VE444: Networks

Yifei Zhu, assistant professor
University of Michigan-Shanghai Jiao Tong University

Acknowledgment:
This lecture's slides are modified from the slides shared by Prof. Jure Leskovec

QR code

VE444 2020FA

该二维码7天内(9月14日前)有效，重新进入将更新

Questions so far?

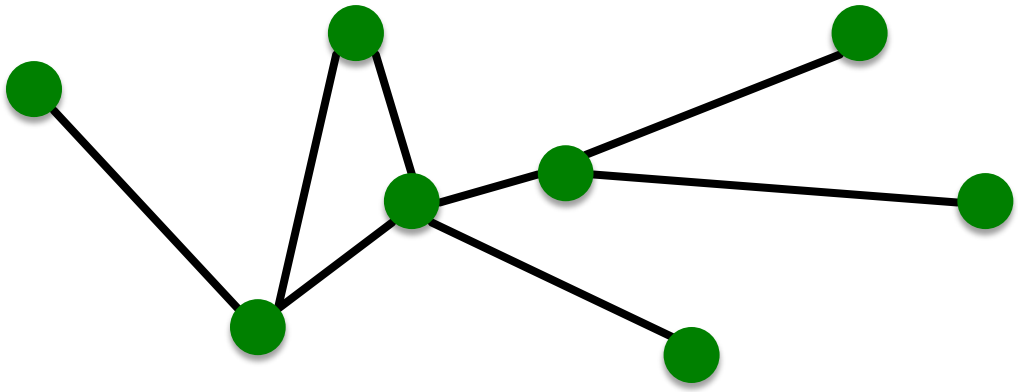
Graph Basics

Structure of Networks?

A network is a collection of objects where some pairs of objects are connected by links

What is the structure of the network?

Components of a Network



- **Objects:** nodes, vertices — N
- **Interactions:** links, edges — E
- **System:** network, graph — $G(N,E)$

Networks or Graphs?

- **Network** often refers to real systems
  - Web, Social network, Metabolic network

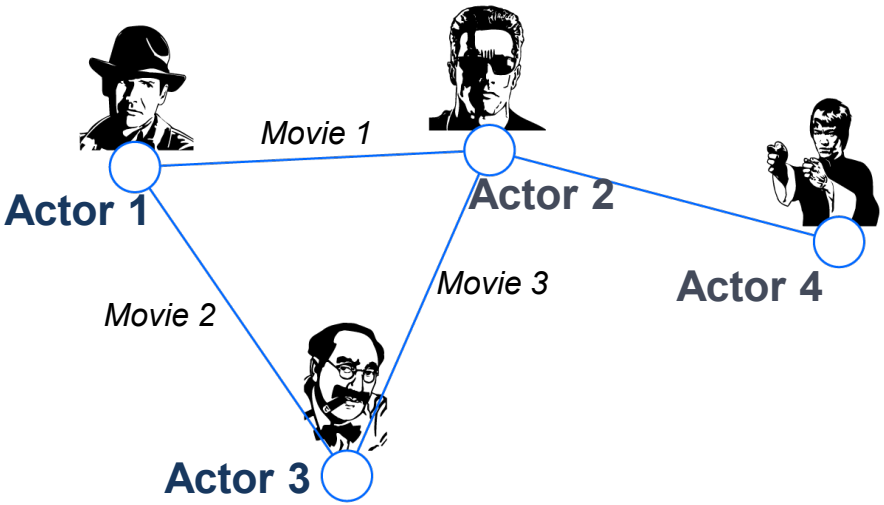
  Language: Network, node, link

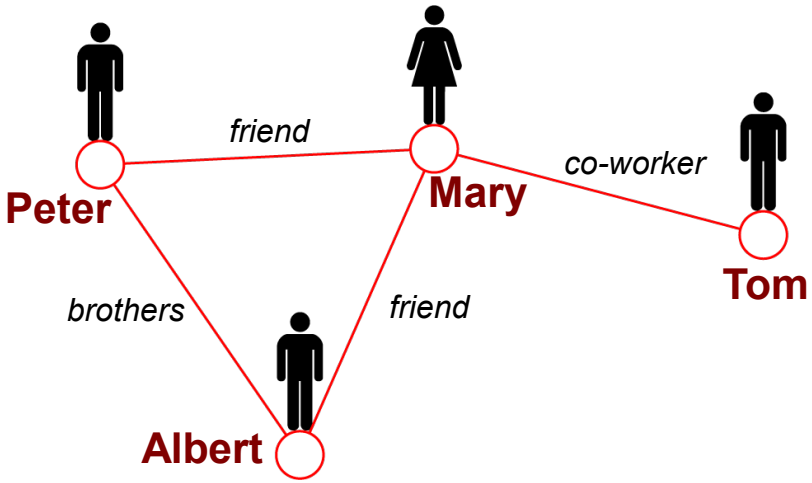
- **Graph** is a mathematical representation of a network
  - Web graph, Social graph, Knowledge Graph

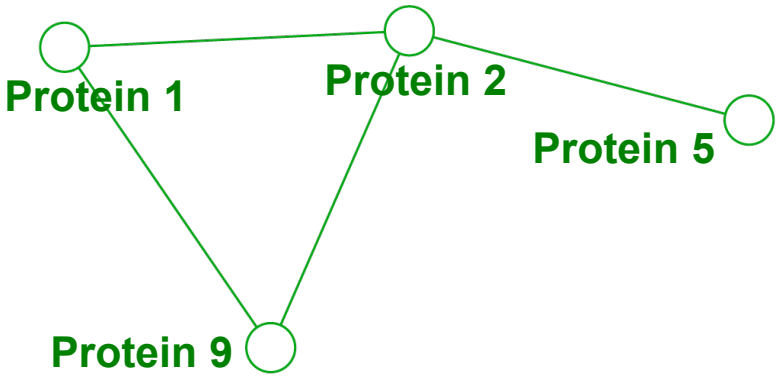
  Language: Graph, vertex, edge

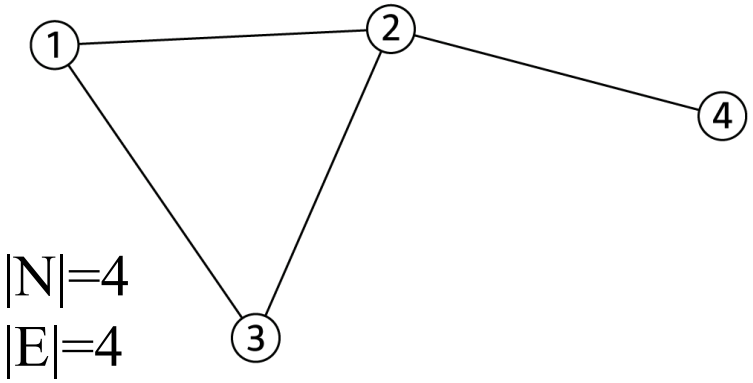
We will try to make this distinction whenever it is appropriate, but in most cases we will use the two terms interchangeably

Networks: Common Language



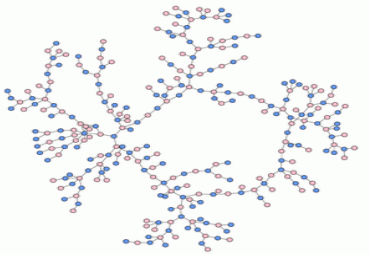


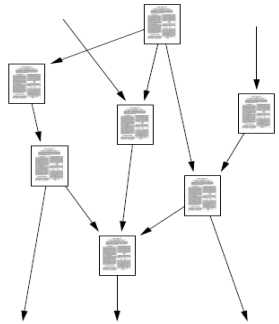




Choosing Proper Representations

- If you connect individuals that work with each other, you will explore a **professional network**
- If you connect those that have a sexual relationship, you will be exploring **sexual networks**
- If you connect scientific papers that cite each other, you will be studying the **citation network**





- **If you connect all papers with the same word in the title, what will you be exploring?** It is a network, nevertheless

How do you define a network?

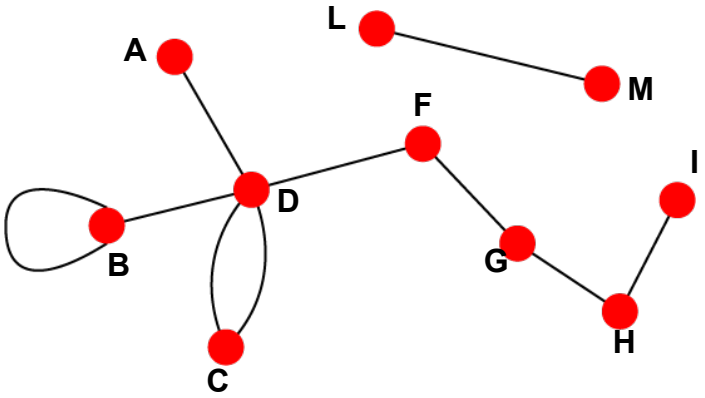
- **How to build a graph:**
  - **What are nodes?**
  - **What are edges?**
- **Choice of the proper network representation of a given domain/problem determines our ability to use networks successfully:**
  - In some cases there is a unique, unambiguous representation
  - In other cases, the representation is by no means unique
  - The way you assign links will determine the nature of the question you can study

Choice of Network Representation

Directed vs. Undirected Graphs

Undirected

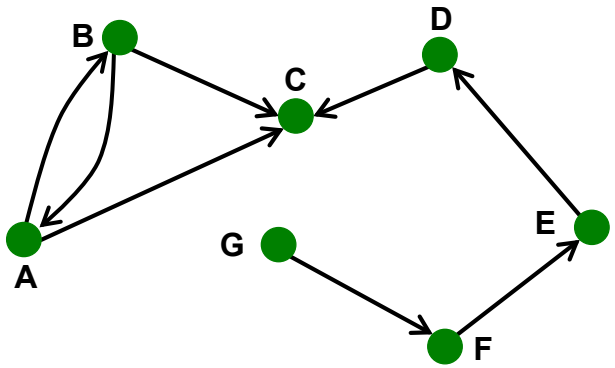
- **Links:** undirected (symmetrical, reciprocal)



- **Examples:**
  - Collaborations
  - Friendship on Facebook

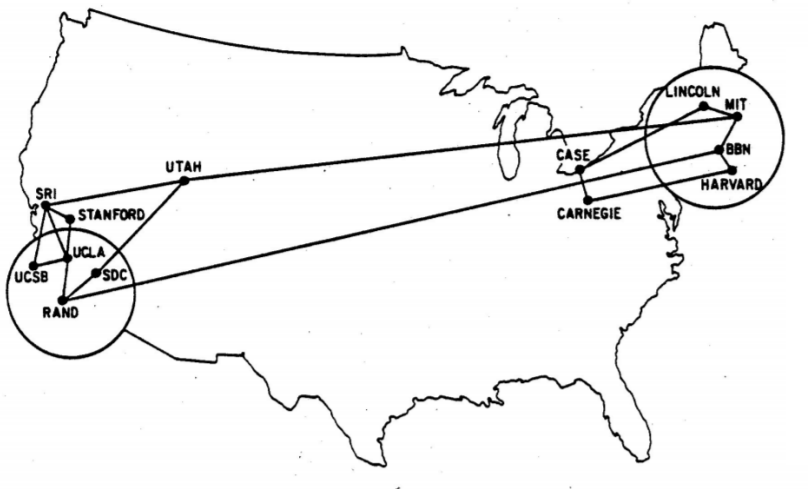
Directed

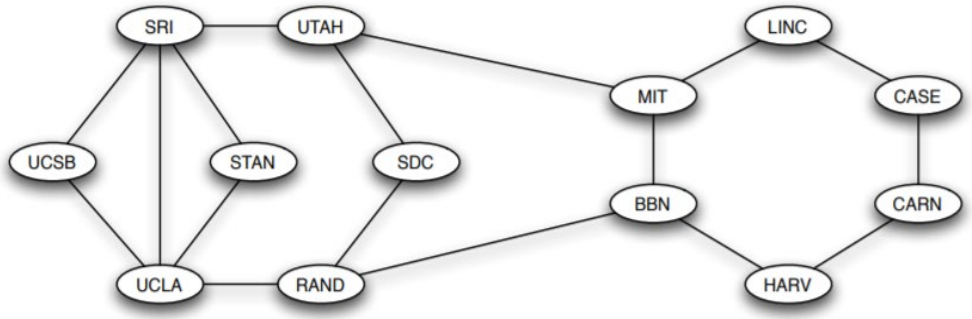
- **Links:** directed (arcs)



- **Examples:**
  - Phone calls
  - Following on Twitter

Node placement





We usually care about the connectedness only.

Edge Attributes

Possible options:

- Weight (e.g. frequency of communication)
- Ranking (best friend, second best friend…)
- Type (friend, relative, co-worker)
- Sign: Friend vs. Foe, Trust vs. Distrust
- Properties depending on the structure of the rest of the graph: number of common friends

Network Representations

Email network >> directed multigraph with self-edges

Facebook friendships >> undirected, unweighted

Citation networks >> unweighted, directed, acyclic

Collaboration networks >> undirected multigraph or weighted graph

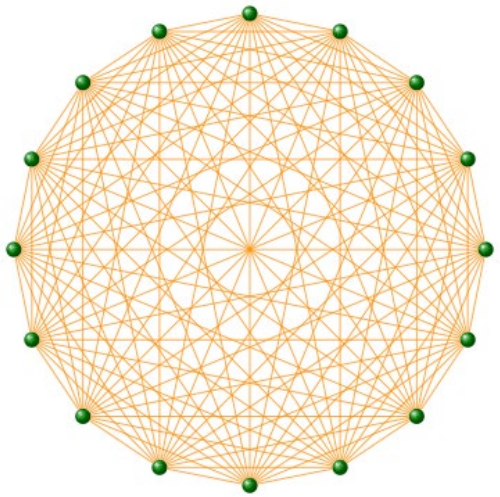
Mobile phone calls >> directed, (weighted?) multigraph

Protein Interactions >> undirected, unweighted with self-interactions

Complete Graph

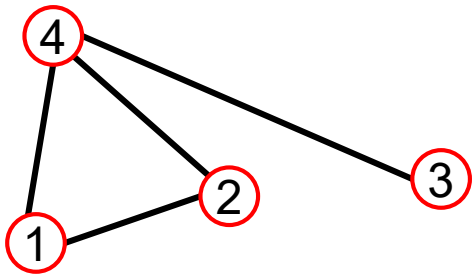
The **maximum number of edges** in an undirected graph on N nodes is

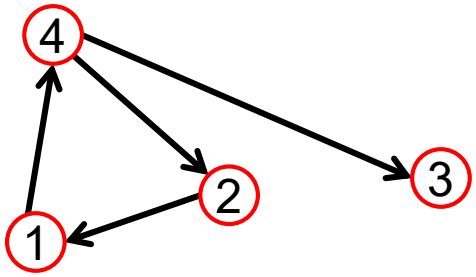
$$E_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$$



An undirected graph with the number of edges $E = E_{max}$ is called a **complete graph**, and its average degree is $N-1$

Representing Graphs: Adjacency Matrix





$A_{ij} = 1$ if there is a link from node ***i*** to node ***j***

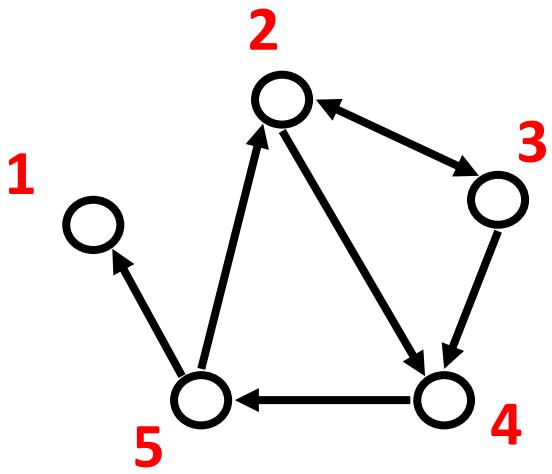
$A_{ij} = 0$ otherwise

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

Note that for a directed graph (right) the matrix is not symmetric.

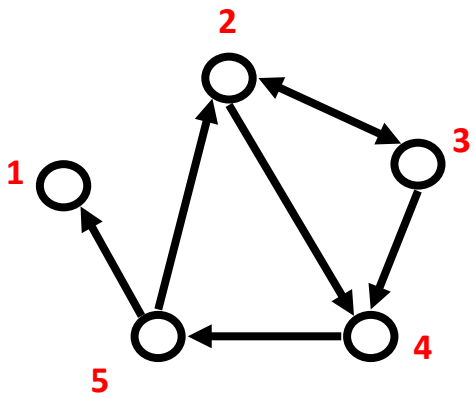
Representing Graphs: Edge list

- **Represent graph as a set of edges:**
  - (2, 3)
  - (2, 4)
  - (3, 2)
  - (3, 4)
  - (4, 5)
  - (5, 2)
  - (5, 1)



Representing Graphs: Adjacency list

- **Adjacency list:**
  - Easier to work with if network is
    - Large
    - Sparse
  - Allows us to quickly retrieve all neighbors of a given node
    - 1:
    - 2: 3, 4
    - 3: 2, 4
    - 4: 5
    - 5: 1, 2



Networks are Sparse Graphs

Most real-world networks are sparse

$$E << E_{max} \quad (\text{or } \bar{k} << N-1)$$

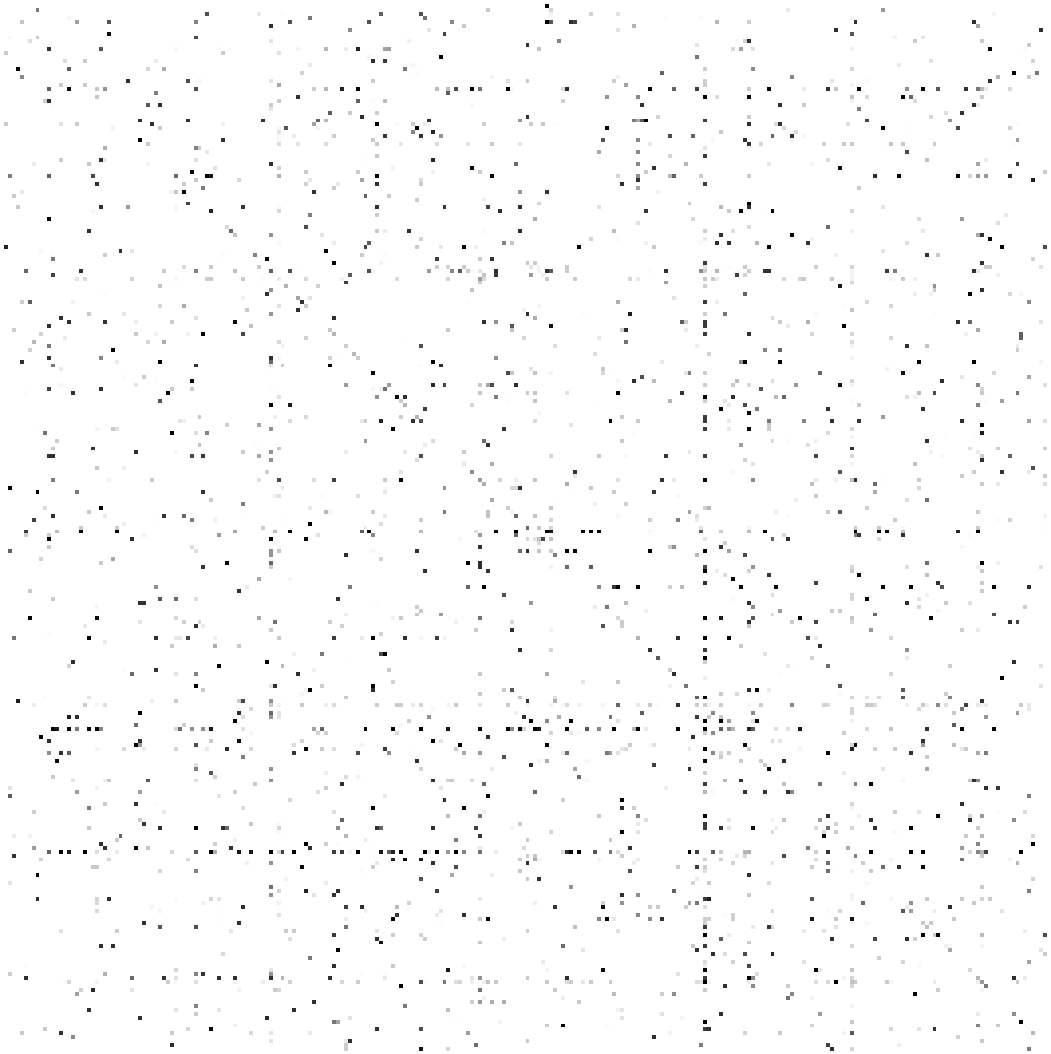
WWW (Stanford-Berkeley):	N=319,717	⟨k⟩=9.65
Social networks (LinkedIn):	N=6,946,668	⟨k⟩=8.87
Communication (MSN IM):	N=242,720,596	⟨k⟩=11.1
Coauthorships (DBLP):	N=317,080	⟨k⟩=6.62
Internet (AS-Skitter):	N=1,719,037	⟨k⟩=14.91
Roads (California):	N=1,957,027	⟨k⟩=2.82
Proteins (S. Cerevisiae):	N=1,870	⟨k⟩=2.39

(Source: *Leskovec et al., Internet Mathematics, 2009*)

Consequence: Adjacency matrix is filled with zeros!

(**Density of the matrix (E/N^2):** WWW=1.51×10^{-5}, MSN IM = 2.27×10^{-8})

Adjacency Matrices are Sparse



Network Properties: How to Measure a Network

Plan: Key Network Properties

Degree distribution: $P(k)$

Path length: h

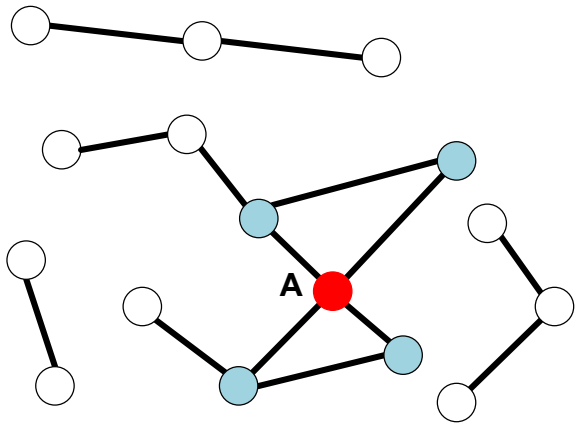
Clustering coefficient: C

Connected components: s

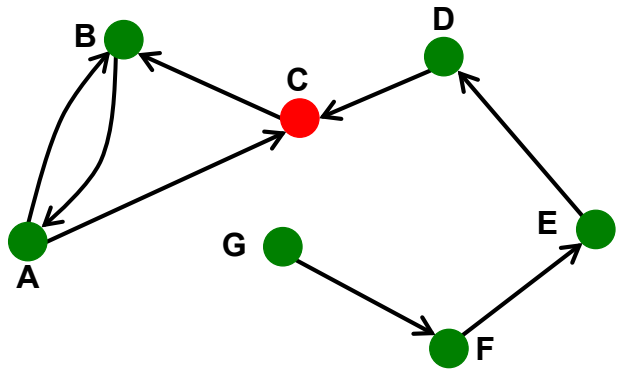
Definitions will be presented for undirected graphs, sometimes we will explicitly mention extensions to directed graphs, and sometimes extensions will be obvious

(1)Node Degrees

Undirected



Directed



Source: Node with $k^{in} = 0$

Sink: Node with $k^{out} = 0$

Node degree, k_i: the number of edges adjacent to node i

$$k_A = 4$$

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

In directed networks we define an **in-degree** and **out-degree.**

The (total) degree of a node is the sum of in- and out-degrees.

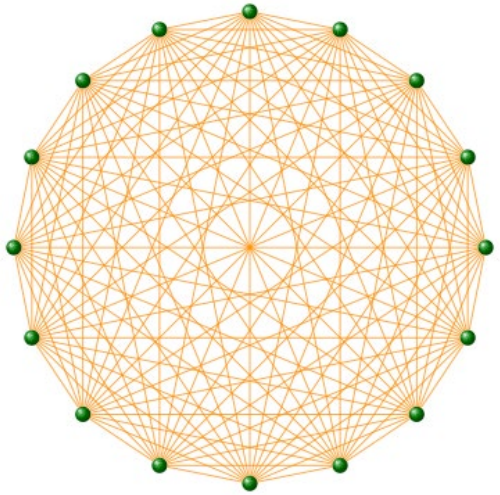
$$k_C^{in} = 2 \qquad k_C^{out} = 1 \qquad k_C = 3$$

$$\overline{k^{in}} = \overline{k^{out}} = \frac{E}{N}$$

Average degree of a complete Graph?

The **maximum number of edges** in an undirected graph on N nodes is

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



An undirected graph with the number of edges $E = E_{max}$ is called a **complete graph**, and its average degree is ?

Directedness & Average Degrees

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

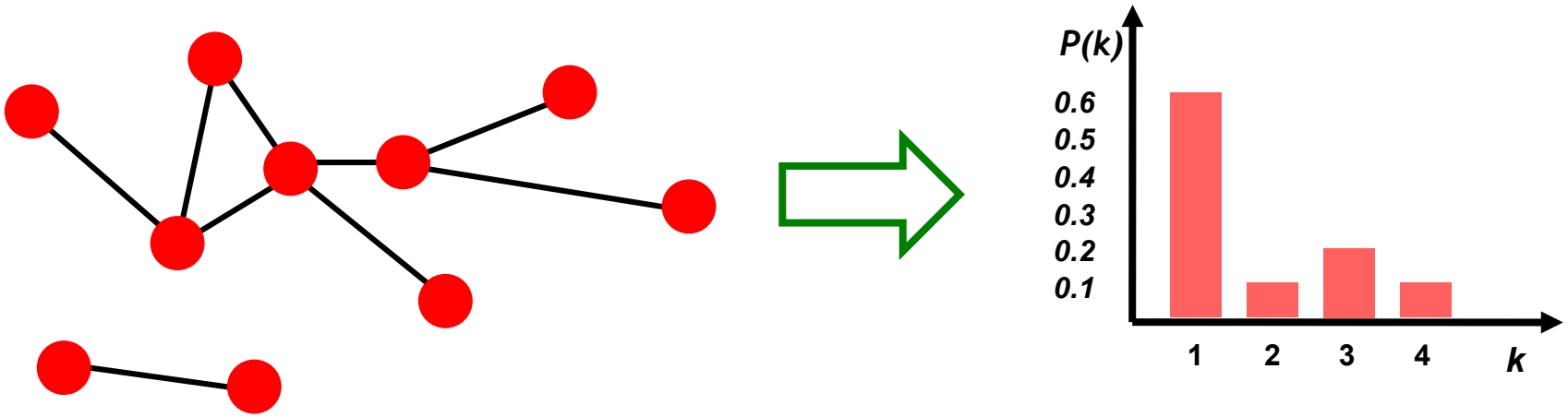
Degree Distribution

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

  N_k = # nodes with degree k

- Normalized histogram:

  $P(k) = N_k / N$ → **plot**

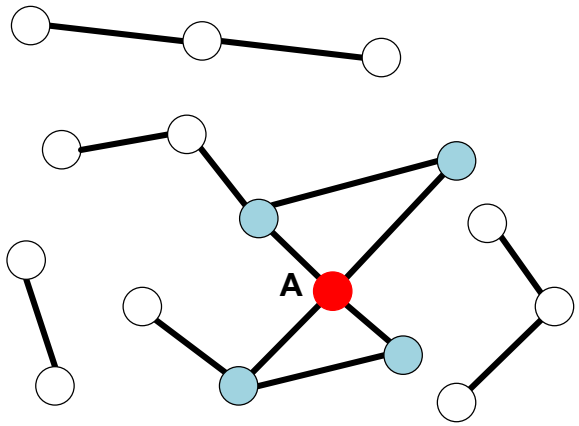


For directed graphs we have separate in- and out-degree distributions.

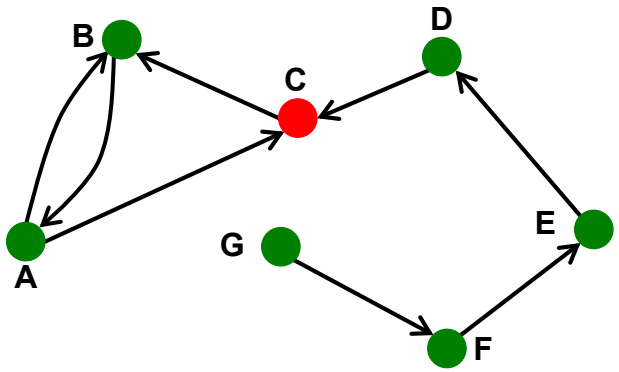
Questions so far?

Why degree

Undirected



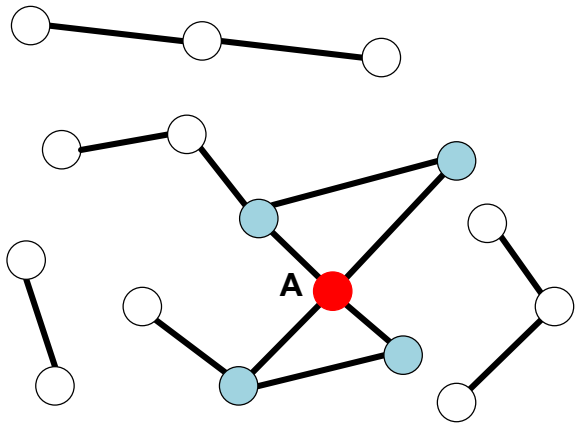
Directed



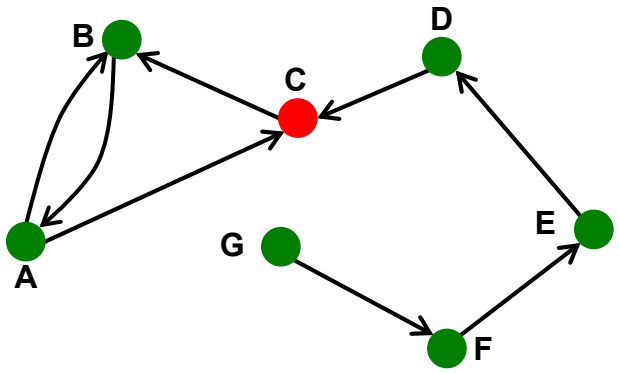
- **Degree** is the simplest, yet very illuminating centrality measure in a network
  - In a social network, the ones who have connections to many others might have more influence, more access to information.
- **Degree** represent the immediate risk/benefit of a node for catching whatever is flowing through the network (virus, information, etc.)

Why degree

Undirected



Directed



- In a social network, the ones who have connections to many others might have more influence, more access to information

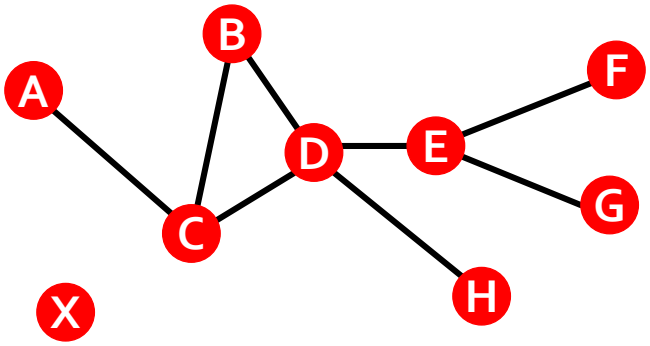
Or is it? ☺ (Spoiler)

(2) Paths in a Graph

- A ***path*** is a sequence of nodes in which each node is linked to the next one

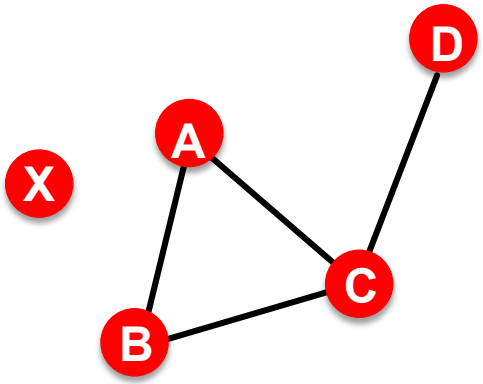
$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$

- A path can intersect itself and pass through the same edge multiple times
  - E.g.: ACBDCDEG

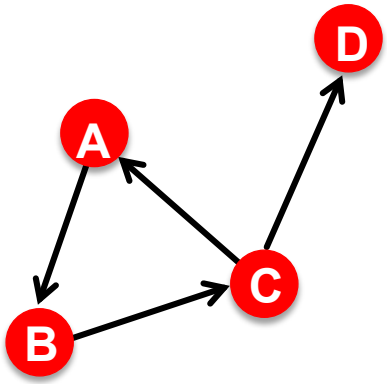


- **Length** is the number of edges in the sequence that comprises a path

Distance in a Graph



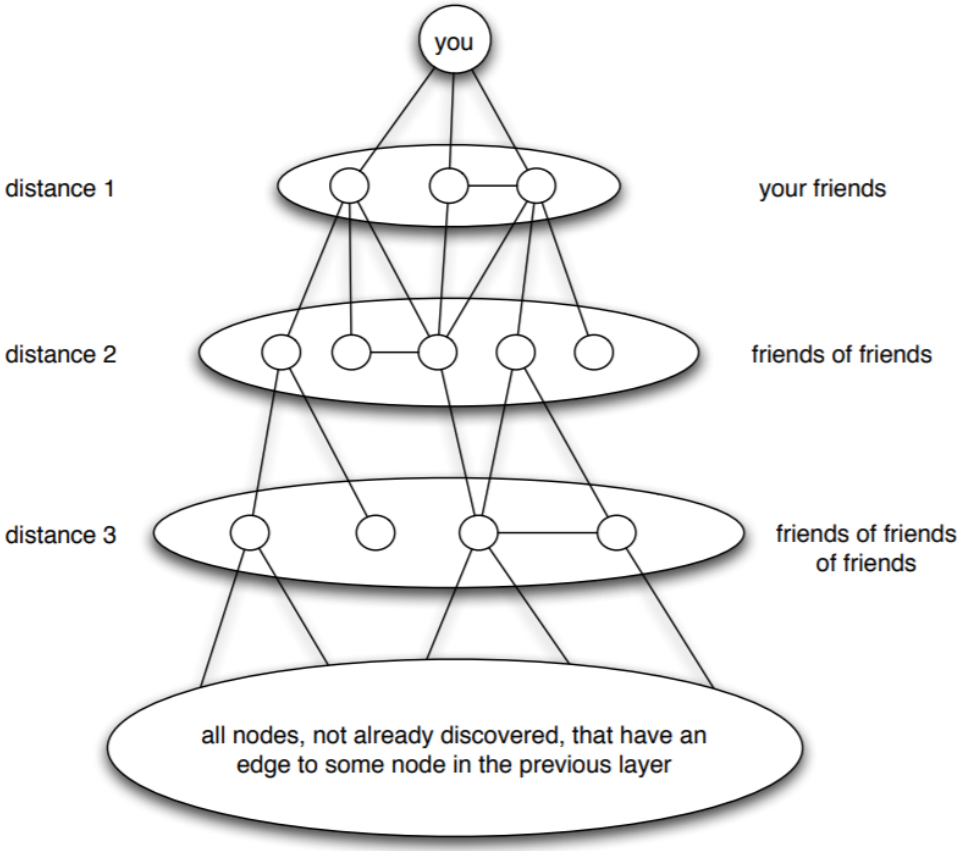
$h_{B,D} = 2$
$h_{A,X} = \infty$



$h_{B,C} = 1, h_{C,B} = 2$

- **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 (or zero)
- In **directed graphs,** paths need to follow the direction of the arrows
  - Consequence: Distance is **not symmetric**: $h_{B,C} \neq h_{C,B}$
- **Diameter:** The maximum (shortest path) distance between any pair of nodes in a graph

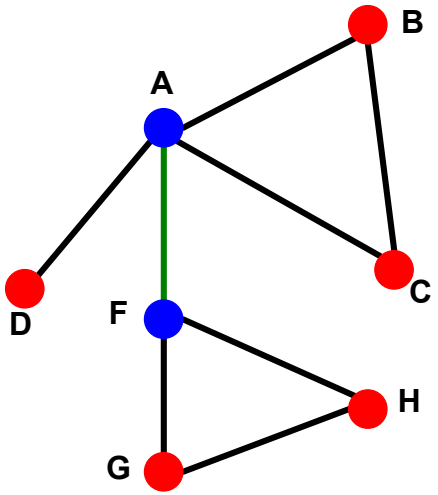
How to calculate the distance of a node to the rest in large scale?

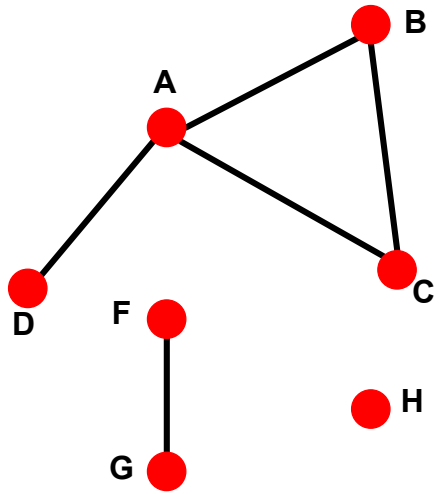


What is this algorithm?

Connectivity of Undirected Graphs

- **Connected (undirected) graph:**
  - Any two vertices can be joined by a path
- A disconnected graph is made up by two or more connected components





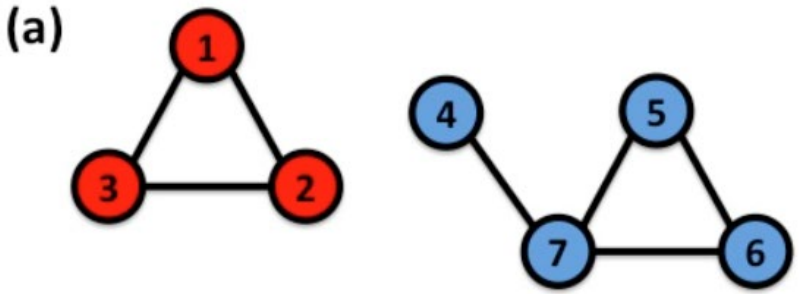
Largest Component:
Giant Component

Isolated node (node H)

Connectivity: Example

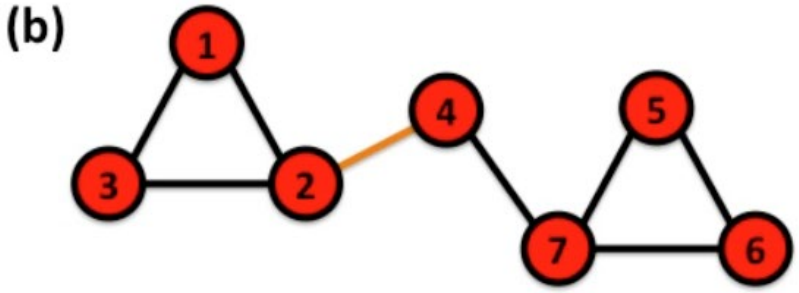
- The adjacency matrix of a network with several components can be written in a block- diagonal form, so that nonzero elements are confined to squares, with all other elements being zero:

Disconnected



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

Connected



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

(3) Clustering Coefficient

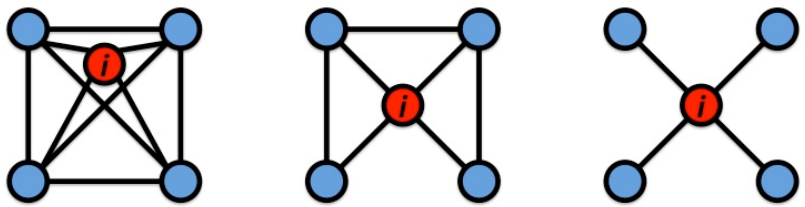
- **Clustering coefficient (for undirected graphs):**
  - How connected are i's neighbors to each other?
  - Clustering coefficient of A = the probability of any two nodes who are neighbors of A are also neighbors = the friend-pairs of A / total pairs
  - Or in a rigorous way…

(3) Clustering Coefficient

- **Clustering coefficient (for undirected graphs):**
  - How connected are ***i***'s neighbors to each other?
  - Node ***i*** with degree k_i
  - $C_i \in [0,1]$
  - $C_i = \dfrac{2e_i}{k_i(k_i-1)}$

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

Note $k_i(k_i-1)/2$ is max number of edges between the k_i neighbors



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

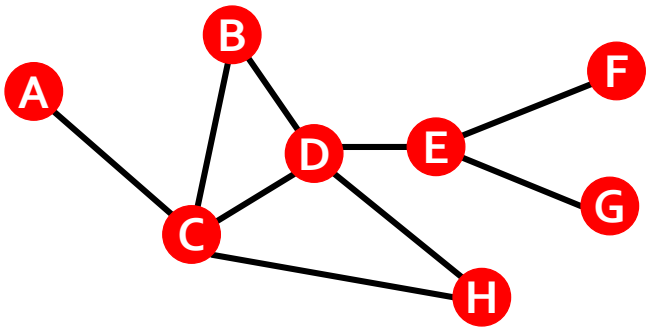
Clustering coefficient is undefined (or defined to be 0) for nodes with degree 0 or 1

- **Average clustering coefficient:** $C = \dfrac{1}{N}\sum_i^N C_i$

Clustering Coefficient

- **Clustering coefficient (for undirected graps):**
  - How connected are ***i***'s neighbors to each other?
  - Node ***i*** with degree k_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



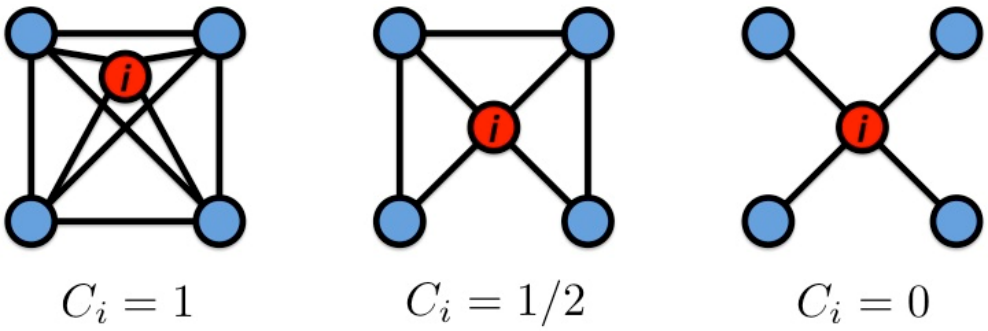
$k_B=2,\ e_B=1,\ C_B=2/2 = 1$

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

Avg. clustering: C=0.33

Why Clustering Coefficient?

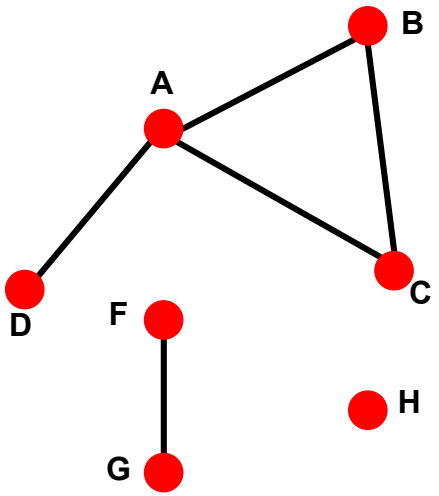
- **Clustering coefficient is**
  - a measure of the degree to which nodes in a graph tend to cluster together
    - from social science point of view, is how coerce this person can be
    - a type of centrality measure how powerful an individual is
  - can be used as a probe for the existence of structural holes in a networks
    - missing links between neighbors of a person



Questions so far?

(4) 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 that BFS visits
- If all nodes are visited, the network is connected
- Otherwise find an unvisited node and repeat BFS

280? 281?

Summary: Key Network Properties

Degree distribution: $P(k)$

Path length: h

Clustering coefficient: C

Connected components: s

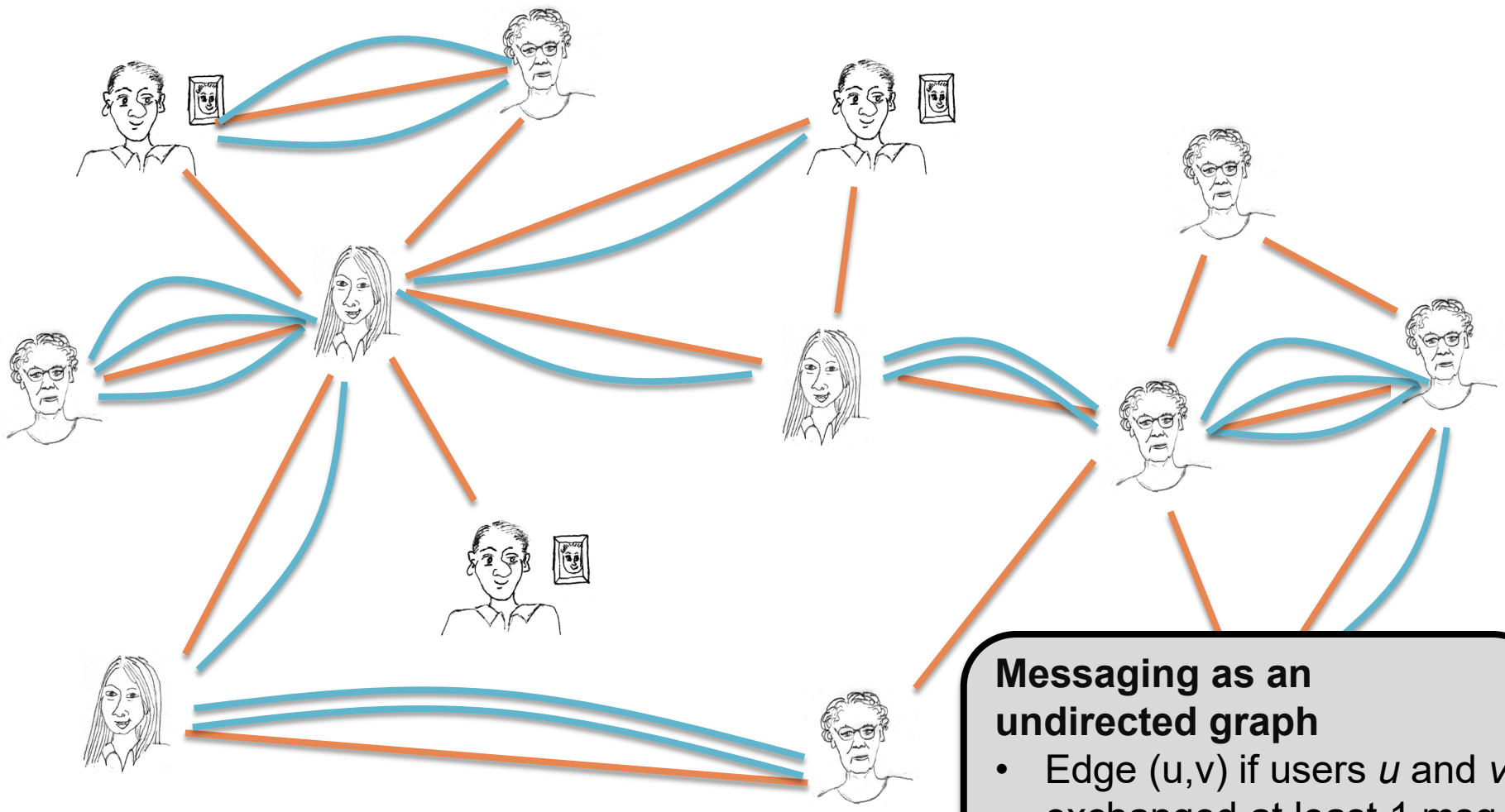
Let's measure these properties on real-world networks!

MSN Messenger

MSN Messenger:

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

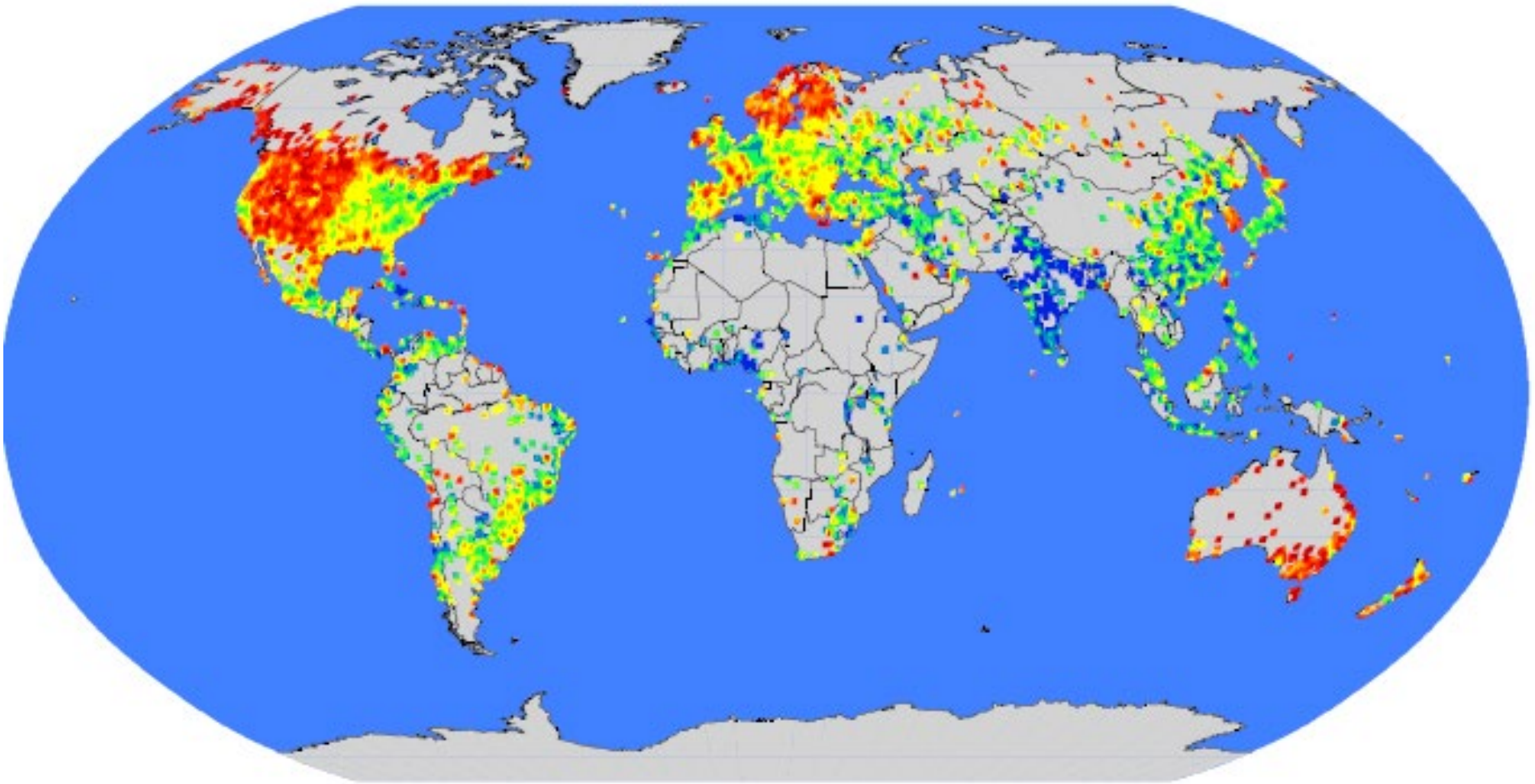
Messaging as a Multigraph



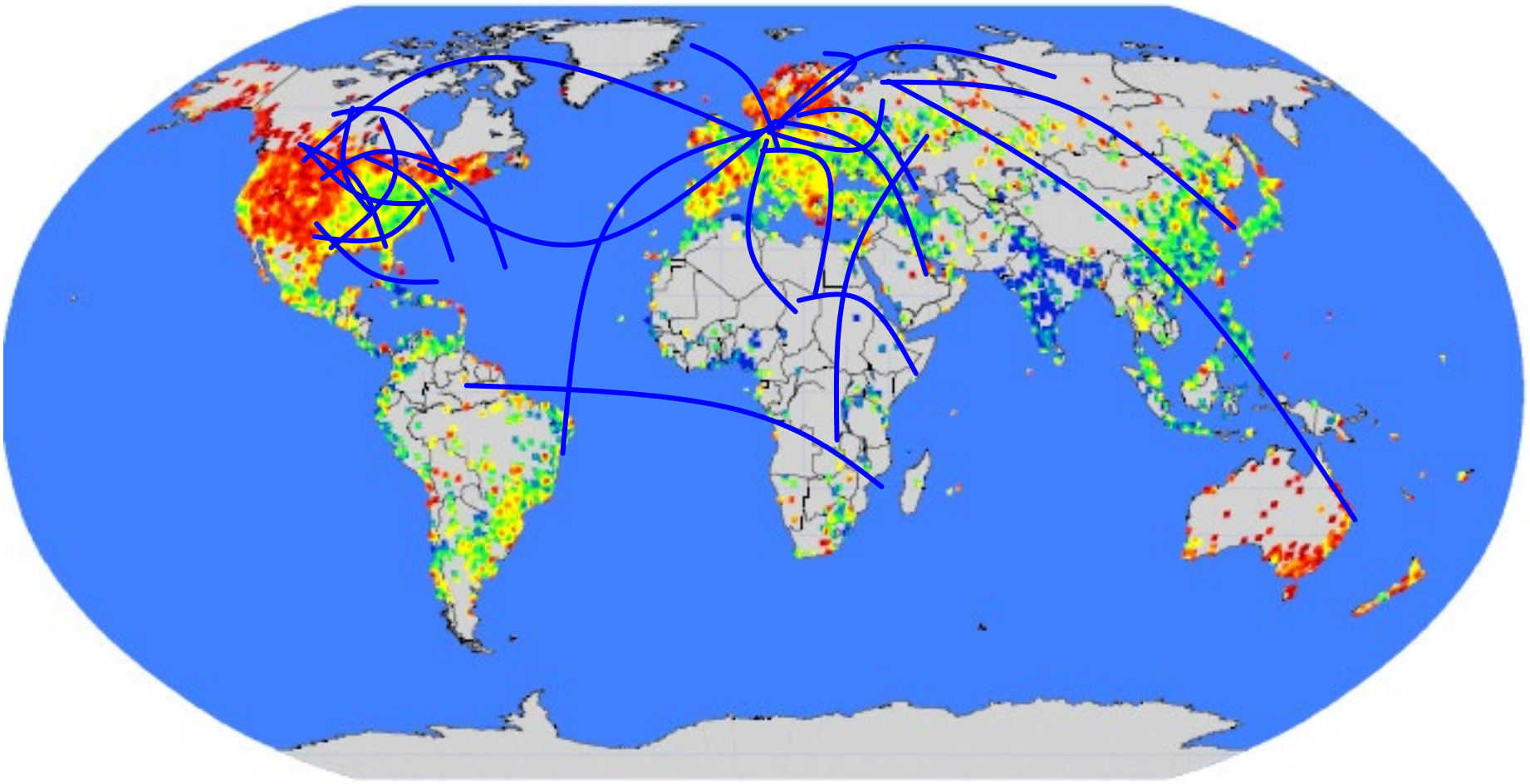
Messaging as an undirected graph

- Edge (u,v) if users *u* and *v* exchanged at least 1 msg
- N=180 million people
- E=1.3 billion edges

Geography of Communication

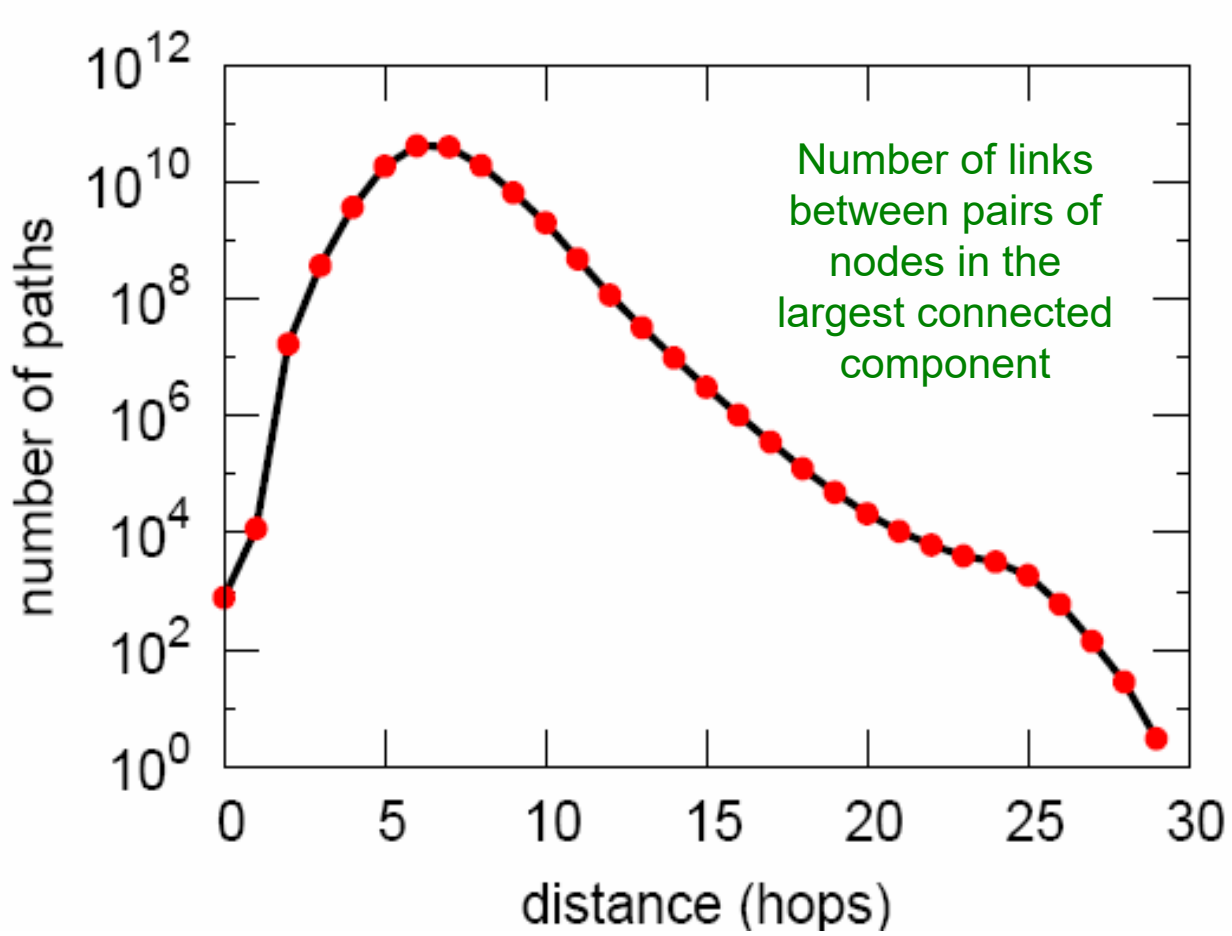


Communication Network



Network: 180M people, 1.3B edges

MSN: 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

So what?

So what?

Erdös-Rényi Random Graph Model

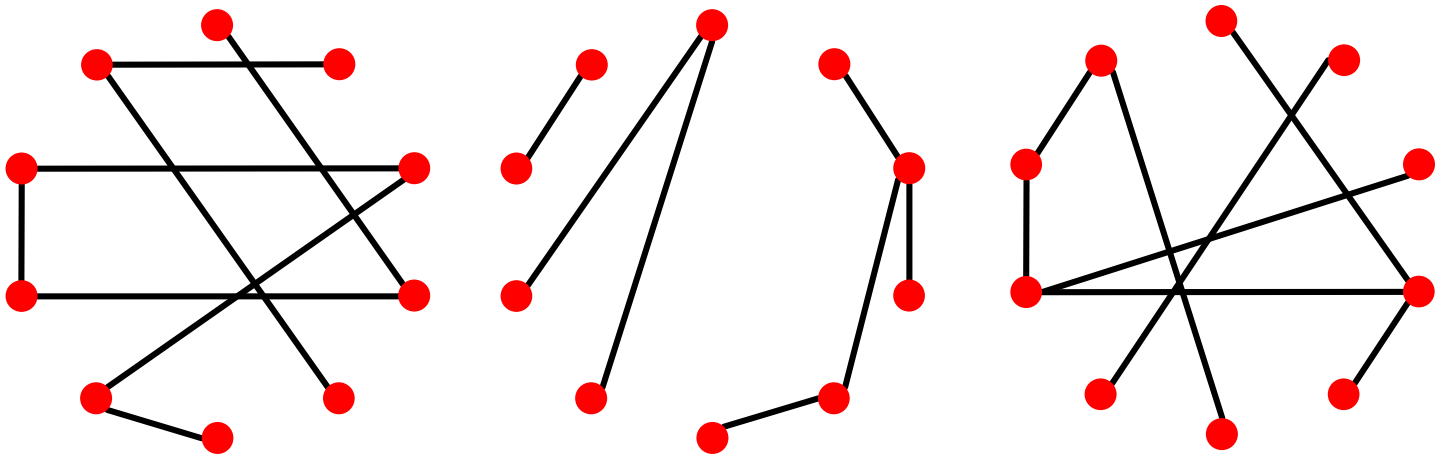
Simplest Model of Graphs

- **Erdös-Renyi Random Graphs** [Erdös-Renyi, '60]
- **Two variants:**
  - G_{np}: undirected graph on n nodes where each edge (u,v) appears i.i.d. with probability p
  - G_{nm} : undirected graph with n nodes, and m edges picked uniformly at random

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 realizations given the same ***n*** and ***p***



n = 10
p= 1/6

Properties of G_{np}

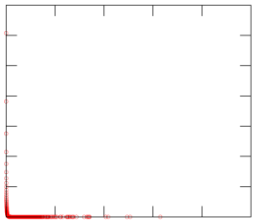
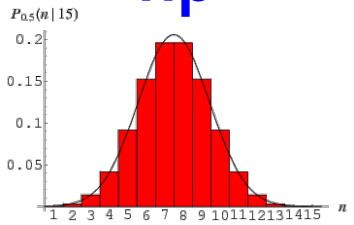
Degree distribution: $P(k)$

Path length: h

Clustering coefficient: C

What are the values of these properties for G_{np}?

Back to MSN vs. G_{np}

	MSN	G_{np} n=18oM	
Degree distribution:			✗
Avg. path length:	**6.6**	$O(\log n)$ *h ≈ 8.2*	✓
Avg. clustering coef.:	*0.11*	$\bar{k} / n$ *C ≈ 8·10⁻⁸*	✗
Largest Conn. Comp.:	*99%*	GCC exists when $\bar{k}>1$. $\bar{k} \approx 14$.	✓

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
  - Giant component in most real networks does NOT emerge through a phase transition
  - No local structure – clustering coefficient is too low
- **Most important: Are real networks random?**
  - The answer is simply: **NO!**

Real Networks vs. G_{np}

- **If G_{np} is wrong, why did we spend time on it?**
  - 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
  - It will help us understand to what degree a particular property is the result of some random process

So, while G_{np} is WRONG, it will turn out to be extremly USEFUL!

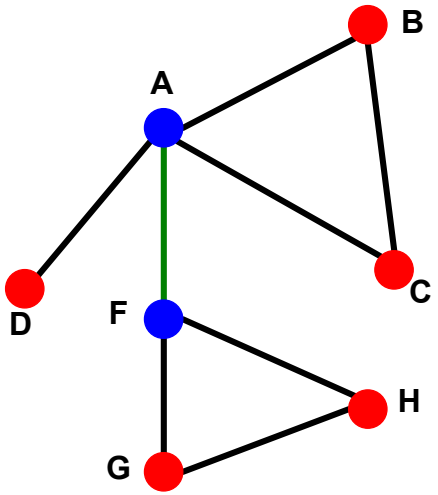
Science to philosophy

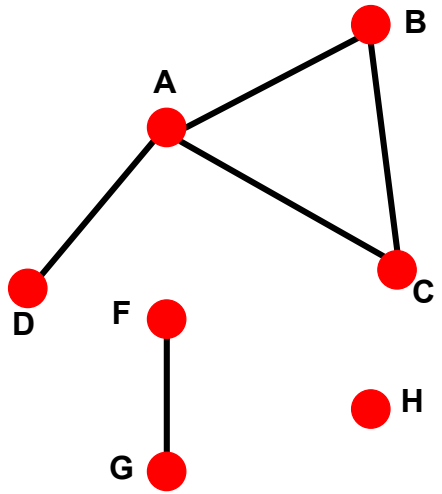
What if we really find one accurate model?

Strong and weak Ties

Connectivity of Undirected Graphs

- **Connected (undirected) graph:**
  - Any two vertices can be joined by a path
- A disconnected graph is made up by two or more connected components





Largest Component:
Giant Component

Isolated node (node H)

Bridge edge: If we erase the **edge**, the graph becomes disconnected

Articulation node: If we erase the **node**, the graph becomes disconnected