# Random Graphs

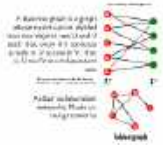
hari sundaram

hs1@illinois.edu

adapted from Jure Leskovec's slides

# Introduction

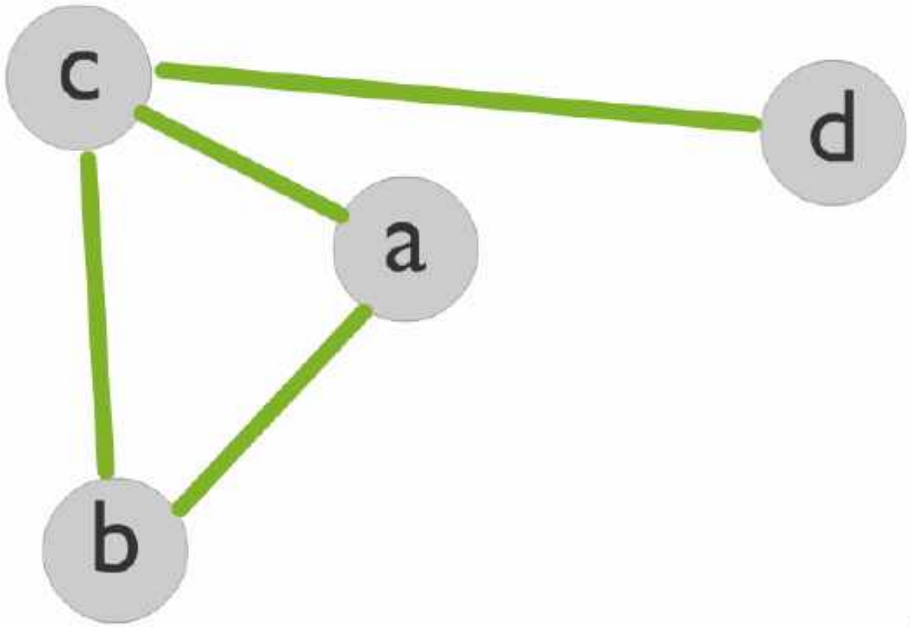






# Adjacency Matrix




undirected graph

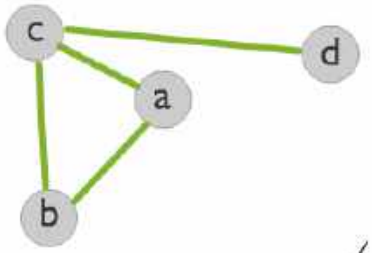
$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

symmetric

directed graphs
do not have
symmetric
adjacency
matrices

What are examples of directed and undirected graphs in real-world datasets?

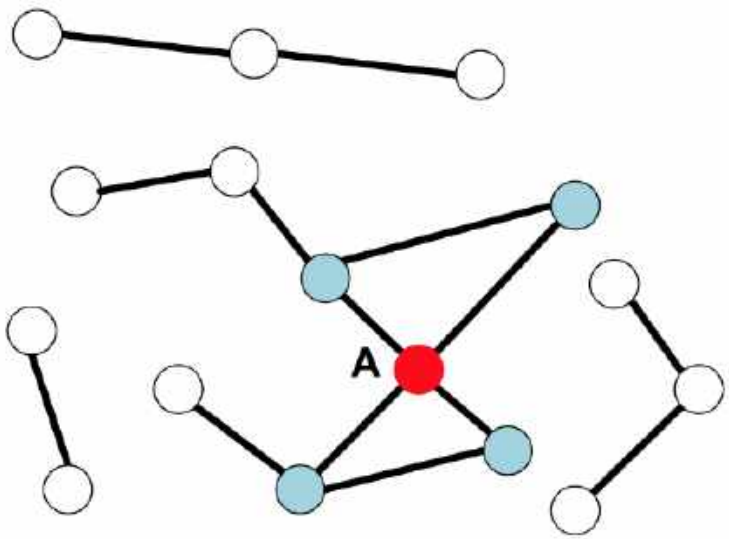
## Adjacency Matrix




undirected graph

$$A = \underset{\text{symmetric}}{} \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

# Node Degrees



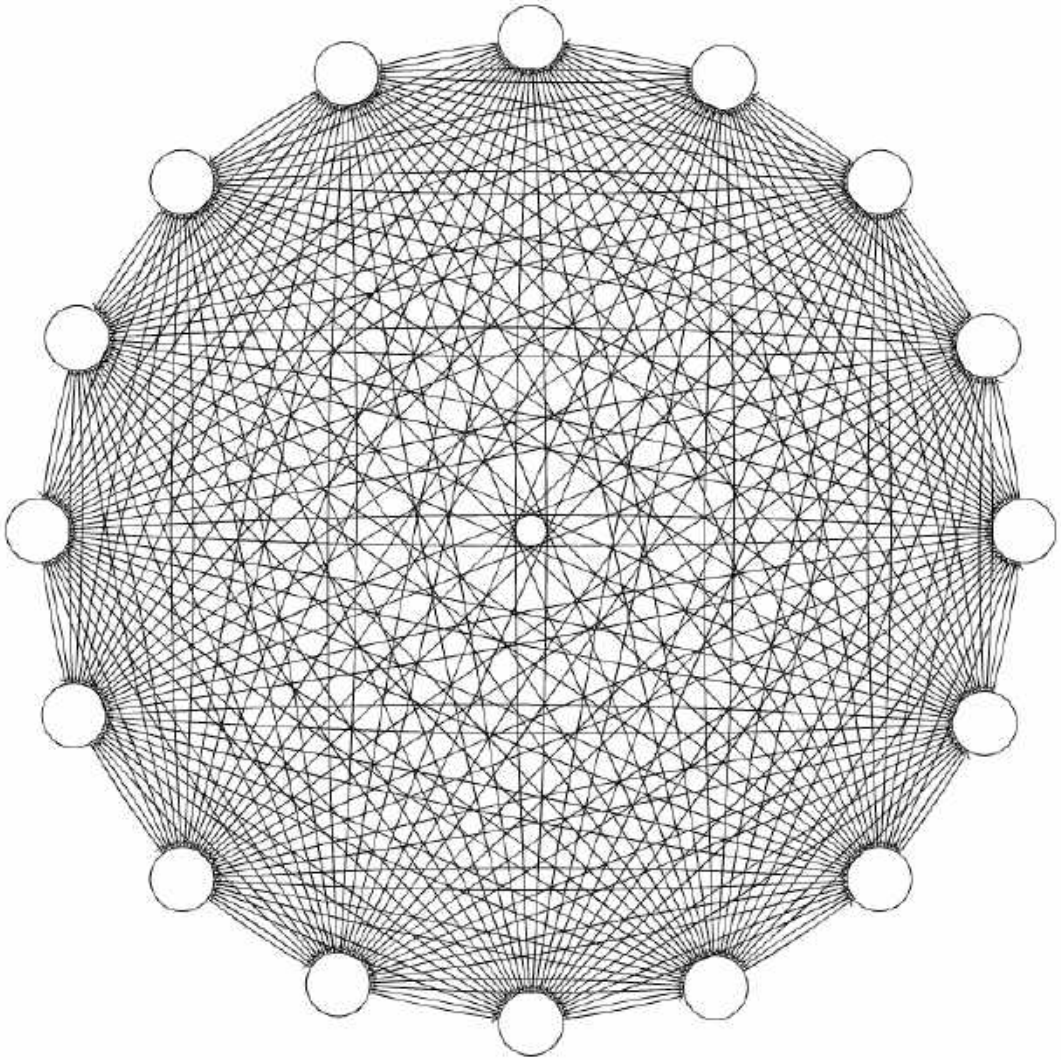

$$\bar{k} = \langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2E}{N}$$

degree of node i

$$k_A = 4$$

We can compute similar quantities for directed graphs too, except that we will have two numbers.

# Complete Graph



$$E_{\max} = \binom{N}{2} = \frac{n(n-1)}{2}$$

degree = n-1

however, most real-world graphs are strikingly different from complete graphs

| Network | Nodes | Average degree |
|---|---|---|
| WWW (Stanford–Berkeley): | 319,717 | 9.65 |
| Social networks (LinkedIn): | 6,946,668 | 8.87 |
| Communication (MSN IM): | 242,720,596 | 11.1 |
| Coauthorships (DBLP): | 317,080 | 6.62 |
| Roads (California): | 1,957,027 | 2.82 |
| Proteins (S. Cerevisiae): | 1,870 | 2.39 |

$$\text{matrix density} \equiv \frac{E}{N^2}$$

$$\text{WWW} : 1.51 \times 10^{-5}$$

$$\text{MSM IM} : 2.27 \times 10^{-8}$$

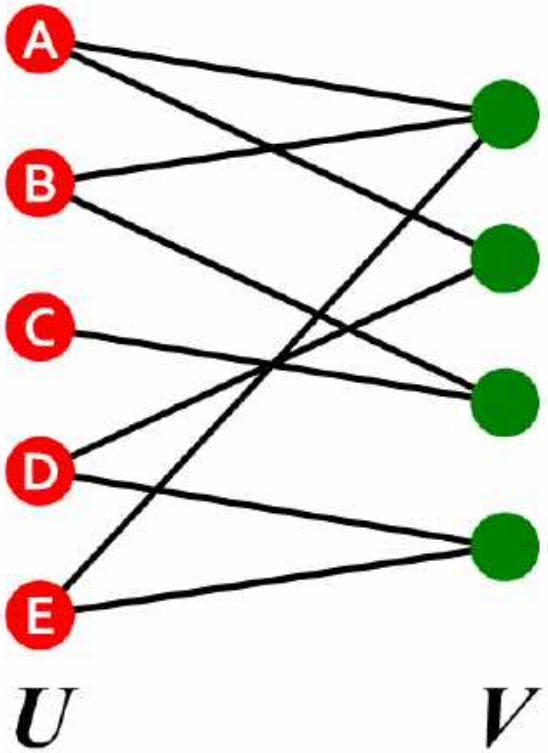
most adjacency matrices are filled with zeros

Most networks are sparse!

$$\bar{k} \ll n - 1$$

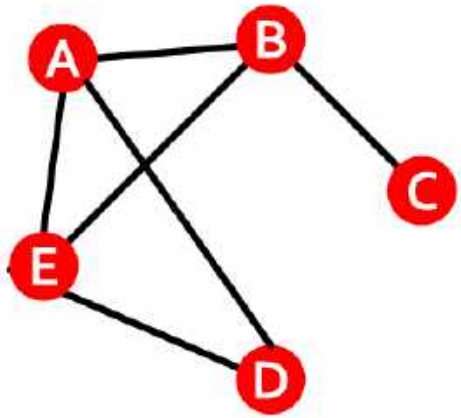
actors-movies; author-papers etc.

A bipartite graph 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




That is, it is a set of vertices S such that for every two vertices in S there is no edge connecting the two.

Author collaboration networks; Movie co rating networks

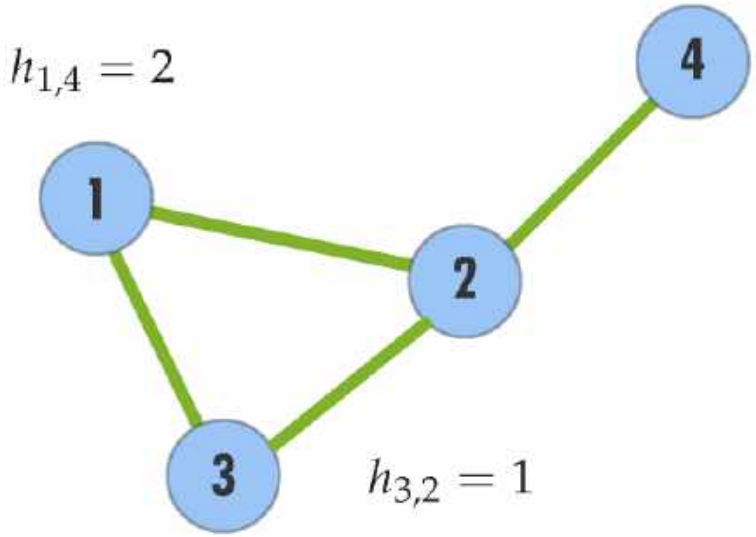



folded graph

# Distance

If the two nodes are disconnected, the distance is usually defined as infinite

Diameter: the maximum (shortest path) distance between any pair of nodes in a graph




**BFS**

Distance (shortest path, geodesic) between a pair of nodes is defined as the number of edges along the shortest path connecting the nodes

$$\bar{h} = \frac{1}{2E_{\max}} \sum_{i,j \neq i} h_{i,j}$$

Average path length for a connected graph (component) or a strongly connected component of a directed graph

Many times we compute the average only over the connected pairs of nodes (we ignore "infinite" length paths)

There are three useful ways to characterize a graph

# Degree Distribution

# nodes with degree k

$$P(k) = \frac{N_k}{N}$$

# 2 Average path length

$$\bar{h} = \frac{1}{2E_{\max}} \sum_{i,j \neq i} h_{i,j}$$

distance between i and j

$$\binom{N}{2}$$

# 3 Clustering coefficient

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

# edges amongst i's neighbors

degree of node i

There are three useful ways to characterize a graph

1

Degree Distribution

# nodes with degree k

$$P(k) = \frac{N_k}{N}$$

2

Average path length

distance between i and j

$$\bar{h} = \frac{1}{2E_{\max}} \sum_{i,j \neq i} h_{i,j}$$

$\binom{N}{2}$

3

Clustering coefficient

# edges in neighbor. i's neighbors

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

degree of node i


# MSN Messenger Dataset

## June, 2006

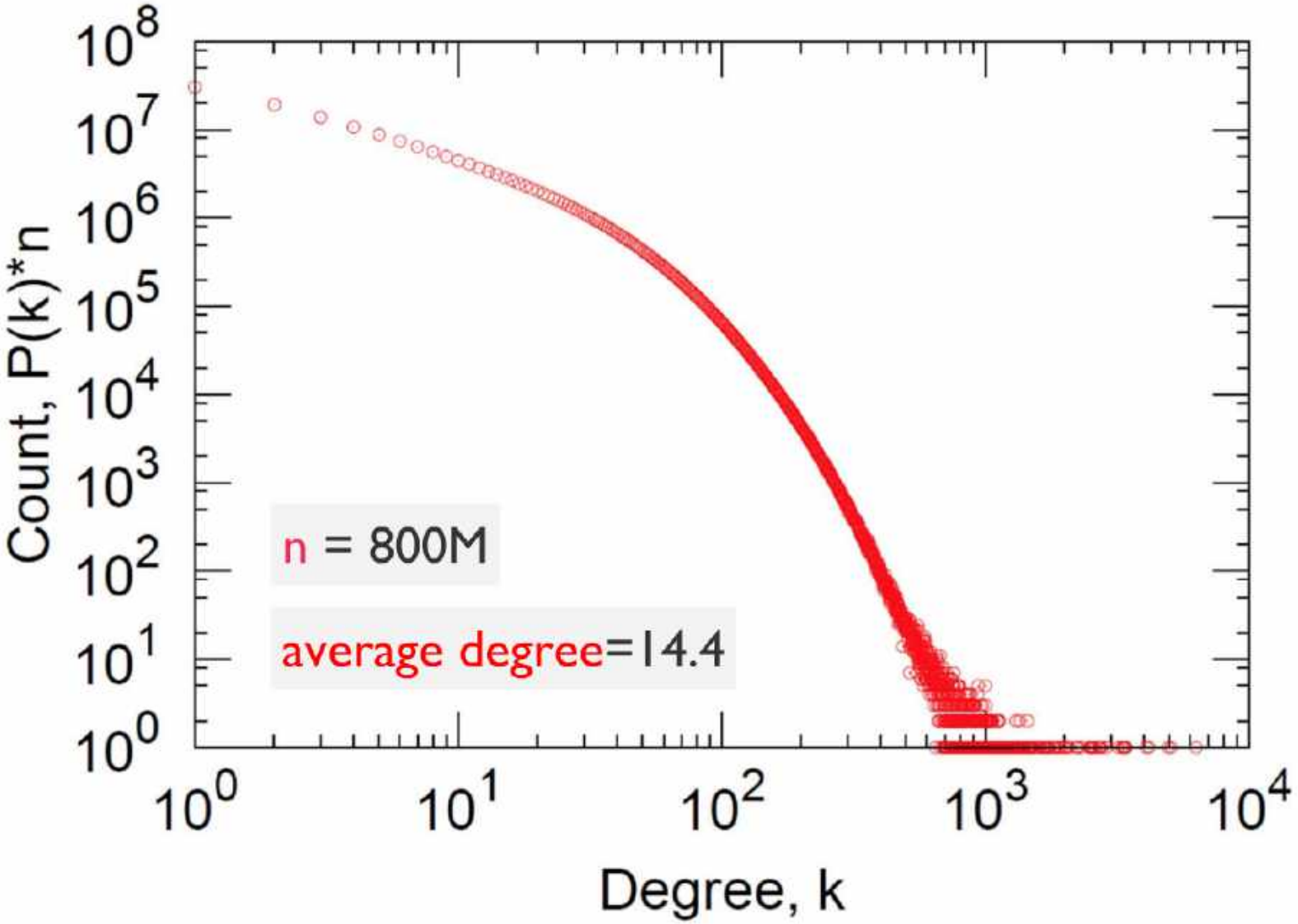
Two individuals are connected if they exchange a message within the last month

- 180M nodes
- 1.3B edges

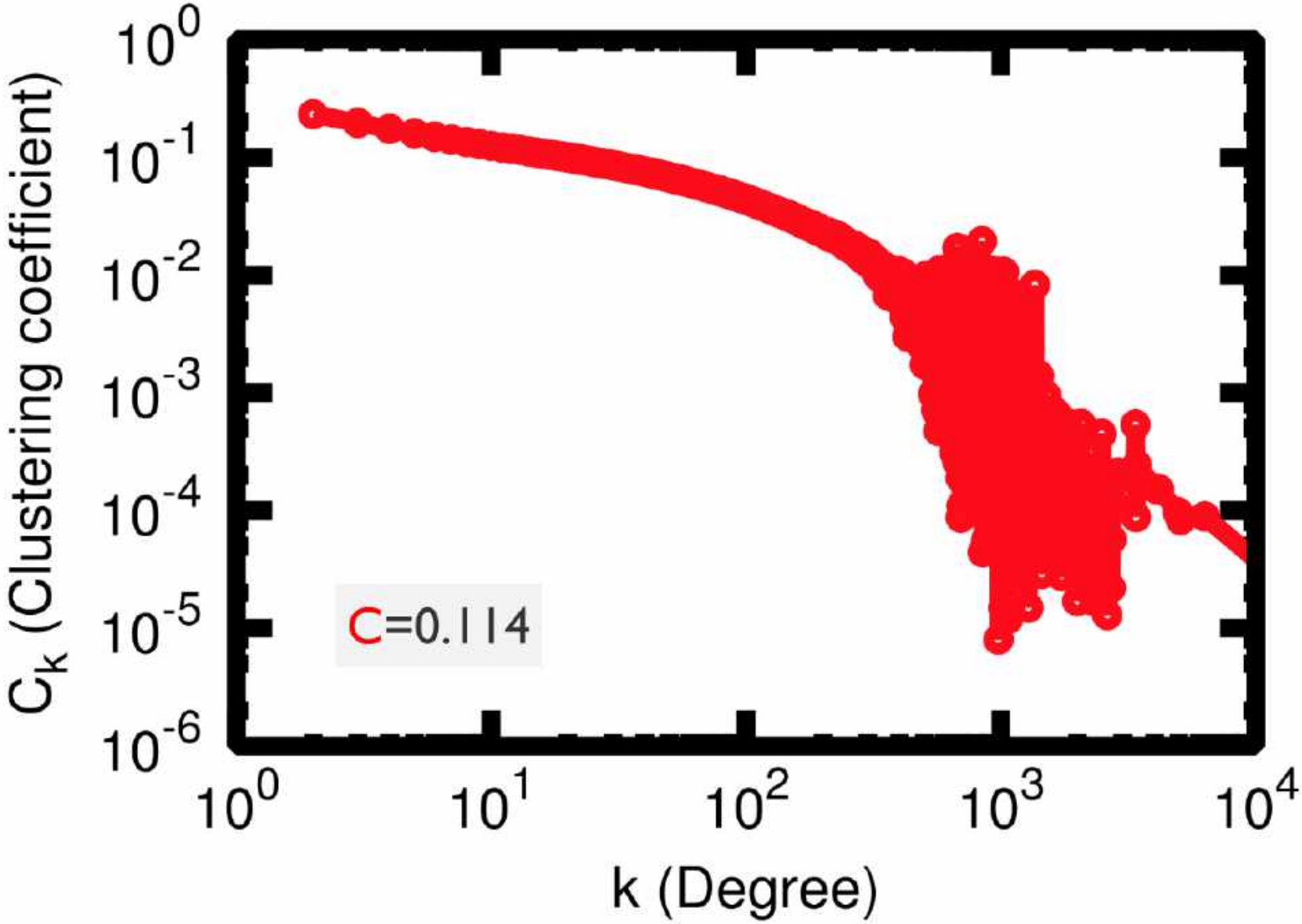
- 4.5Tb / month
- 245 million users logged in
- 180 million users engaged in conversations
- More than 30 billion conversations
- More than 255 billion exchanged messages

J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In Proceedings of the 17th international conference on World Wide Web, pages 915–924. ACM, 2008.

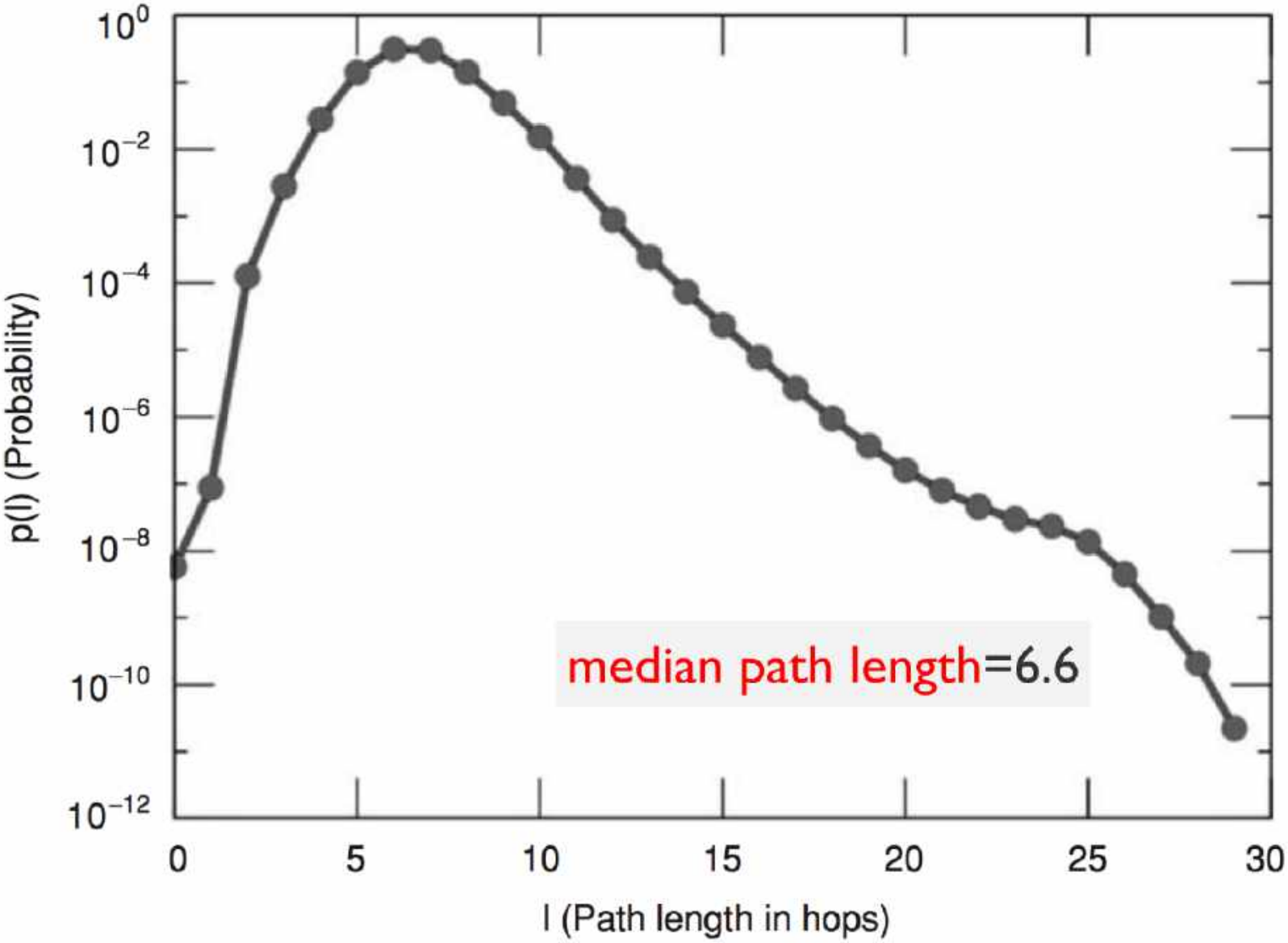
# Degree distribution



# Clustering Coefficient



# Path Length


$10^{0}$
$10^{-2}$
$10^{-4}$
$10^{-6}$
$10^{-8}$
$10^{-10}$
$10^{-12}$
p(l) (Probability)
0
5
10
15
20
25
30
l (Path length in hops)
median path length=6.6

# MSN Messenger Dataset

**June, 2006**

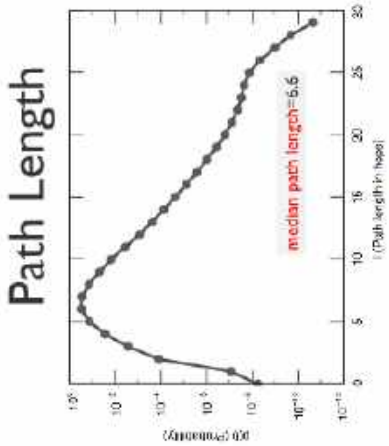
Two individuals are connected if they exchange a message within the last month

- 180M nodes
- 1.3B edges

- 4.5Tb / month
- 245 million users logged in
- 180 million users engaged in conversations
- More than 30 billion conversations
- More than 255 billion exchanged messages

J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In Proceedings of the 17th international conference on World Wide Web, pages 915–924. ACM, 2008.

## Path Length




There are three useful ways to characterize a graph

1 Degree Distribution

$$P(k) = \frac{N_k}{N}$$

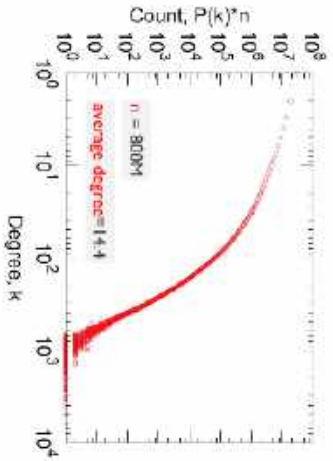
2 Average path length

$$\bar{h} = \frac{1}{2E_{max}} \sum_{i,j \neq i} h_{ij}$$

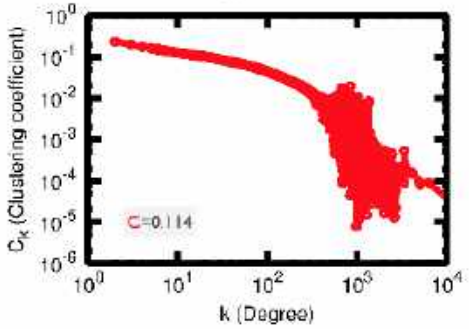
3 Clustering coefficient

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

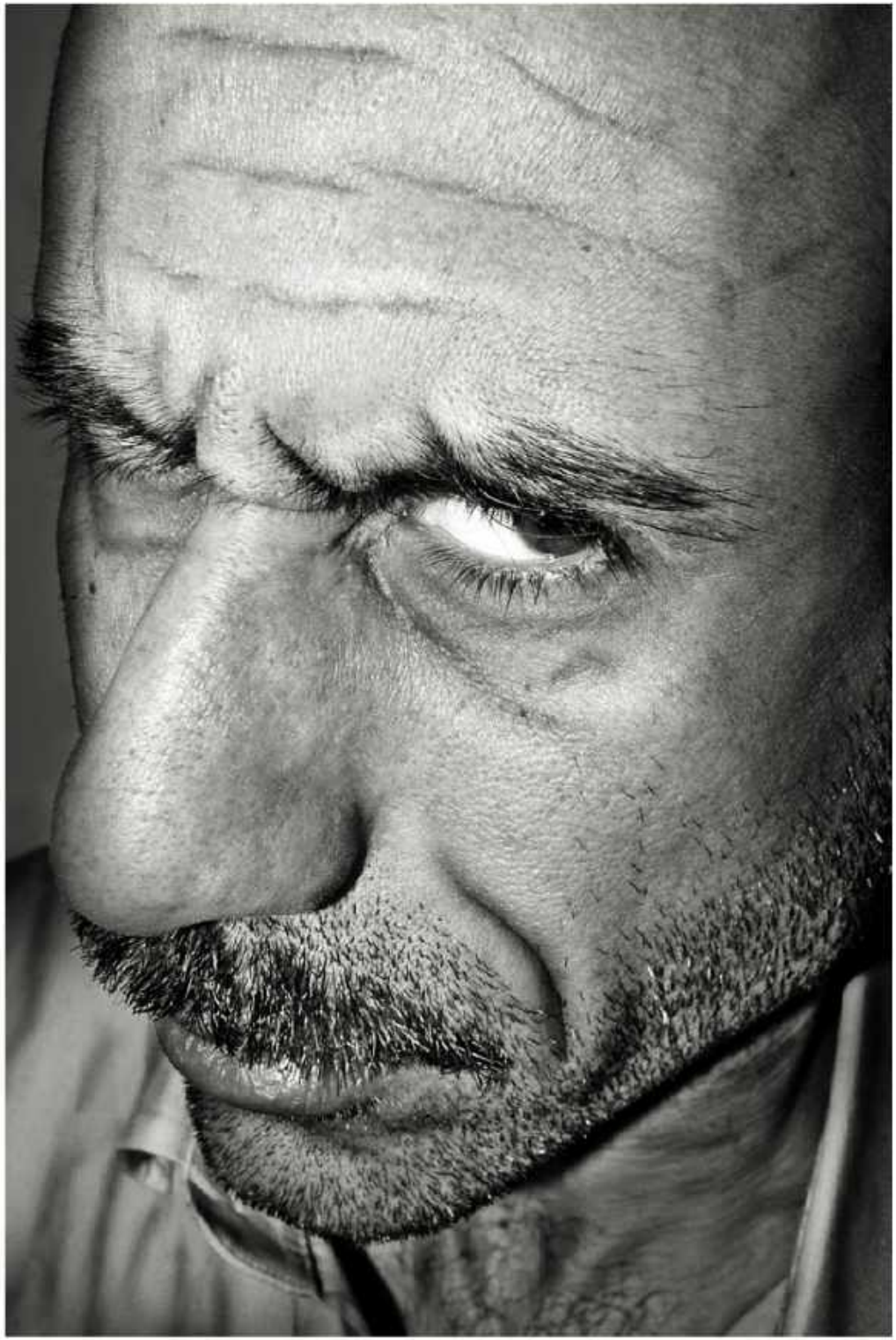
## Degree distribution



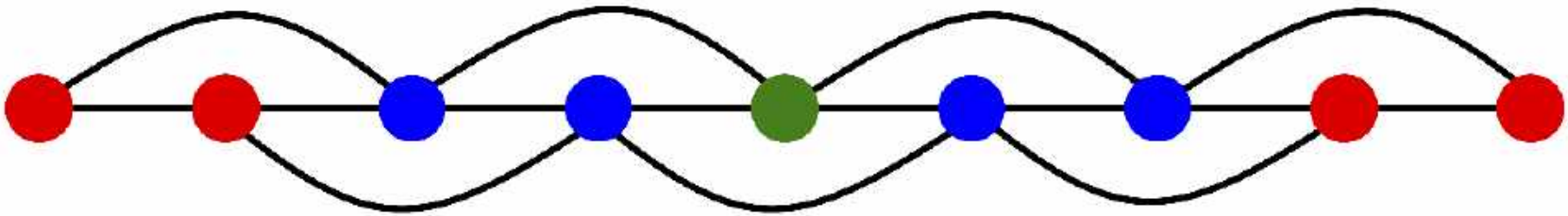

## Clustering Coefficient



Are these numbers surprising? Unexpected?



# Is the MSN network like a "Chain"?



$$P(k) = \delta(k-4)$$

P(k): the degree distribution of a node

$$C = \frac{1}{2}$$

$$\bar{h} = O(N)$$

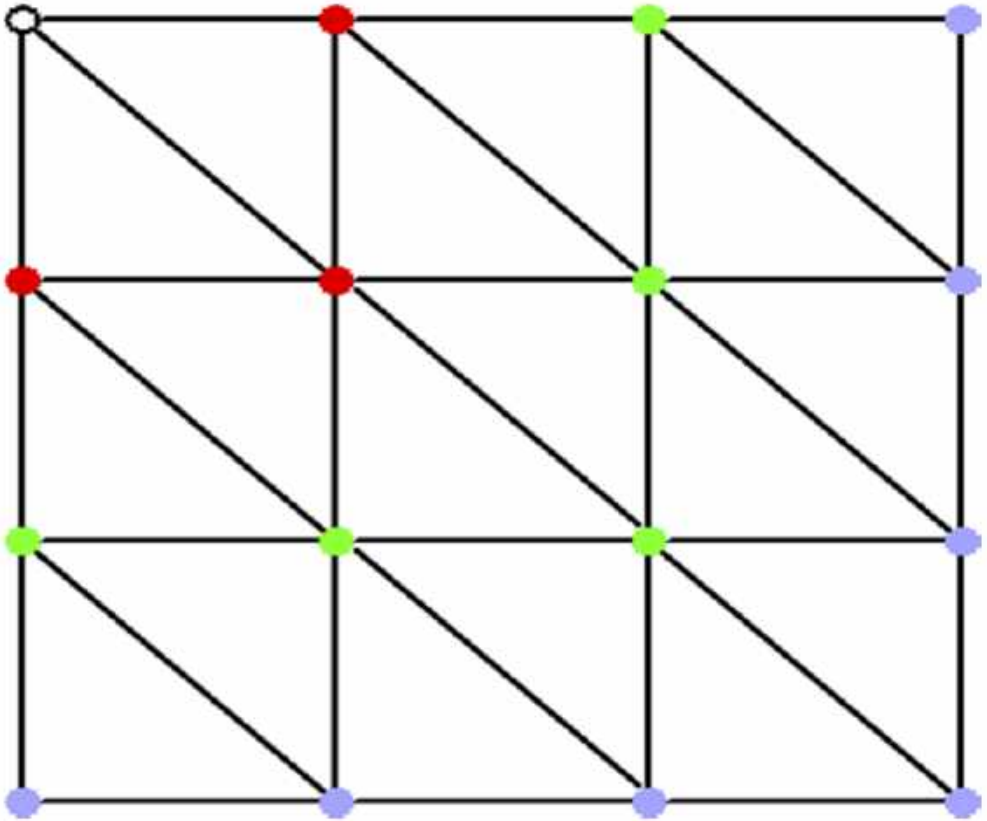
Constant degree; constant average clustering coefficient; average path length linear in N

# Is the MSN network like a "Grid"?

$$P(k) = \delta(k-6)$$

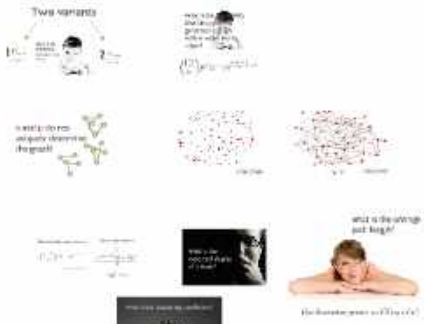
$$C = \frac{6}{15}$$

$$\bar{h} = O(\sqrt{N})$$



Constant degree, constant clustering coefficient

# Erdös Rényi models








to what degree is a network random?

# Two variants

# $G_{n,p}$

n nodes, edges are i.i.d
with probability p

# 2 $G_{n,m}$

n nodes, m edges are picked at random

# Two variants

1 $G_{n,p}$

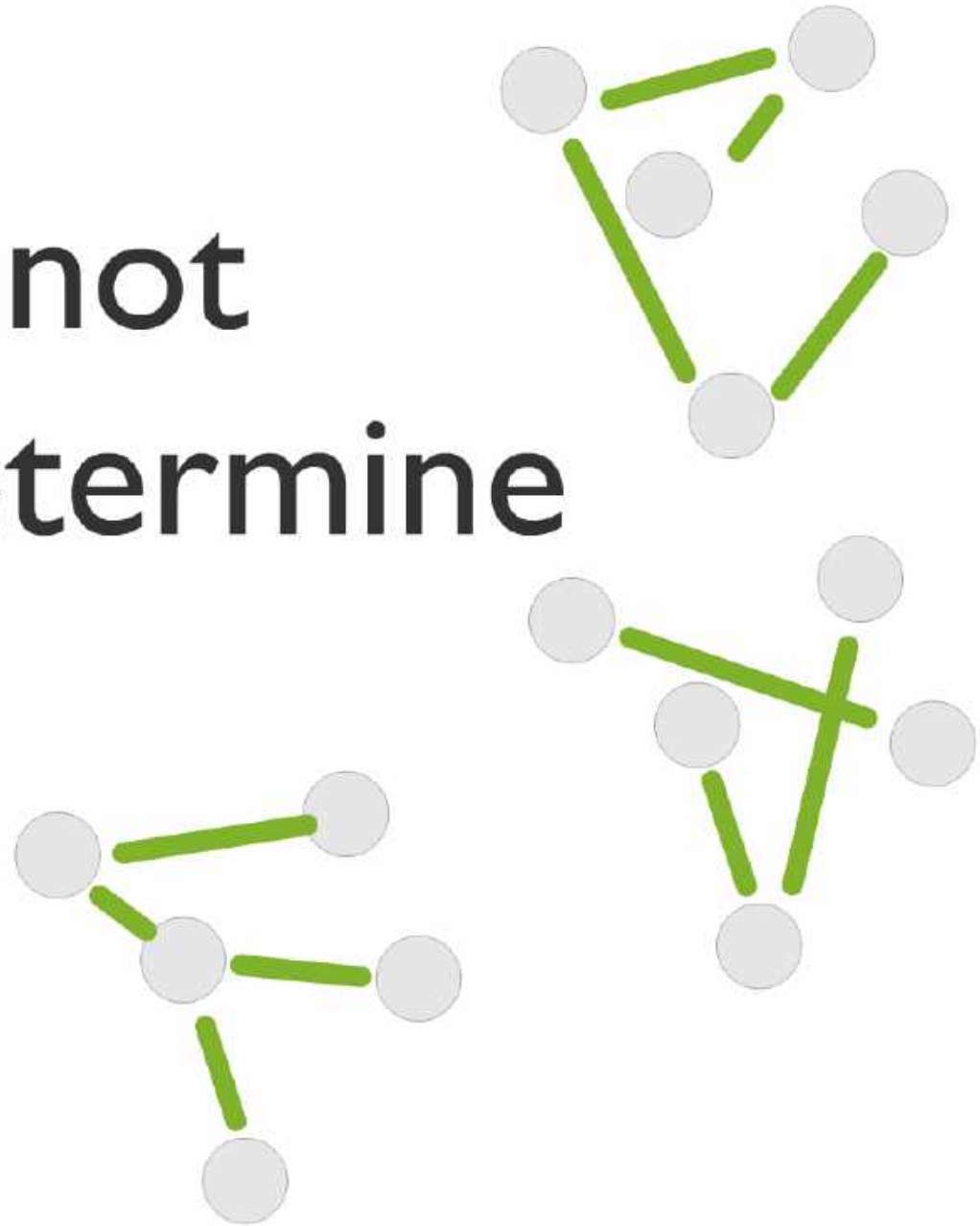
n nodes, edges are i.i.d with probability p

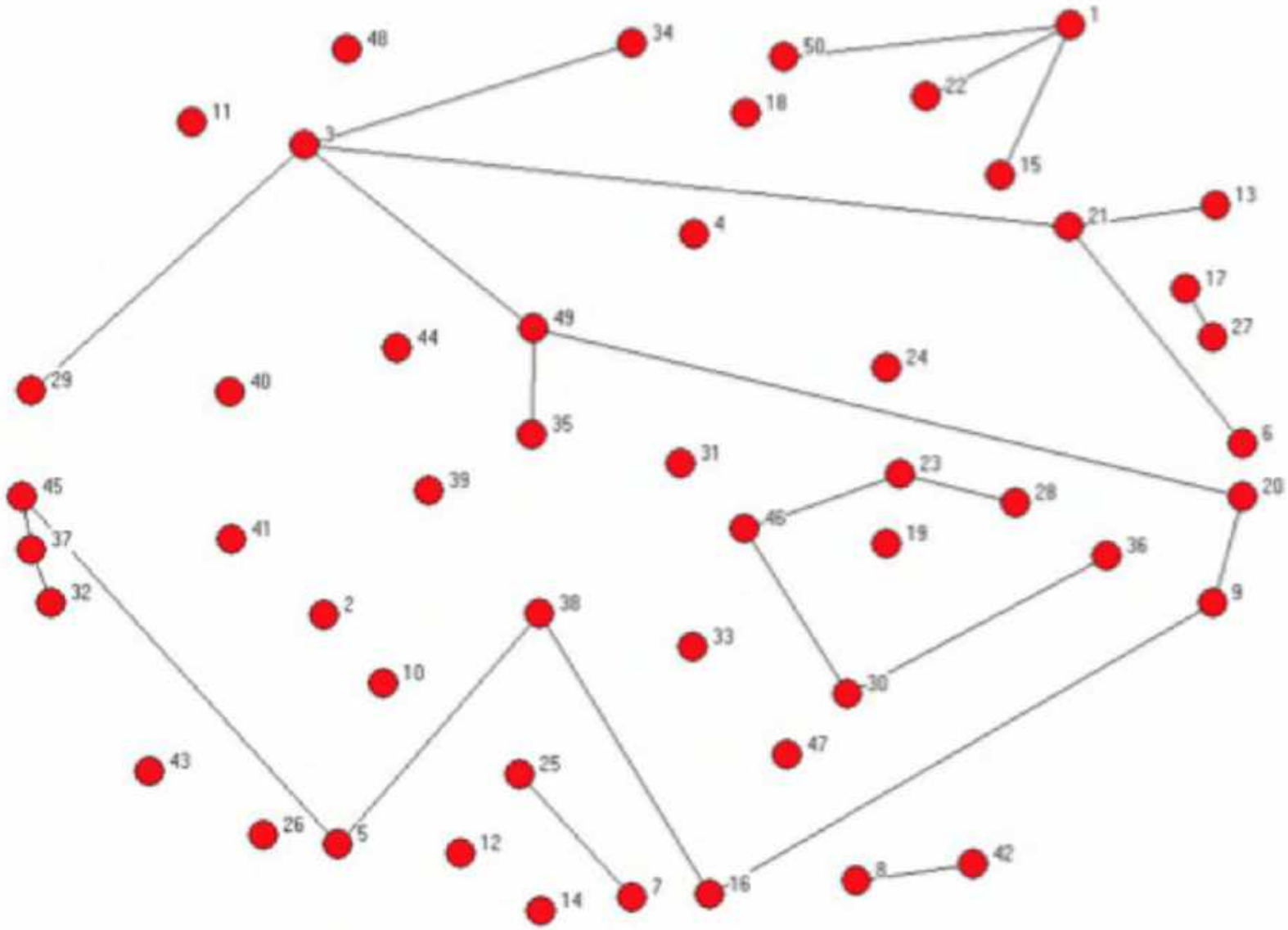
what is the difference between the two?

2 $G_{n,m}$

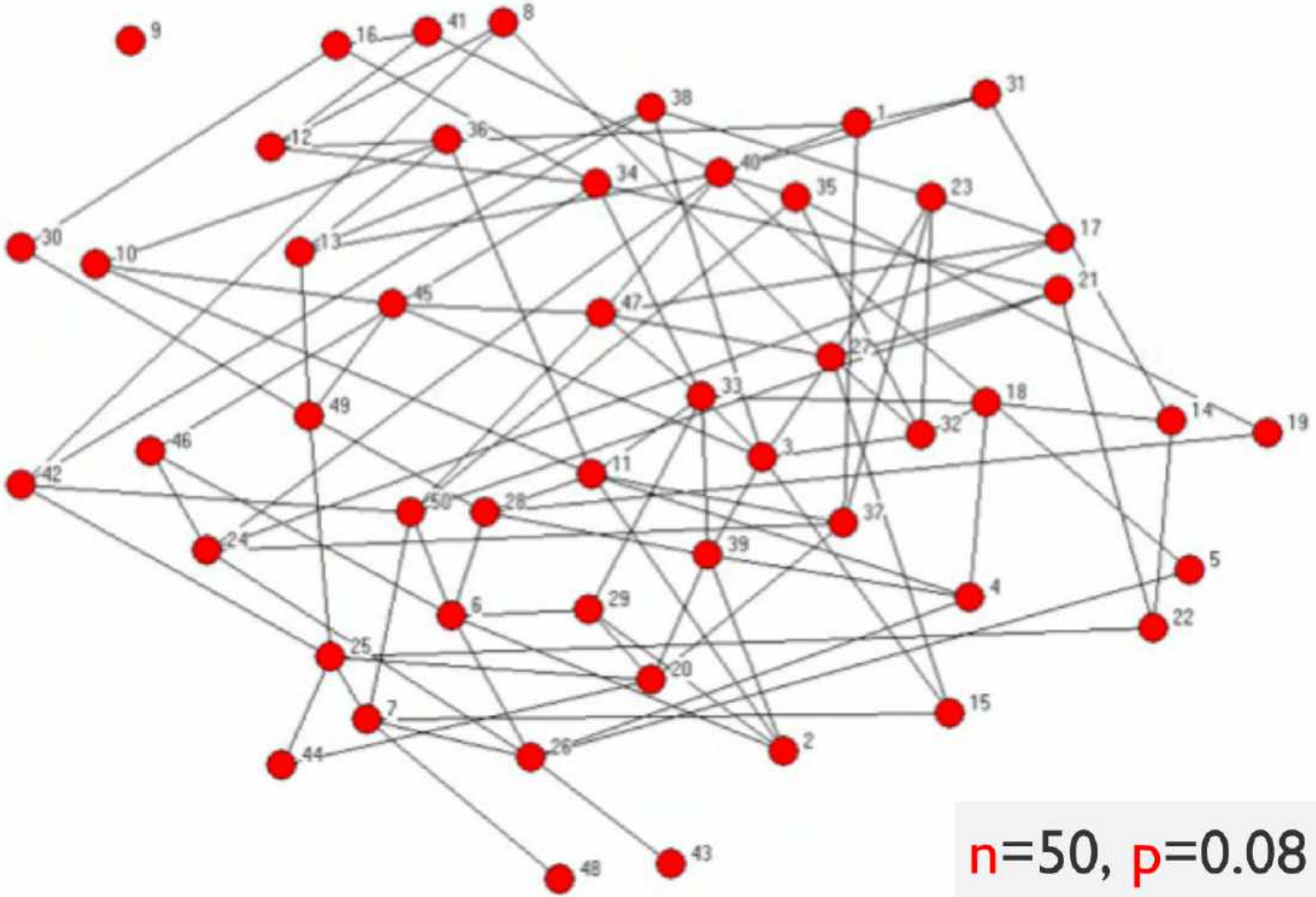
n nodes, m edges are picked at random

n and p do not uniquely determine the graph!



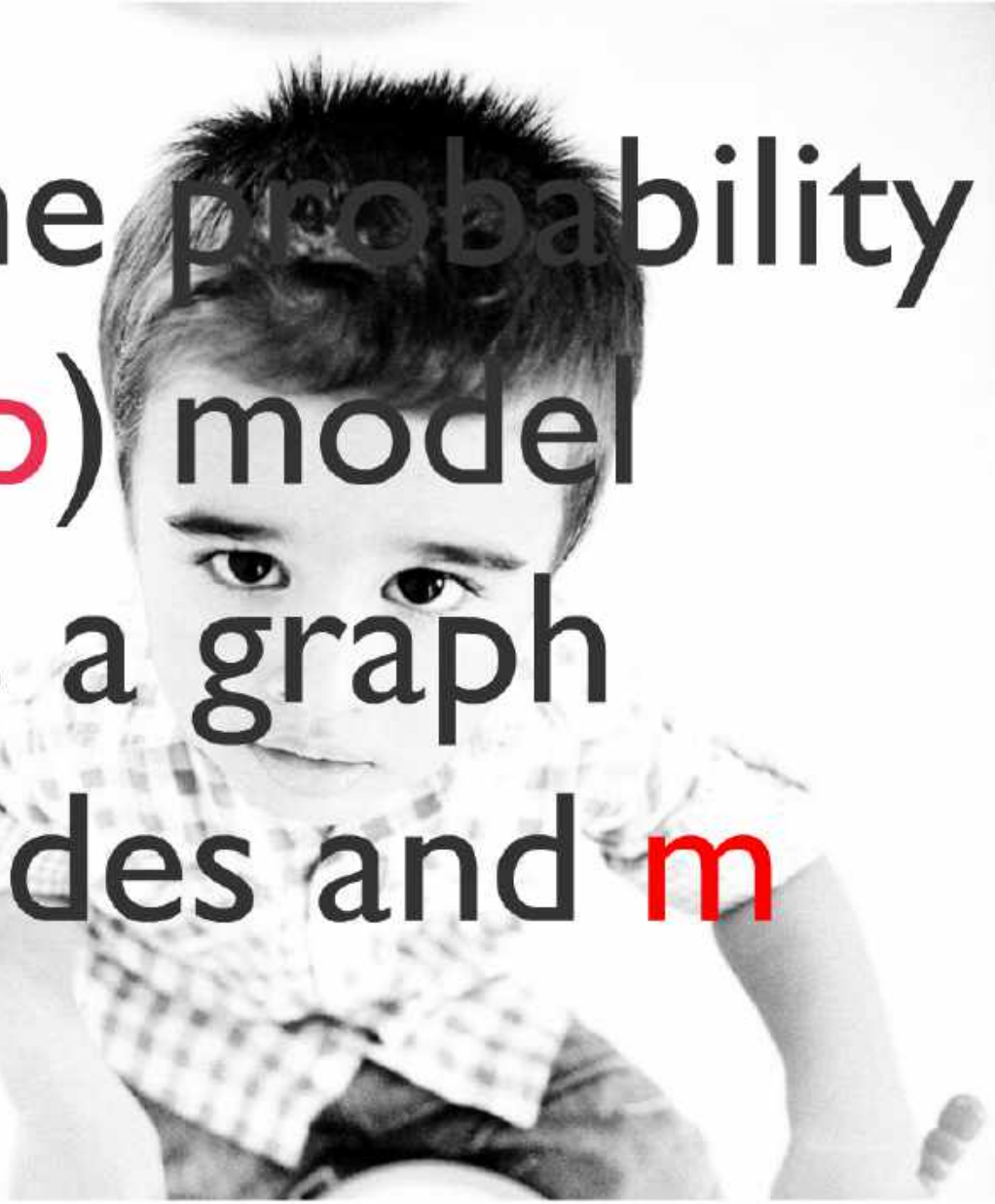

1
48
34
50
22
18
11
3
15
21
13
4
17
27
49
44
24
29
40
35
6
31
23
45
39
28
20
46
19
41
37
36
32
2
38
9
33
10
30
43
47
25
26
5
12
8
42
14
7
16

n=50, p=0.02

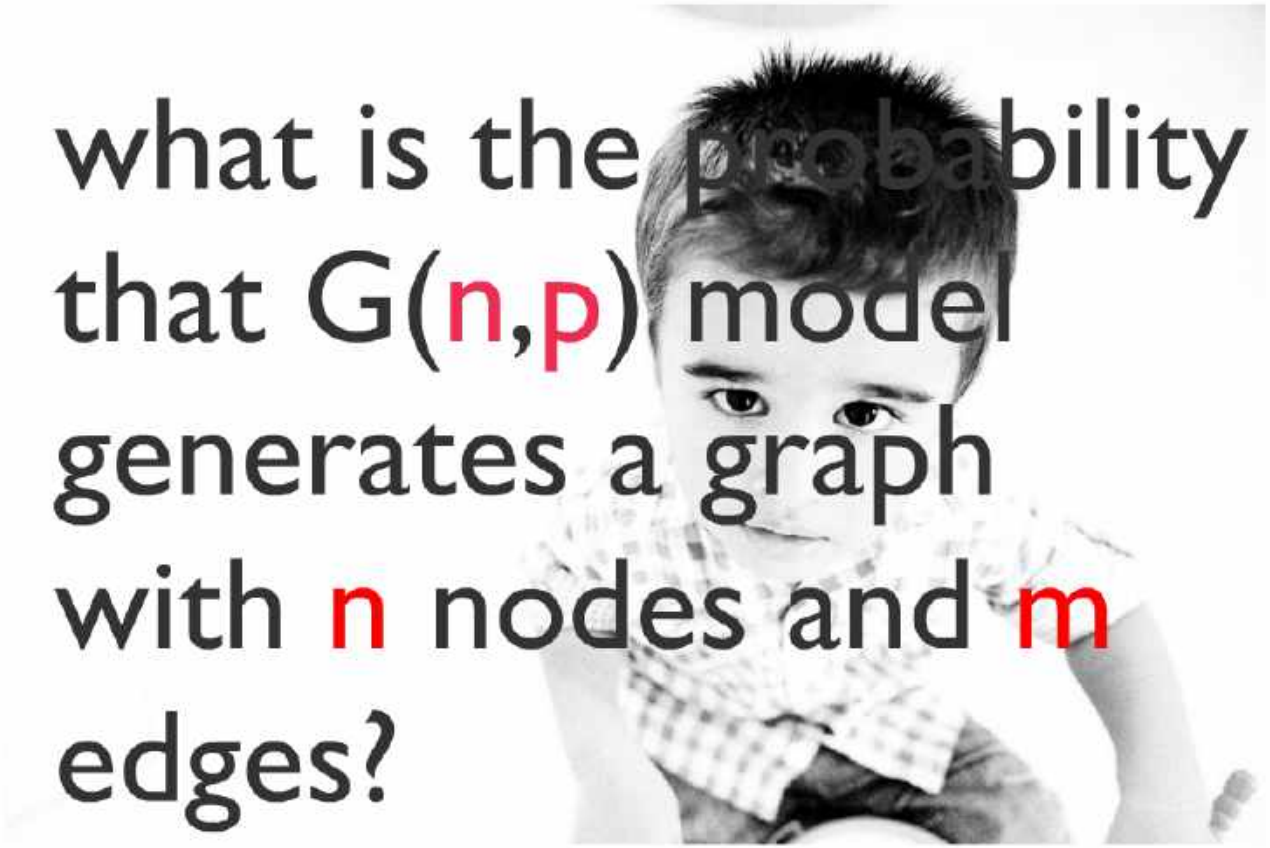

9
16
41
8
38
31
1
12
36
40
34
35
23
30
10
13
17
21
45
47
27
33
18
14
19
49
32
46
3
42
11
50
26
37
24
39
5
4
6
29
22
25
20
7
15
44
26
2
48
43

n=50, p=0.08

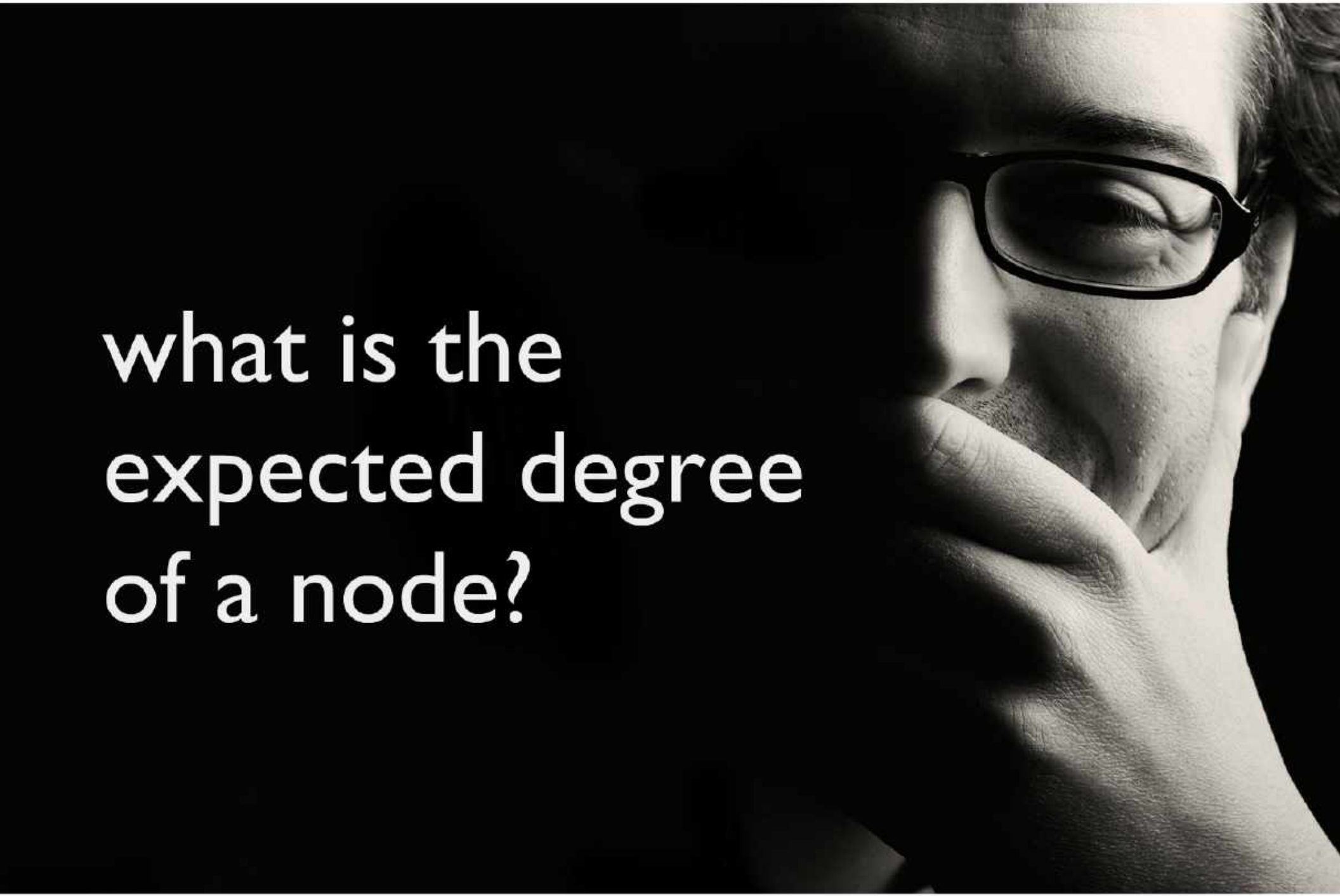
what is the probability that G(n,p) model generates a graph with n nodes and m edges?



what is the probability that G(n,p) model generates a graph with n nodes and m edges?



$$\binom{\binom{n}{2}}{m} p^m (1-p)^{\frac{n(n-1)}{2}-m}$$


what is the
expected degree
of a node?

# The probability that a node has degree d

$$\binom{n-1}{d} p^d (1-p)^{n-1-d}$$

P(d): the degree distribution of a node

**The probability that a node has degree d**

$$\binom{n-1}{d} p^d (1-p)^{n-1-d}$$

P(d): the degree distribution of a node

Poisson random networks

$$\approx \frac{e^{-(n-1)p}((n-1)p)^d}{d!}$$

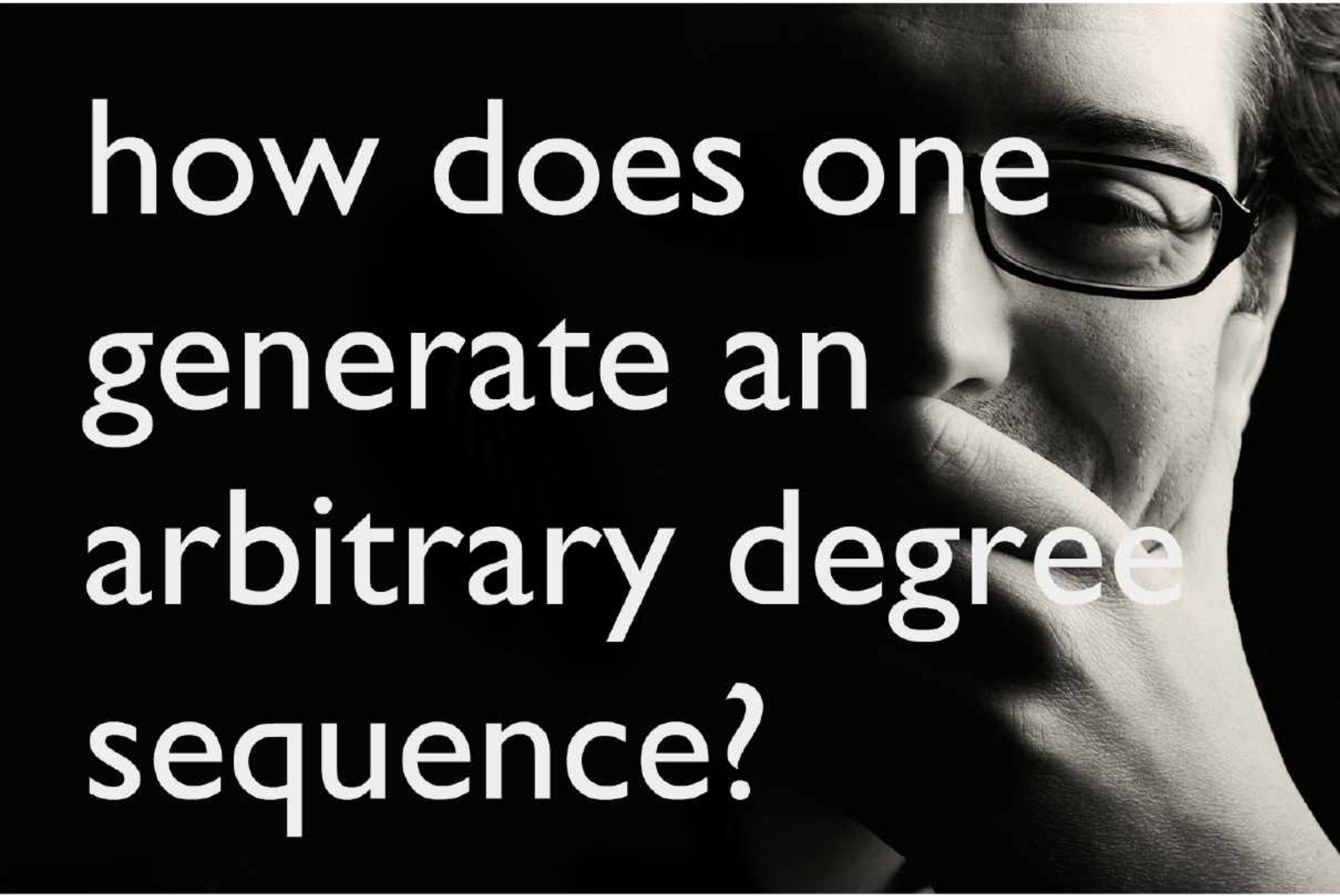
$$\frac{e^{-\lambda}\lambda^d}{d!} \quad \lambda = np$$

# what is the clustering coefficient?

what is the average path length?

what is the average path length?

the diameter grows as $O(\log n/\alpha)$


how does one
generate an
arbitrary degree
sequence?

$$d_1, d_2, d_3, \ldots, d_n$$

i=3

configuration model

1 assume that each node has the corresponding number of half edges

2 simply pair up edges

$$d_1, d_2, d_3, \ldots, d_n \qquad \sum_i d_i \text{ is even}$$

i=3

configuration model

1 assume that each node has the corresponding number of half edges

2 simply pair up edges

delete self loops and multiple edges

# What are some assumptions here?

$$d_1, d_2, d_3, \ldots, d_n$$

# expected degree model

$$p_{ij} \propto \frac{d_i d_j}{\sum_k d_k}$$

where $\left(\max_i d_i\right)^2 \leq \sum_k d_k$

# Two random graph generators

$d_1, d_2, d_3, \ldots, d_n$ $\quad \sum_i d_i$ is even

configuration model

i=3

1 assume that each node has the corresponding number of half edges

2 simply pair up edges

delete self loops and multiple edges

$d_1, d_2, d_3, \ldots, d_n$

expected degree model

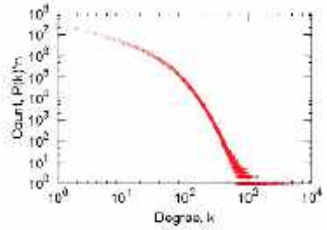
$$p_{ij} \propto \frac{d_i d_j}{\sum_k d_k}$$

where $\left(\max_i d_i\right)^2 \leq \sum_k d_k$

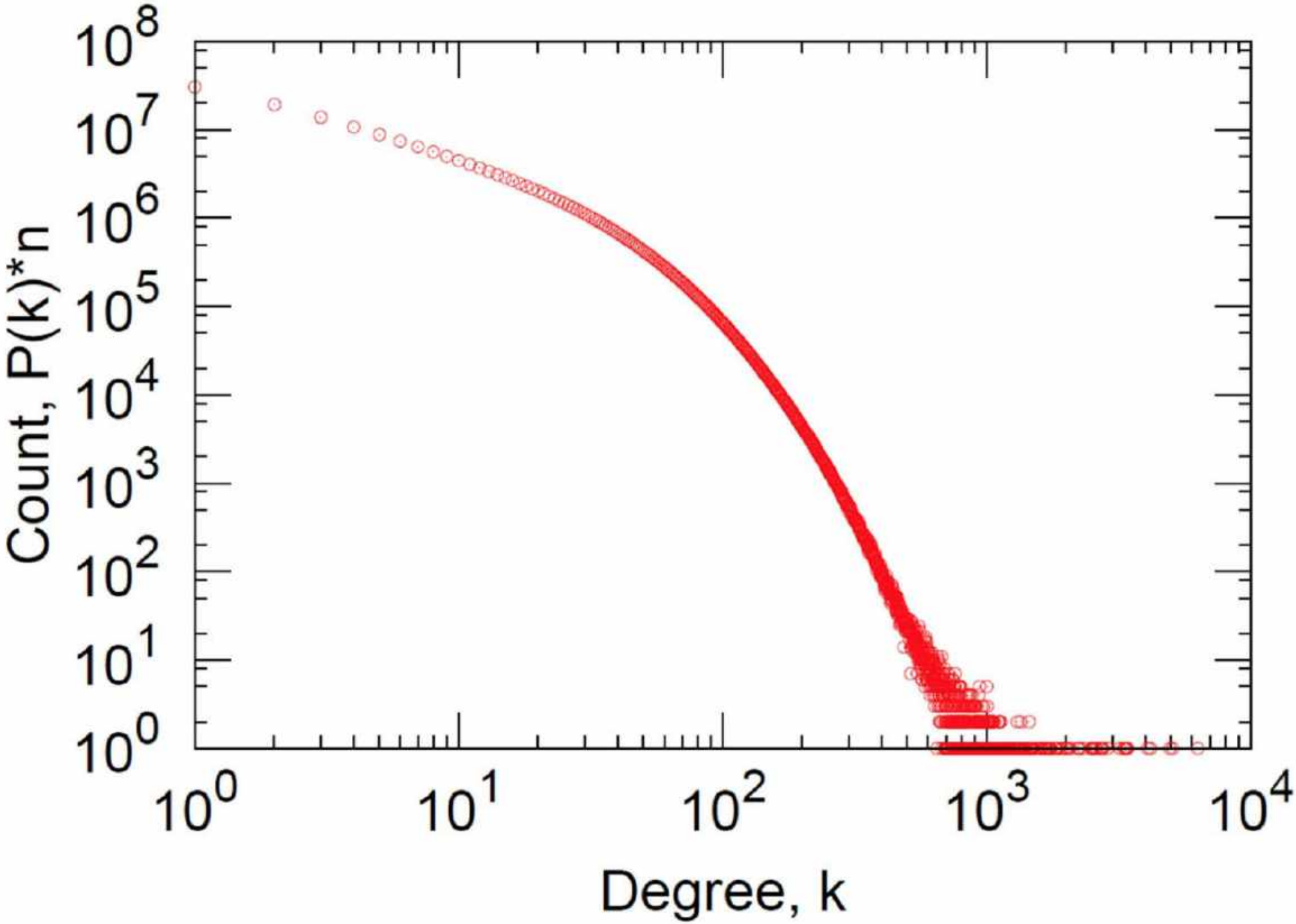
What are some assumptions here?

**What are the differences between the two?**

how does the Poisson model stack up to reality?



| Property | MSN | Poisson |
|---|---|---|
| Degree distribution | Power law* | Binomial |
| Path Length | 6.6 | log(n) ≅ 8.8 |
| Clustering Coefficient | 0.11 | p |


Count, P(k)*n
Degree, k
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7
10^8

# Poisson models: a reality check

Giant component

Clustering coefficient is way off

Average path length is comparable

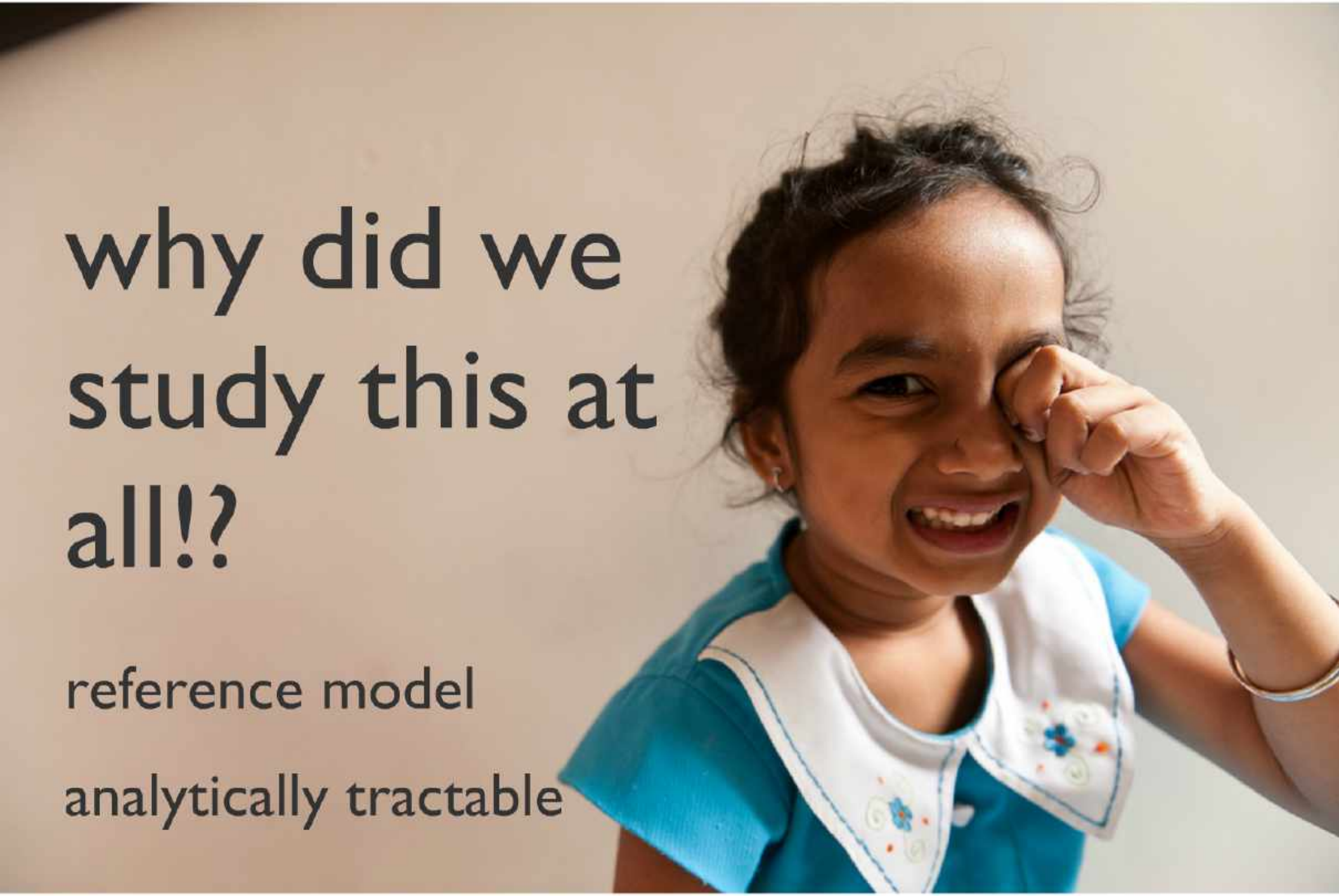
Degree distributions are unrealistic

**Real networks are not Poisson!**

# why did we study this at all!?

reference model

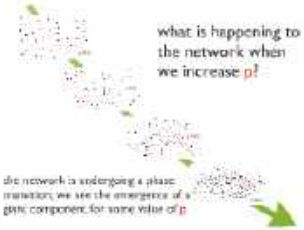
analytically tractable




to what degree is a network random?

# Giant Components

The distribution is now P(d)



what is happening to the network when we increase p?

this network is undergoing a phase transition; we see the emergence of a giant component for some value of p


A property A(N) is often specified by simply listing the networks that satisfy the property

$$A(N) = \{g \mid N_1(g) \neq \emptyset \; \forall i \in N\}$$



A large component is said to exist when the number of nodes in the component exceeds

$$> n^{2/3}$$

As a concrete example, consider the case that node 1 has at least one link.

$$A(N) = \{g \mid d_1 \geq 1\}$$

probability that A(N) is true

$$1 - (1 - p(n))^{n-1}$$

$$\frac{1}{n^2}$$

the first links emerge

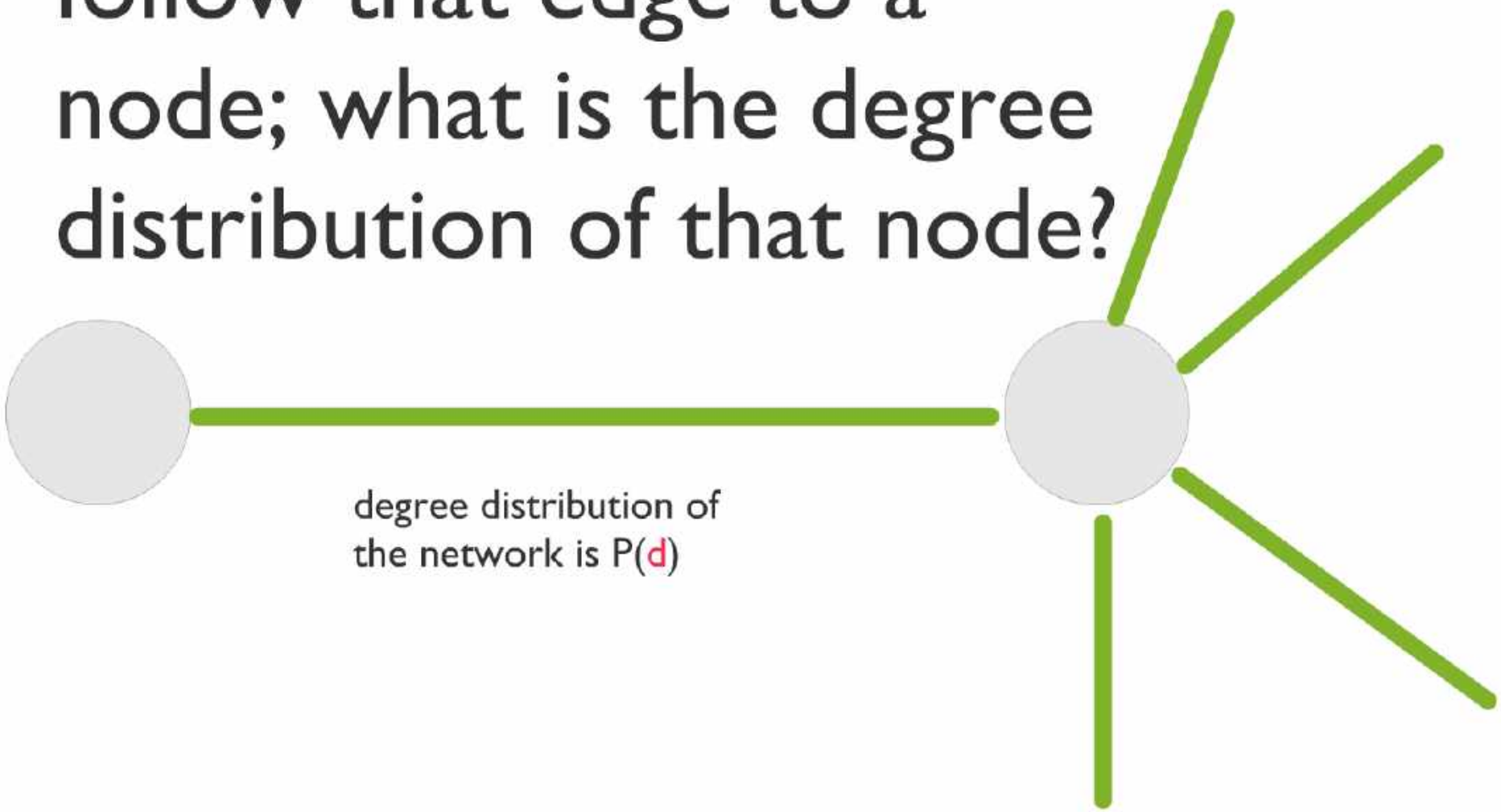
$$\frac{\log(n)}{n}$$

the network becomes connected

$$\frac{1}{n}$$

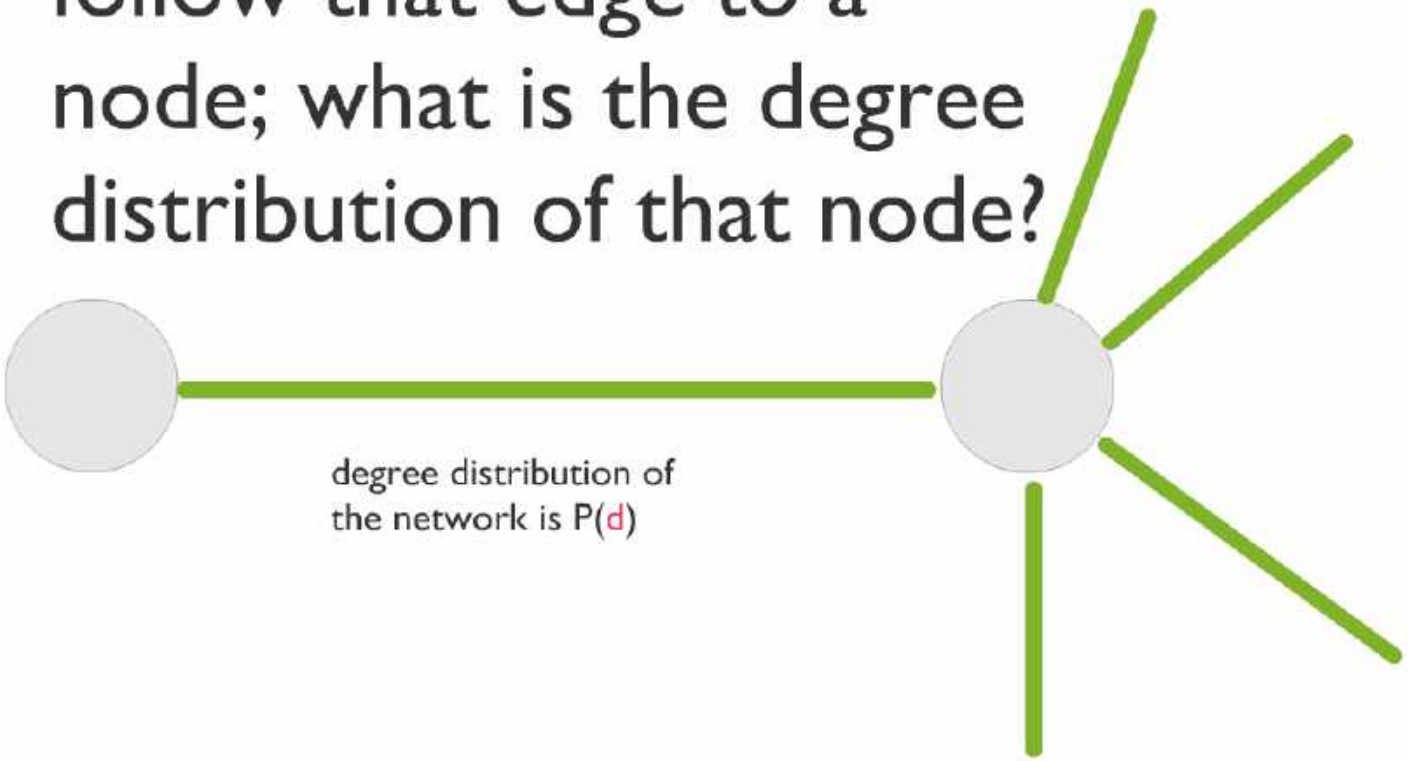
the first cycle emerges

Let's assume that you pick a random edge and follow that edge to a node; what is the degree distribution of that node?



degree distribution of the network is P(d)

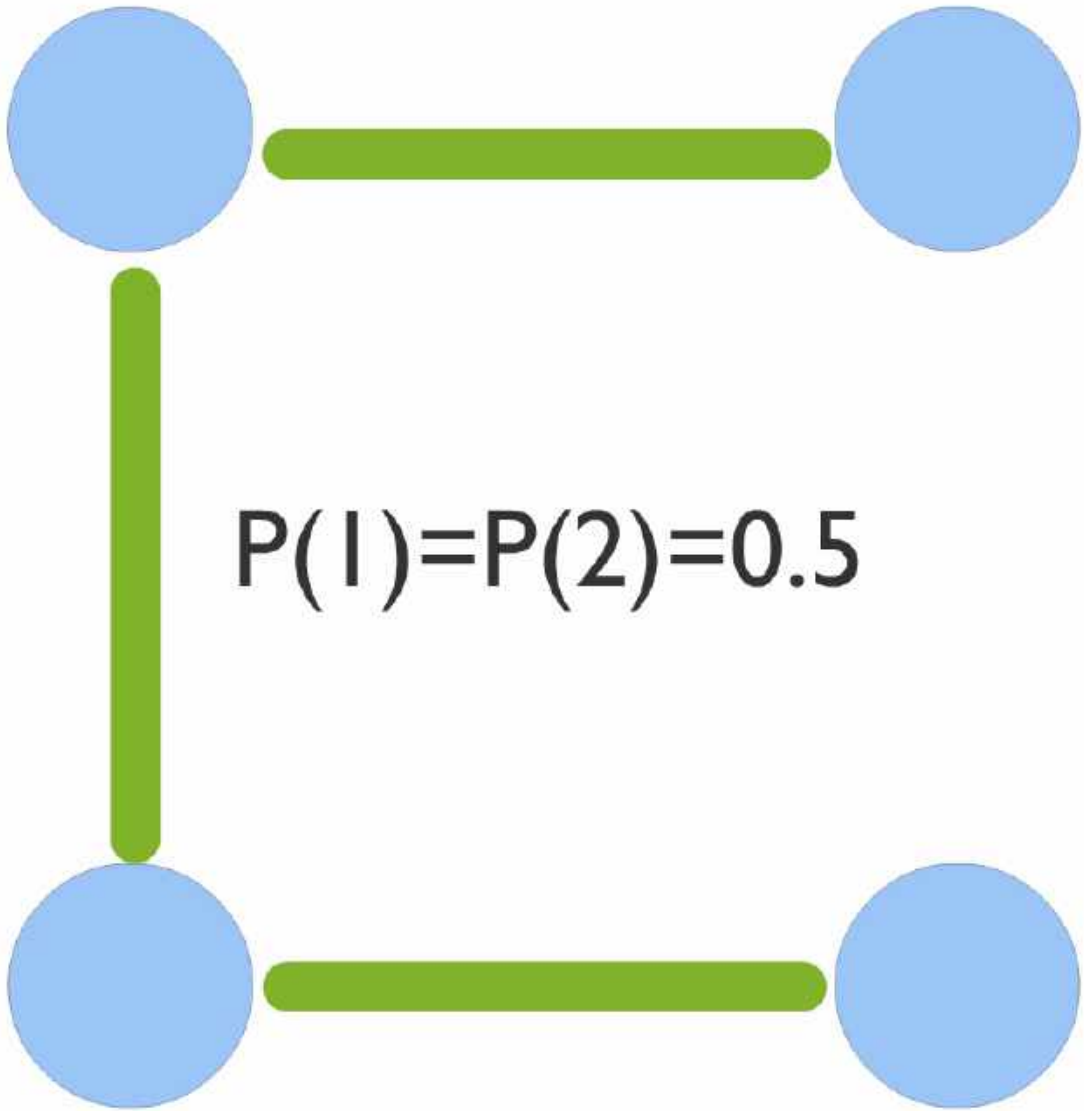
Let's assume that you pick a random edge and follow that edge to a node; what is the degree distribution of that node?



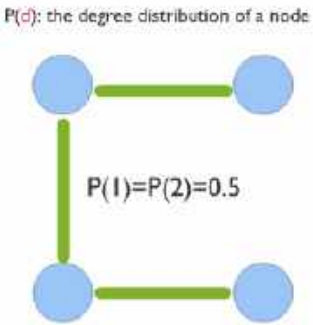
degree distribution of the network is P(d)

The distribution is **not** P(d)

P(d): the degree distribution of a node



what is the probability of finding a node with degree 2?

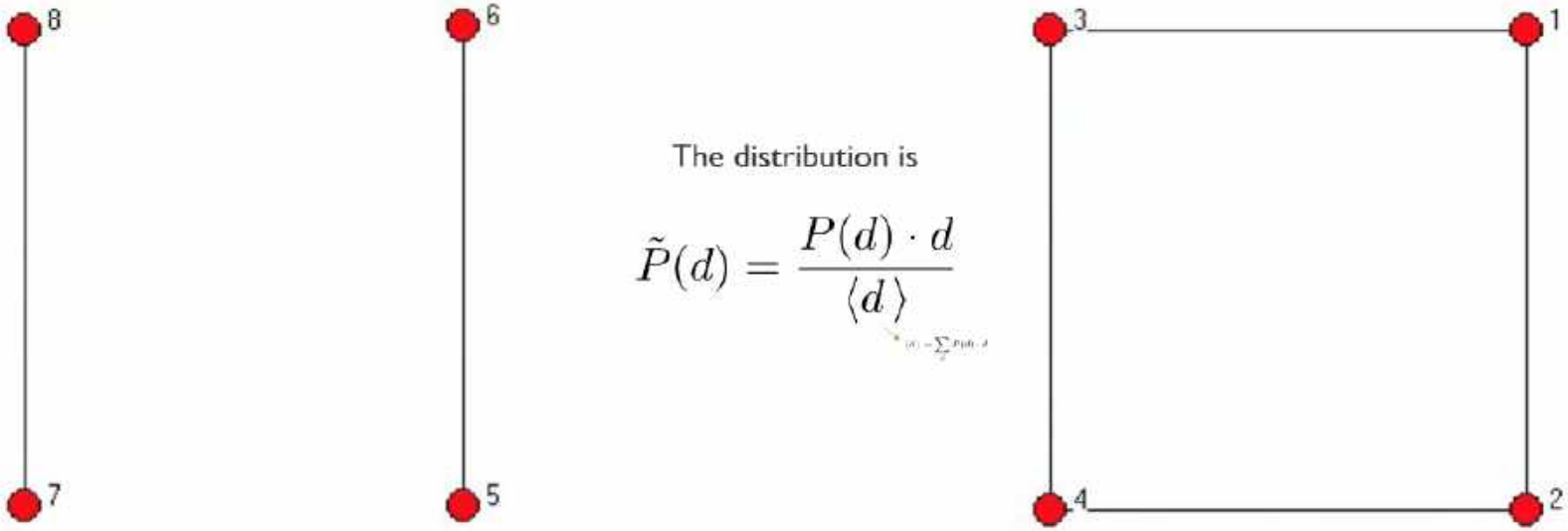



# The distribution is

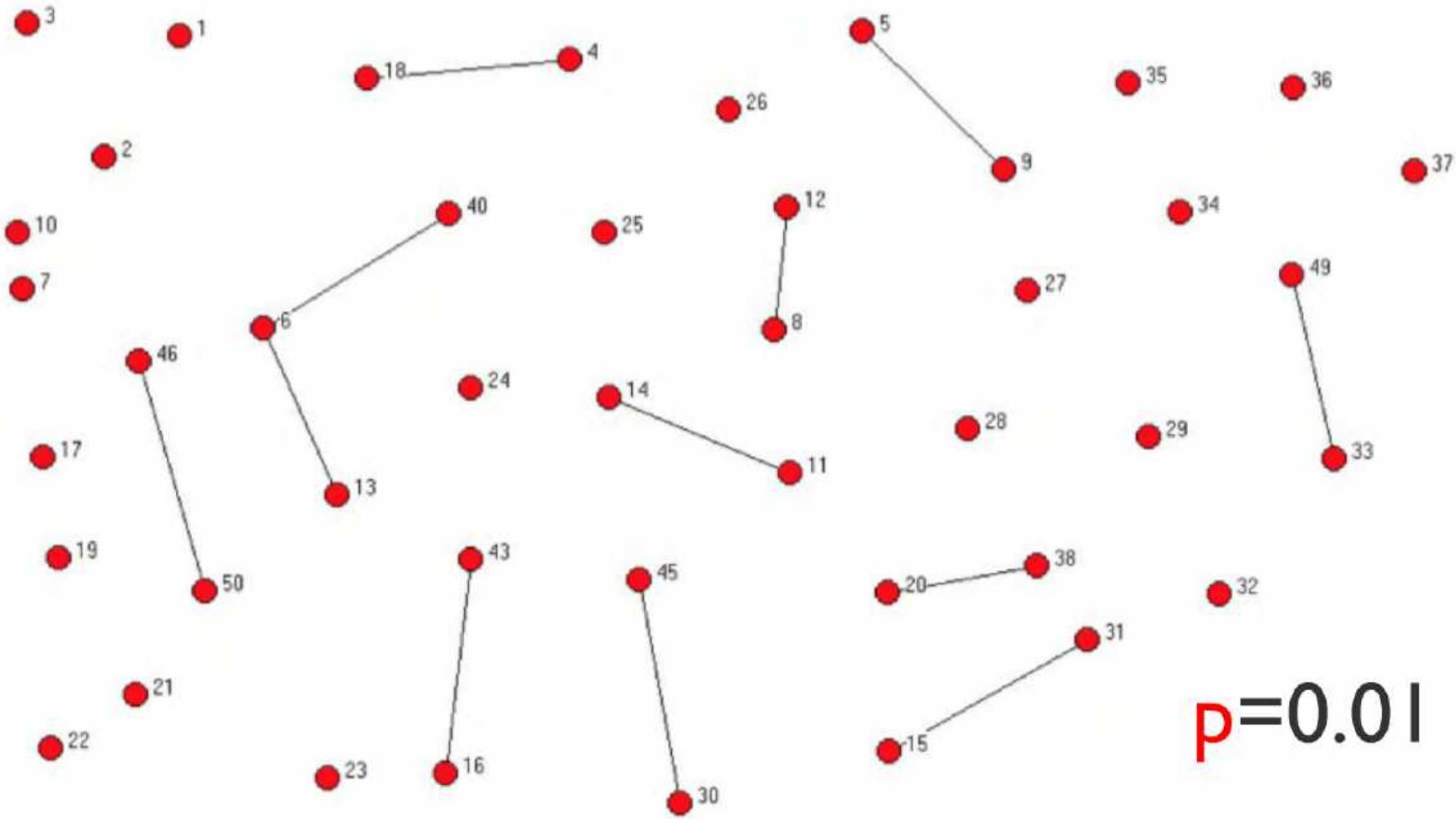
$$\tilde{P}(d) = \frac{P(d) \cdot d}{\langle d \rangle}$$

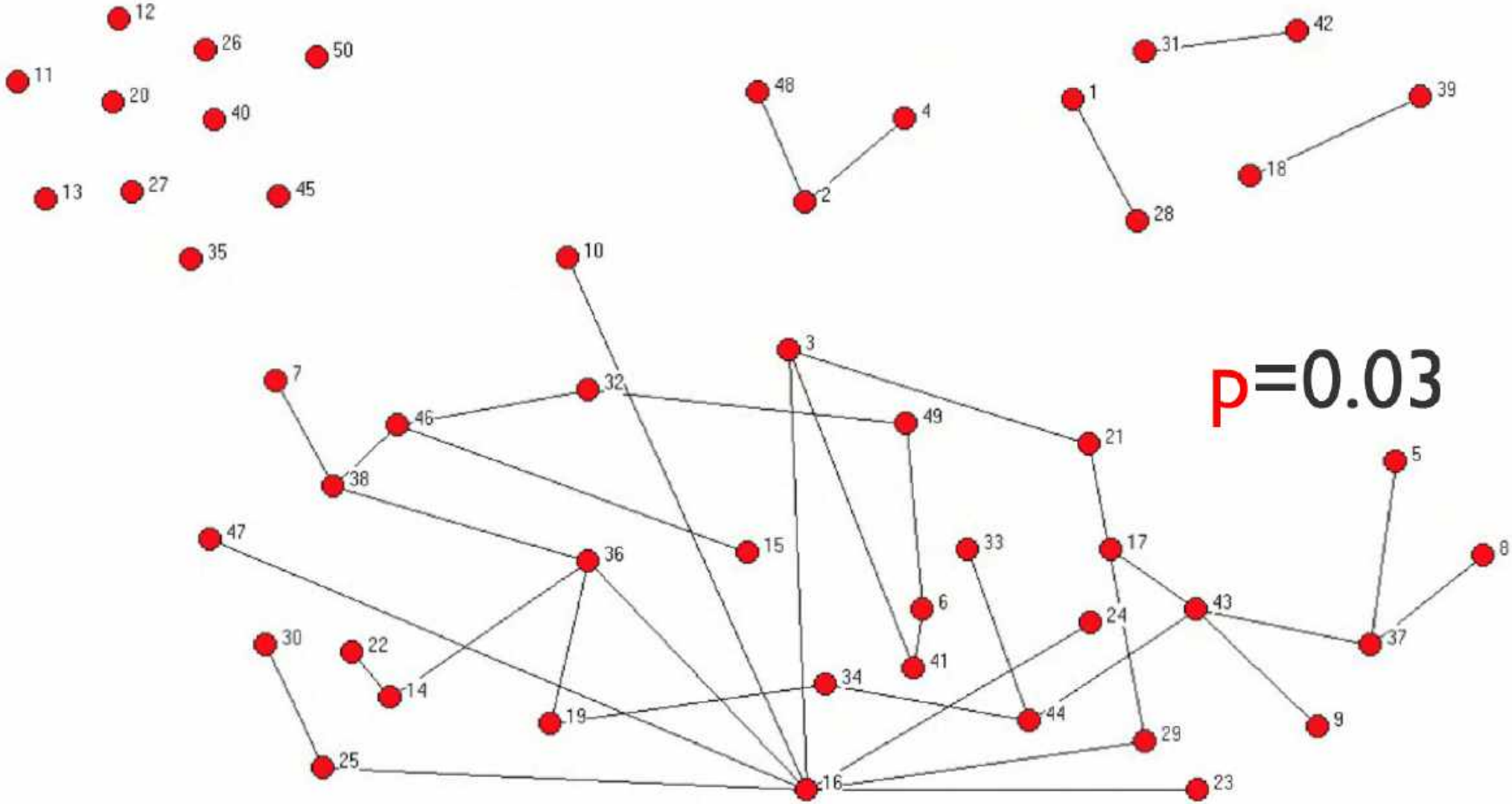
$$\langle d \rangle = \sum_d P(d) \cdot d$$

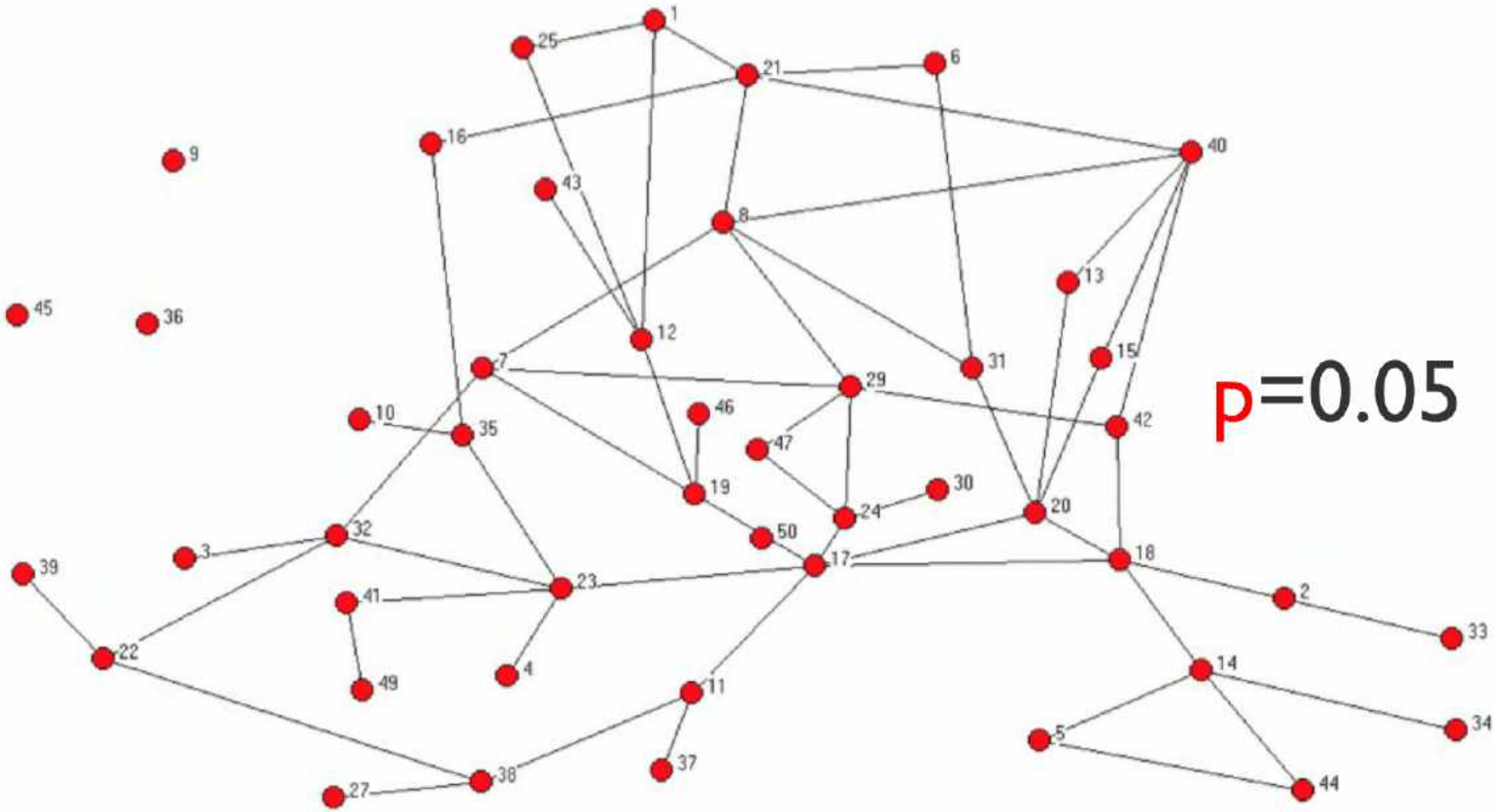
Doesn't hold when the neighboring degrees are highly correlated



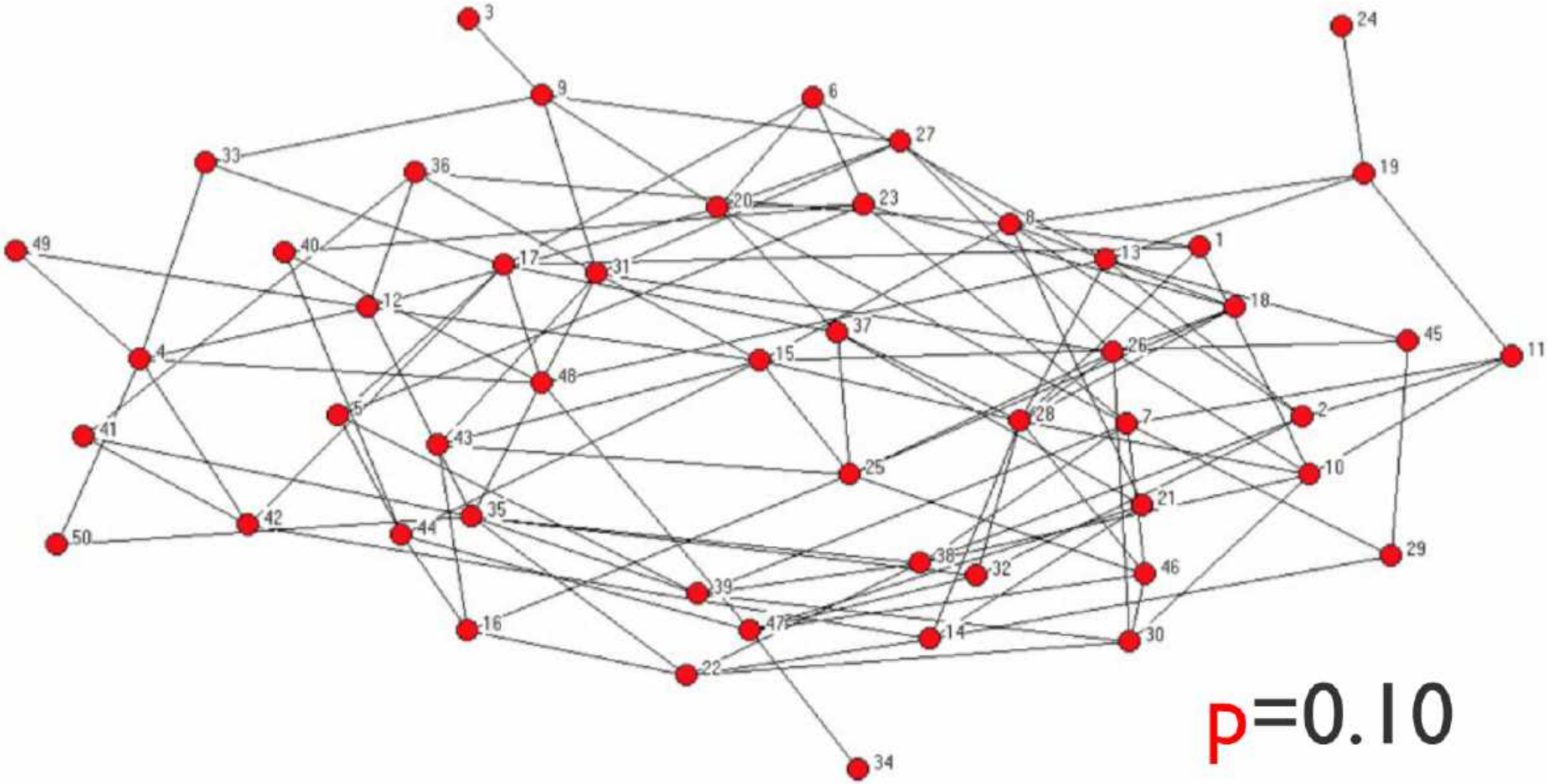

Networks that form due to preferential attachment will also not exhibit this behavior


3
1
18
4
26
5
35
36
2
9
37
10
40
25
12
34
7
27
49
6
8
46
24
14
28
29
17
11
33
13
19
43
45
38
50
20
32
31
21
22
23
16
30
15
p=0.01

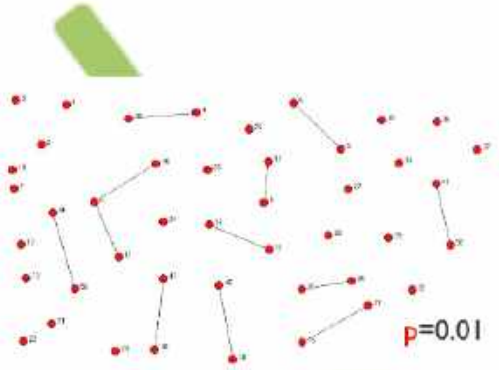

12
26
50
11
20
40
13
27
45
35
48
4
2
31
42
1
39
18
28
10
3
7
32
46
49
21
38
5
47
15
36
33
17
8
6
43
24
30
22
41
37
14
34
44
9
19
29
25
16
23
p=0.03

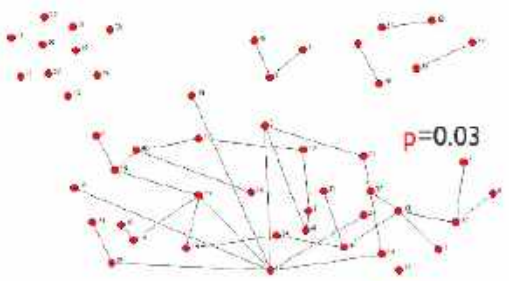

1
25
21
6
16
40
9
43
8
13
45
36
12
15
7
31
29
42
10
46
35
47
19
30
24
20
32
50
3
17
18
39
23
2
41
33
22
14
49
4
11
34
5
37
38
44
27

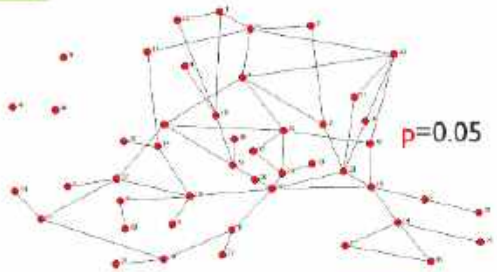
p=0.05

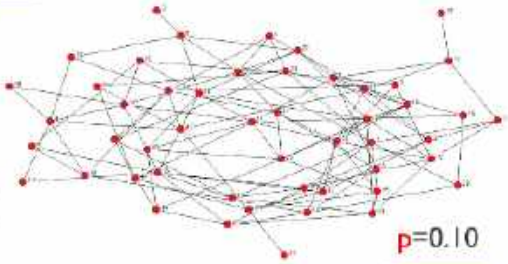

3
24
9
6
27
33
36
19
20
23
8
49
40
1
17
31
13
12
18
37
45
4
15
26
11
48
28
7
2
41
43
25
10
42
35
21
50
44
38
32
46
29
39
16
47
14
30
22
34

p=0.10



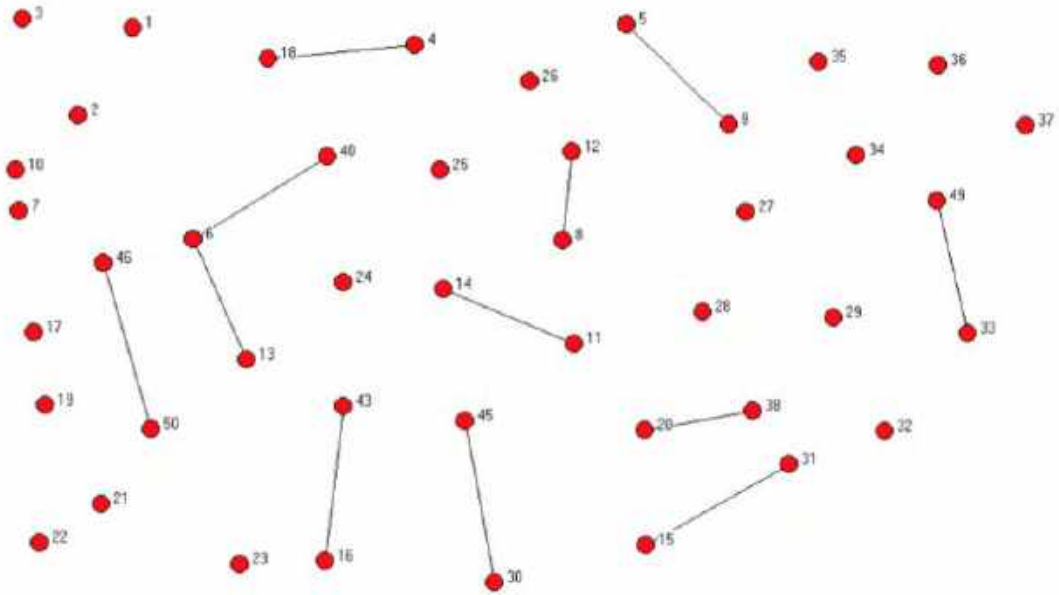




what is happening to the network when we increase p?

the network is undergoing a phase transition; we see the emergence of a giant component for some value of p

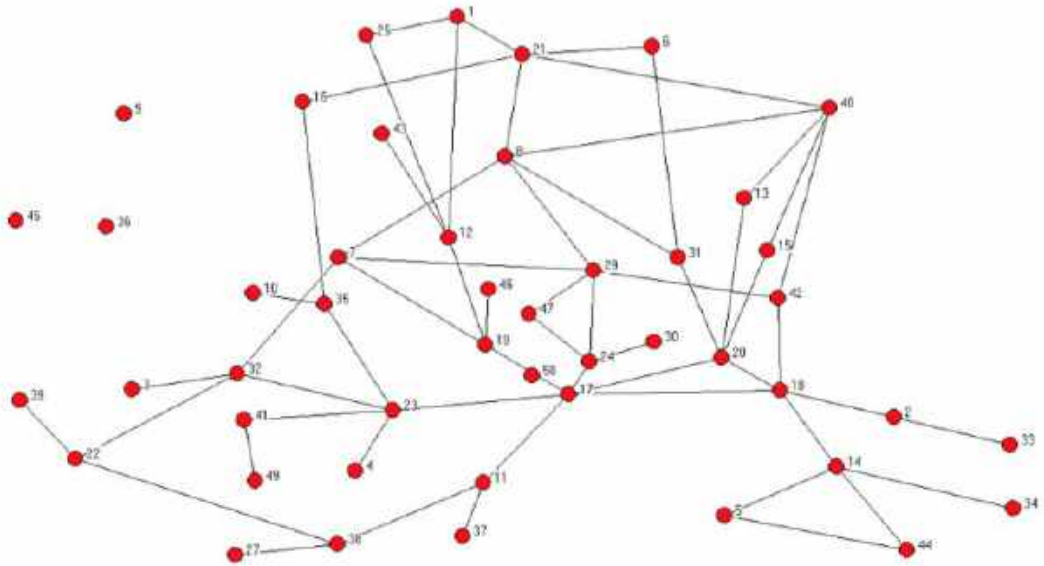
$$\frac{1}{n^2}$$

# the first links emerge



$$\frac{1}{n}$$

the first cycles emerge; giant component begins to form



$$\frac{\log(n)}{n}$$

the network becomes connected