Social and Information Network Analysis
Network Structure and Properties

Cornelia Caragea

Department of Computer Science and Engineering
University of North Texas

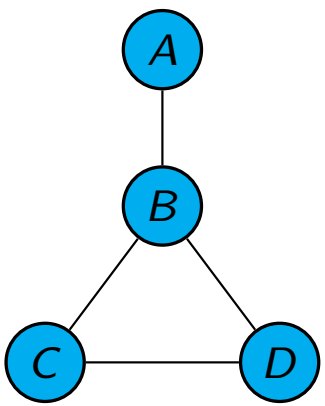
Acknowledgement: Lada Adamic and Jure Leskovec

June 10, 2016

What is the Structure of a Network?

What is a Network?

- A *network* is a way of specifying relationships among a collection of items.
- Components of a network:
  - It consists of a set of objects, called *nodes*, with certain pairs of these objects connected by *links*, called edges.
  - Two nodes are *neighbors* if they are connected by an edge.



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

Networks in Real-World Domains

- Social networks:
  - nodes are people or groups of people
  - edges are some kind of social interaction
- Information networks:
  - nodes are information resources such as Web pages or documents
  - edges are logical connections such as hyperlinks, citations, or cross-references
- Communication networks:
  - nodes are computers or other devices that can relay messages
  - edges are direct links along which messages can be transmitted
- Transportation networks:
  - nodes are destinations
  - edges are direct connections

Networks or Graphs?

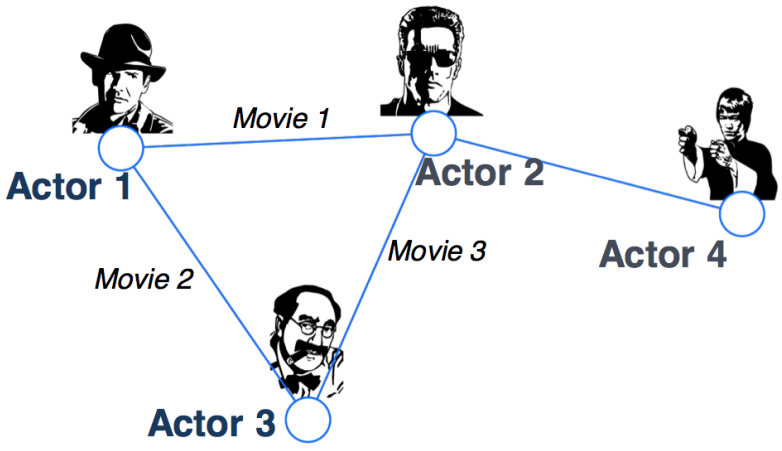
- Network: often refers to real systems
  - Web, Social network, Metabolic network

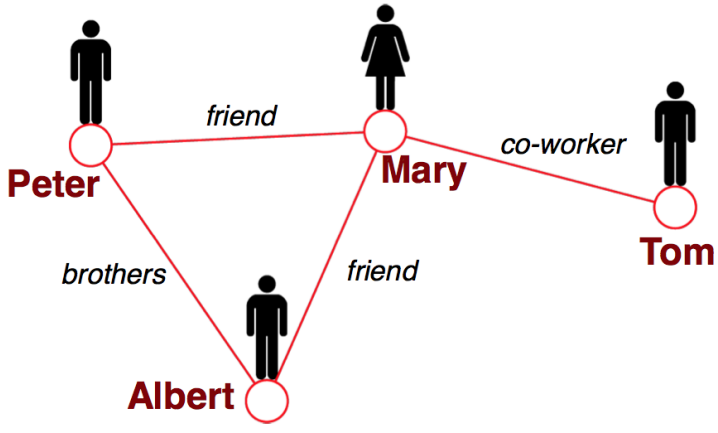
  Language: Network, node, link
- Graph: mathematical representation of a network
  - Web graph, Social graph (a Facebook term)

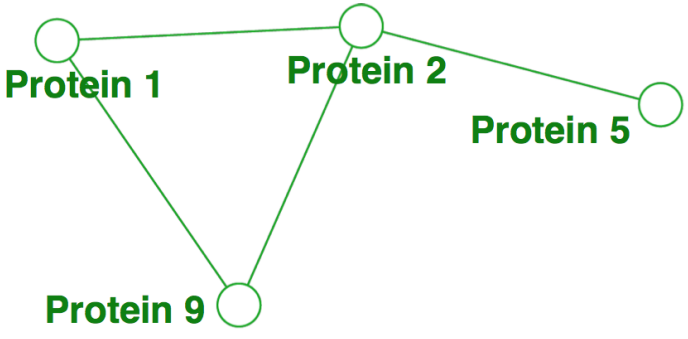
  Language: Graph, vertex, edge

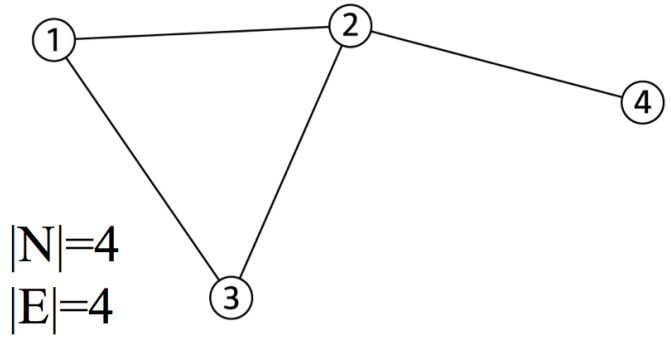
In most cases, we will use the two terms interchangeably

Networks: Common Language









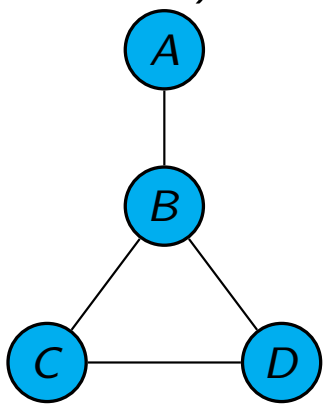
Choosing a Proper Representation

- Choice of the proper network representation determines our ability to use networks successfully:
  - In some cases there is a unique, unambiguous representation
    - E.g., citation networks
  - In other cases, the representation is by no means unique
    - E.g., communication networks; phone calls
  - The way you assign links will determine the nature of the question you can study

Undirected vs. Directed Networks

Undirected

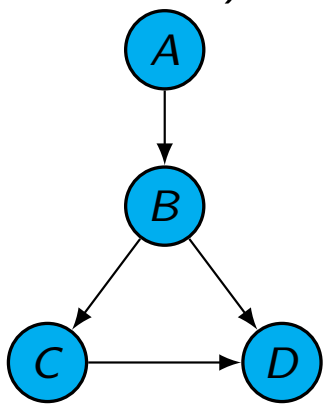
- Links: undirected (symmetrical)



- Examples:
  - Collaborations Networks
  - Friendship on Facebook
  - Professional Networks

Directed

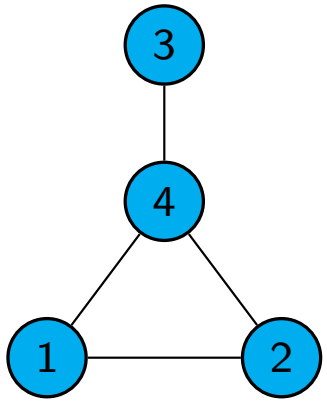
- Links: directed (asymmetrical)



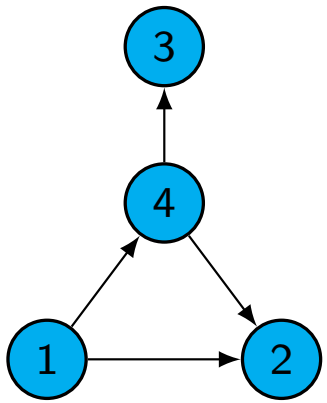
- Examples:
  - Citations Networks
  - Follower on Twitter
  - The Web

Representation - Adjacency Matrix

$A_{ij} = 1$ if there is a link from i to j and $A_{ij} = 0$, otherwise.



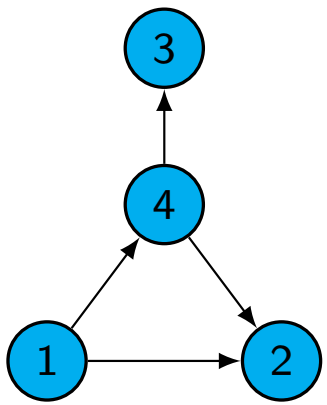
$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$



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

Representation - Adjacency List

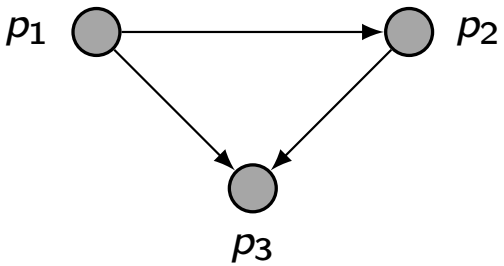
- Easier to work with adjacency lists when the network is very large and sparse.
- Fast to retrieve all neighbors for a node.



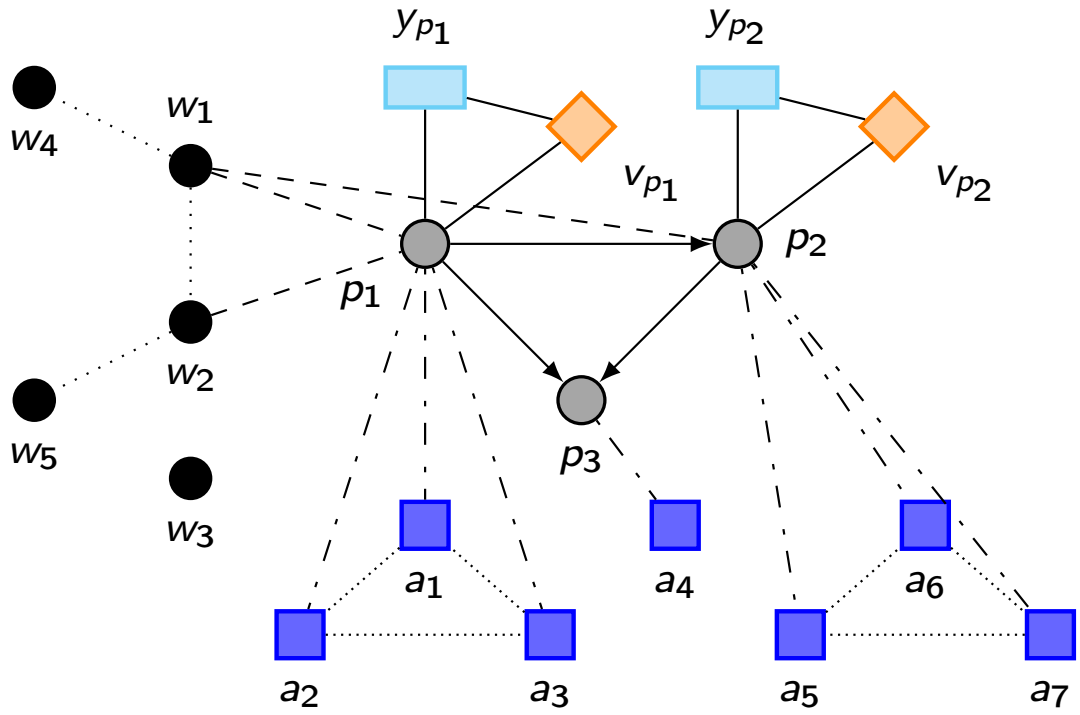
- 1: 2, 4
- 2:
- 3:
- 4: 3

Homogeneous vs. Heterogeneous Networks

Homogeneous:



Heterogeneous:



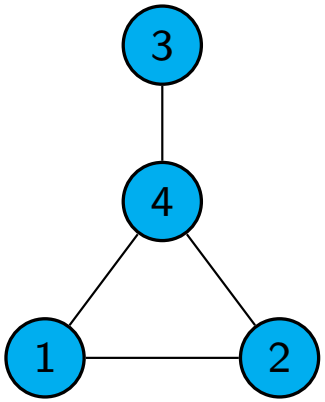
Citation network

Scientific network

Other Network Structures

Unweighted vs. Weighted Networks

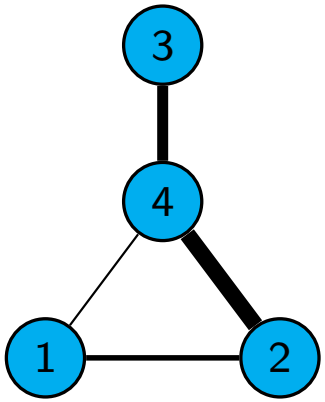
Unweighted (undirected)



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

- Examples:
  - Friendship on Facebook
  - Citation Networks

Weighted (undirected)

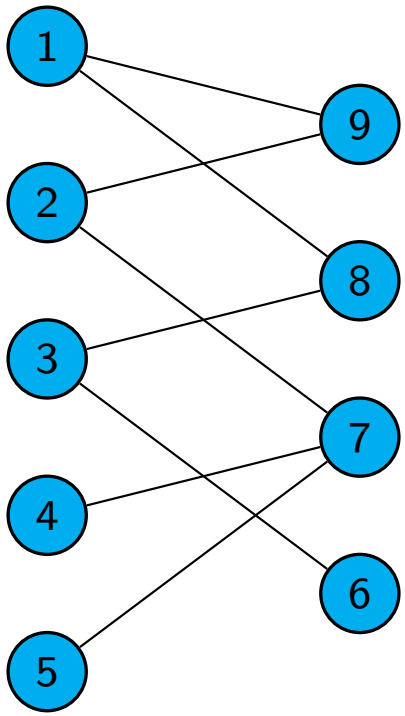


$$\begin{pmatrix} 0 & 1 & 0 & 0.5 \\ 1 & 0 & 0 & 3 \\ 0 & 0 & 0 & 2 \\ 0.5 & 3 & 2 & 0 \end{pmatrix}$$

- Examples:
  - Collaboration Networks
  - Document Networks

Bipartite Graphs

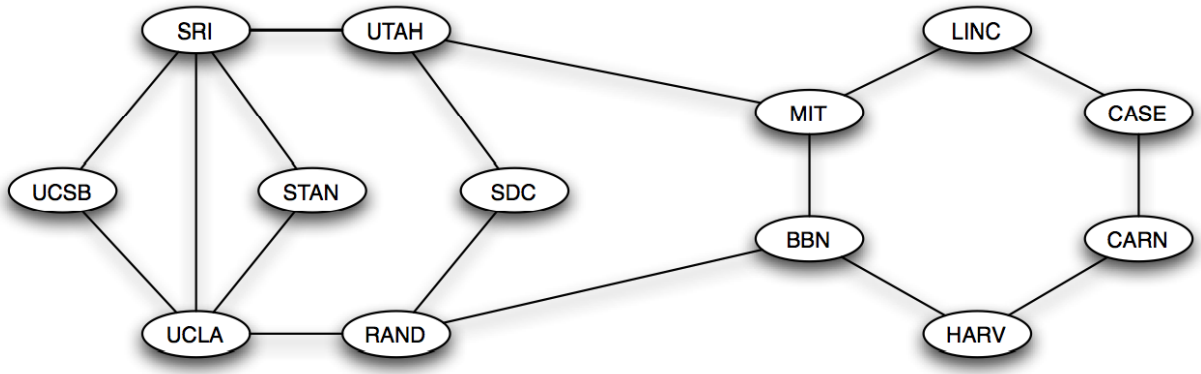
- Bipartite graph is a graph whose nodes can be divided into two disjoint sets U and V such that every link connects a node in U to one in V; that is, U and V are independent sets.
- Examples:
  - Authors-to-Papers
  - Actors-to-Movies
  - Users-to-Movies
  - Papers-to-Reviewers
  - Authors-to-Conferences



Paths and Connectivity in Graphs

Paths in Graphs

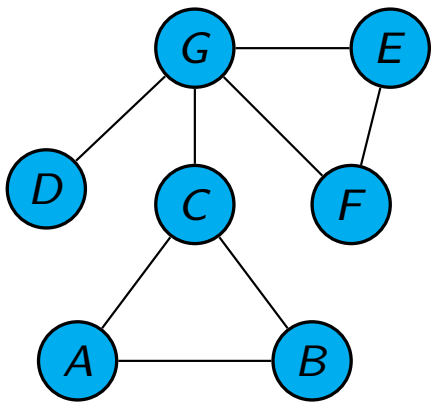
- A *path* is simply a sequence of nodes with the property that each consecutive pair in the sequence is connected by an edge
- A *simple path* is a path that does not repeat nodes
- A *cycle* is a path with at least three edges, where first and last nodes are the same, but otherwise all nodes are distinct
- Example: A 13-node Internet graph:

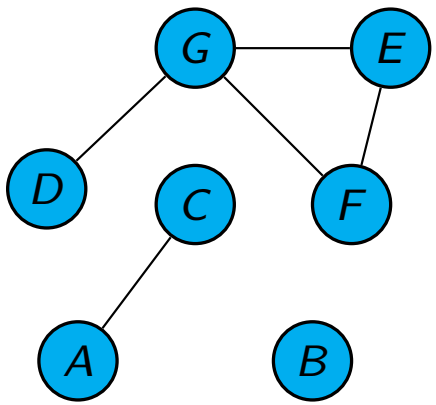


  - *Observation*: every edge belongs to a cycle, by design.
  - In communication and transportation networks, cycles are often present to allow for redundancy or alternate routings.

Connectivity in Graphs

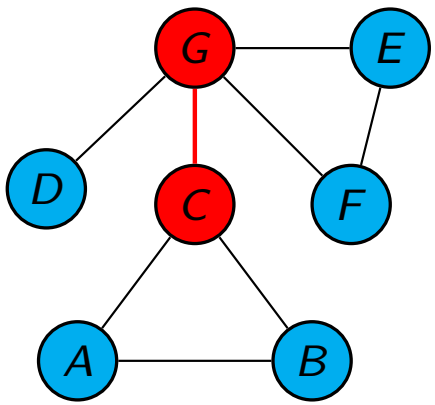
- A graph is *connected* if for every pair of nodes, there is a path between them
- A *connected component* (or simply, a *component*) is a subset of the nodes such that:
  - every node in the subset has a path to every other node
  - the subset is not part of some larger set with the property that every node can reach every other node
- A *disconnected* graph is made up by two or more connected components
  - Connected versus disconnected graphs:

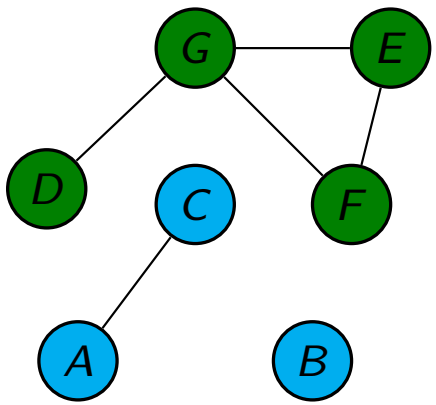




Connectivity in Graphs

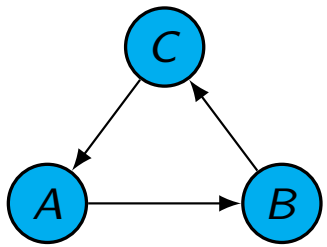
- *Giant component:* the largest component in a graph
- *Bridge edge*: If erased, the graph becomes disconnected
- *Articulation point*: If erased, the graph becomes disconnected



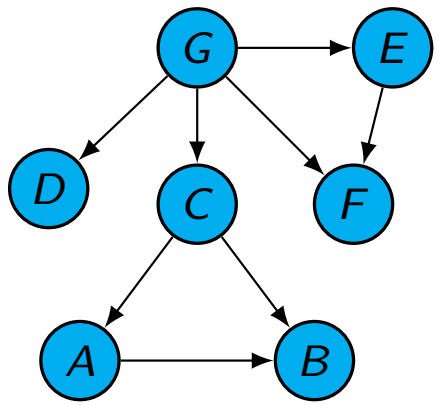


Connectivity in Directed Graphs

- Strongly connected directed graph
  - has a path from each node to every other node and vice versa (e.g., A-B path and B-A path)



- Weakly connected directed graph
  - is connected if we disregard the edge directions

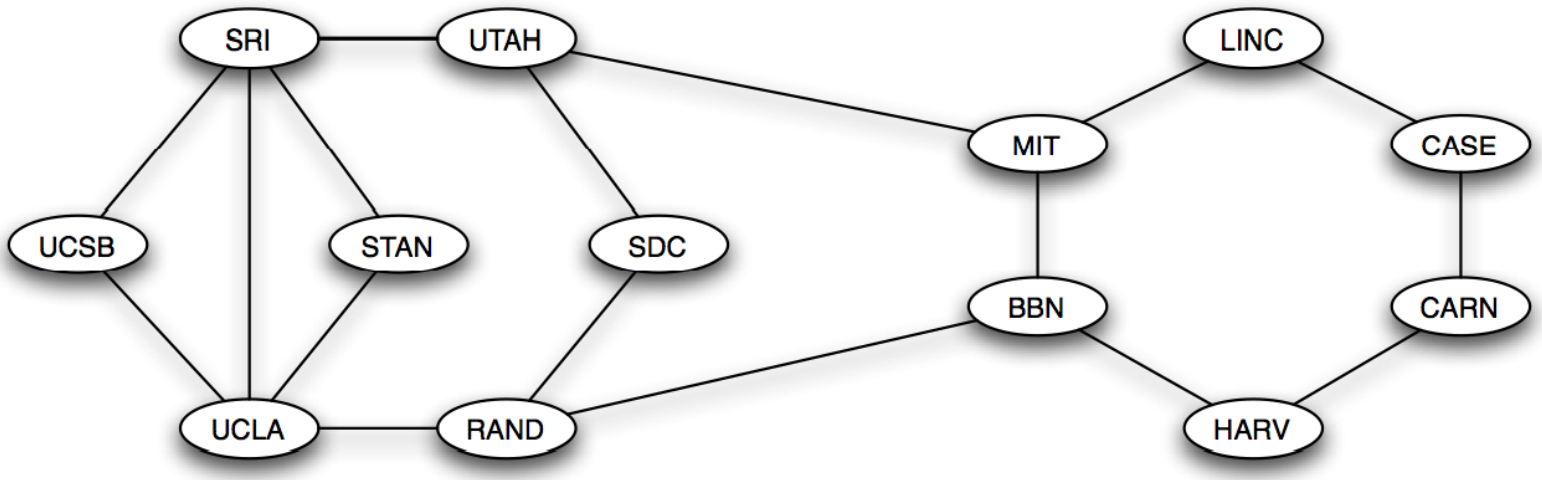


Connectivity in Graphs - Real-World Examples

Airline routes:

Subway map:

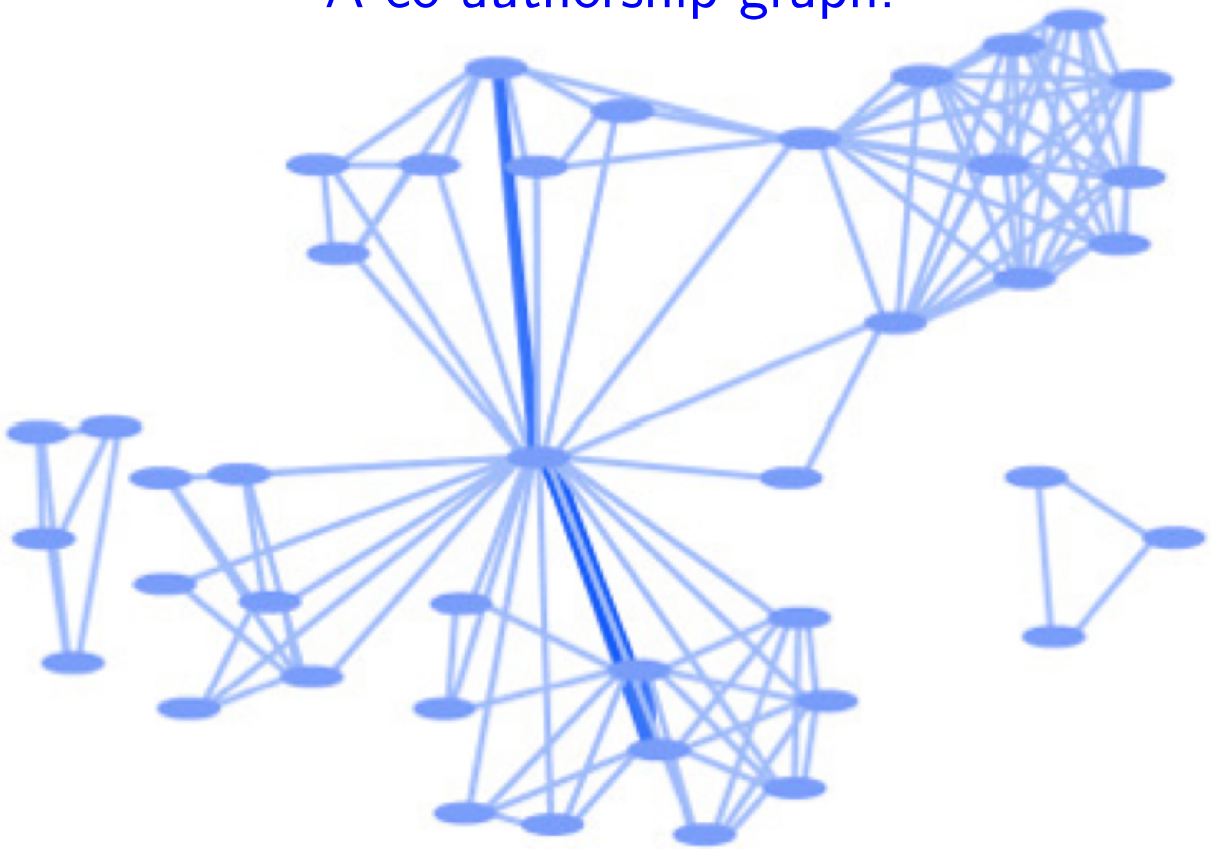
Connectivity in Graphs - Real-World Examples



Connectivity in Graphs - Real-World Examples

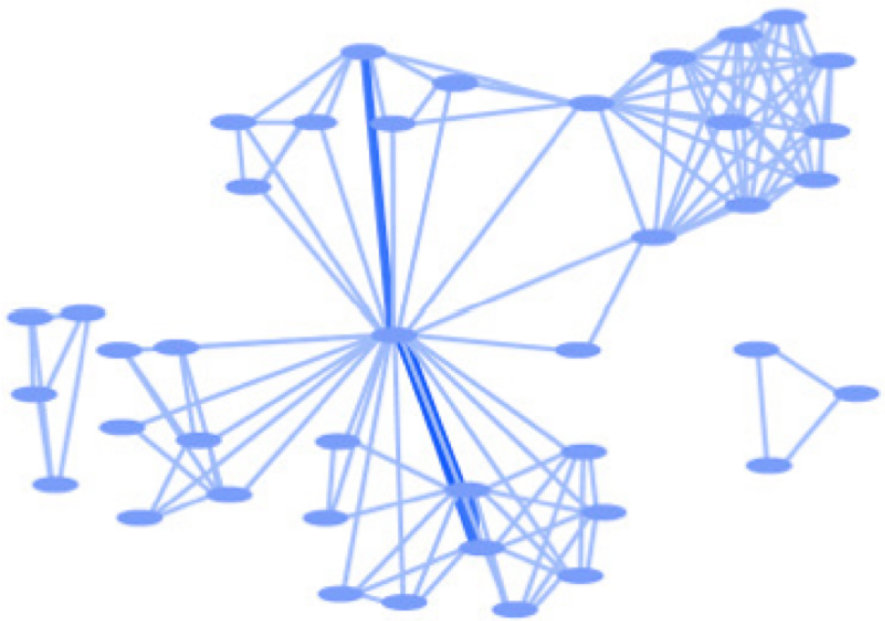
No *a priori* reason to expect graphs in other settings to be connected.

A co-authorship graph:



Why Does Connectivity Matter?

- Connectivity: helps describe the network structure
  - Within a component, there may be richer internal structure that is important for the interpretation of the network
  - Example:
    - A prominent node at the center in the graph below, and tightly-knit groups linked to this node, but not to each other
    - If removed, the largest connected component would break apart into three distinct components



How to Characterize a Network Structure?

Giant Component

What can we say about connected components of large networks?

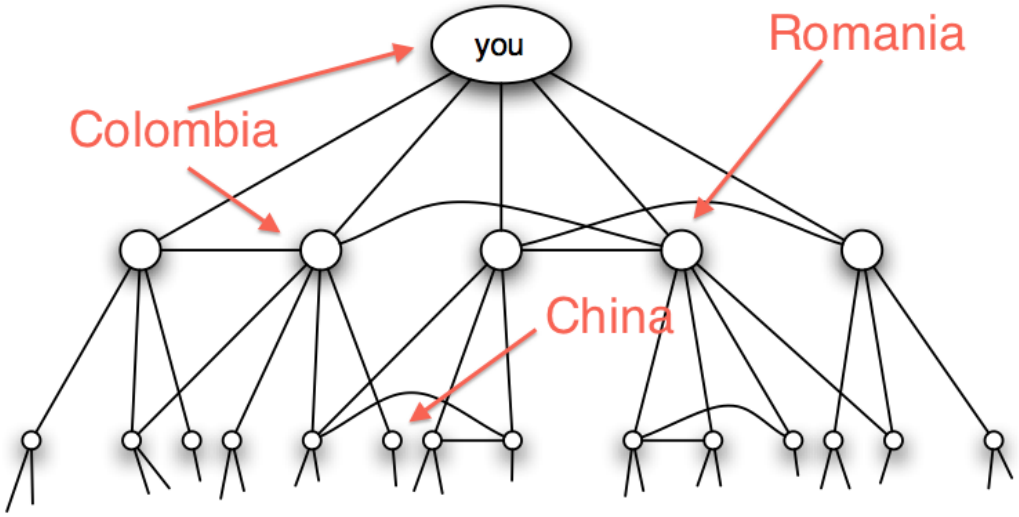
- Example: the friendship network of the entire world [Not explicitly recorded anywhere, need to use our intuition to answer basic questions]
  - Is the global friendship network connected?

Giant Component

What can we say about connected components of large networks?

- Example: the friendship network of the entire world [Not explicitly recorded anywhere, need to use our intuition to answer basic questions]
  - Is the global friendship network connected? Presumably not
    - A person with no friends = a one-node component → disconnected graph [Too extreme!]
  - **Is there a giant component?**
    - A connected component that contains a significant fraction of all the nodes

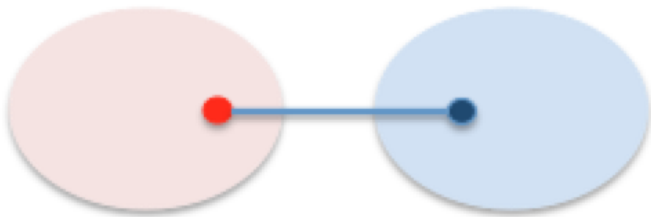
You on the Global Friendship Network



- You seem to belong to a component that contains a significant fraction of the world's population

Giant Component in Large Networks

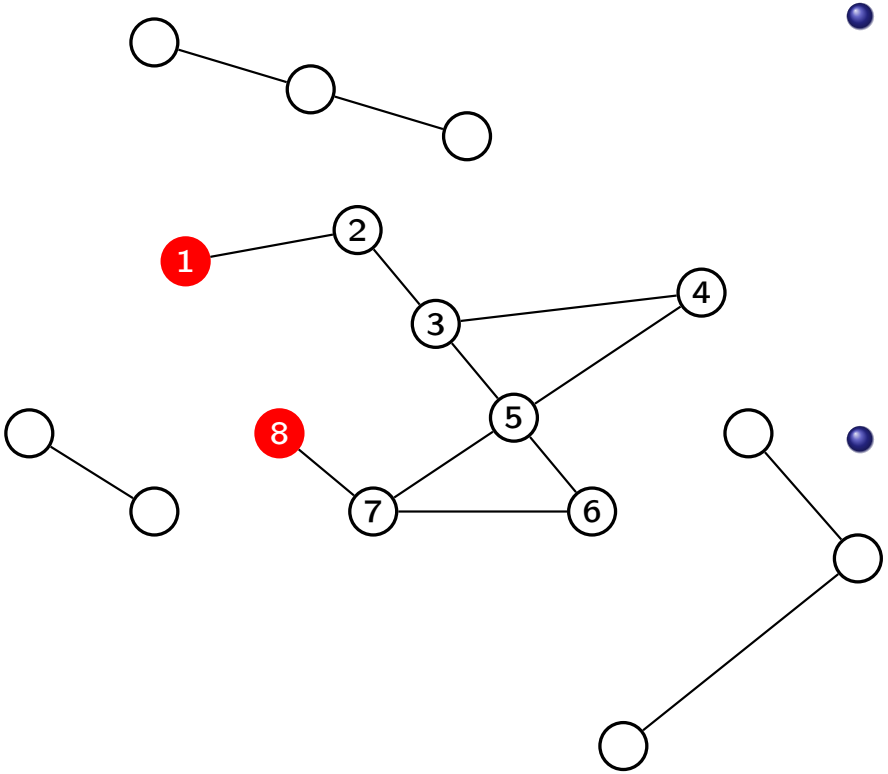
- Large, complex networks generally have a giant component - true fact when one looks across a range of network datasets
- When a network contains a giant component, it almost always contains only one:
  - Heuristic argument:
    - It just takes a single edge from someone in the first of these components to someone in the second, and the two giant components would merge into a single component



    - If the two connected components have millions of people, the likelihood of this not happening is very very small

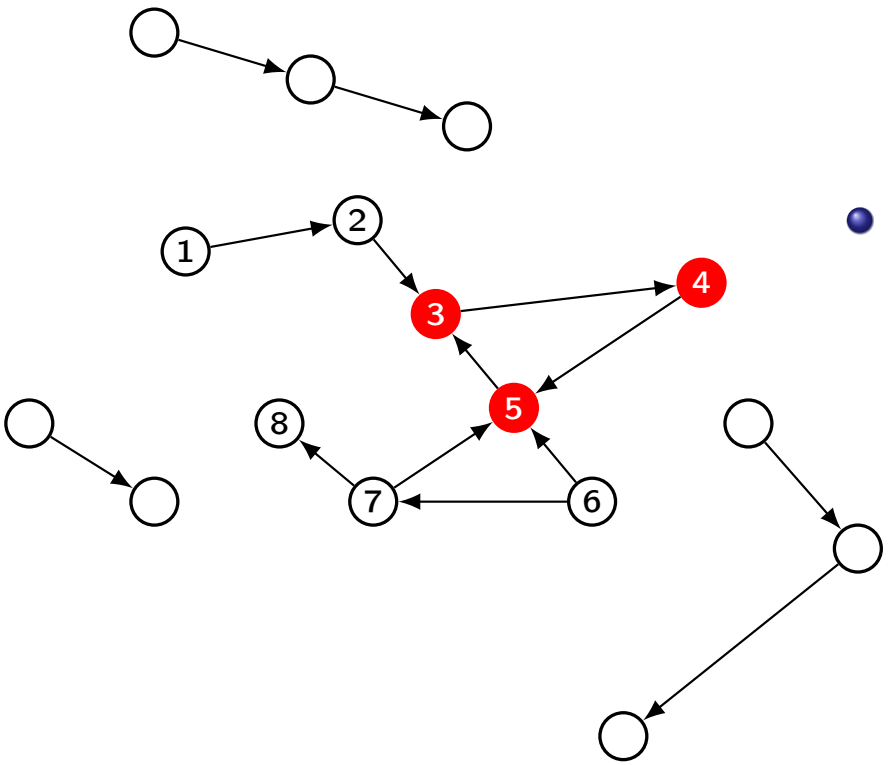
Distance in an Undirected Graph

Between two nodes connected by a path, how long such a path is?



- **Recall...** A path is a sequence of nodes in which each node is linked to the next one
  - $P_{18} = \{1, 2, 3, 4, 5, 6, 7, 8\}$,
    $P_{18} = \{1, 2, 3, 5, 6, 7, 8\}$,
    $P_{18} = \{1, 2, 3, 5, 7, 8\}$
- Distance or shortest path between a pair of nodes is defined as the number of edges along the shortest path connecting the nodes
  - $P_{18} = \{1, 2, 3, 5, 7, 8\}$

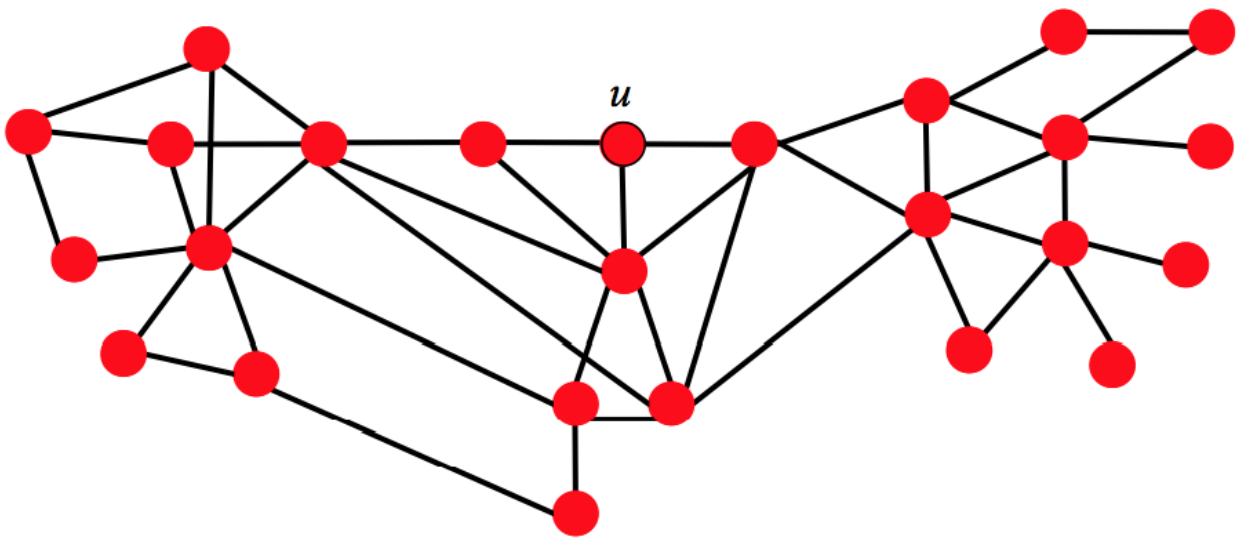
Distance in a Directed Graph



- In directed graphs, paths need to follow the direction of the arrows
  - Consequence: Distance is not symmetric (consider a cycle)
  - $d_{53} = 1$, $d_{35} = 2$

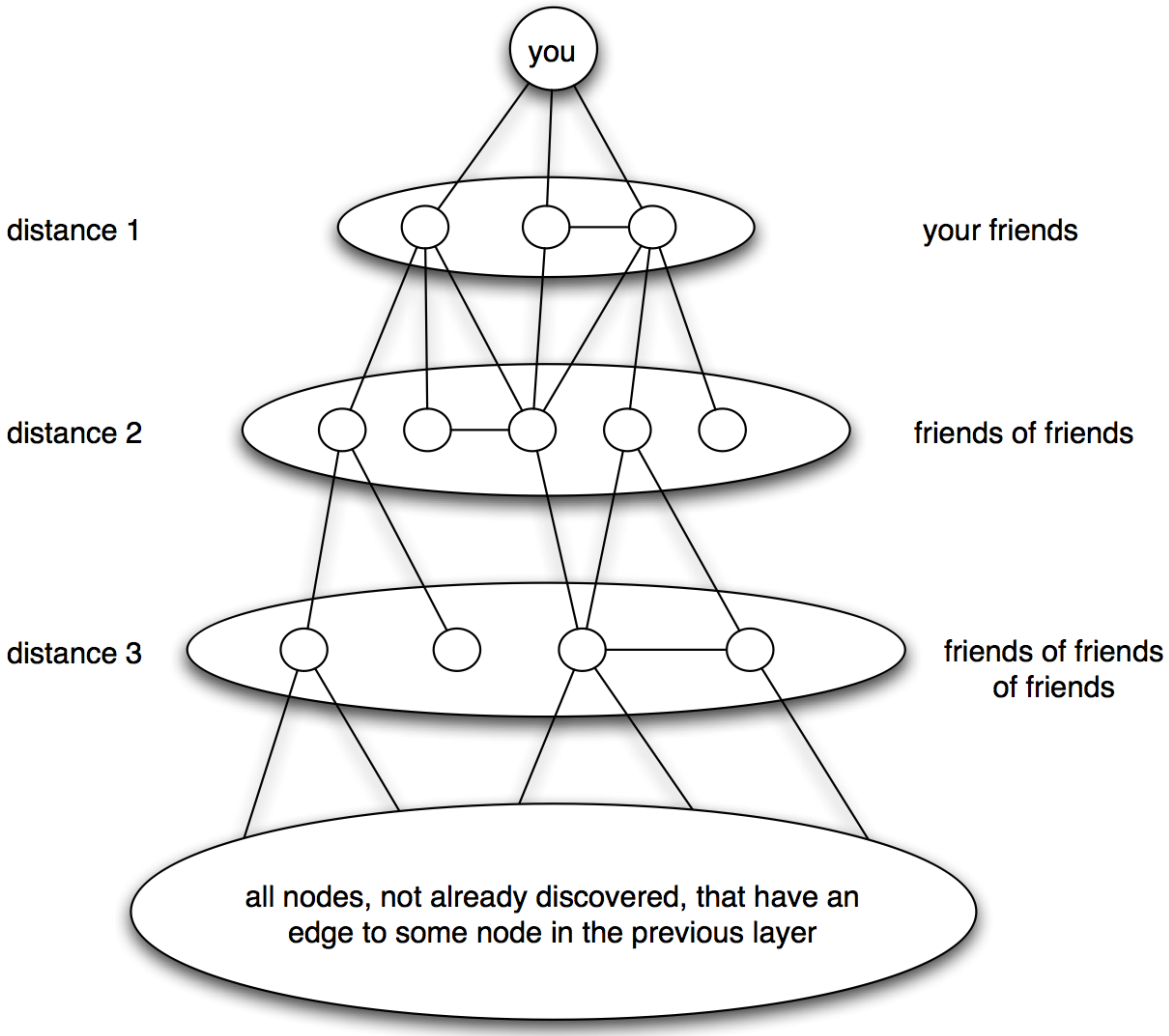
Finding Shortest Paths

- Breath-First Search:
  - Start with node u, mark it to be at distance $d_u(u) = 0$, add u to the queue
  - While the queue not empty:
    - Take node v off the queue, put its unmarked neighbors w into the queue and mark $d_u(w) = d_u(v) + 1$.



Finding Shortest Paths

Breath-First Search



Network Diameter

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

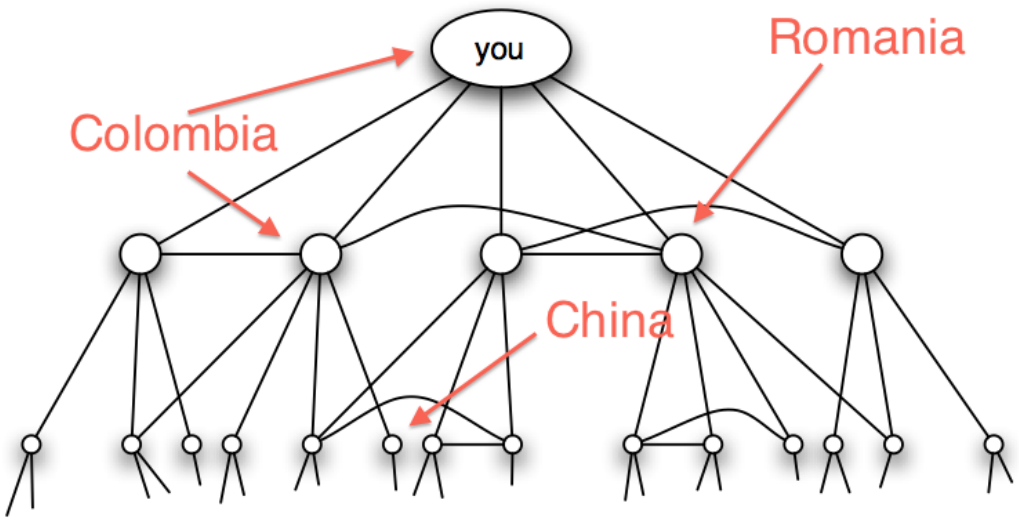
$$< d >= \frac{1}{2E_{max}} \sum_{i,j \neq i} d_{ij}$$

where d_{ij} is the distance from node i to node j, and $E_{max} = \frac{N(N-1)}{2}$, with N = number of nodes in the graph

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

The Small-World Phenomenon

- Back to the question "How long are the paths between two connected nodes?"
- Research has been shown that:
  - Not only do you have paths of friends connecting you to a large fraction of the world's population, but these paths are surprisingly short = small-world phenomenon (or the six degrees of separation)

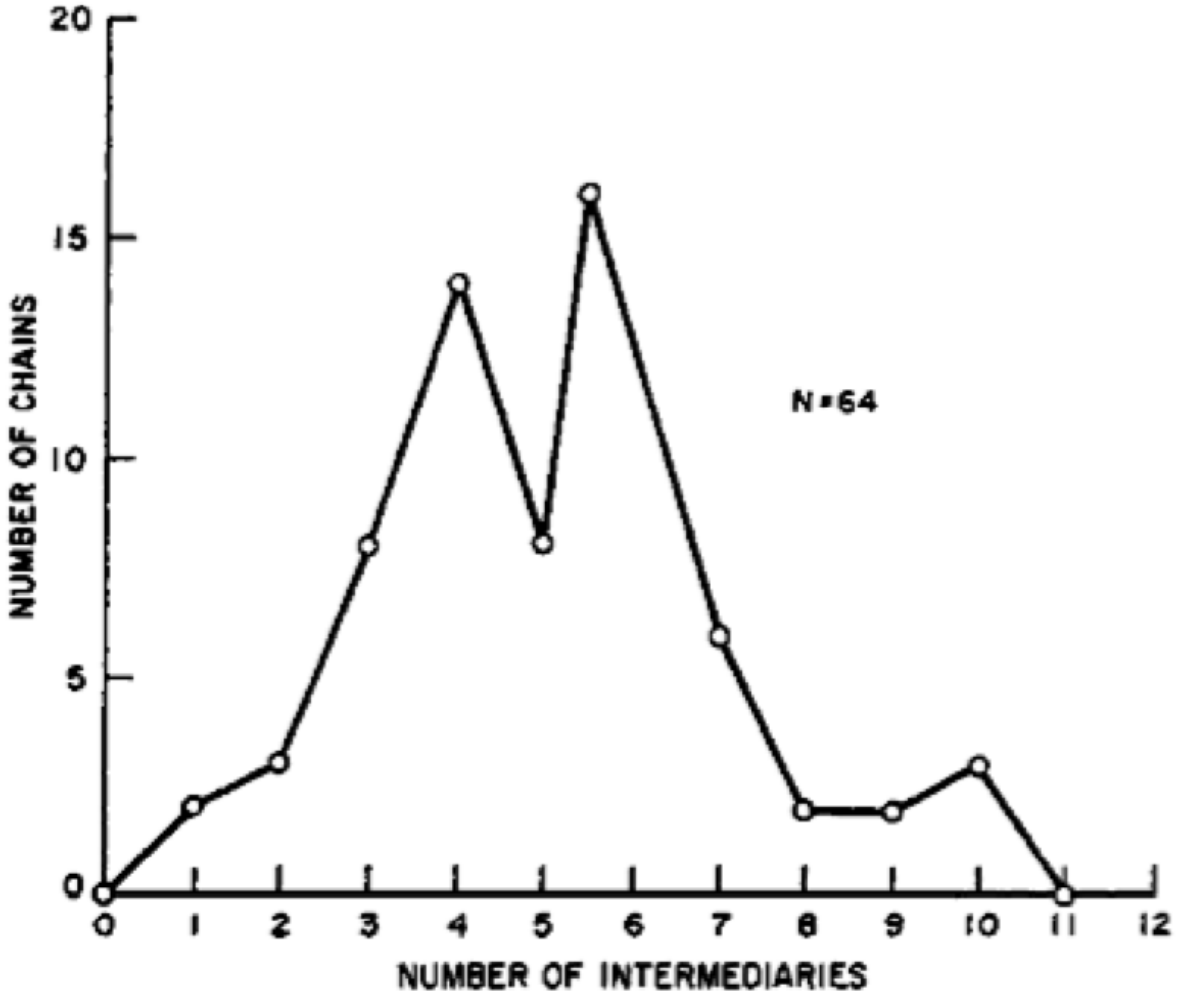


The Small-World Phenomenon

Research experiments:

- Milgram (1960s)
  - Wanted to test the idea that people are really connected in the global friendship by short chains of friends
  - Picked 296 randomly chosen “starters" and asked them to forward a letter to a “target" person, i.e., a stockbroker who lived in a suburb of Boston
  - The starters were given information about the target person, and were asked to forward the letter to someone they knew, with the same instructions, in order to reach the target as soon as possible
  - 64 chains succeeded in reaching the target

The Distribution of Path Lengths in Milgram's Experiment



Caveats about Milgram's Experiment

- We cannot necessarily draw the conclusion that there are only "six degrees of separation between us and everyone else on this planet"
- Even if you can reach someone through a short chain of friends, is this useful to you?

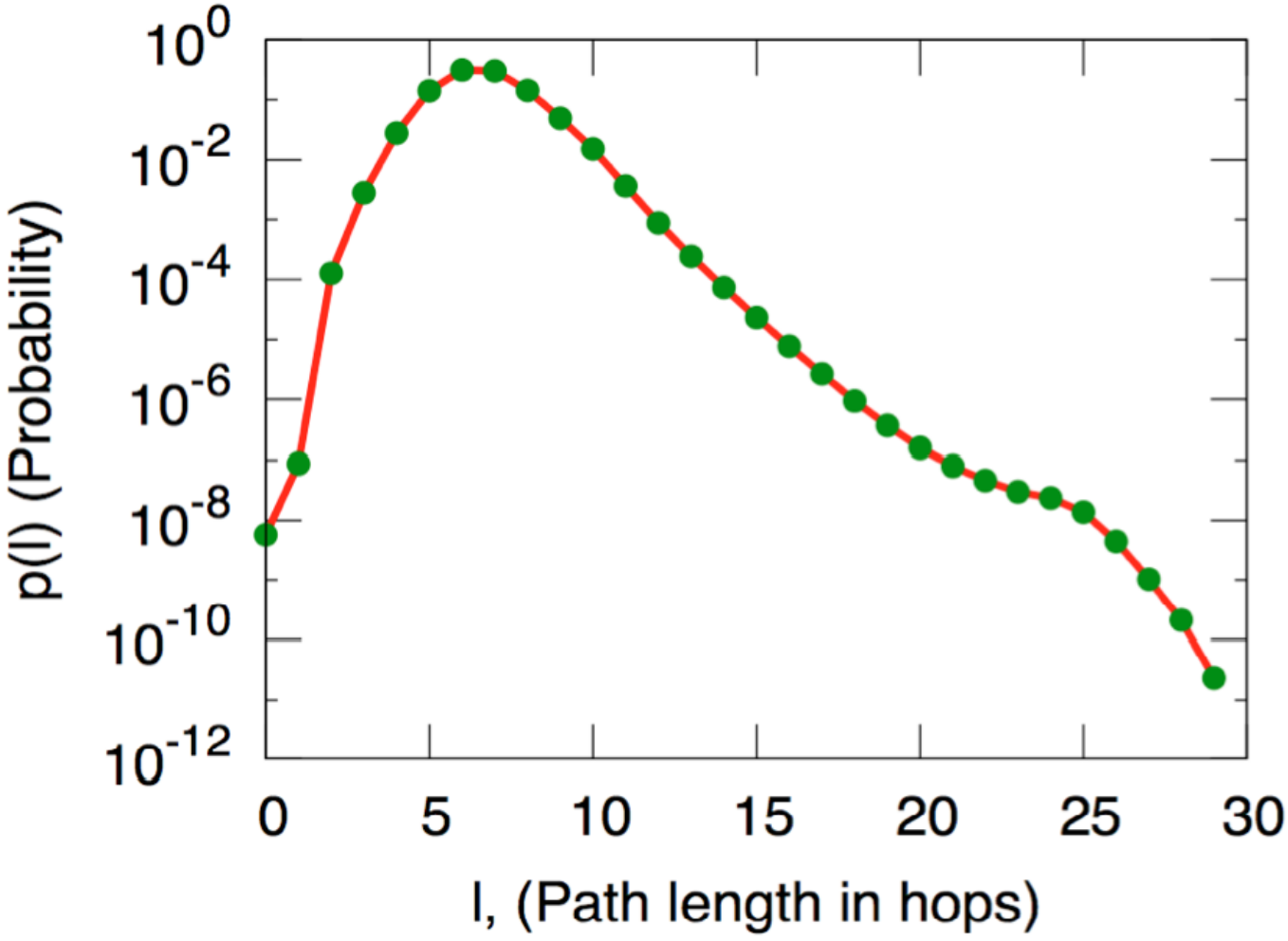
Caveats about Milgram's Experiment

- We cannot necessarily draw the conclusion that there are only "six degrees of separation between us and everyone else on this planet"
- Even if you can reach someone through a short chain of friends, is this useful to you?
- The existence of short paths implies high speed with which information or diseases spread through a network

Instant Messaging

- Leskovec and Horvitz analyzed the 240 million active user accounts on Microsoft Instant Messenger
  - Nodes correspond to users
  - Edges exist between two users engaged in a two-way conversation during a month
  - The graph has a giant component, consisting of the majority of nodes, with an estimated average distance of 6.6

The Distribution of Distances in Leskovec's Experiment



Caveats about Leskovec's Experiment

- It only tracks people who are technologically-endowed enough to have access to instant messaging
- Not based on the graph of who is truly friends with whom
- It analyzes who talks to whom during an observation period

Caveats about Leskovec's Experiment

- It only tracks people who are technologically-endowed enough to have access to instant messaging
- Not based on the graph of who is truly friends with whom
- It analyzes who talks to whom during an observation period
- However, these experiments provide good approximation of the real-world friendship networks

A Collaboration Network - The Erdös Number

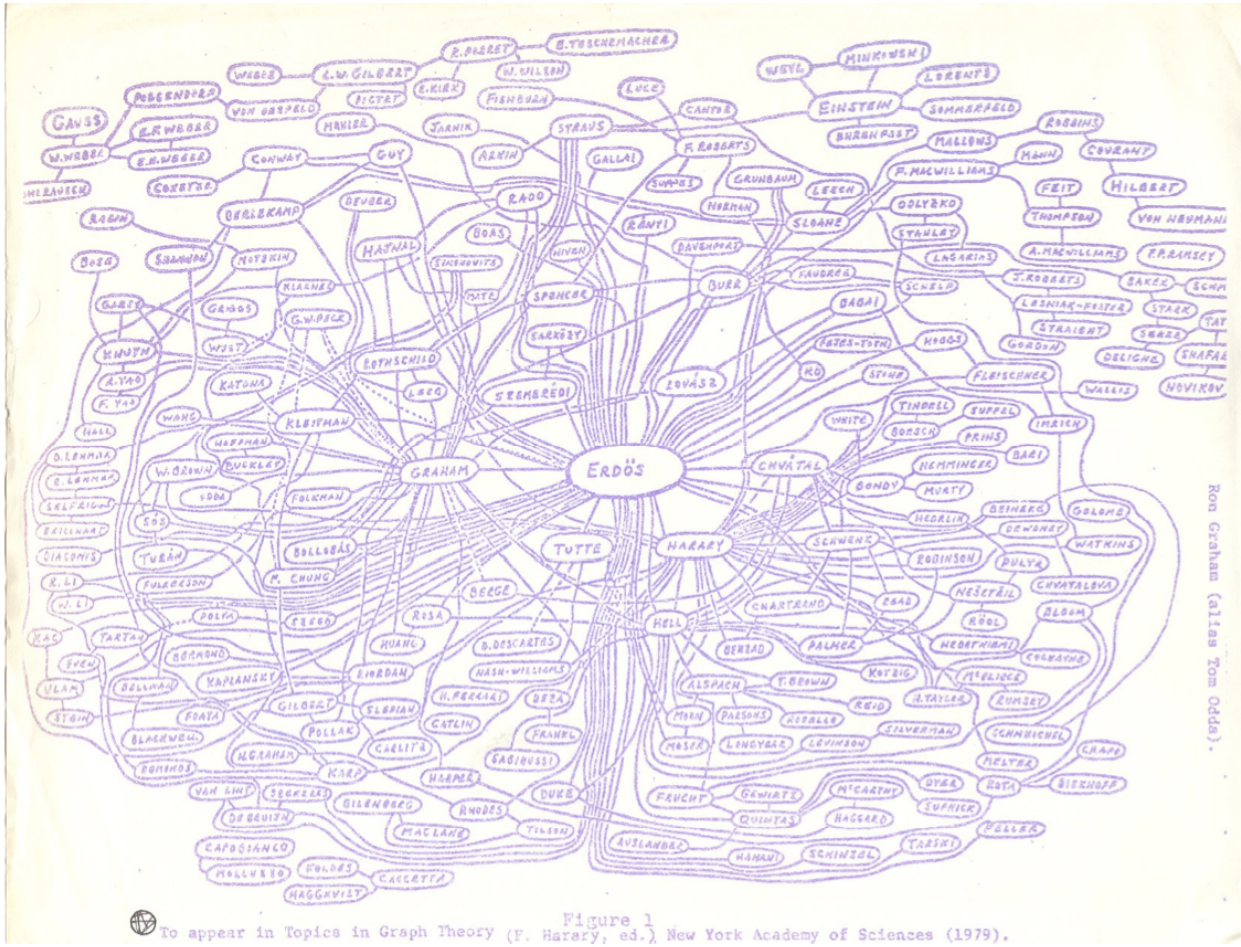
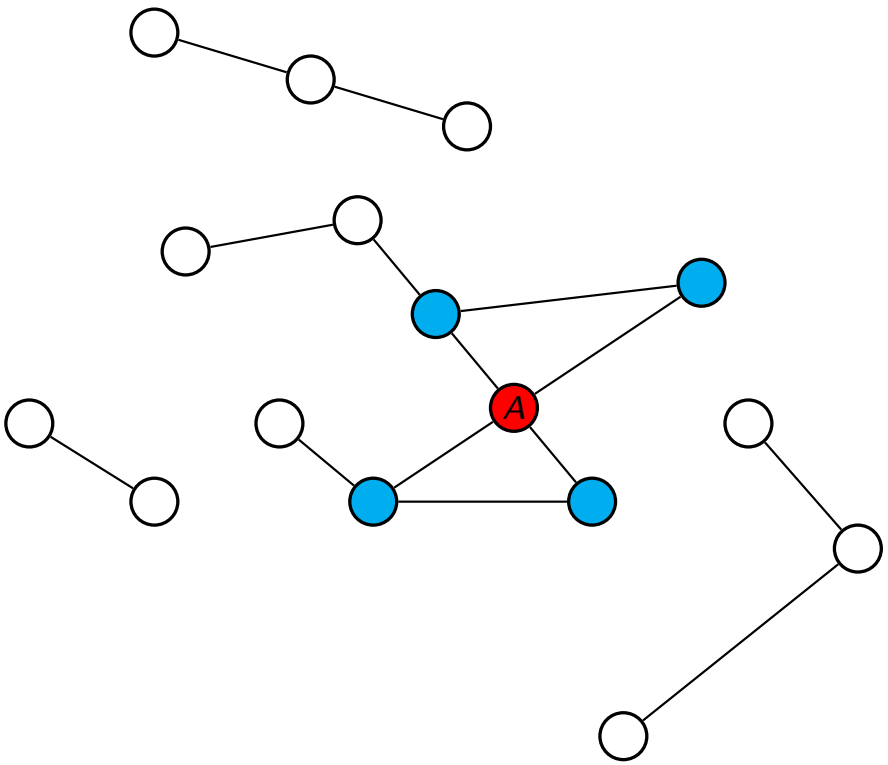


Figure 1
To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

Most mathematicians have Erdös numbers of at most 4 or 5

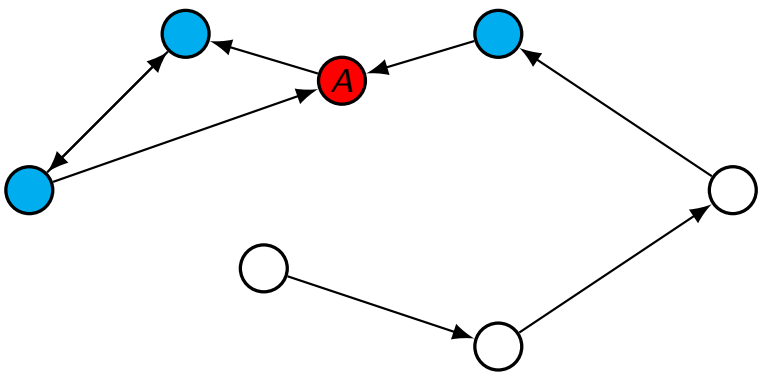
Further Analysis on How to Characterize a Network Structure

Node Degrees - Undirected Graphs



- Node degree, k_i: the number of edges adjacent to node i, $k_A = 4$
- Average degree, $< k >= \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2E}{N}$

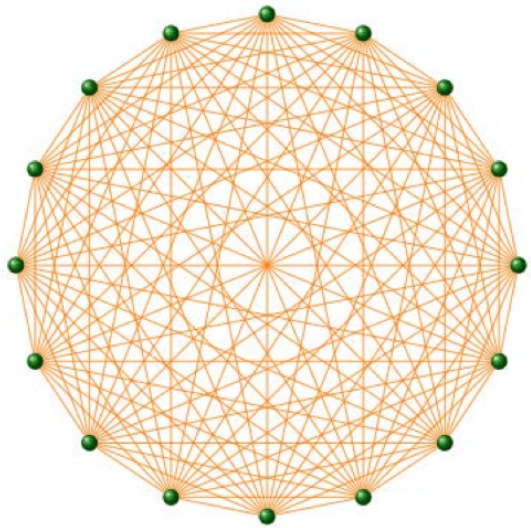
Node Degrees - Directed Graphs



- Here, we define in-degree and out-degree.
- The total degree of a node is the sum of in- and out- degrees.
  $k_A^{in} = 2$, $k_A^{out} = 1$, $k_A = 3$.
  $< k >= \frac{E}{N}$.
  $< k^{in} >=< k^{out} >$.
- A node with $k^{in} = 0$: source
- A node with $k^{out} = 0$: sink

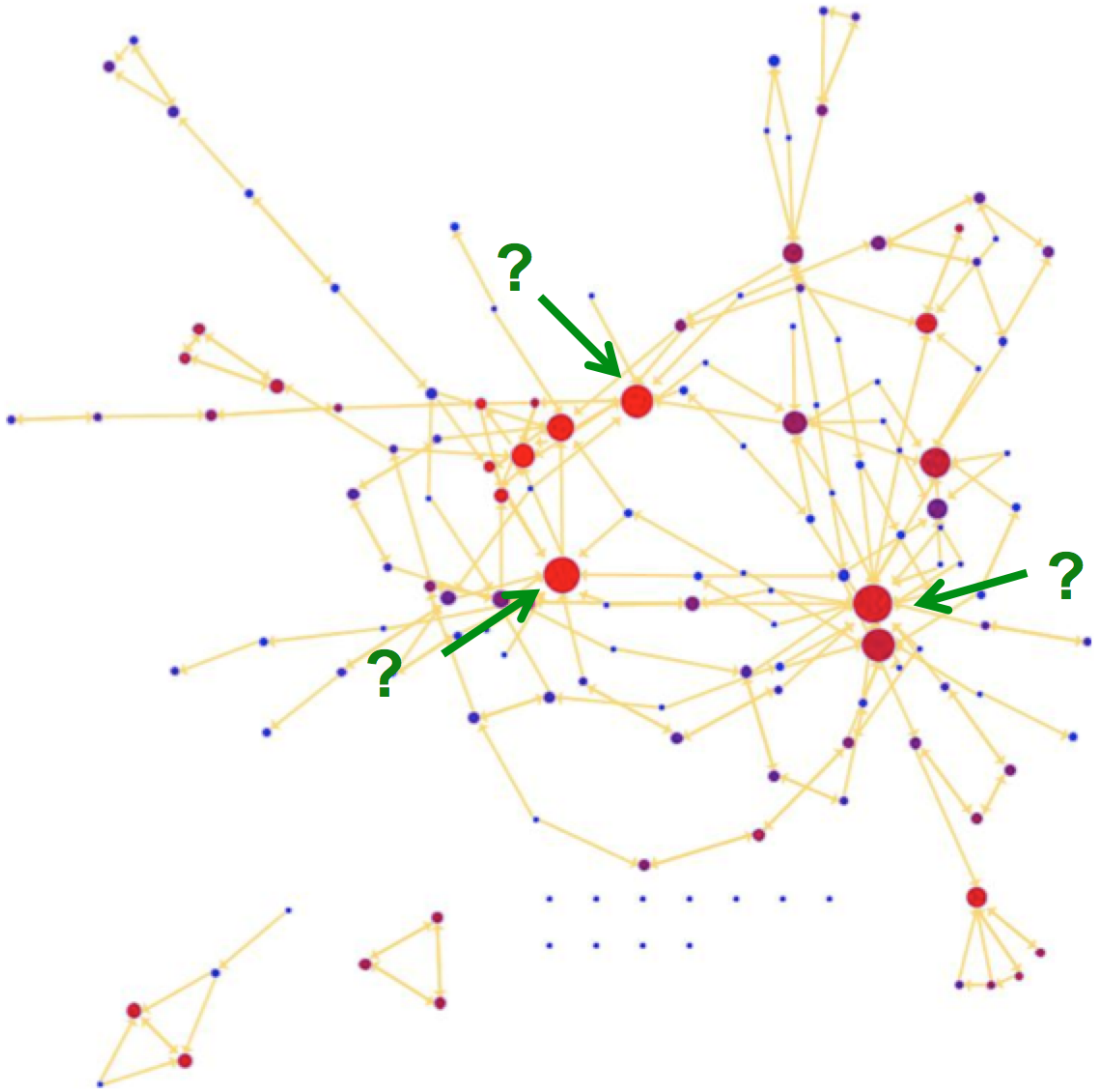
Complete Graphs

- The maximum number of edges in an undirected graph on N nodes is $E_{max} = \frac{N(N-1)}{2}$
- A graph with the number of edges $E = E_{max}$ is a complete graph, and its average degree is $N - 1$.

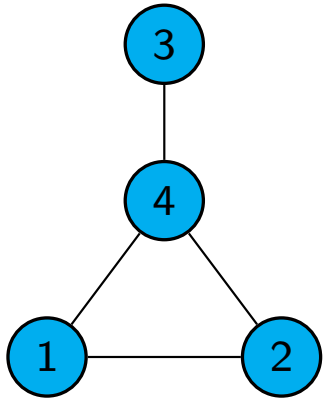


However, networks are generally sparse graphs, i.e., $E << E_{max}$ or $< k >>> N - 1$. Thus, the adjacency matrix is filled with many zeros!

Degree: what node has the most edges?



Node Degree from the Adjacency Matrix



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

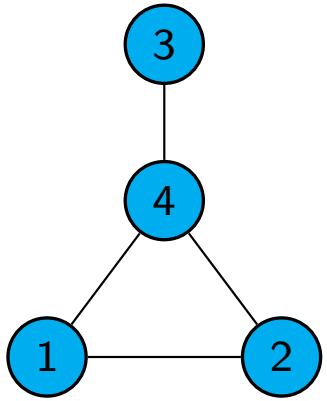
In-degree: $\sum_{i=1}^{N} A_{ij}$

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

Out-degree: $\sum_{j=1}^{N} A_{ij}$

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

Node Degree from the Adjacency Matrix



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

In-degree: $\sum_{i=1}^{N} A_{ij}$

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

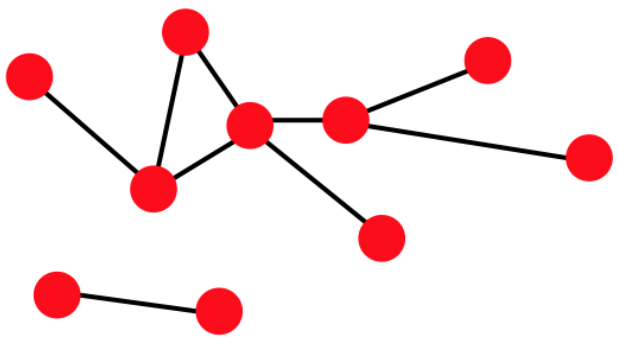
Out-degree: $\sum_{j=1}^{N} A_{ij}$

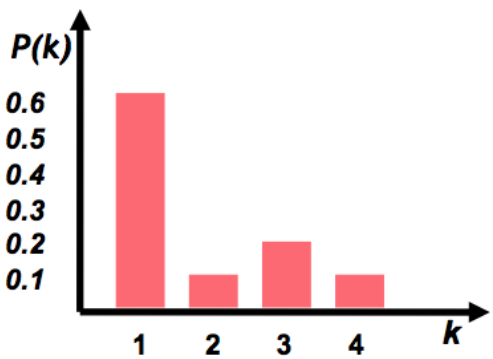
$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

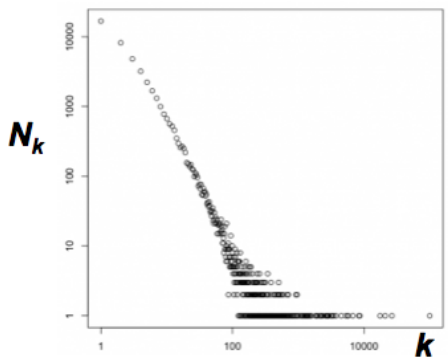
- How about node degree from adjacency lists?

Degree Distribution

- Node degree gives node network properties from immediate connections of nodes.
- Degree distribution $P(k)$: Probability that a randomly chosen node has degree k
  - $N_k = \#$ nodes with degree k
- Normalized histogram:
  - $P(k) = N_k/N$





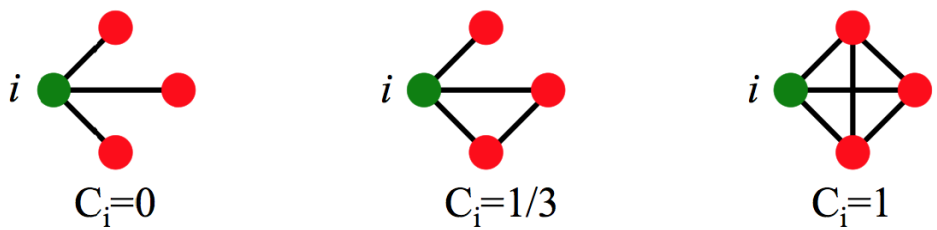


Clustering Coefficient

- Clustering coefficient:
  - What portion of i's neighbors are connected?
  - Node i with degree k_i
  - $C_i \in [0, 1]$
  - Given by:

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

    where e_i = number of edges between the neighbors of node i
  - Example:



- Average Clustering Coefficient:

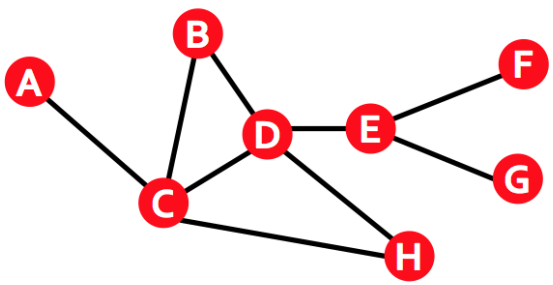
$$C = \frac{1}{N} \sum_{i=1}^{N} C_i$$

Clustering Coefficient

- Clustering coefficient:
  - What portion of i's neighbors are connected?
  - Node i with degree k_i
  - $C_i \in [0, 1]$
  - Given by:

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

    where e_i = number of edges between the neighbors of node i
  - Another example:



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

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

Key Network Properties

- Degree distribution: $P(k)$
- Path length: d_{ij}
- Clustering coefficient: C