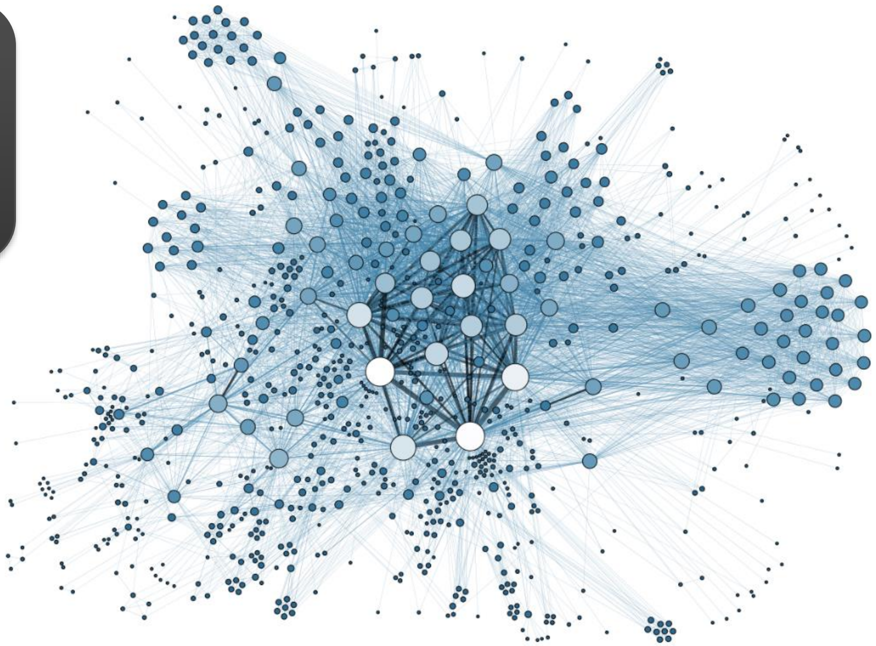
Social Media Analytics

INFS7450

Network Measures And Models



Prof. Hongzhi Yin

School of EECS

The University of Queensland

Network Properties: How to Measure a Network?

Degree Distribution
Average Shortest Path Length
Average Clustering Coefficient
Size of Giant Component

Key Network Properties

Degree distribution: $P(k)$

Average Shortest Path length: h

Average Clustering coefficient: C

Size of Giant Component : s

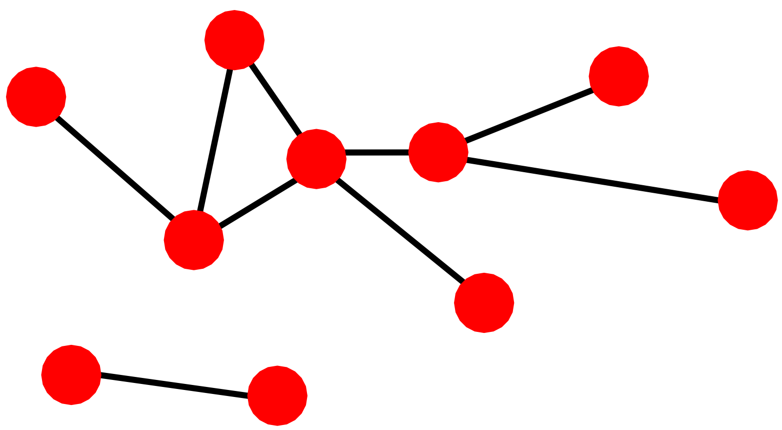
Degree Distribution

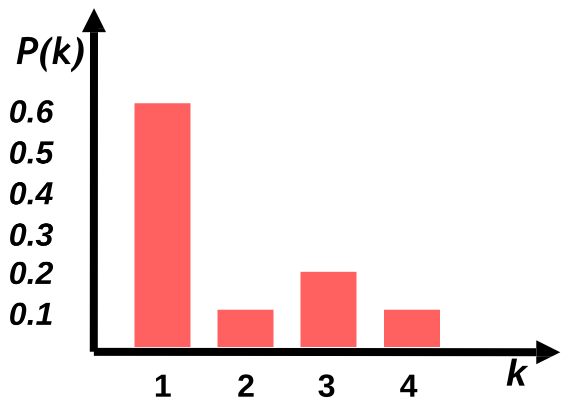
- **Degree distribution $P(k)$:** Probability that a randomly chosen node has degree k

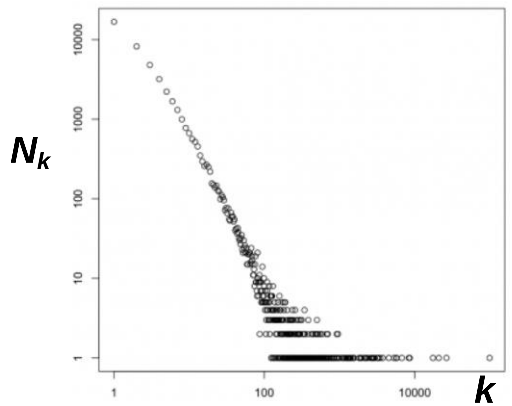
  N_k = # nodes with degree k

- Normalized histogram:

  $$P(k) = N_k / N \quad \rightarrow \quad \text{plot}$$







Power-Law Degree Distribution

The frequency of degree *d* follows a **power-law**

Power-law intercept

The power-law exponent and its value is typically in the range of **[2, 3]**

$$p_d = ad^{-b}$$

Fraction of users with degree d

Node degree

$$\boxed{\ln p_d} = -b \boxed{\ln d} + \ln a$$

Power-Law Distribution: Examples

- **Call networks:**
  - The fraction of telephone numbers that receive k calls per day is roughly proportional to $\mathbf{1/k^2}$
- **Book Purchasing:**
  - The fraction of books that are bought by k people is roughly proportional to $\mathbf{1/k^3}$
- **Scientific Papers:**
  - The fraction of scientific papers that receive k citations in total is roughly proportional to $\mathbf{1/k^3}$
- **Social Networks:**
  - The fraction of users that have in-degrees of k is roughly proportional to $\mathbf{1/k^2}$

Power-Law Distribution

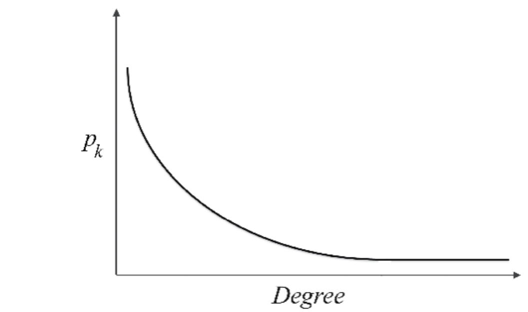
- Many real-world networks exhibit a *power-law* distribution.

- Power-laws appear
  - When the quantity being measured can be viewed as a type of popularity.

- In a power-law distribution
  - **Small occurrences**: common
  - **Large instances**: extremely rare

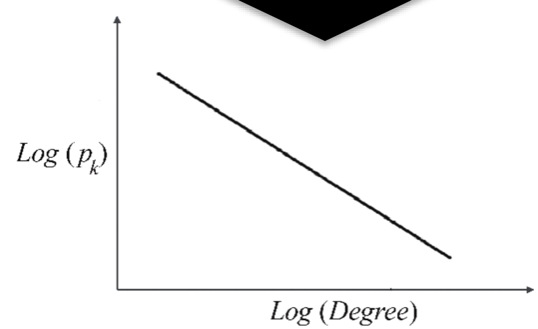
Small numbers are common, while large numbers are rare.

A typical shape of a power-law distribution



(a) Power-Law Degree Distribution

Log-Log plot



(b) Log-Log Plot of Power-Law Degree Distribution

Power-law Distribution: An Elementary Test

To test whether a network exhibits a power-law distribution

1. Pick a popularity measure and compute it for the whole network
   – Example: number of friends for all nodes

2. Compute p_k, the fraction of individuals having popularity k.

3. Plot a log-log graph, where the x-axis represents $\ln k$ and the y-axis represents $\ln p_k$.

4. If a power-law distribution exists, we should observe a straight line

This is not a systematic approach!

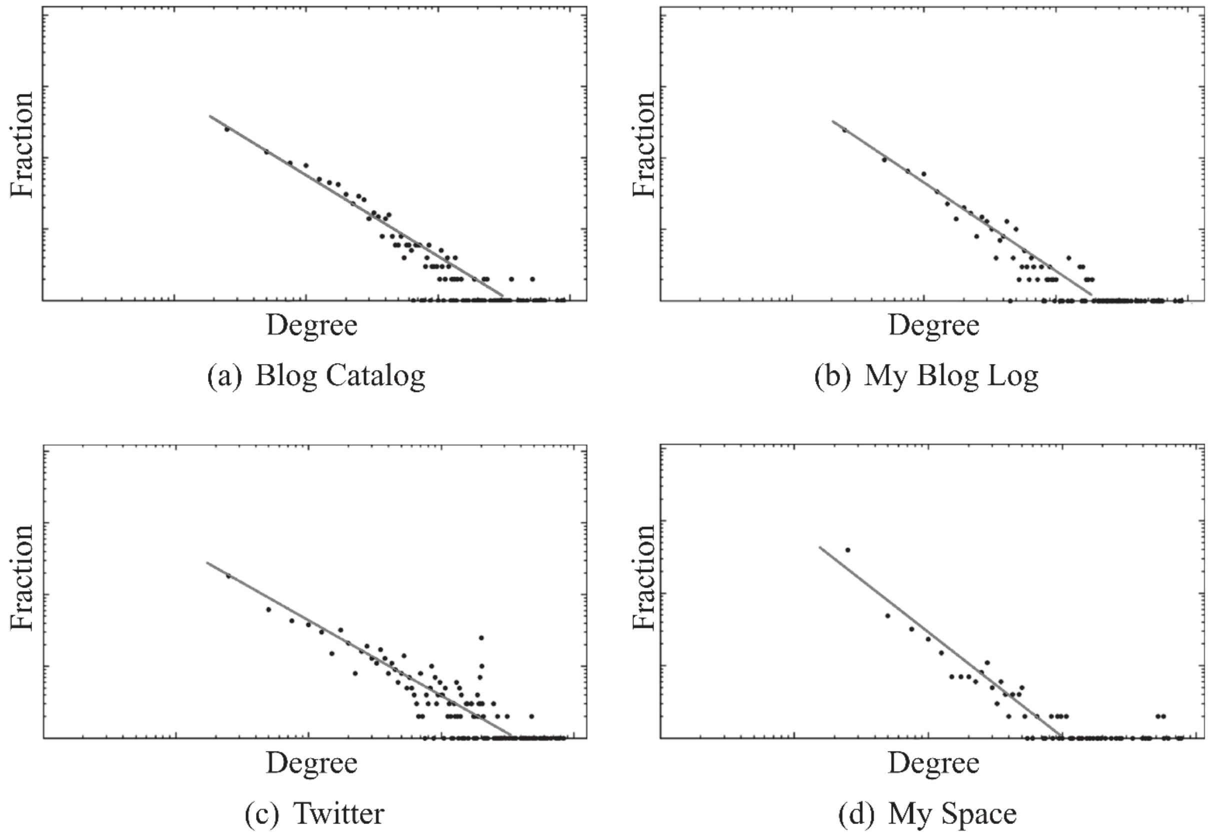
1. Other distributions could also exhibit this pattern

For a systematic approach see:

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. "Power-law distributions in empirical data." *SIAM review 51(4)* (2009): 661-703.

Power-Law Distribution: Real-World Networks

Real-world social networks follow power-law distributions (called **Scale-Free** networks)



Key Network Properties

Degree distribution: $P(k)$

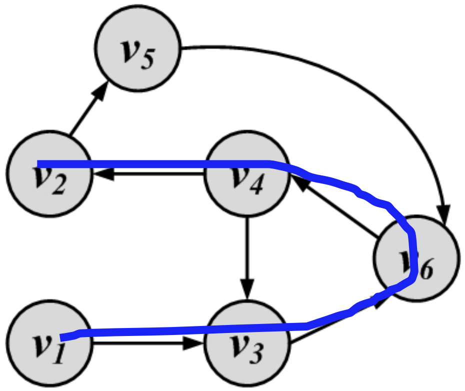
Average Shortest Path length: h

Average Clustering coefficient: C

Connected components: s

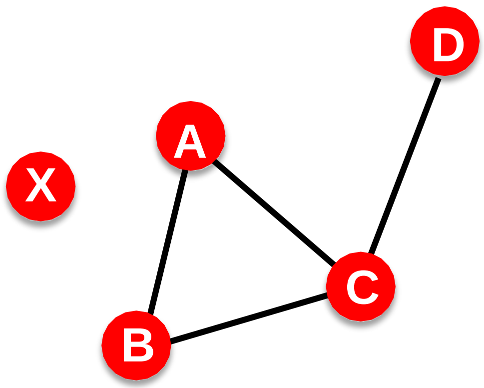
Path

- A walk where **nodes and edges** **are distinct** is called a **path**
- A **cycle** is a special path with the same starting point and ending point
- The length of a path on unweighted graphs is the number of edges visited in the path

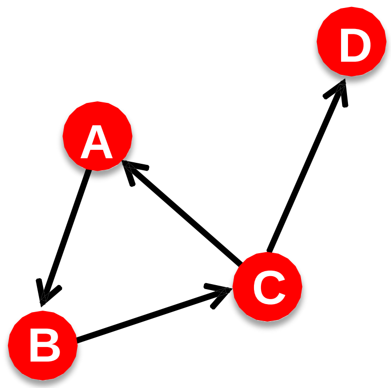


Length of path= 4

Distance in an Unweighted Graph



$h_{B,D} = 2$
$h_{A,X} = \infty$



- **Distance (shortest path, geodesic)** between a pair of nodes is defined as the number of edges along the shortest path connecting the nodes
  - *If the two nodes are not connected, the distance is usually defined as infinite
- In **directed graphs** paths need to follow the direction of the arrows
  - Consequence: Distance is **not symmetric**: $h_{B,C} \neq h_{C,B}$

Network Diameter

- **Diameter:** The maximum (shortest path) distance between any pair of nodes in a graph
- **Average shortest path length** for a connected graph (component) or a strongly connected (component of a) directed graph

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

where h_{ij} is the distance from node i to node j
E_{max} is max number of edges (total number of node pairs) = $n(n-1)/2$

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

Web	Facebook	Flickr	LiveJournal	Orkut	YouTube
16.12	4.7	5.67	5.88	4.25	5.10

Key Network Properties

Degree distribution: $P(k)$

Average Shortest Path length: h

Average Clustering coefficient: C

Size of Giant component: s

Clustering Coefficient for Undirected Graphs

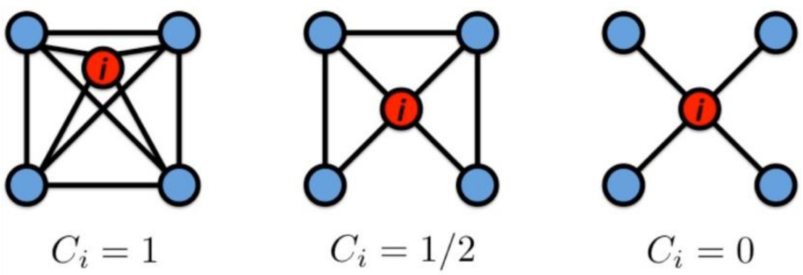
- **Clustering coefficient:**
  - What portion of ***i***'s neighbors are linked?
  - Node ***i*** with degree $\boldsymbol{k_i}$
  - $C_i \in [0,1]$

$k_i(k_i - 1)/2$ The maximum number of edges between the neighbors of node i

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

where e_i is the number of edges between the neighbors of node i

$C_i = 0$ If the degree of node i is **1**.

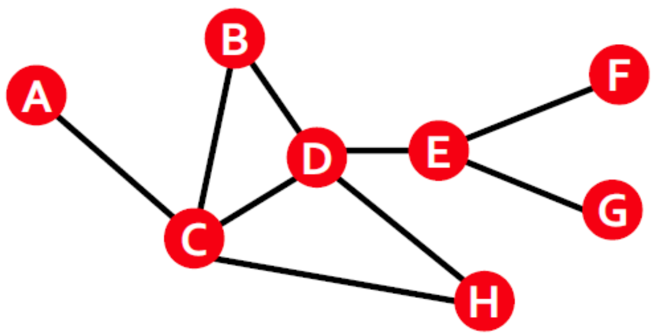


- **Average clustering coefficient:** $C = \frac{1}{N}\sum_{i}^{N} C_i$

Clustering Coefficient

- **Clustering coefficient:**
  - What portion of i's neighbors are connected?
  - Node i with degree k_i
  - $C_i = \dfrac{2e_i}{k_i(k_i - 1)}$ where e_i is the number of edges between the neighbors of node i

$C_i = 0$ If the degree of node i is **1**.



$k_B=2,\ e_B=1,\ C_B=2/2 = 1$

$k_D=4,\ e_D=2,\ C_D=4/12 = 1/3$

Avg. clustering: $C=0.33$

Web	Facebook	Flickr	LiveJournal	Orkut	YouTube
0.081	0.14	0.31	0.33	0.17	0.13

Key Network Properties

Degree distribution: $P(k)$

Average Shortest Path length: h

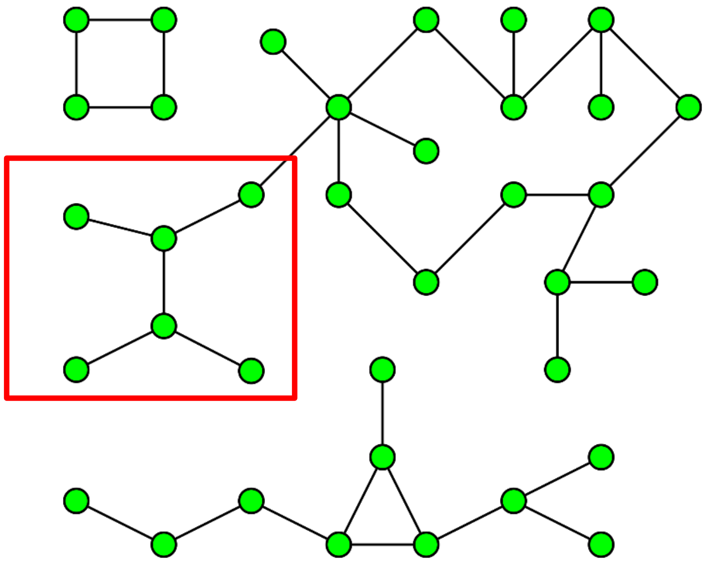
Average Clustering coefficient: C

Size of Giant component: s

Connectivity

- A component of an undirected graph is a subgraph
  - in which any two vertices are connected to each other by paths,
  - and which is connected to no additional vertices in the original graph.

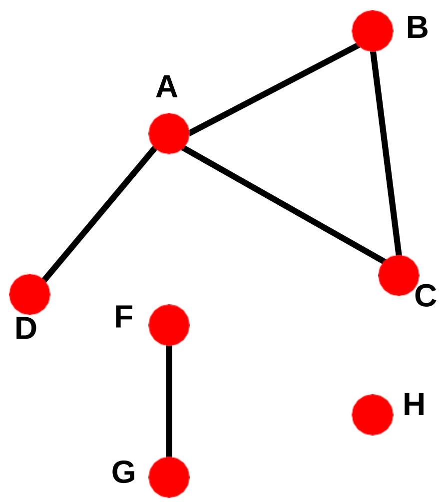
It is not a component



A Graph with 3 Components

Connectivity

- **Size of the largest connected component**
  - Largest set where any two vertices can be joined by a path
- **Largest component = Giant component**



How to find connected components:

- Start from random node and perform Breadth First Search (BFS)
- Label the nodes BFS visited
- If all nodes are visited, the network is connected
- Otherwise find an unvisited node and repeat BFS

Measuring a real-world network using these measures

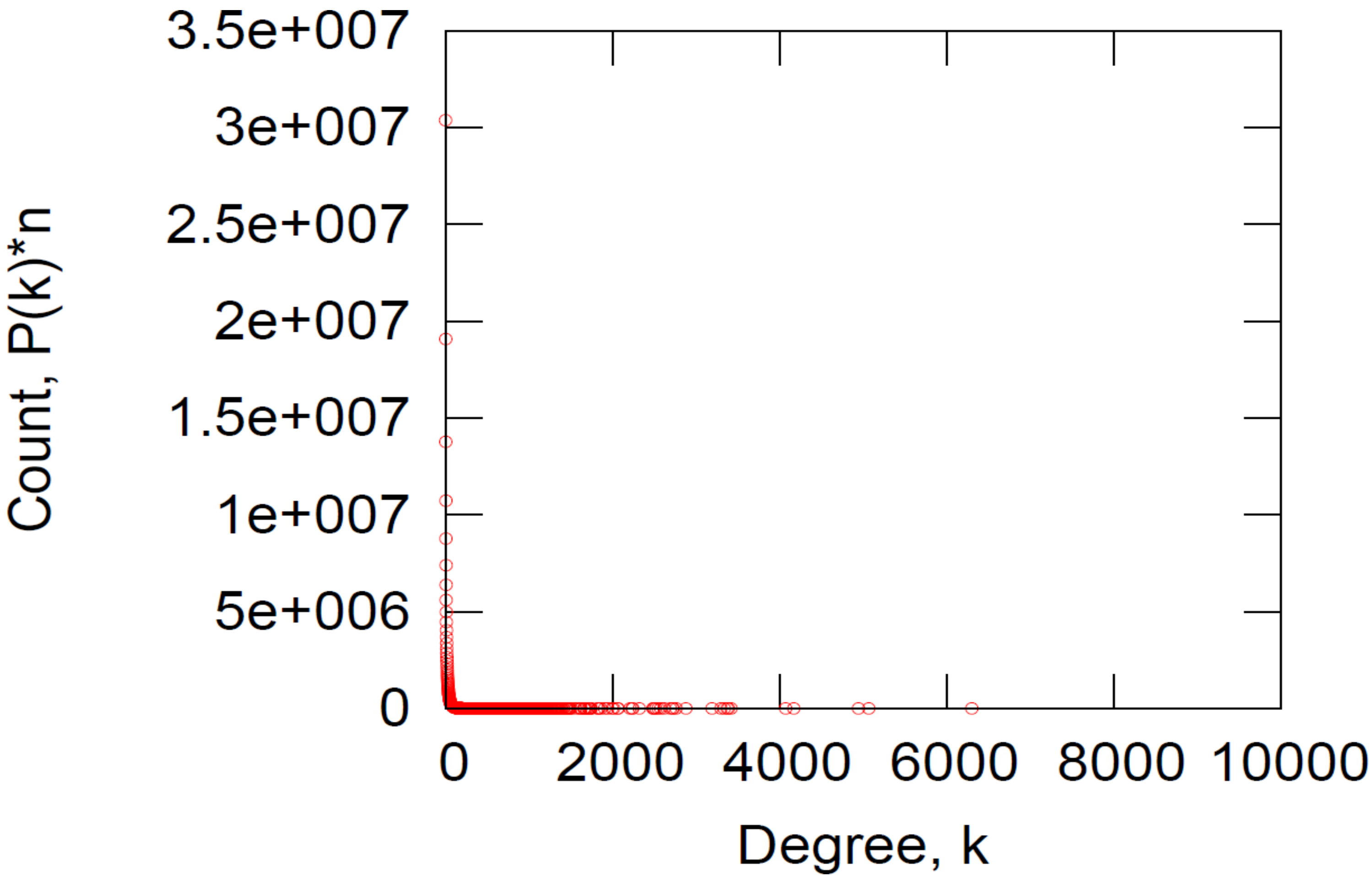
Degree Distribution
Average Shortest Path Length
Average Clustering Coefficient
Size of Giant Component

MSN Messenger

MSN Messenger.

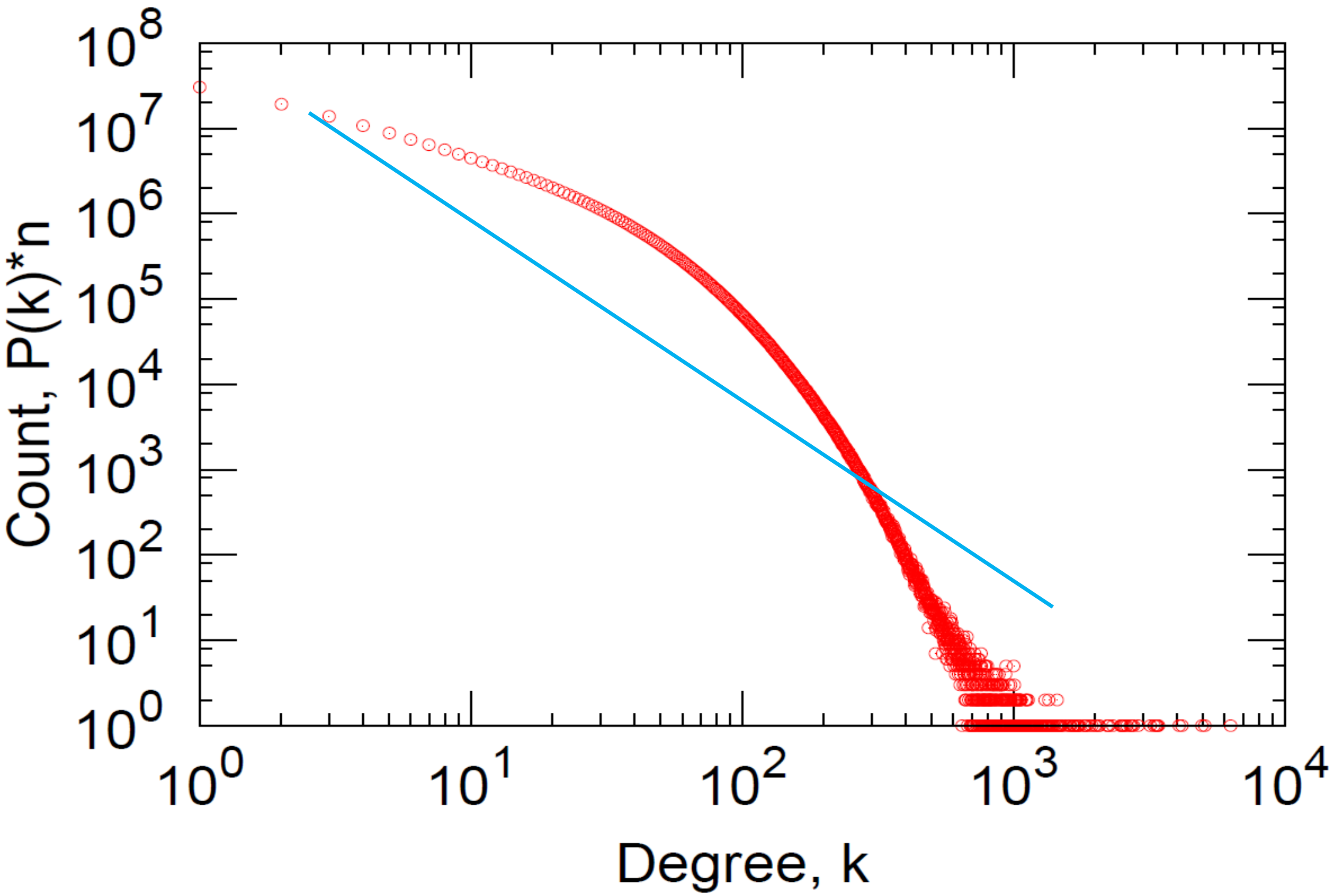
- **1 month activity**
  - 245 million users logged in
  - 180 million users engaged in conversations
  - More than 30 billion conversations
  - More than 255 billion exchanged messages

MSN (1) : Visualization of Degree Distribution



Linear Scale – Linear Plot

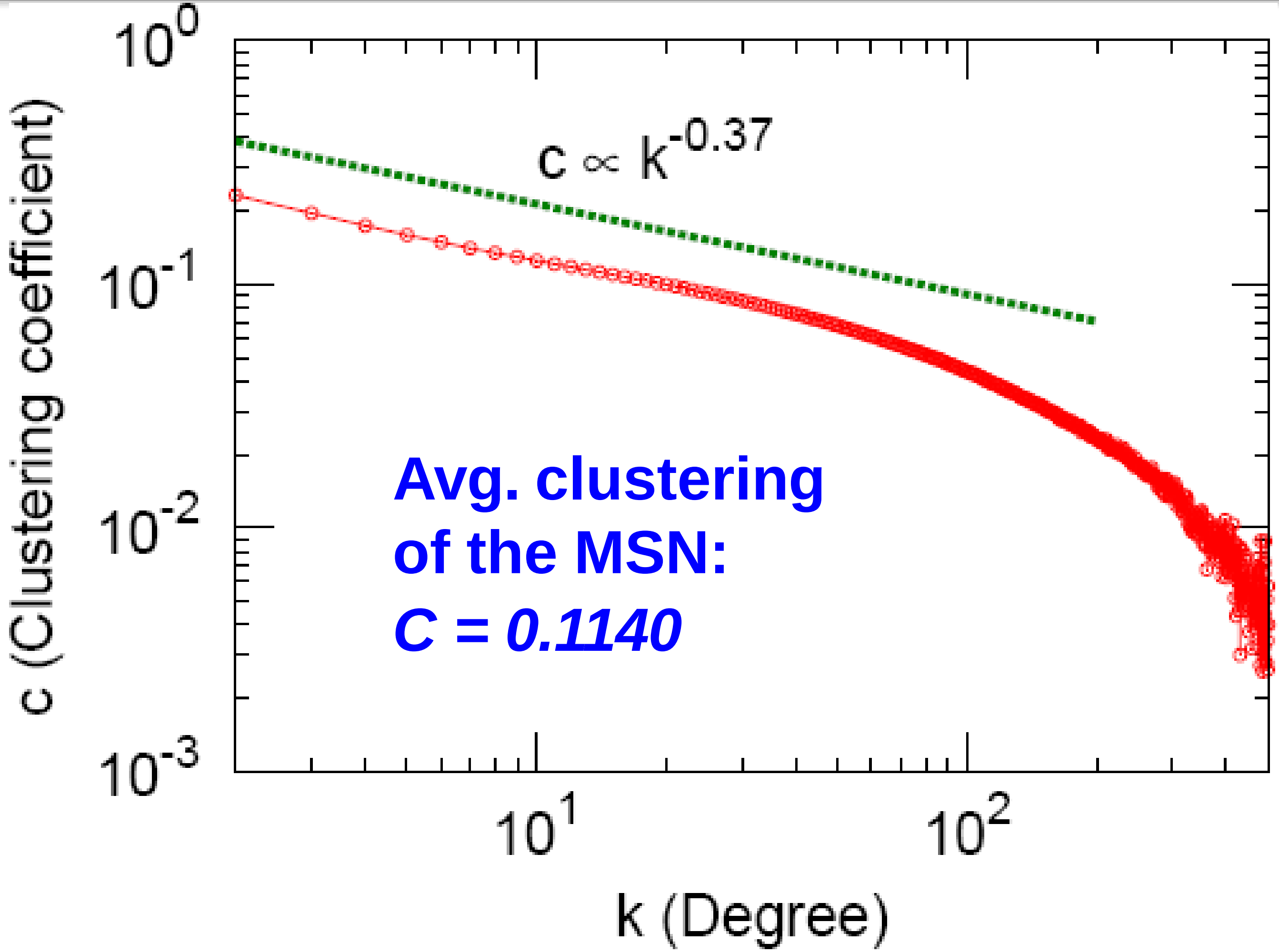
MSN (1) : Visualization of Degree Distribution



Log-Log Scale – Log-Log Plot

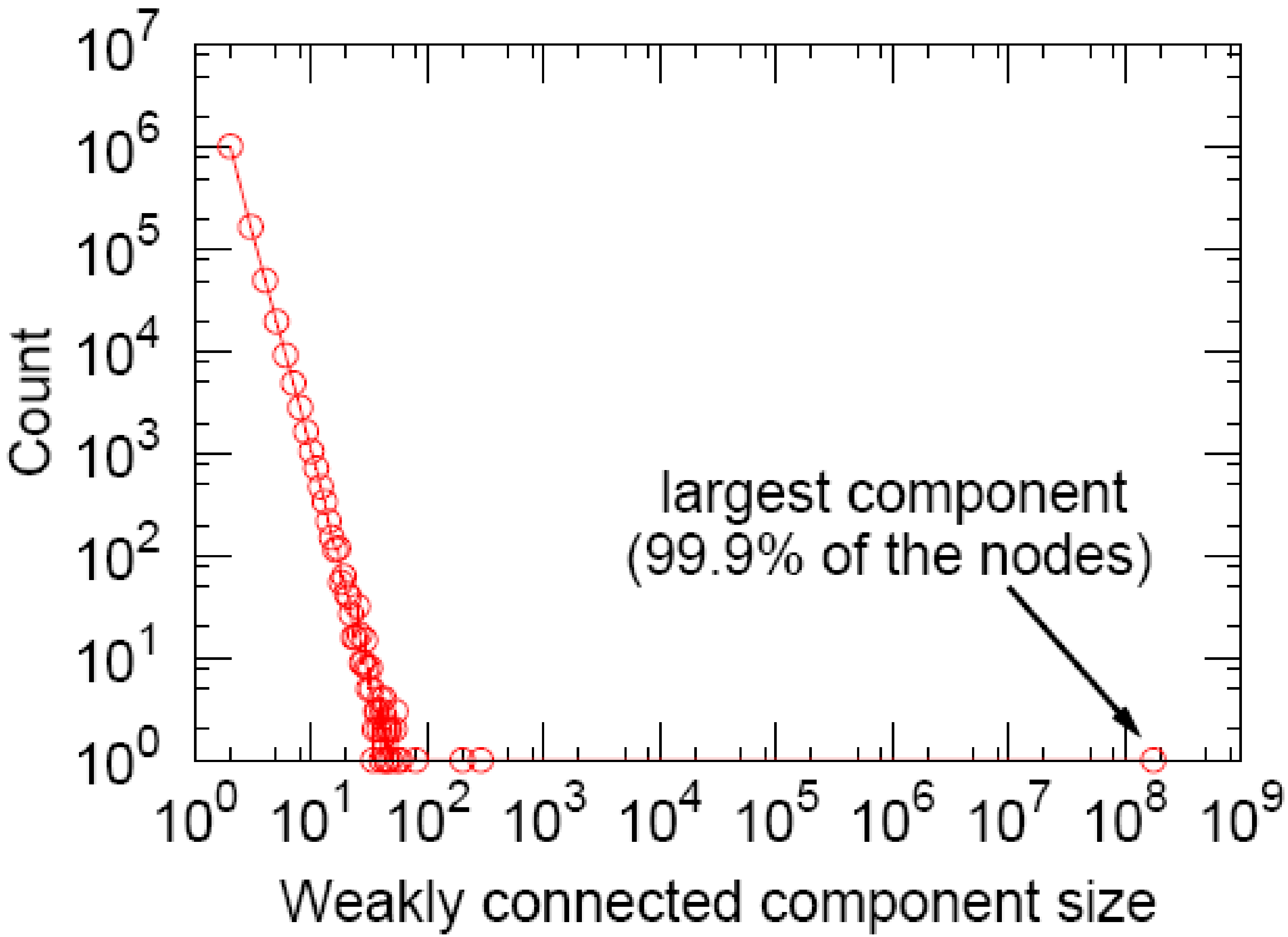
https://en.wikipedia.org/wiki/Log-log_plot

MSN (2) : Clustering

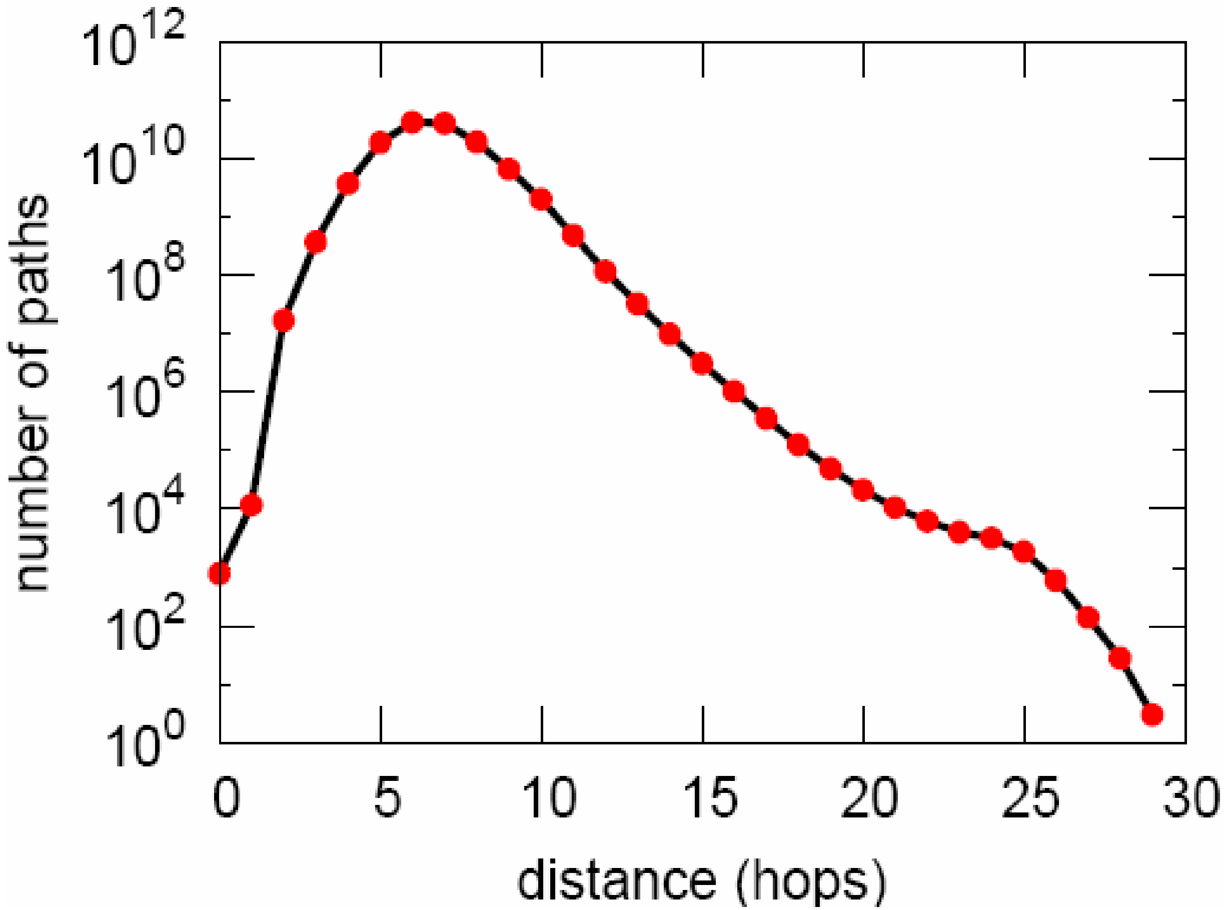


C_k: average C_i of nodes i of degree k: $C_k = \frac{1}{N_k} \sum_{i:k_i=k} C_i$

MSN (3) : Connected Components



MSN (4) : Diameter of WCC



Avg. path length **6.6**
90% of the nodes can be reached in < 8 hops

nodes as we do BFS out of a random node

Steps	#Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

MSN: Key Network Properties

Degree distribution: *Heavily skewed avg. degree= 14.4*

Path length: *6.6*

Clustering coefficient: *0.11*

Connectivity: *giant component*

Are these values "expected"?
Are they "surprising"?

To answer this we need a null-model!

Network Models

Random Graph Model
Small-World Model

Why should I use network models?

Facebook

May 2011:

- **721 millions** users.
- Average number of friends: **190**
- A total of **68.5 billion** friendships

September 2015:

- **1.35 Billion** users

1. What are the principal processes that help initiate these friendships?
2. How can these seemingly independent friendships form this complex friendship network?
3. In social media there are many networks with millions of nodes and billions of edges.
   - **They are complex and it is difficult to analyze them**

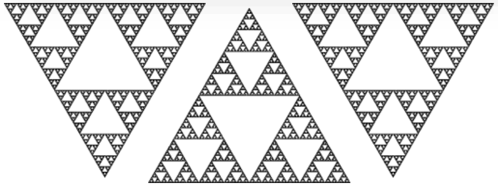
So, what do we do?

Design models that generate graphs

- The generated graphs should be similar to real-world networks.

If we can guarantee that generated graphs are similar to real-world networks in terms of graph properties:

1. We can analyze simulated graphs instead of real-networks (**cost-efficient**)
2. We can better understand real-world networks by providing concrete mathematical explanations; and
3. We can perform controlled experiments on synthetic networks when real-world networks are unavailable or sensitive.



Basic Intuition:

Hopefully! The complex output [social network] is generated by a simple process

Simplest Graph Model

- **Random Graph Model** [Erdös-Renyi, ‘60]
- **Two variants:**
  - $G_{n,p}$: undirected graph on n nodes and each edge (u,v) appears i.i.d. with probability p

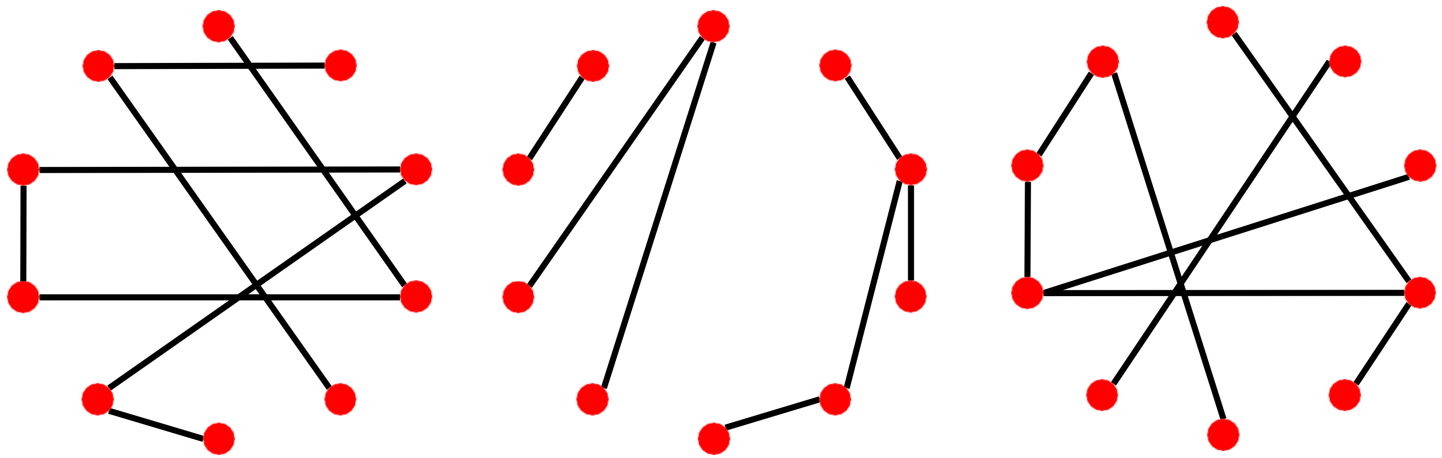
independently and identically distributed

  - $G_{n,m}$: undirected graph with n nodes, and m uniformly at random picked edges

What kind of networks do such models produce?

Random Graph Model

- ***n* and *p* do not uniquely determine the graph!**
  - The graph is a result of a random process
- We can have many different graphs given the same ***n*** and ***p***



n = 10
p= 1/6

Properties of G_{np}

Degree distribution: $P(k)$

Average Shortest Path length: h

Average Clustering coefficient: C

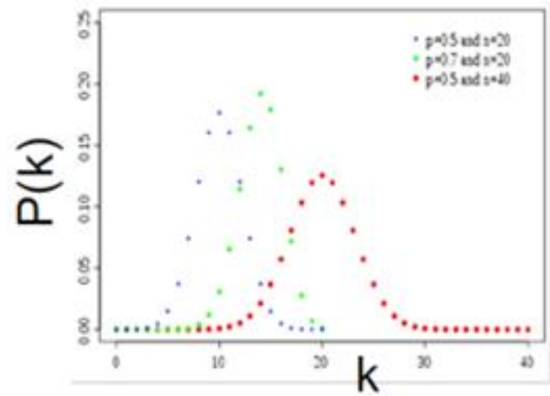
Size of giant component

What are the property values of G_{np}?

Degree Distribution of G_{np}

- **Fact:** **Degree distribution of G_{np} is binomial.**
- Let $P(k)$ denote the fraction of nodes with degree k:

$$P(k) = \underbrace{\binom{n-1}{k}}_{\text{Select } k \text{ nodes out of } n-1} \underbrace{p^k}_{\text{Probability of having } k \text{ edges}} \underbrace{(1-p)^{n-1-k}}_{\text{Probability of missing the rest of the } n-1-k \text{ edges}}$$



Mean, variance of a binomial distribution

$$c = \bar{k} = p(n-1)$$

$$\sigma^2 = p(1-p)(n-1)$$

https://en.wikipedia.org/wiki/Binomial_distribution

Clustering Coefficient of G_{np}

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

Where e_i is the number of edges between i's neighbors

- Edges in G_{np} appear i.i.d. with prob. p
- **So, expected E[e_i] is:** $= p\frac{k_i(k_i-1)}{2}$

Each pair is connected with prob. p

Number of distinct pairs of neighbors of node i of degree k_i

- **Then *E[C]*:** $\frac{2\,\mathrm{E}[e_i]}{k_i(k_i-1)} = \frac{p \cdot k_i(k_i-1)}{k_i(k_i-1)} = p = \frac{\bar{k}}{n-1} \approx \frac{\bar{k}}{n}$

Clustering coefficient of a random graph is small.
If we generate bigger and bigger graphs with fixed avg. degree k (that is we set $p = k \cdot 1/n$), then C decreases with the graph size n.

Network Properties

Degree distribution: $$P(k)=\binom{n-1}{k}p^k(1-p)^{n-1-k}$$

Clustering coefficient: $$C=p=\bar{k}/n$$

Path length: *next!*

Connectivity:

The Average Shortest Path Length

The average path length in a random graph is $h = O(\ln |V|)$

Proof.

- Assume D is the expected diameter of the graph
- Starting with any node and the expected degree c,
  - one can visit approximately c nodes by traveling one edge
  - c^2 nodes by traveling 2 edges, and
  - c^D nodes by traveling diameter number of edges
- We should have visited all nodes $c^{\mathcal{D}} \approx |V|$
- The expected diameter size tends to be twice of the average path length h

$$c^{\mathcal{D}} \approx c^{2h} \approx |V| \quad \Rightarrow \quad 2h \approx \frac{\ln |V|}{\ln c} \qquad h = O(\ln |V|)$$

Network Properties

Degree distribution: $$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

Path length: $$O(\log n)$$

Clustering coefficient: $$C = p = \bar{k} / n$$

Connected components: *next!*

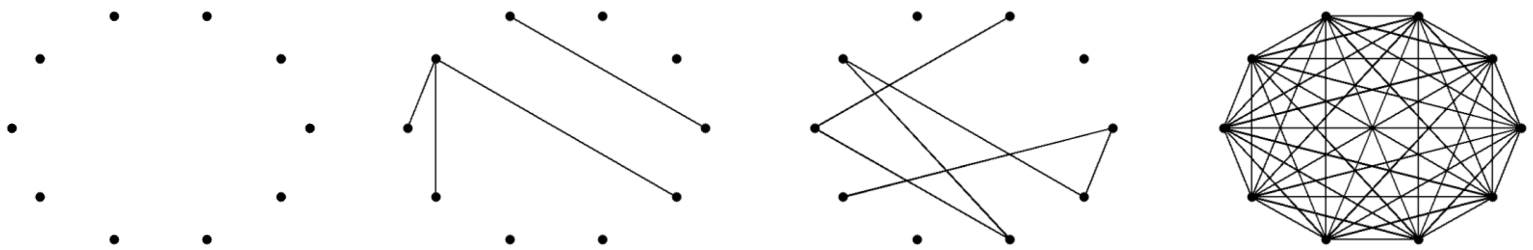
The Giant Component

- In random graphs, as we increase p, a large fraction of nodes start getting connected
  - i.e., we have a path between any pair

- This large fraction forms a connected component:
  - **Largest connected component**, also known as the **Giant component, will appear when p is big enough**

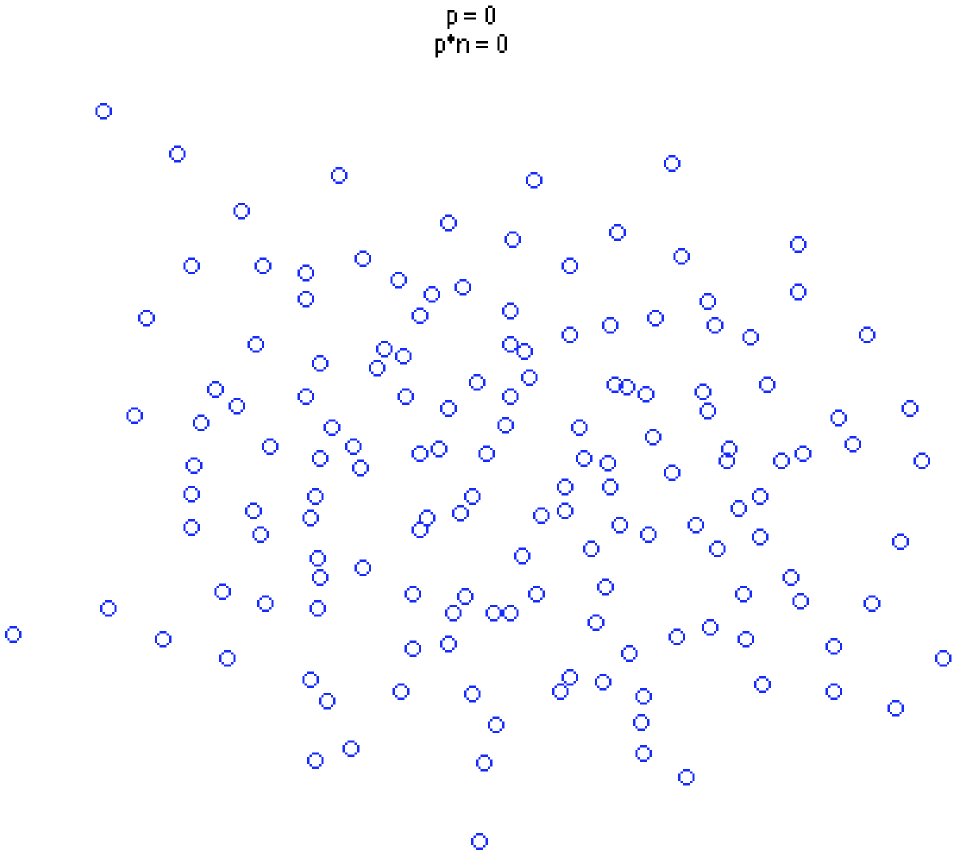
- In random graphs:
  - $p = 0$
    - the size of the giant component is 0
  - $p = 1$
    - the size of the giant component is n

The Giant Component



Probability (p)	0.0	0.088	0.11 (=1/(n-1)=1/9)	1.0
Average Node Degree (c)	0.0	0.8	≈1	n-1=9
Diameter	0	2	6	1
Giant Component Size	0	4	7	10
Average Path Length	0.0	1.5	2.66	1.0

Demo ($n = 150$)



When p reaches ~ 1/149, the giant component appears

From *David Gleich*

1st Phase Transition (Rise of the Giant Component)

- **Phase Transition**: the point where diameter value starts to shrink in a random graph

- The phase transition we focus on happens when
  - average node degree $c = 1$ (or when $p = 1/(n-1)$)

- At this Phase Transition:
  1. The giant component, which just started to appear, starts to grow, and
  2. The diameter, which *just* reached its maximum value, starts decreasing.

Random Graphs

c – average degree

If $c < 1$:

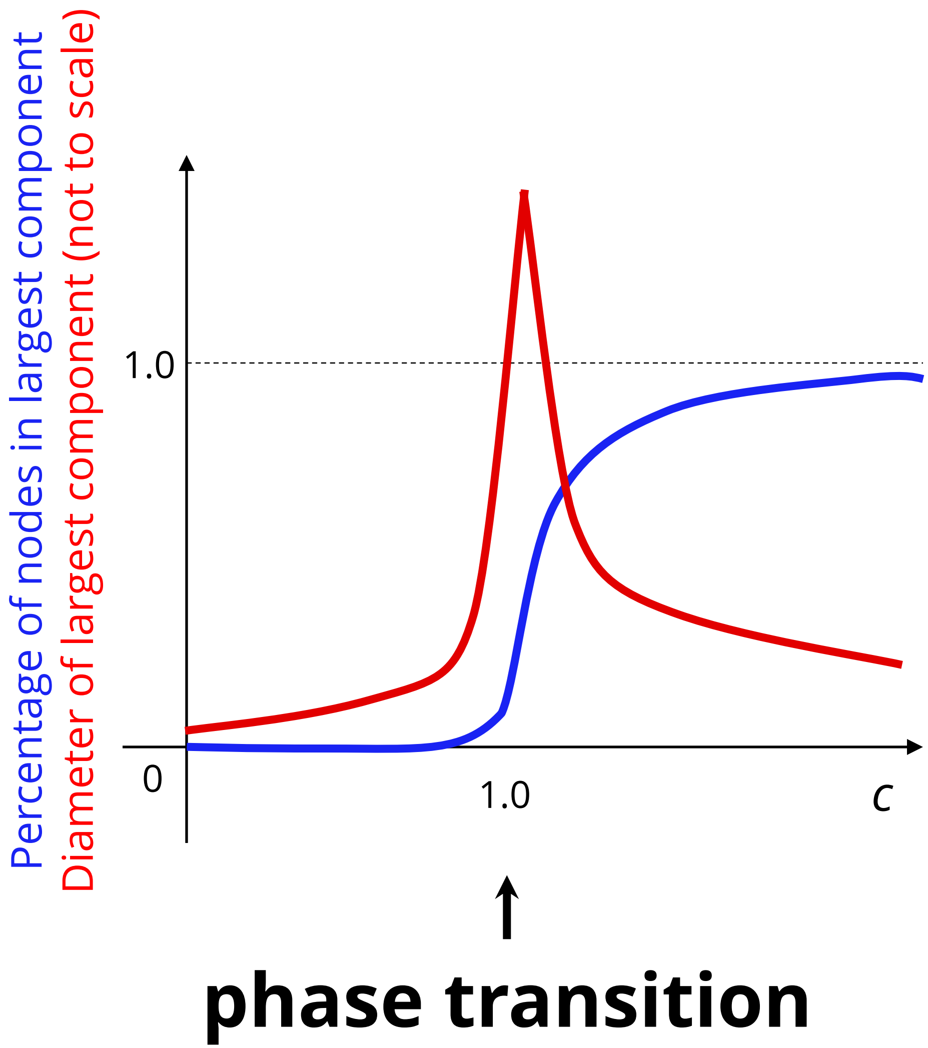
- **small**, isolated clusters
- **small** diameters
- **short** path lengths

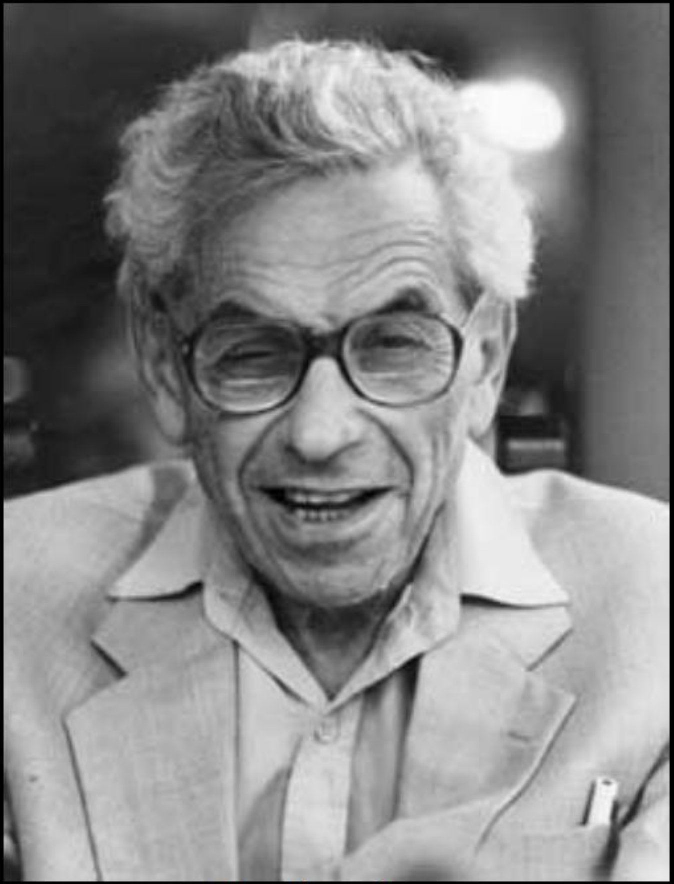
At $c = 1$:

- a ***giant component*** appears
- diameter **peaks**
- path lengths are **long**

For $c > 1$:

- almost **all** nodes **connected**
- diameter **shrinks**
- path lengths **shorten**

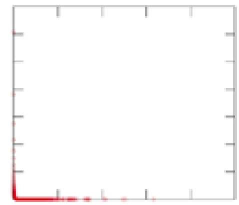
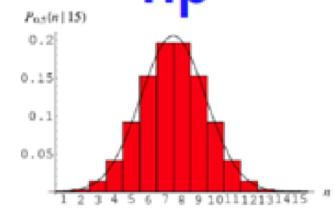




Paul Erdös

G_{np} is so cool!
Let’s compare it to real networks.

Back to MSN Vs. G_{np}

	MSN	G_{np}	n=180M
Degree distribution:			✗
Avg. path length:	**6.6**	$O(\log n)$ $h \approx 8.2$	✓
Avg. clustering coef.:	*0.11*	$\bar{k} / n$ $C \approx 8 \cdot 10^{-8}$	✗
Largest Conn. Comp.:	*99%*	GCC exists when $\bar{k}$>1. $\bar{k} \approx 14$.	✓

Note: the average degree of the random graph is equal to the average degree in MSN.

Real Networks Vs. G_{np}

- **Are real networks like random graphs?**
  - Giant connected component: ☺
  - Average path length: ☺
  - Clustering Coefficient: ☹
  - Degree Distribution: ☹
- **Problems with the random networks model:**
  - Degree distribution differs from that of real networks
  - No local structure – clustering coefficient is too low
- **Most important: Are real networks random?**
  - The answer is simply: **NO!**

References

- R. Zafarani, M. A. Abbasi, and H. Liu, Social Media Mining: An Introduction, Cambridge University Press, 2014.
- http://socialmediamining.info/
- Stanford CS224W Analysis of Networks