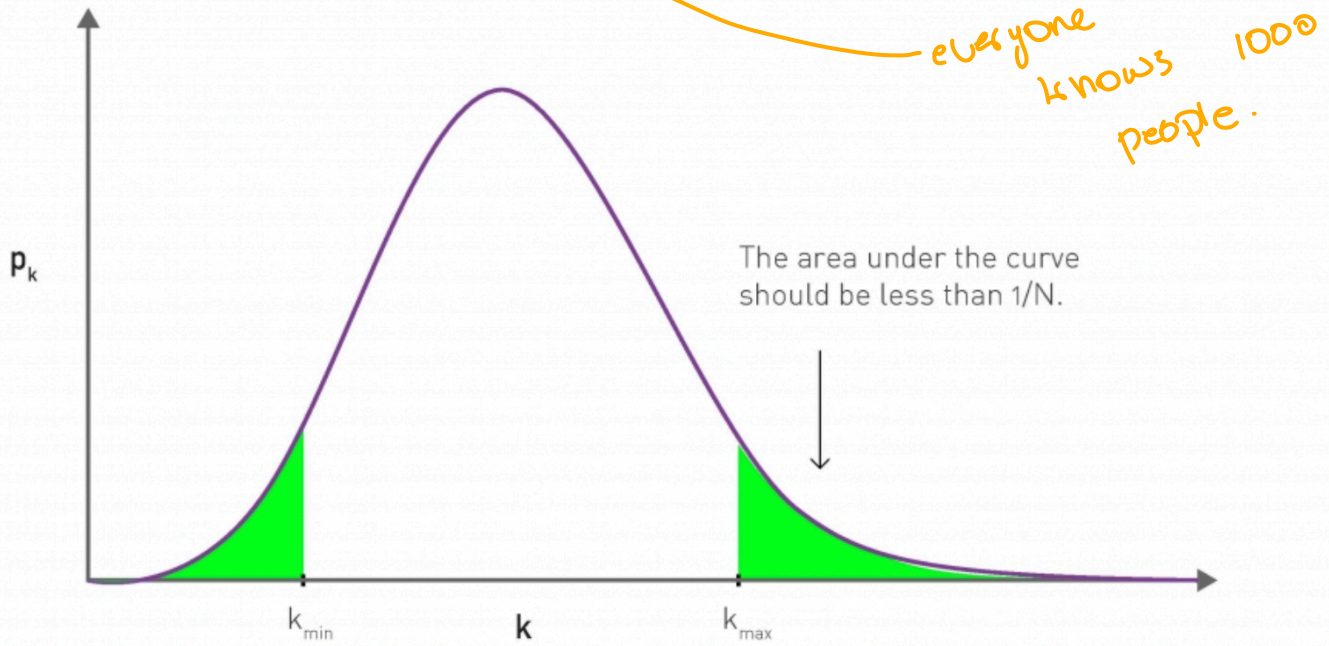
Section 3.5 Maximum and minimum degree

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

Toy model

almost all ppl on earth

everyone knows 1000 people.



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

$k_{max}=1,185$

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

$k_{min}=816$

NO OUTLIERS IN A RANDOM SOCIETY

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

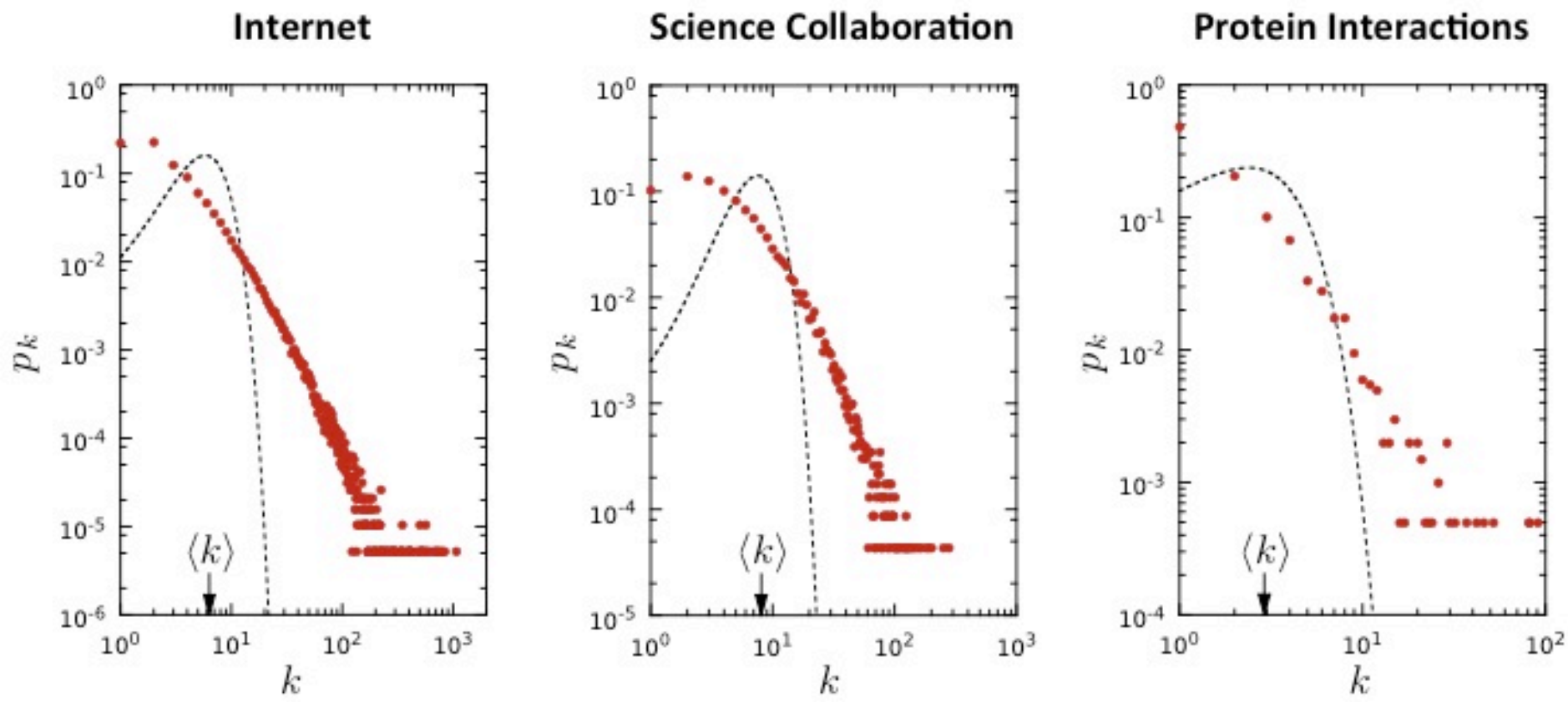
→The most connected individual has degree k_{max}~1,185
→The least connected individual has degree k_{min} ~ 816

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.

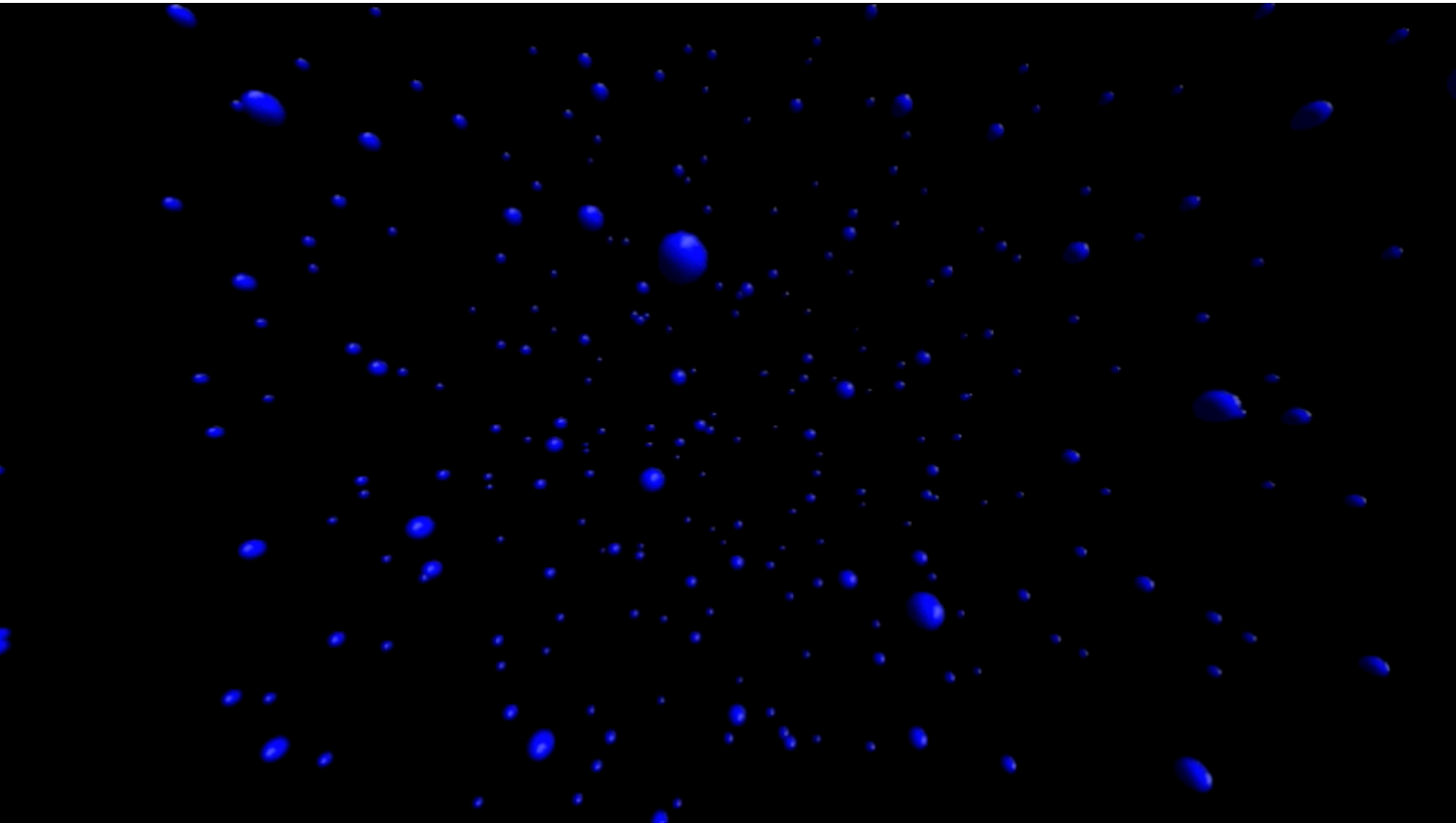
FACING REALITY: Degree distribution of real networks

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



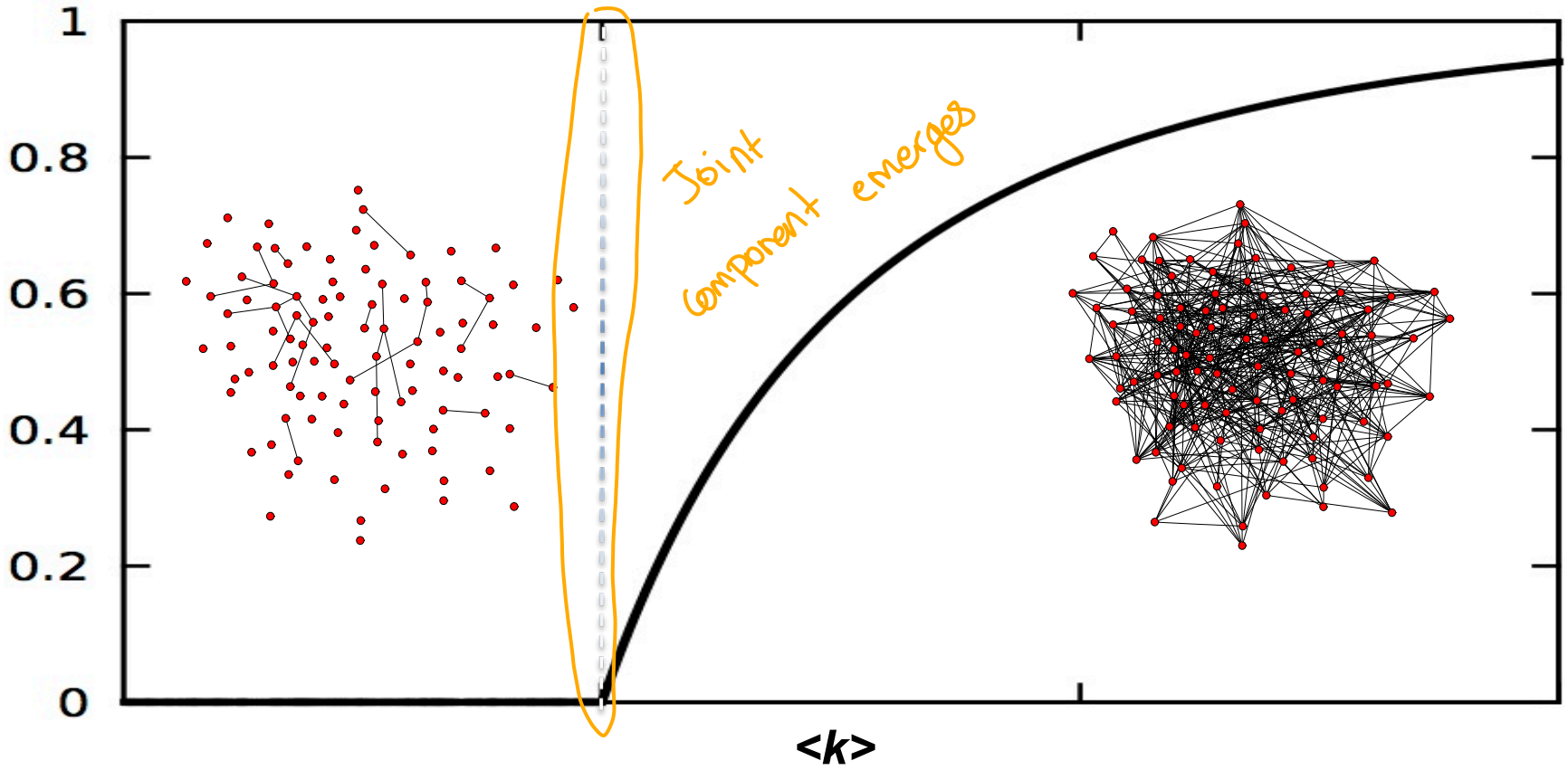
Section 6

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**.

$\langle k_c \rangle = 1$ (Erdös and Renyi, 1959)

exactly 1 → the fact transition.

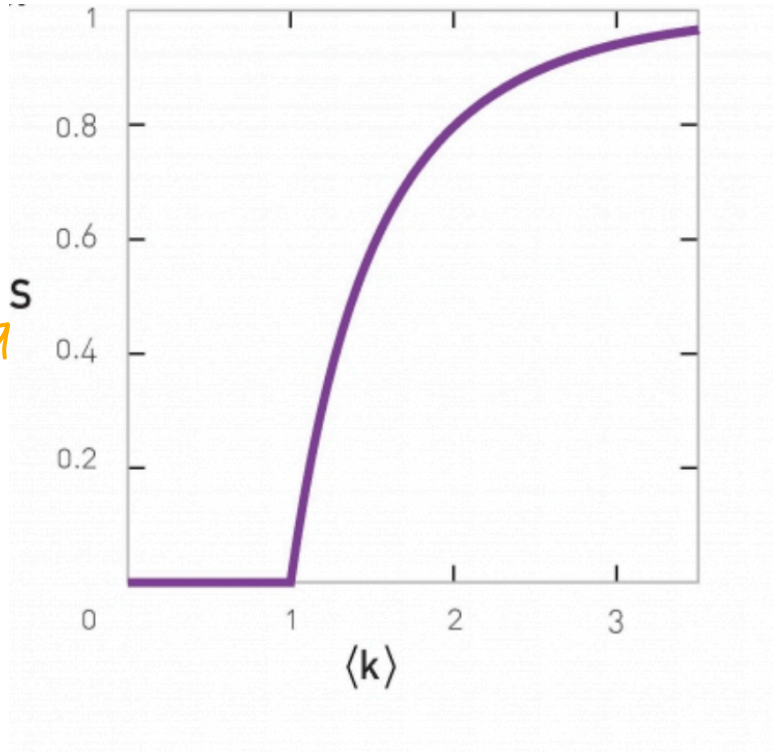
The fact that at least one link per node is *necessary* to have a giant component is not unexpected. Indeed, for a giant component to exist, each of its nodes must be linked to at least one other node.

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 a second order phase transition at <k>=1.

Section 3.4

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

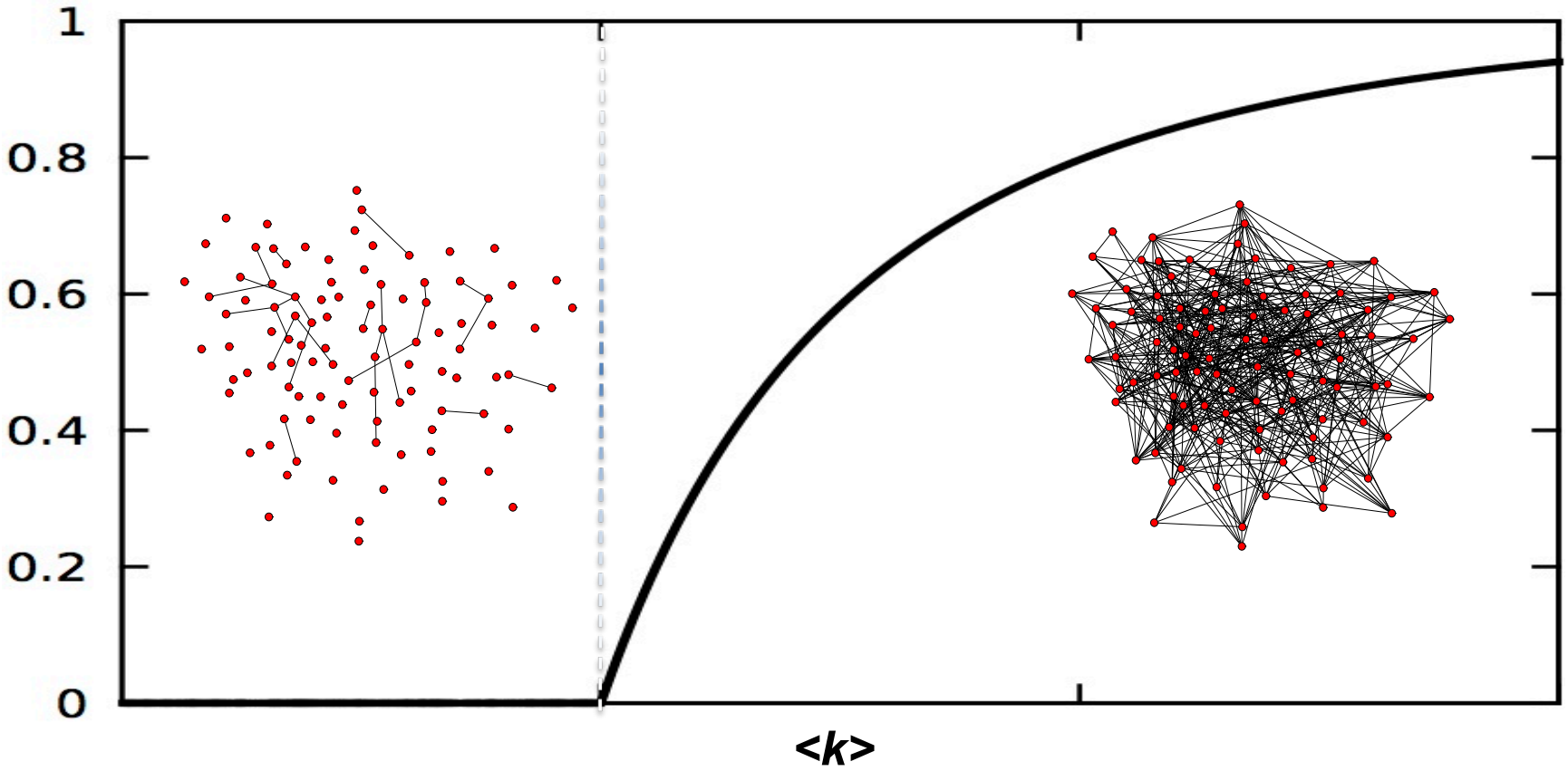


Fraction of nodes

S grows with ⟨k⟩

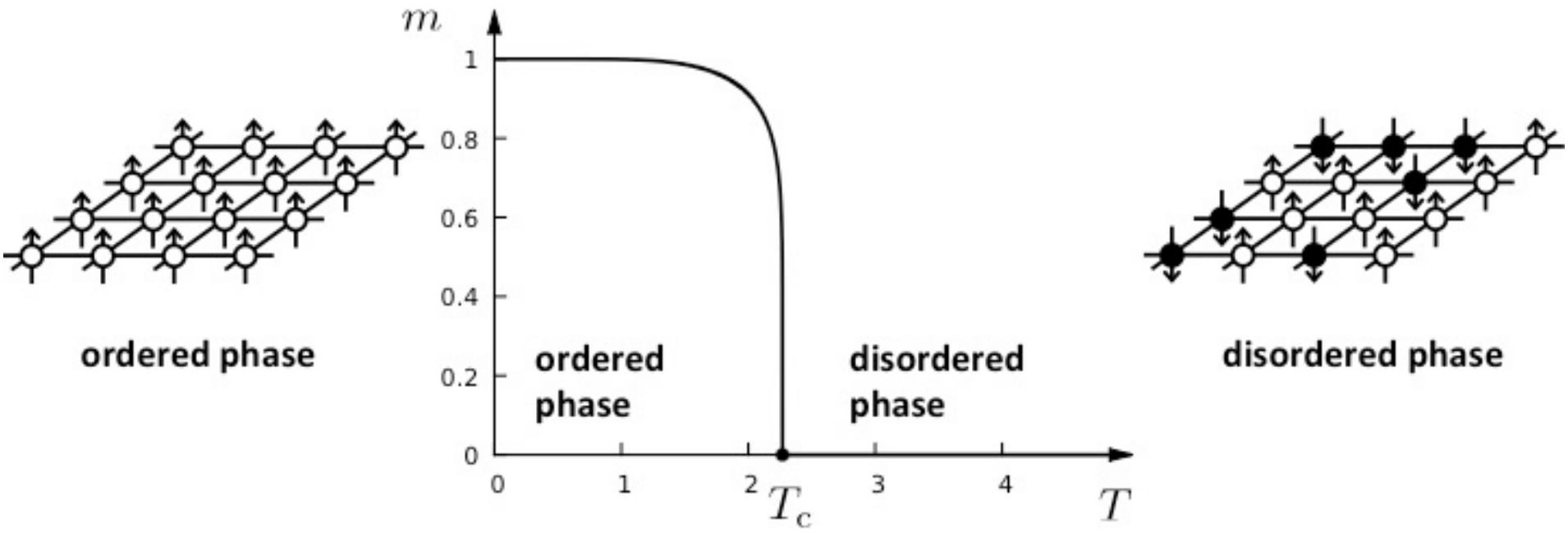
EVOLUTION OF A RANDOM NETWORK

disconnected nodes ➔ **NETWORK**.

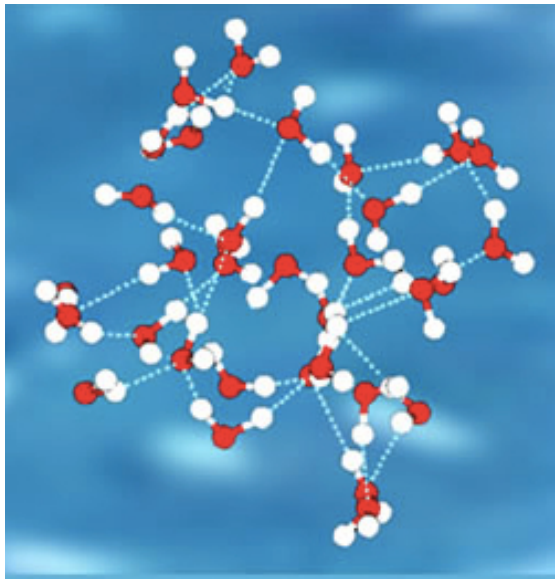


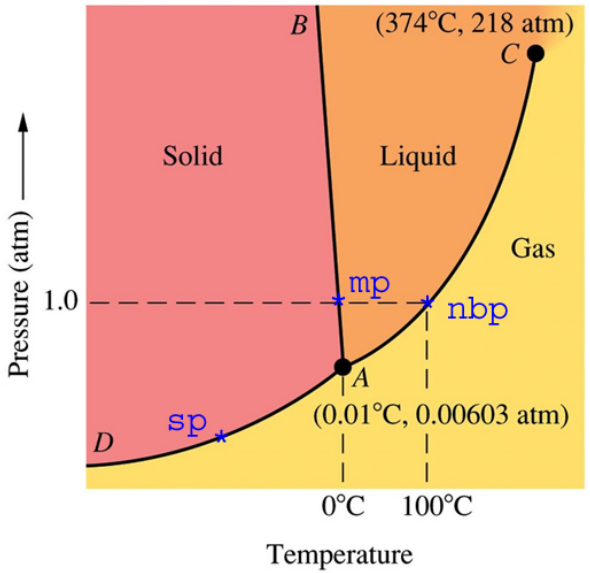
How does this transition happen?

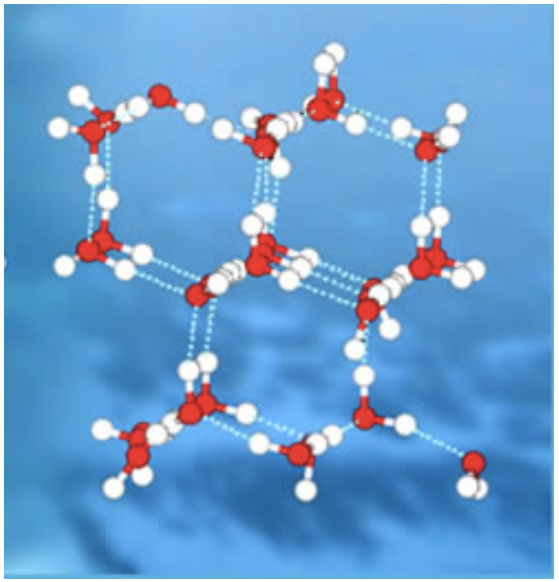
Phase transitions in complex systems I: Magnetism



Phase transitions in complex systems II: liquids



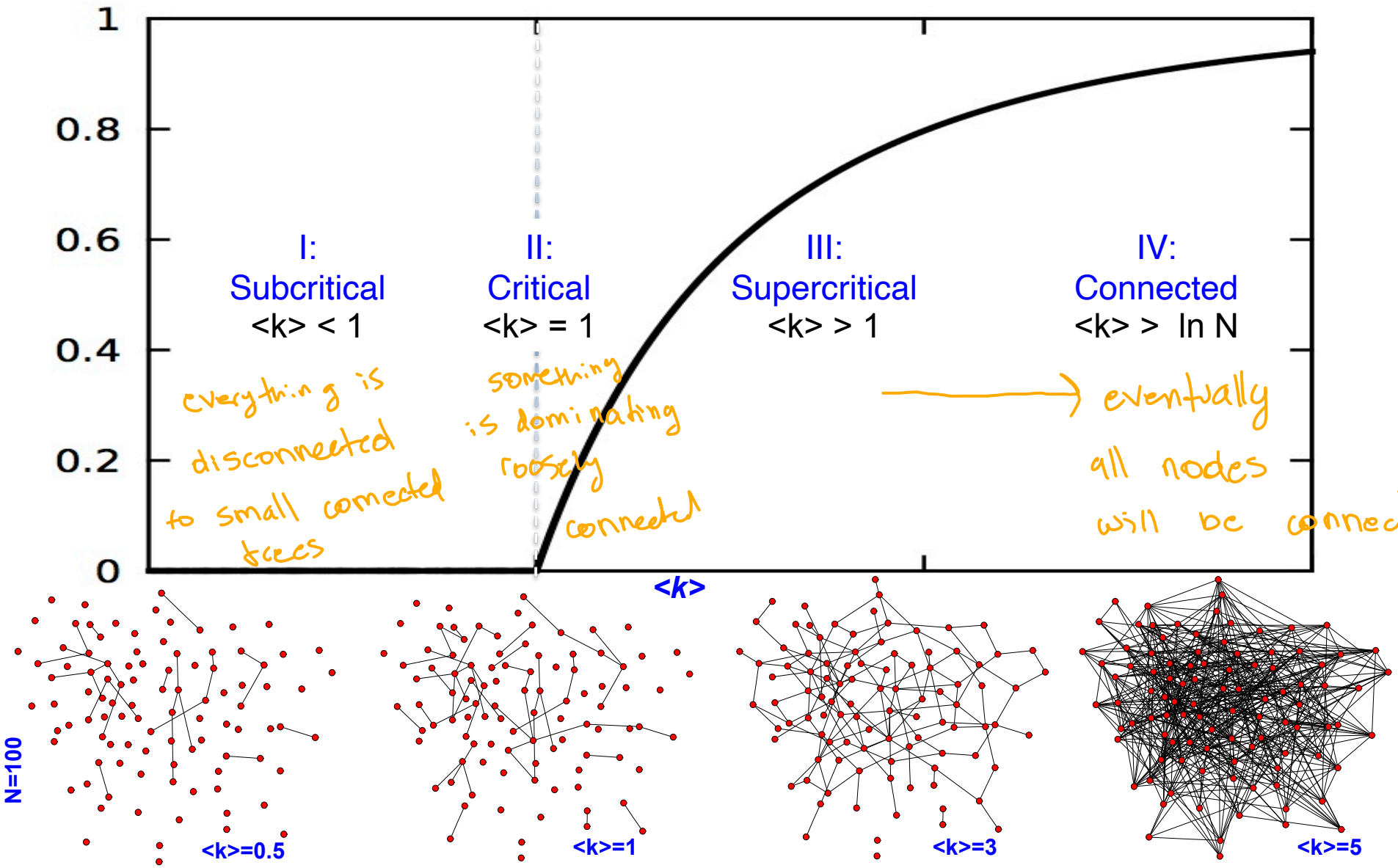


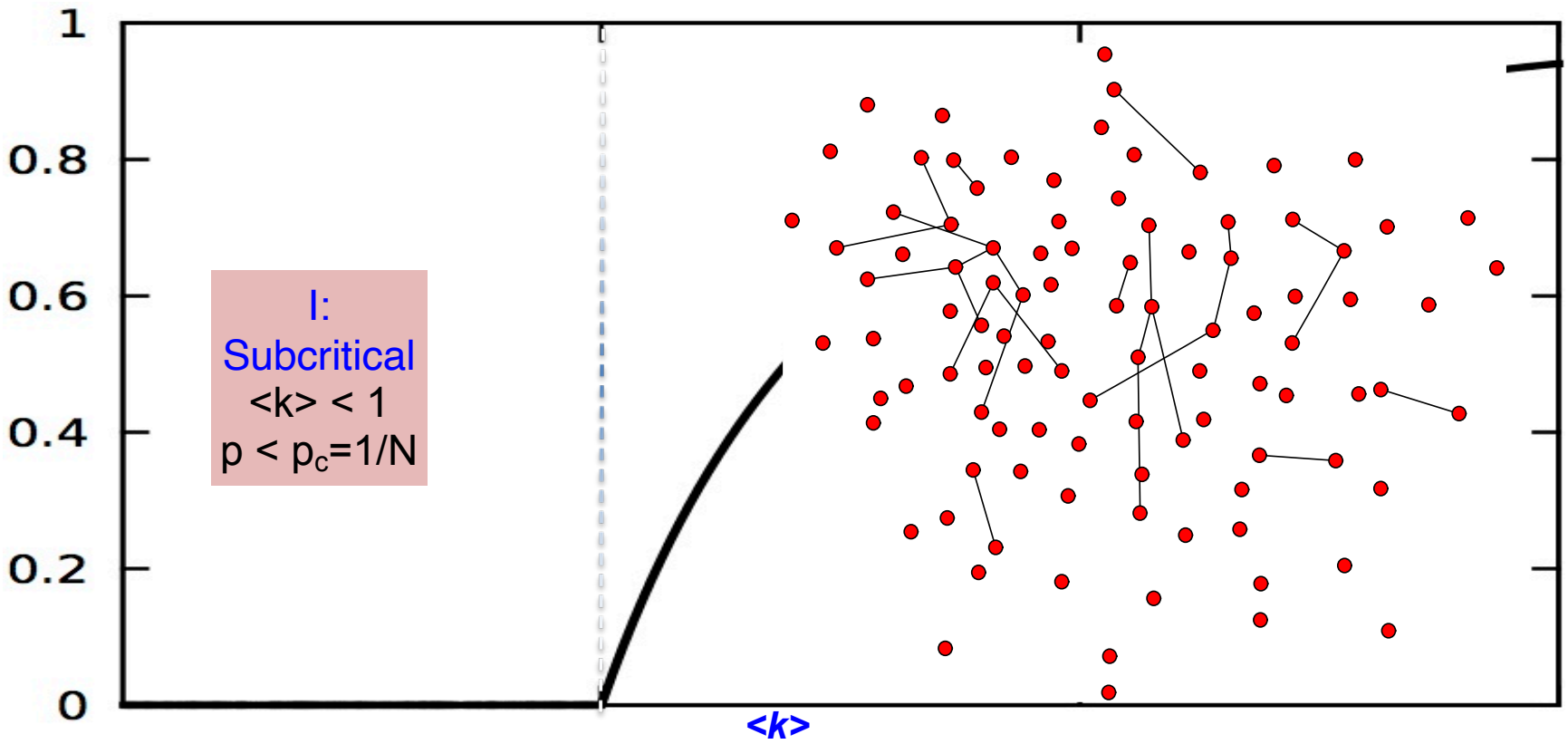


Water **Ice**

just like water where it suddenly changes

IMPORTANT

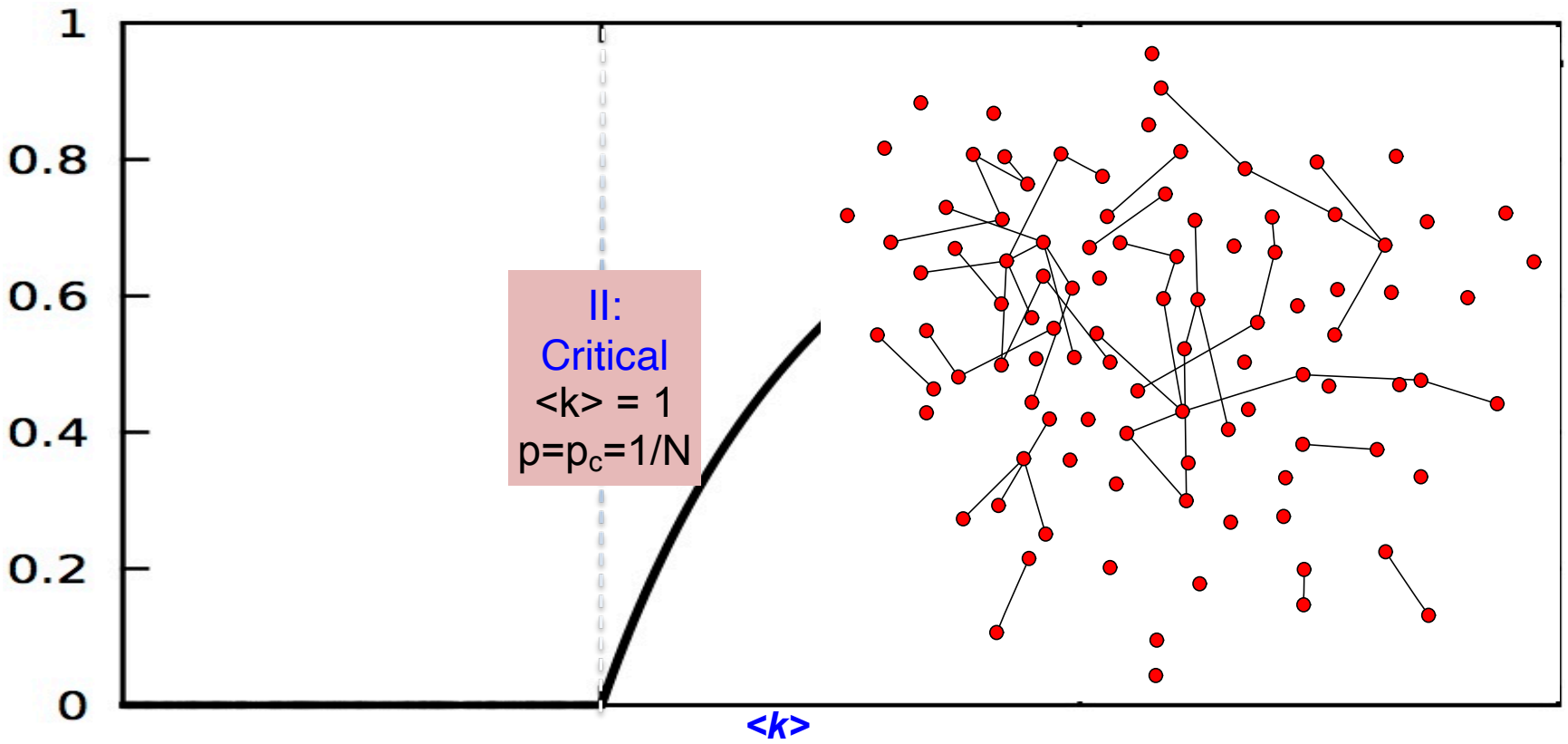

1
0.8
0.6
0.4
0.2
0
I:
Subcritical
<k> < 1
II:
Critical
<k> = 1
III:
Supercritical
<k> > 1
IV:
Connected
<k> > ln N
everything is disconnected to small comected trees
something is dominating loosely connected
eventually all nodes will be connected.
<k>
N=100
<k>=0.5
<k>=1
<k>=3
<k>=5



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, $N_G/N \sim N^{-1/3}$

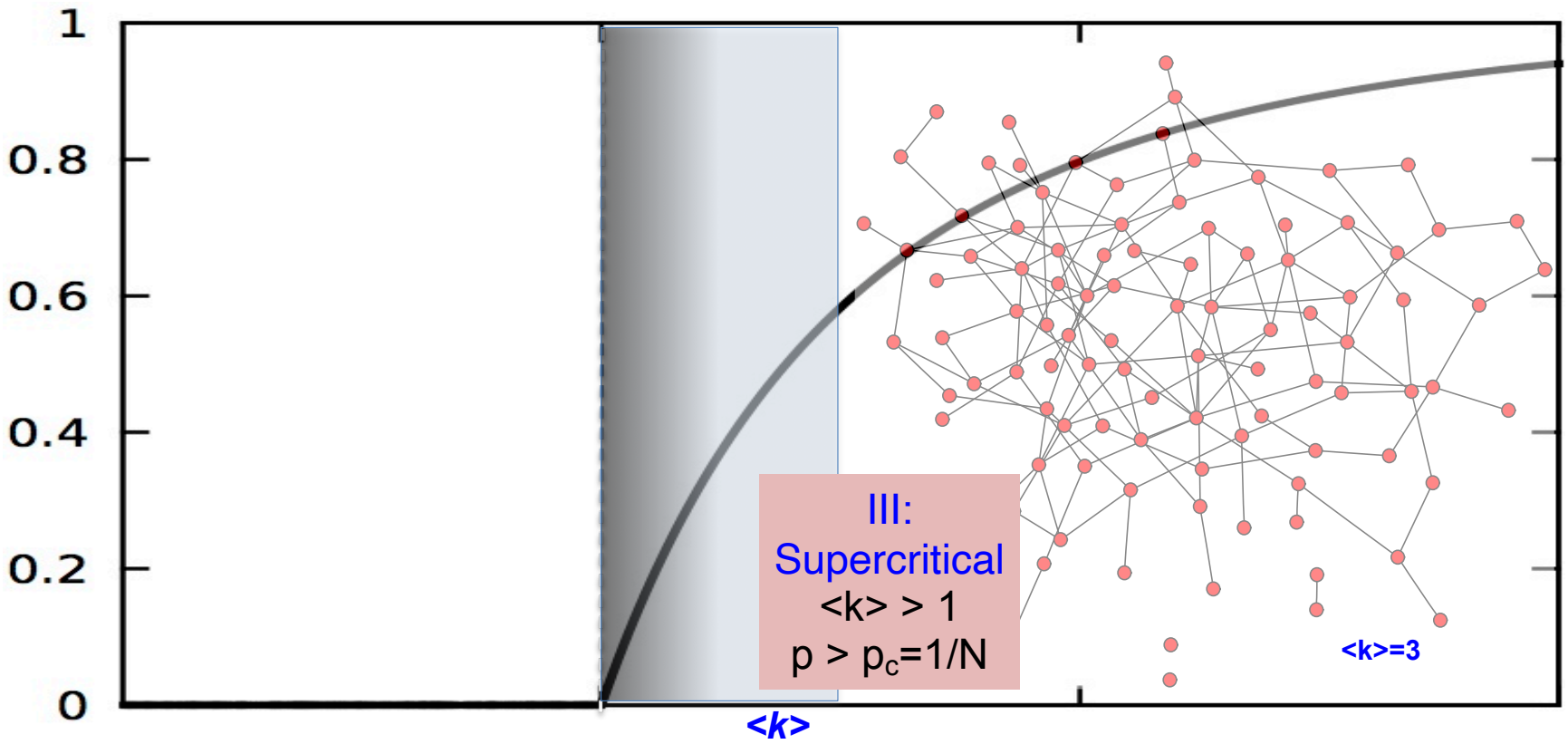
→Small components are trees, GC has loops.

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

A jump in the cluster size:

N=1,000 → ln N~ 6.9; $N^{2/3} \sim 95$

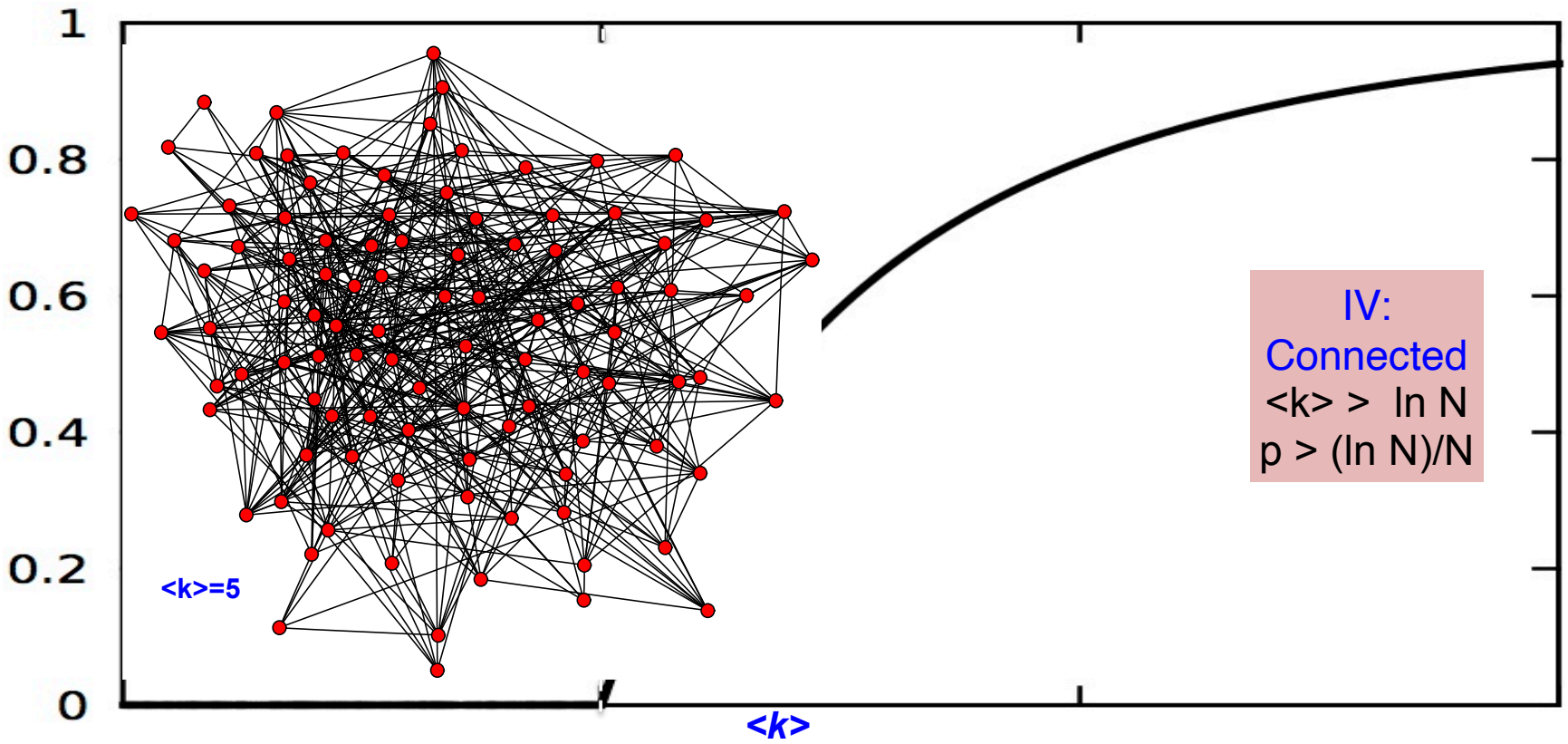
$N=7\ 10^9$ → ln N~ 22; $N^{2/3} \sim 3{,}659{,}250$



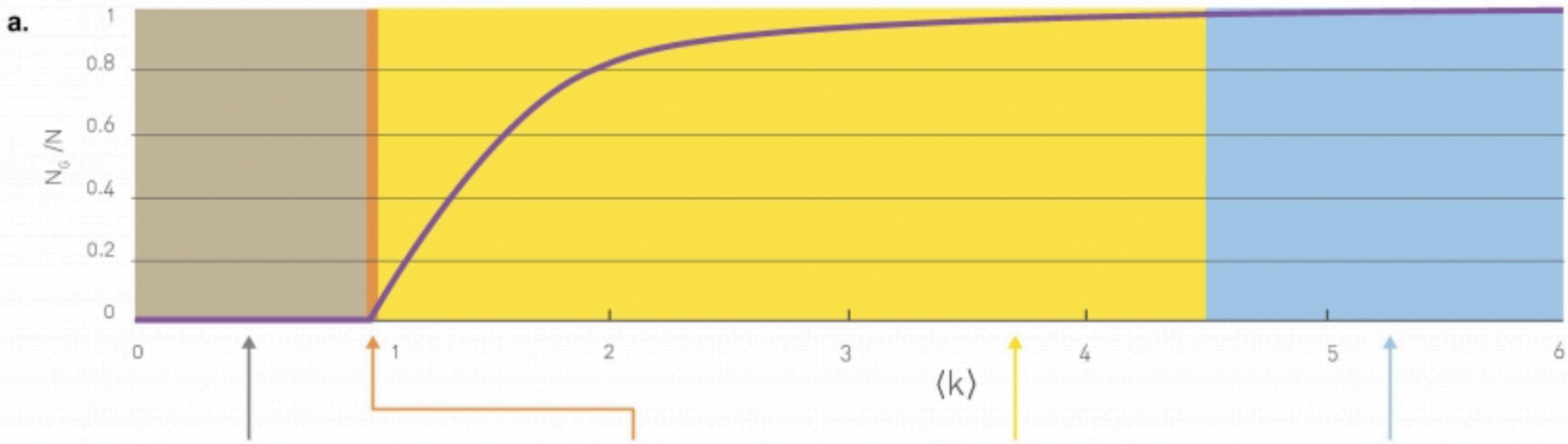
Unique giant component: $N_G \sim (p-p_c)N$

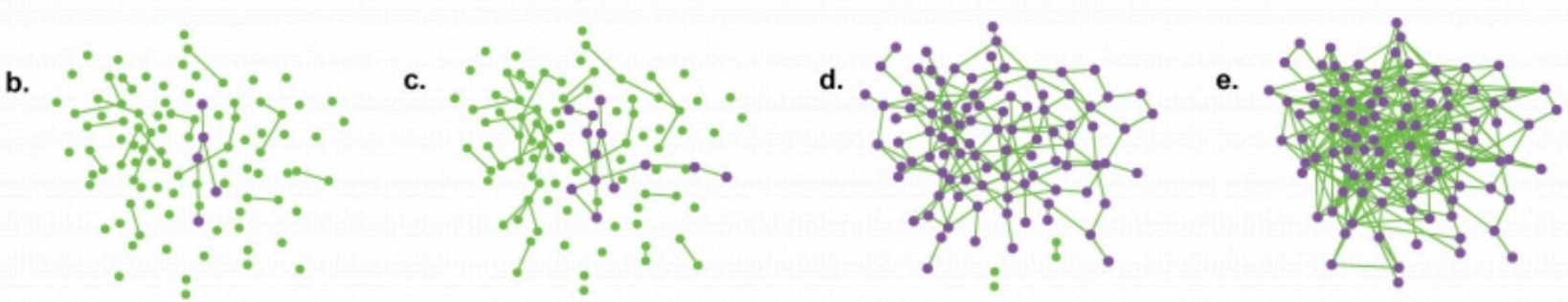
→GC has loops.

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



Only one cluster: N_G=N
→GC is dense.
Cluster size distribution: None





$\langle k \rangle < 1$ | $\langle k \rangle = 1$ | $\langle k \rangle > 1$ | $\langle k \rangle \geq \ln N$

(b) Subcritical Regime
- No giant component
- Cluster size distribution: $p_s \sim s^{-3/2} e^{-s/s_c}$
- 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

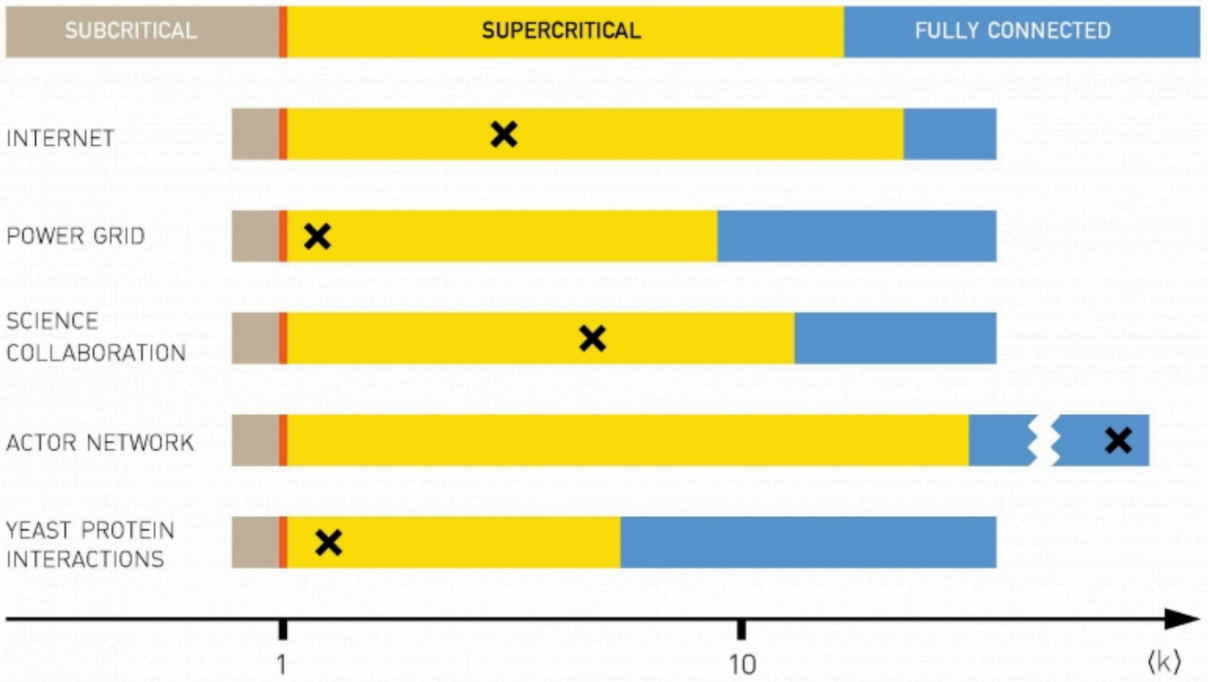
(d) Supercritical Regime
- Single giant component
- Cluster size distribution: $p_s \sim s^{-3/2} e^{-s/s_c}$
- Size of the giant component: $N_G \sim (p - p_c)N$
- The small clusters are trees
- Giant component has loops

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

Section 7

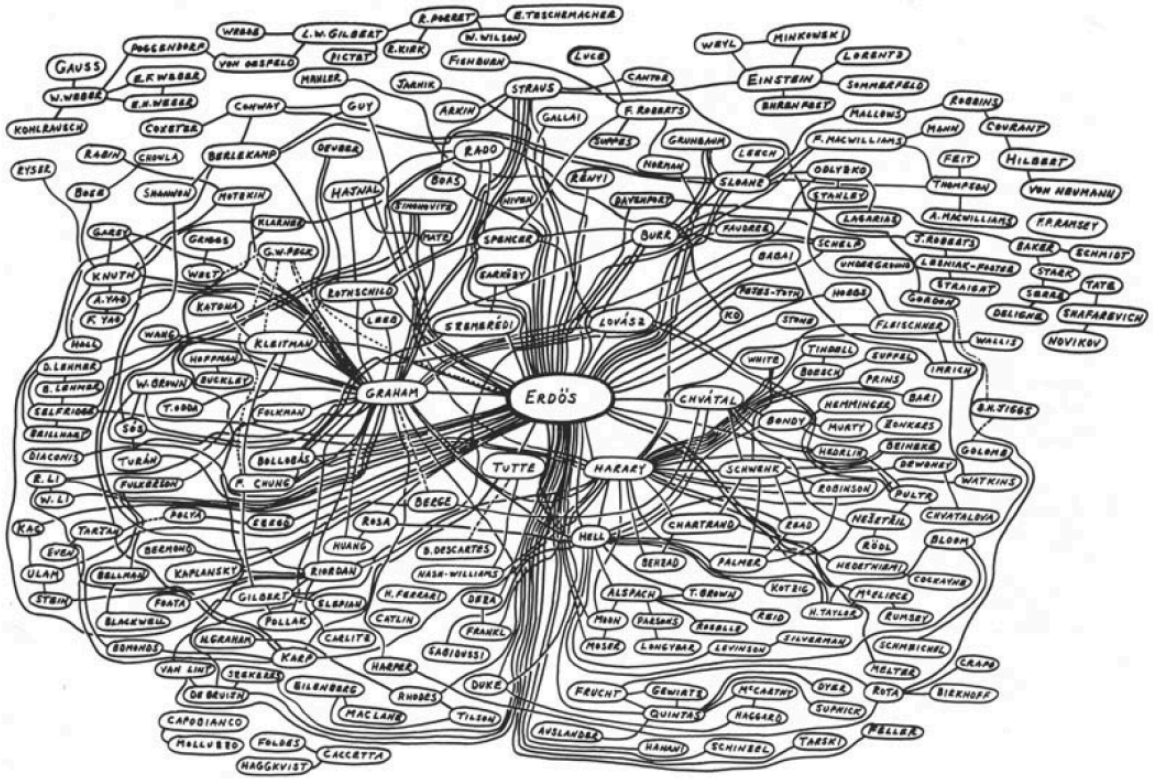
Real networks are supercritical

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

Erdös number



Erdös number

For those who have published a scientific paper: what is your Erdös number?

For those that have not published a scientific paper: what are some of your profs' Erdös numbers? Your friends'?

How did you find it?

Erdös number:

0 Paul Erdös
1 Robert James McEliece
2 Jonathan Harel
3 Christof Koch
4 Albert-László Barabási
5 Emma Towlson

Bacon number:

0 Kevin Bacon
1 James Cromwell
2 Charlotte Le Bon
3 Xavier Lafitte
4 Emma Towlson

https://en.wikipedia.org/wiki/List_of_people_by_Erd%C5%91s_number