Social Media Analytics

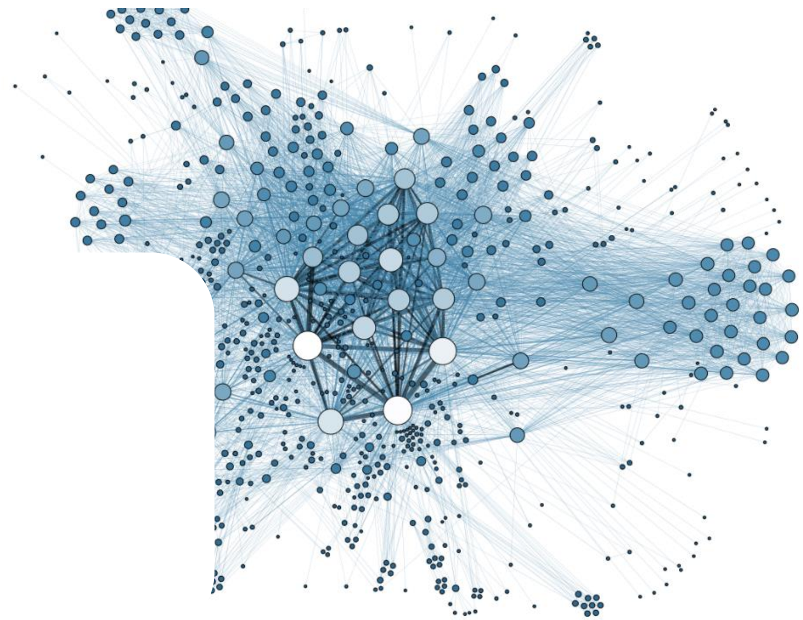
INFS7450

Graph Essentials

Prof. Hongzhi Yin

School of EECS

The University of Queensland



Why should I care about networks or graphs in this course?

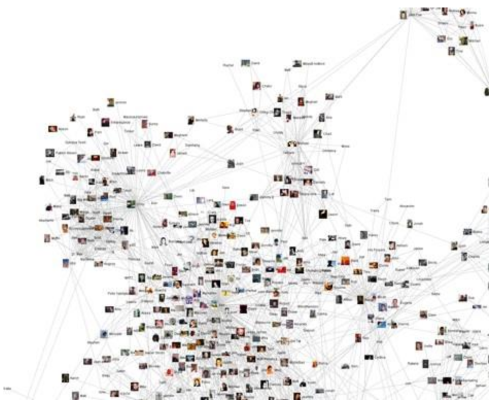
Ways to Analyze Social Media

- Social Media is a complex system consisting of
  - individuals (also called users)
  - information (e.g., reviews, posts, photos, short videos, video, live stream)
  - and their interactions
- Networks are a general language for describing such complex systems

We will never be able to model and predict the social media system unless we understand the networks behind it!

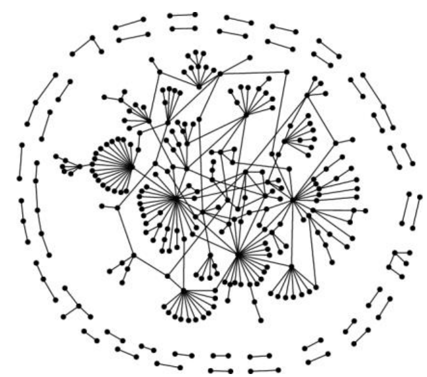
Why Networks?

- **Universal language for describing complex data**
  - Networks from science, nature, and technology are more similar than one would expect
- **Shared vocabulary between fields**



Social networks

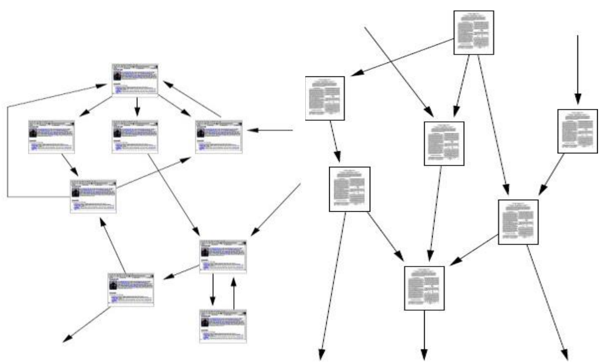
Road networks



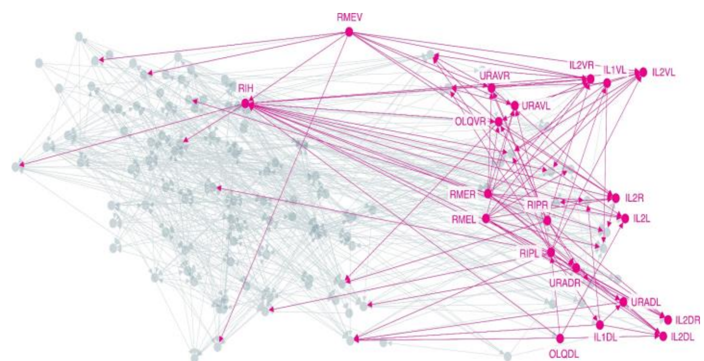
Communication graphs

Why Networks?

- **Universal language for describing complex data**
  - Networks from science, nature, and technology are more similar than one would expect
- **Shared vocabulary between fields**

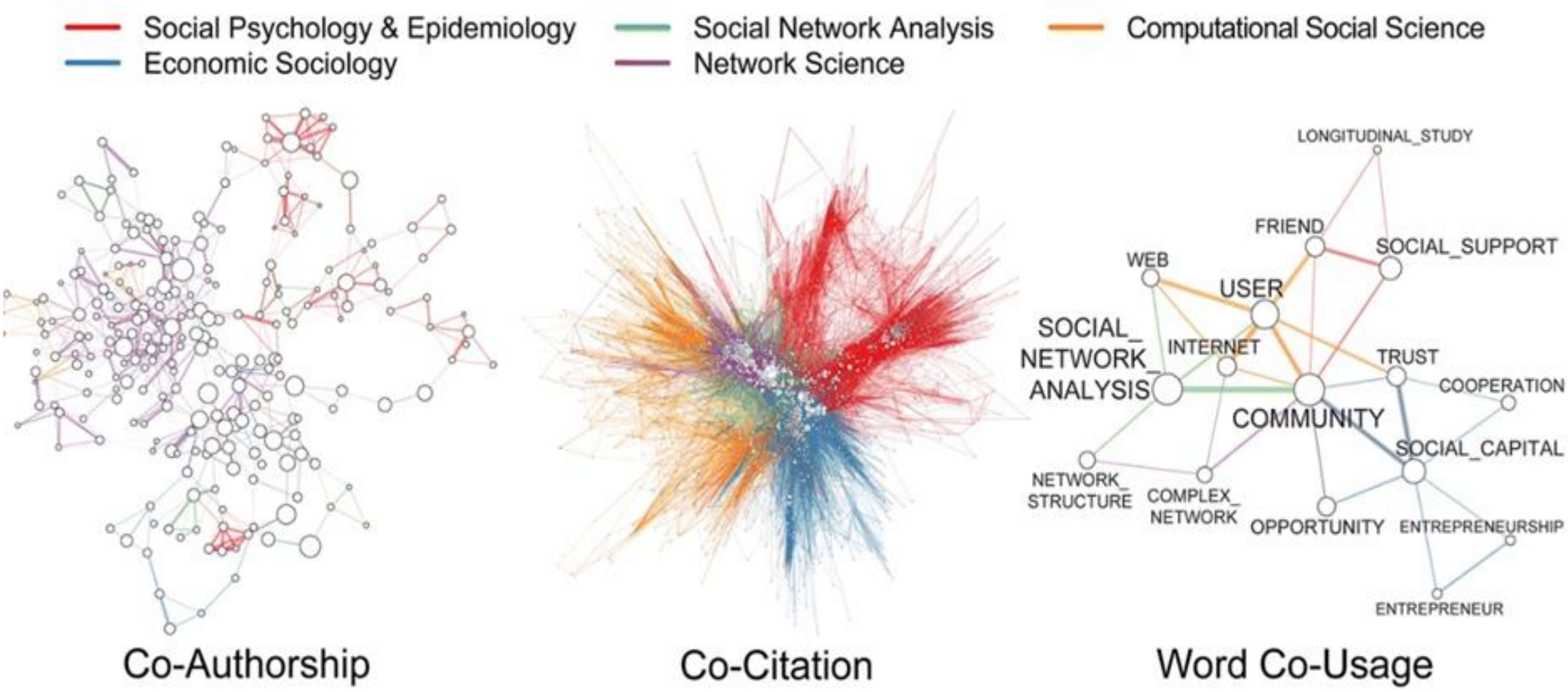


Information networks:
Web & citations



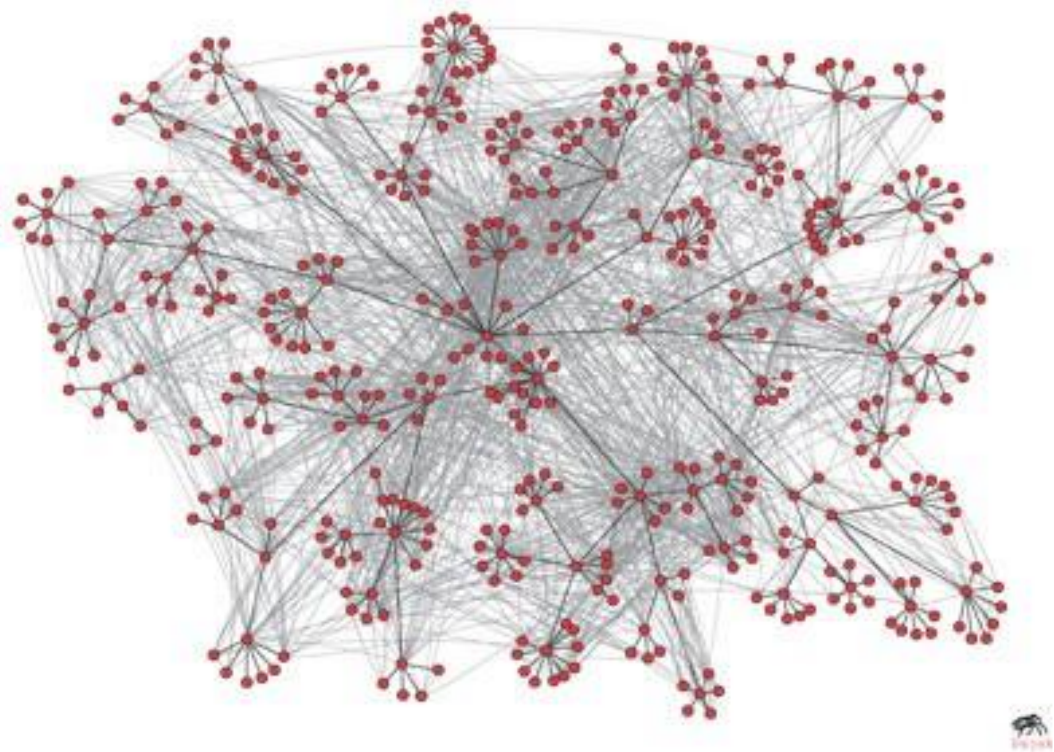
Networks of neurons

Many Types of Data are Networks



Graph Basics

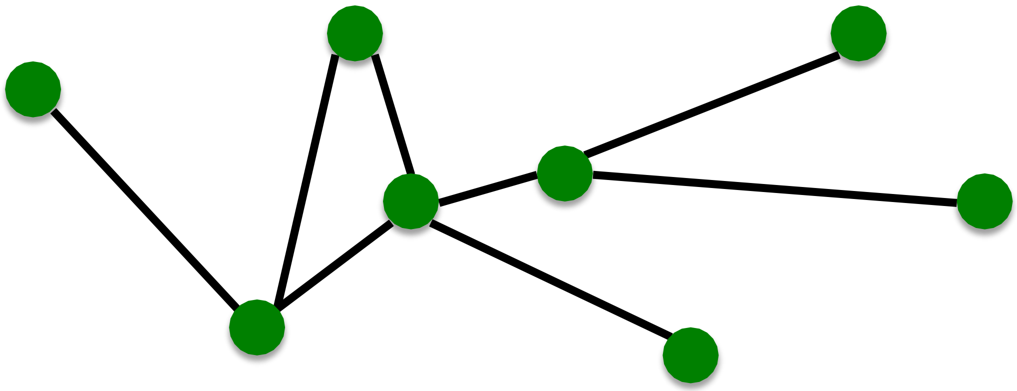
Structure of Networks



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

What are components of a network?

Components of a Network



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

Networks or Graphs?

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

  Language: Network, node, link

- **Graph** is a mathematical representation of a network

  Language: Graph, vertex, edge

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

Nodes or Actors

- In a friendship social graph, nodes are users and a link denotes the friendship between two users
- Depending on the context, these nodes have different names
  - In a web graph, "*nodes*" represent sites and the connection between nodes indicates web-links between them
  - In a social setting, these nodes are called actors

$$V = \{v_1, v_2, \ldots, v_n\}$$

  - The number of nodes is

$$|V| = \mathbf{n}$$

Edges

- Links/edges that connect user nodes are also known as **ties** or **relationships in the social setting**

- In a social setting, where nodes represent social entities such as people, edges indicate social relationships, therefore known as social ties

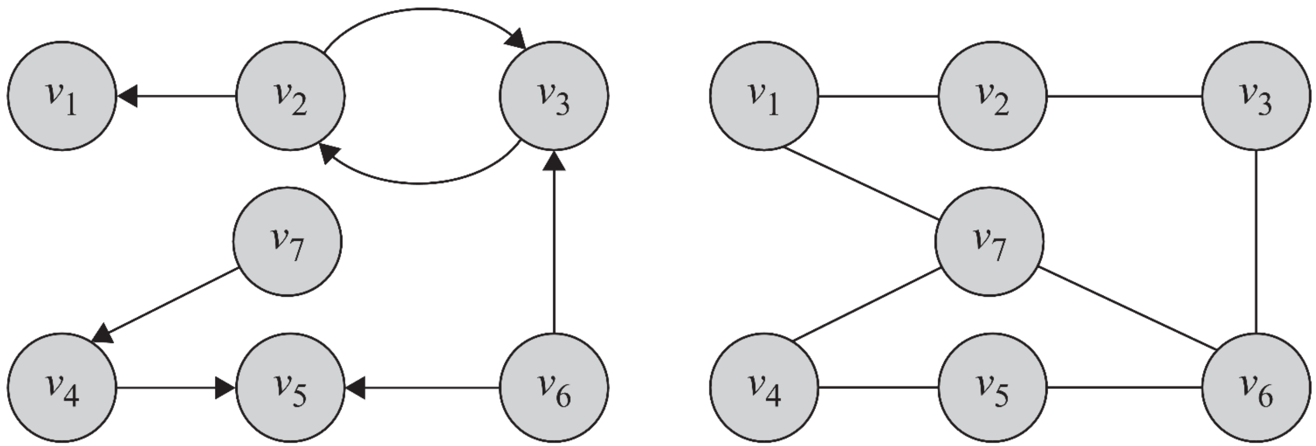
$$E = \{e_1, e_2, \ldots, e_m\}$$

- The number of edges (size of the edge-set) is denoted as

$$|E| = \mathbf{m}$$

Directed Edges and Directed Graphs

- Edges can have directions. A directed edge is sometimes called an arc



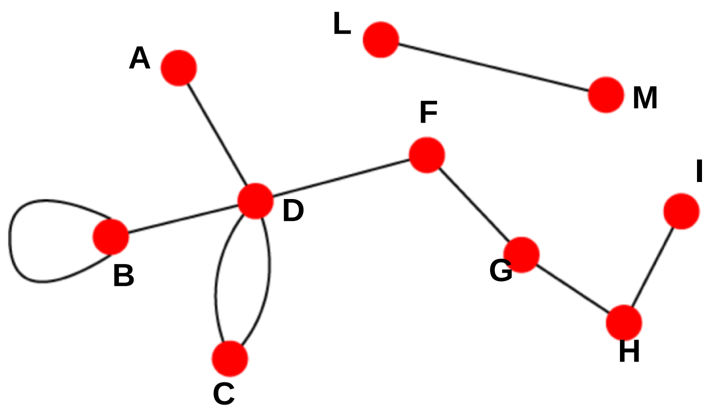
(a) Directed Graph (b) Undirected Graph

- An edge is represented by a starting - end node pair $e(v_2, v_1)$
- In undirected graphs both representations $e(v_2, v_1)$ and $e(v_1, v_2)$ are the same, referring to the same edge.

Directed Graphs vs. Undirected Graphs

Undirected

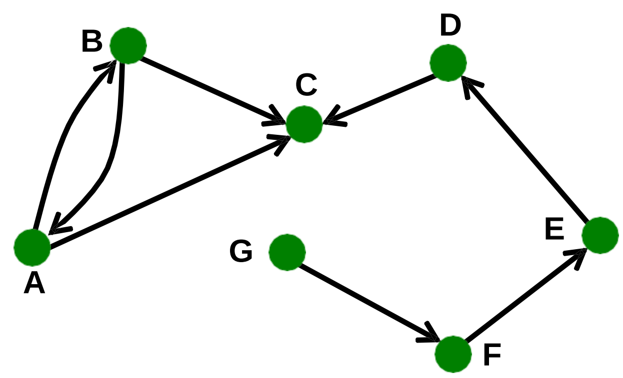
- **Links:** undirected (symmetrical, reciprocal)



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

Directed

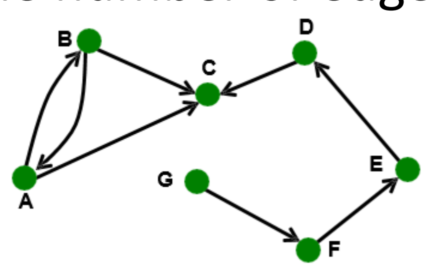
- **Links:** directed (arcs)



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

Neighbourhood and Degree

- For any node v, in an undirected graph, the set of nodes it is directly connected to is called its neighbourhood and is represented as $N(v)$
  - Directed graphs have incoming neighbors $\mathrm{N_{in}(v)}$ (nodes that point to v) and outgoing neighbors $\mathrm{N_{out}(v)}$ (nodes that are pointed by v).
- The number of edges connected to one node is the degree of that node (the size of its neighborhood)
  - Degree of a node i is usually presented using notation d_i
- In directed graphs:
  - d_i^{in} in-degrees is the number of edges pointing towards a node
  - d_i^{out} out-degree is the number of edges pointing away from a node



Degree and Degree Distribution

- **Theorem 1.** The summation of node degrees in an undirected graph is twice the number of edges

$$\sum_i d_i = 2|E|$$

What is the average degree of an undirected graph?

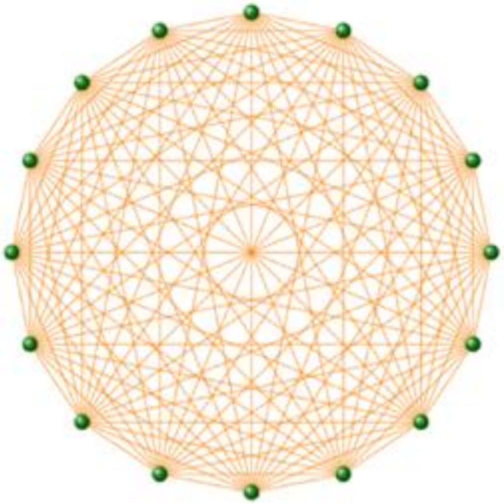
- **Lemma 1.** In any directed graph, the summation of in-degrees is equal to the summation of out-degrees,

$$\sum_i d_i^{out} = \sum_j d_j^{in}$$

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 $N\text{-}1$

Degree Distribution

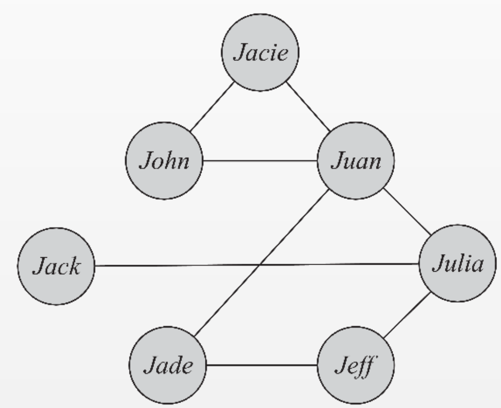
When dealing with very large graphs, the nodes' degree distribution is an important property of a network.

$$\pi(d) = \{d_1, d_2, \ldots, d_n\}$$ (Degree sequence)

$$p_d = \frac{n_d}{n}$$ (Probability of degree *d*, i.e, the fraction of nodes having degree *d*)

n_d is the number of nodes with degree d

$$\sum_{d=0}^{\infty} p_d = 1$$



$$p_1 = \frac{1}{7}, p_2 = \frac{4}{7}, p_3 = \frac{1}{7}, p_4 = \frac{1}{7}$$

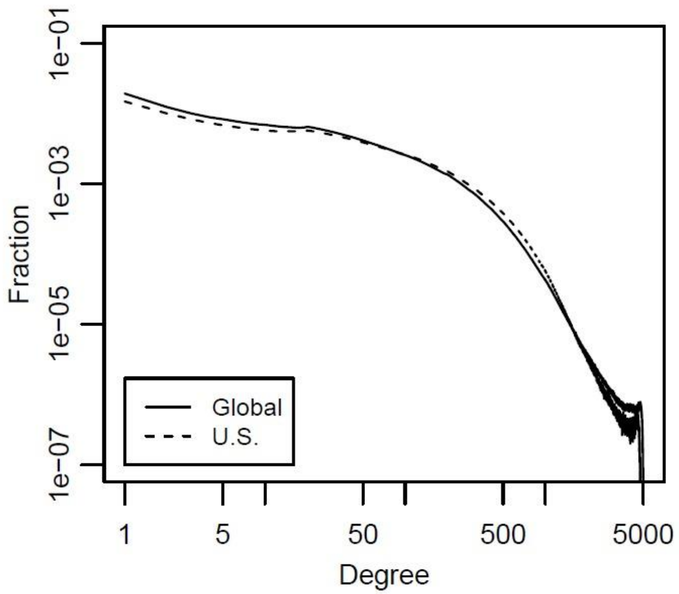
Degree Distribution Plot

The x-axis represents the degree and the y-axis represents the fraction of nodes having that degree

– On social networking sites

There exist many users with few connections and there exist a handful of users with very large numbers of friends.

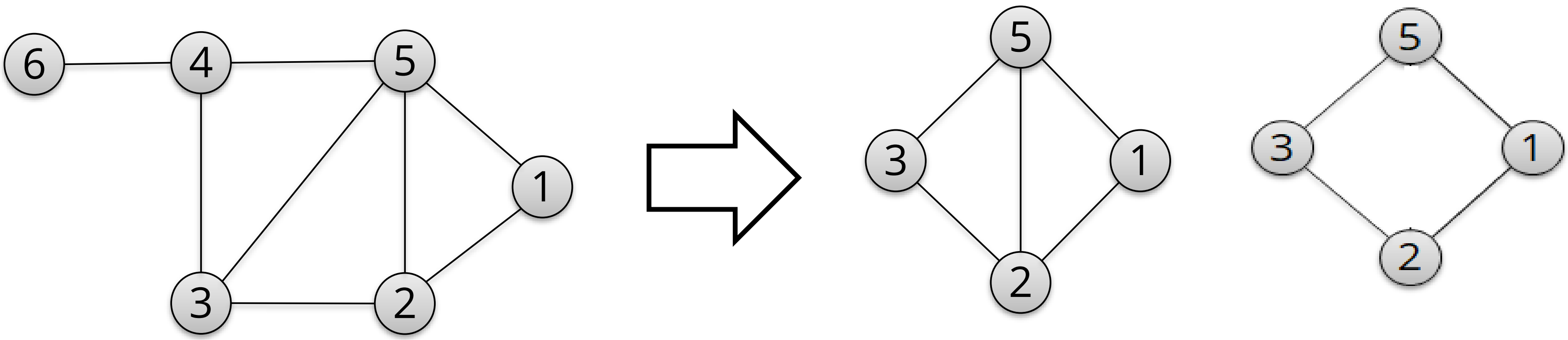
(**Power-law degree distribution)**



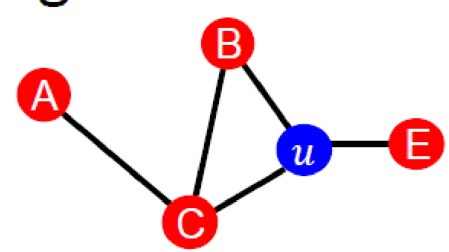
Facebook
Degree Distribution

Subgraph

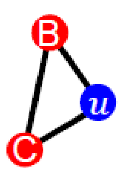
- A subgraph S of a graph G is another graph formed from a subset of the vertices and edges of G.



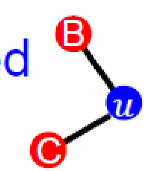
- **Def: Induced subgraph** is another graph, formed from a subset of vertices and *all* of the edges connecting the vertices in that subset.



Induced subgraph:

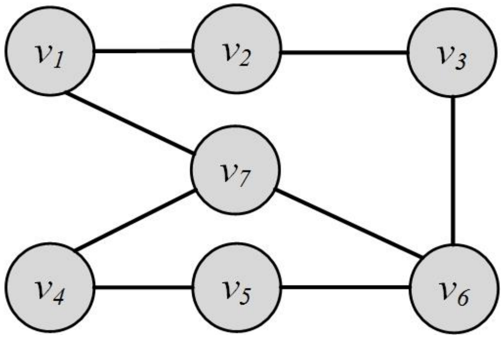


Not induced subgraph:

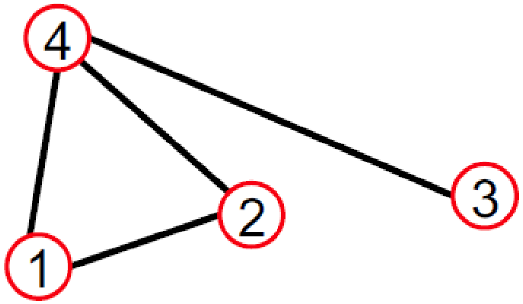


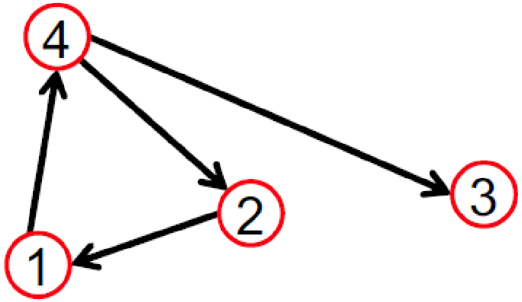
Graph Representation

- Adjacency Matrix
- Edge List
- Adjacency List
- Embedding



Adjacency Matrix





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

$A_{ij} = 0$ otherwise

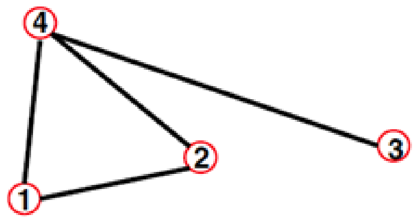
$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \qquad A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

$(\boldsymbol{A} = \boldsymbol{A}^{\boldsymbol{T}})$

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

Adjacency Matrix

Undirected



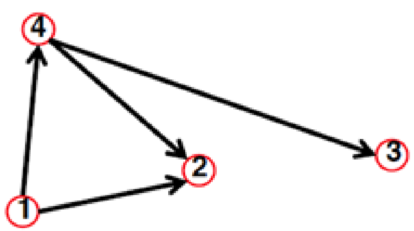
$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A_{ij} = A_{ji}$$
$$A_{ii} = 0$$

$$\mathrm{d}_i = \sum_{j=1}^{N} A_{ij}$$

$$\mathrm{d}_j = \sum_{i=1}^{N} A_{ij}$$

Directed



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

$$A_{ij} \neq A_{ji}$$
$$A_{ii} = 0$$

$$\mathrm{d}_i^{out} = \sum_{j=1}^{N} A_{ij}$$

$$\mathrm{d}_j^{in} = \sum_{i=1}^{N} A_{ij}$$

Social media networks have very sparse Adjacency matrices

Networks are Sparse Graphs

Most real-world networks are sparse

$$|E| \ll E_{max} \quad (\text{or } \bar{d} \ll N-1)$$

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

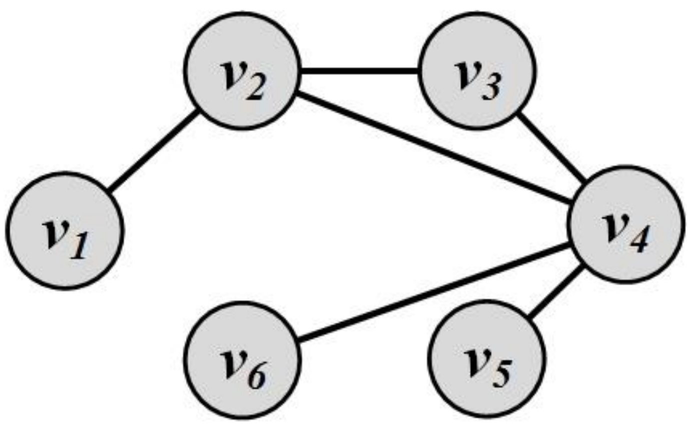
(Source: *Leskovec et al., Internet Mathematics, 2009*)

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})

Edge List

- In this representation, each element is an edge and is usually represented as (u, v), denoting that node u is connected to node v

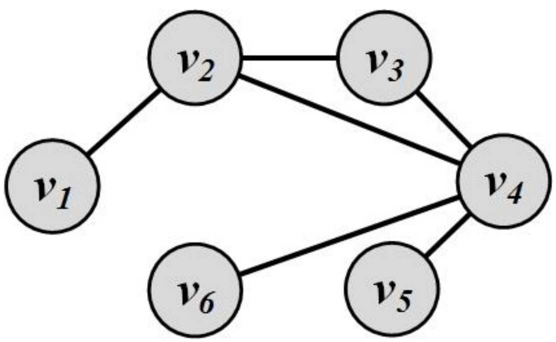


(v_1, v_2)
(v_2, v_3)
(v_2, v_4)
(v_3, v_4)
(v_4, v_5)
(v_4, v_6)

Given a node v, how to find all its neighbors?
What is the time complexity?

Adjacency List

- In an adjacency list, for every node, we maintain a list of its neighbors
- The list is usually sorted based on the node order or other preferences

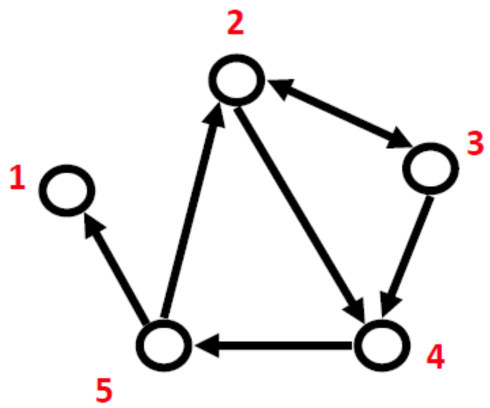


Key | **Value**

Node	Connected To
v_1	v_2
v_2	v_1, v_3, v_4
v_3	v_2, v_4
v_4	v_2, v_3, v_5, v_6
v_5	v_4
v_6	v_4

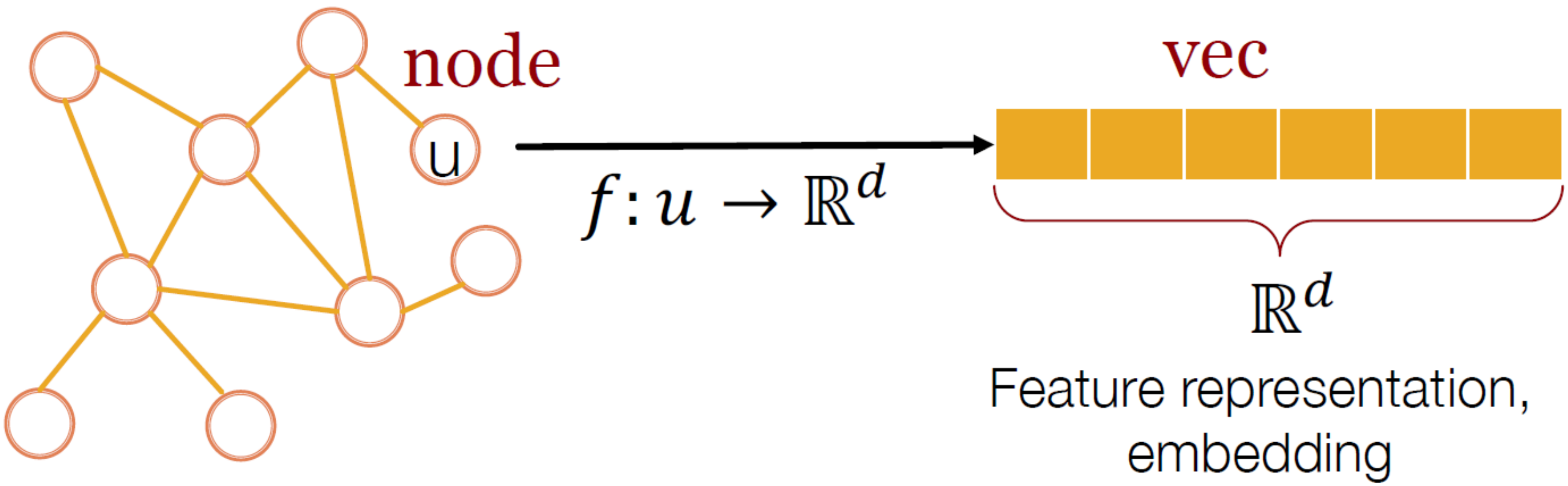
Adjacency List

- **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



Network Embedding/Graph Representation Learning

- We map each node in a network into a low dimensional space so that the network structure information can be effectively preserved.



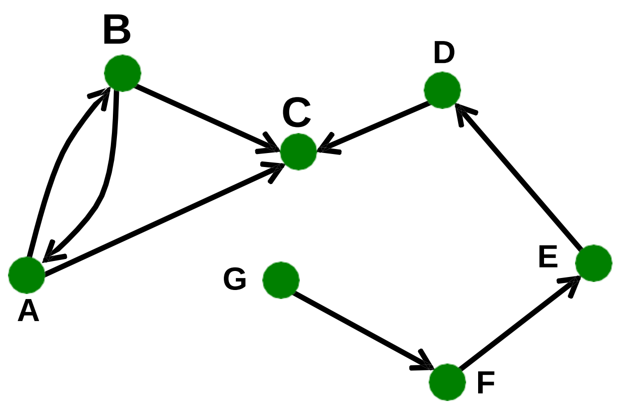
We use a low-dimensional vector to represent a node.
Why is the vector called "low dimensional"?

Connectivity in Graphs

- **Adjacent nodes/Incident Edges, Walk/Path/Trail/Tour/Cycle**

Adjacent nodes and Incident Edges

- Two nodes are adjacent if they are connected via an edge.
- Two edges are incident, if they share one node
- When the graph is directed, edge directions must match for edges to be incident
  - Incident edges should have the same direction in a directed graph



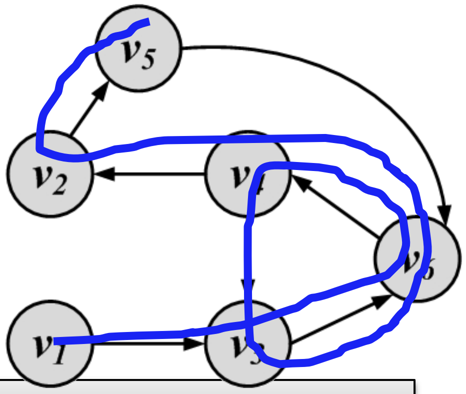
Walk, Path, Trail, Tour, and Cycle

Walk: A walk is a sequence of incident edges visited one after another
- **Open walk**: A walk does not end where it starts
- **Closed walk**: A walk returns to where it starts

- Representing a walk:
  - A sequence of edges: $e_1, e_2, \ldots, e_n$
  - A sequence of nodes: $v_1, v_2, \ldots, v_n$

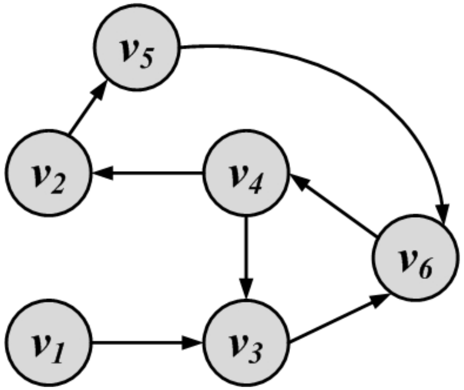
- Length of walk:
  the number of visited edges

Length of walk= 8



Random walk

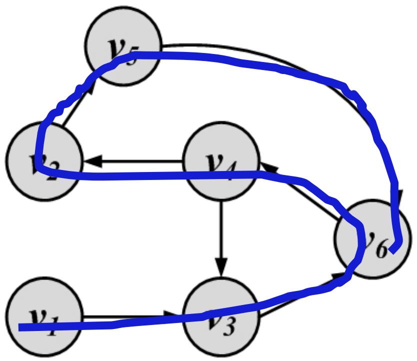
- A walk in which the next node is selected randomly among the neighbors each time



- A random walk starting from v_1
  - v_1, v_3, v_6, v_4, v2, v5…
  - v_1, v_3, v_6, v_4, v_3, v6…

Trail

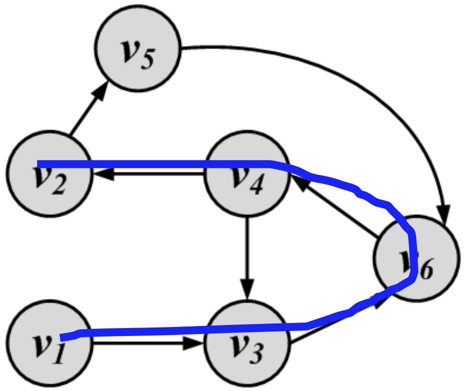
- A trail is a walk where **no edge is visited more than once** and all walk edges are distinct



- A closed trail (one that ends where it starts) is called a **tour** or **circuit**

Path

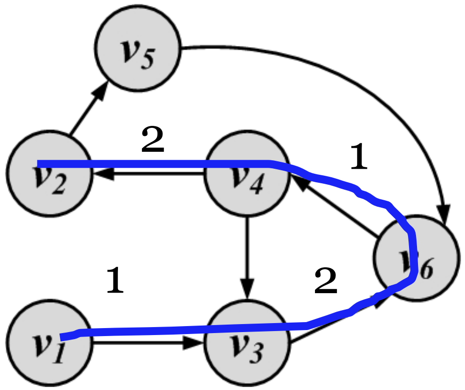
- A walk where **nodes and edges are distinct** is called a **path**
- One special case – the starting node and end node can be the same one. In this case, it is called a **cycle**.
- <u>The length of a path in an unweighted graph</u> is the number of edges visited in the path



Length of path= 4

Path

- A walk where **nodes and edges** **are distinct** is called a **path**
- One special case – the starting node and end node can be the same one. In this case, it is called a **cycle**.
- <u>The length of a path in a weighted graph</u> is the sum of the weights of the edges visited in the path



Length of path= 6

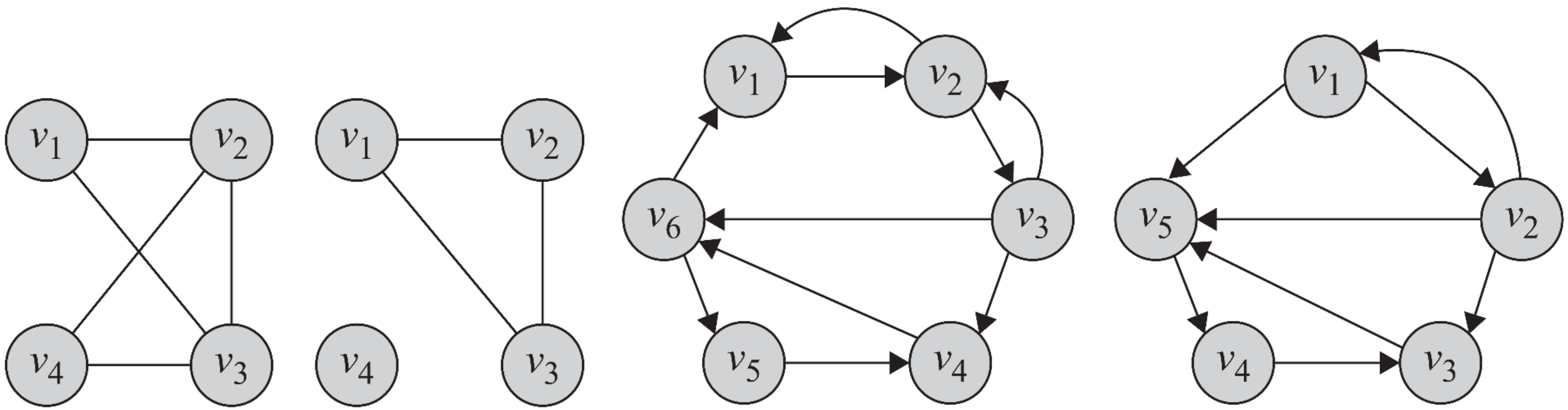
Connectivity

- **A node v_i is connected to node v_j** (or reachable from v_j) if they are adjacent or there exists a **path** from v_i to v_j.

- **A graph is connected**, if there exists a path between any pair of nodes in it
  - In a directed graph, **a graph is strongly connected** if there exists **a directed path** between any pair of nodes
  - In a directed graph, **a graph is weakly connected** if there exists a path between any pair of nodes, without considering the edge directions

- A graph is **disconnected,** if it not connected.

Connectivity: Example

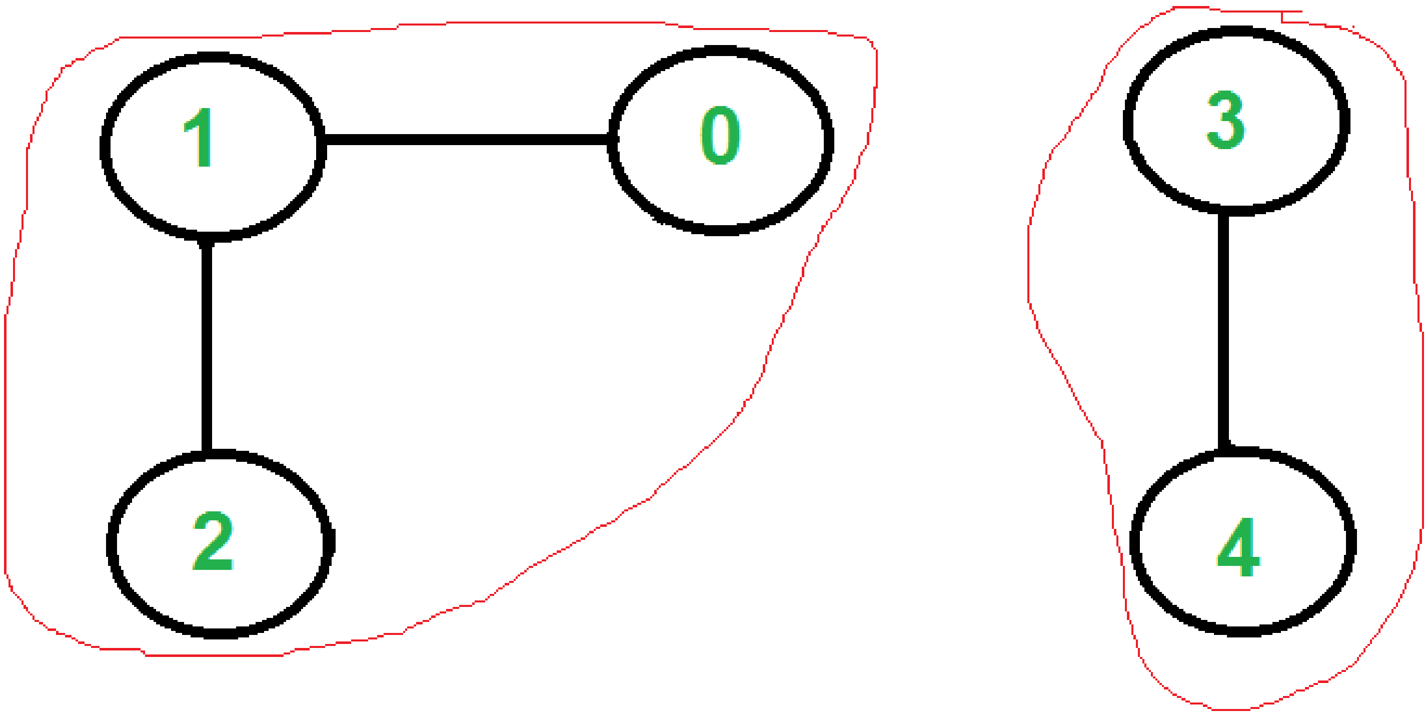


(a) Connected (b) Disconnected (c) Strongly connected (d) Weakly connected

Component

- A **component** of an undirected graph is **a subgraph**
  - where any two nodes are connected to each other, and
  - which is connected to no additional nodes in the **supergraph**
- A component is **strongly connected** in a directed graph if there exists **a directed path** from any node u to any other node v in the component.
- A component is **weakly connected** if it is connected without considering the edge directions

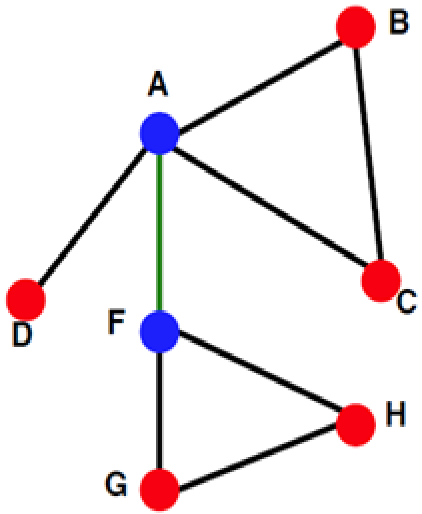
Component

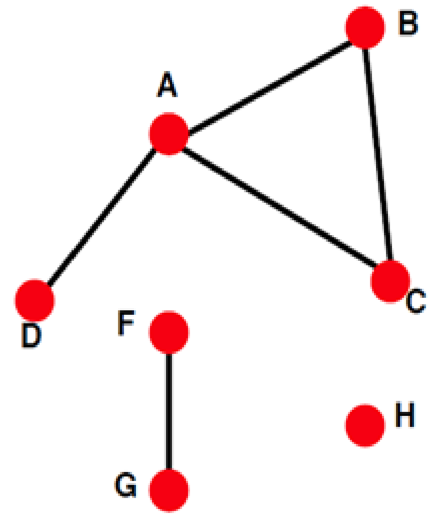


There are two connected components in above undirected graph
0 1 2
3 4

Connectivity of Undirected Graphs

- 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

Shortest Path

- **Shortest Path** is the path between two nodes that has the shortest length.
  - We denote the length of the shortest path between nodes v_i and v_j as $l_{i,j}$
- The length of the shortest path is called **network distance** or distance on a graph from v_i to v_j .

Diameter

The diameter of a graph is the **length of the longest shortest path** between any pair of nodes in the graph

$$diameter_G = \max_{(v_i, v_j) \in V \times V} l_{i,j}$$

- How big is the diameter of the social graph in Facebook?

Types of Graphs

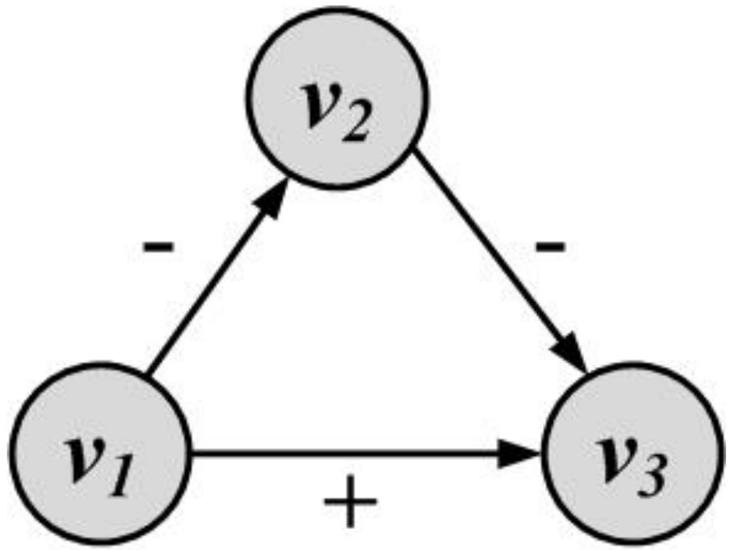
Edge Attributes

Possible options:

- Weight (e.g. frequency of communication)
- 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

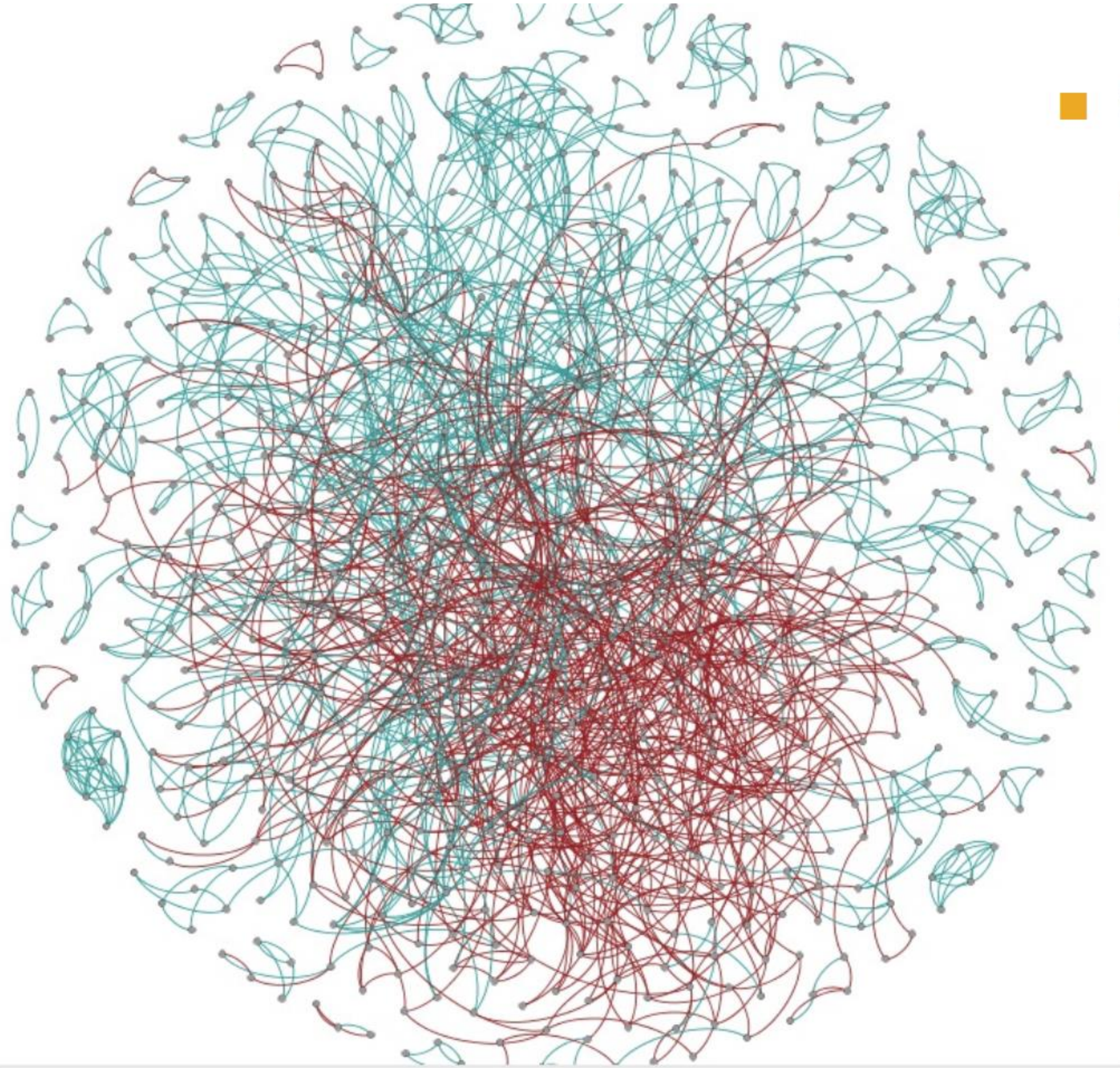
Signed Graphs

- When edge weights are binary (0/1, -1/1, +/-) we have a signed graph



- It is used to represent friends or foes

Signed Graphs

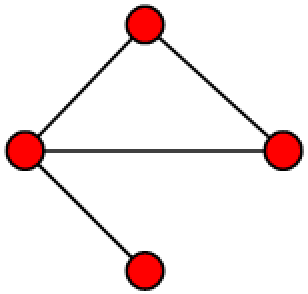


sample of positive & negative ratings from Epinions network

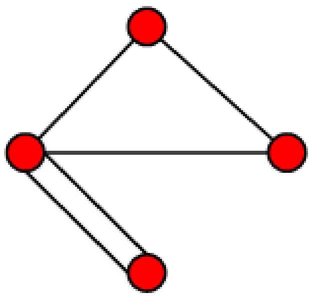
- **One person trusting/distrusting another**
  - Research challenge: How does one 'propagate' negative feelings in a social network? Is my enemy's enemy my friend?

Simple Graphs and Multigraphs

- Simple graphs are graphs where only a single edge is allowed to exist between any pair of nodes
- Multigraphs are graphs where you can have multiple edges between two nodes (loops are allowed)



Simple graph

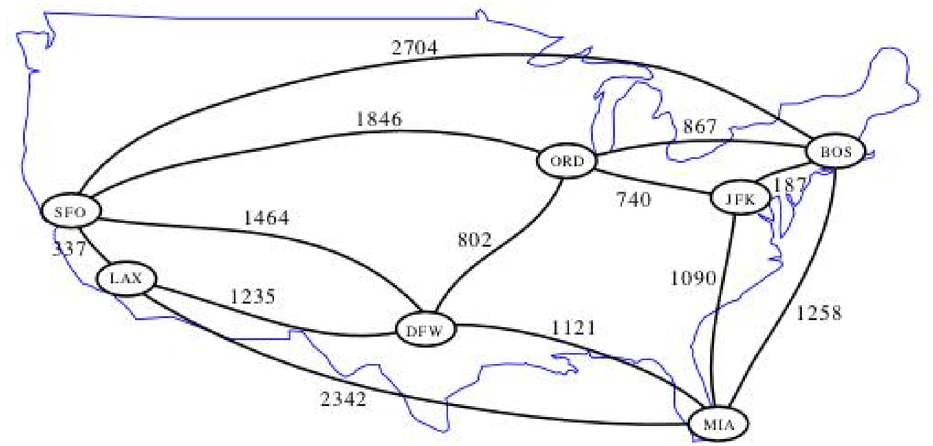


Multigraph

- The adjacency matrix for multigraphs can include elements larger than one, indicating multiple edges between nodes;
  - A_ij denotes the number of edges between node i and node j.

Weighted Graph

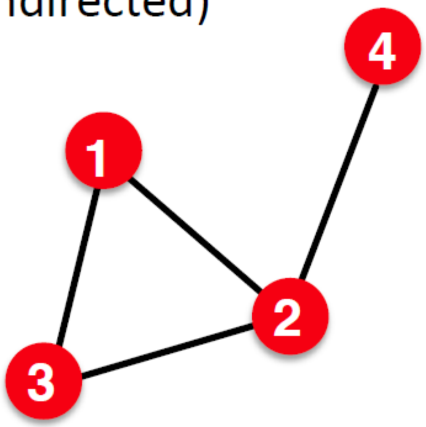
- A weighted graph $\boldsymbol{G(V,E,W)}$ is one where edges are associated with weights
  - For example, a graph could represent a map where nodes are cities and edges are roads between them
    - The weight associated with each edge could represent the distance between the corresponding cities



$$A_{ij} = \begin{cases} w_{ij} \text{ or } w(i,j), w \in R \\ 0, \text{There is no edge between } v_i \text{ and } v_j \end{cases}$$

Weighted Graph

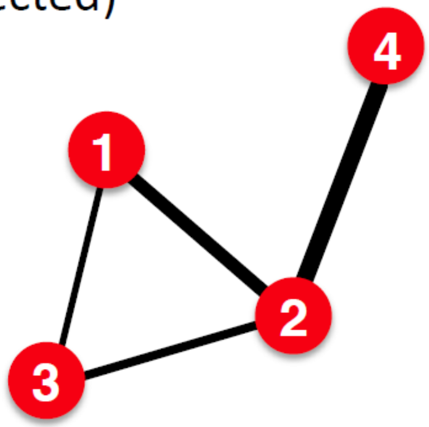
- **Unweighted**
  (undirected)



$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$A_{ii} = 0 \qquad A_{ij} = A_{ji}$

- **Weighted**
  (undirected)



$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

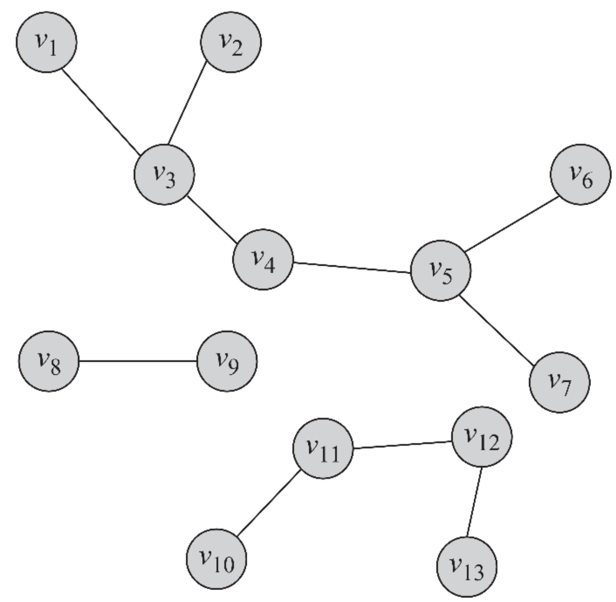
$A_{ii} = 0 \qquad A_{ij} = A_{ji}$

How to compute the degree of each node in a weighted graph?

Trees and Forests

- **Trees** are special cases of undirected graphs
- A tree is a graph structure that has no cycle in it
- In a tree, there is exactly one path between any pair of nodes
- In a tree: $|V| = |E| + 1$

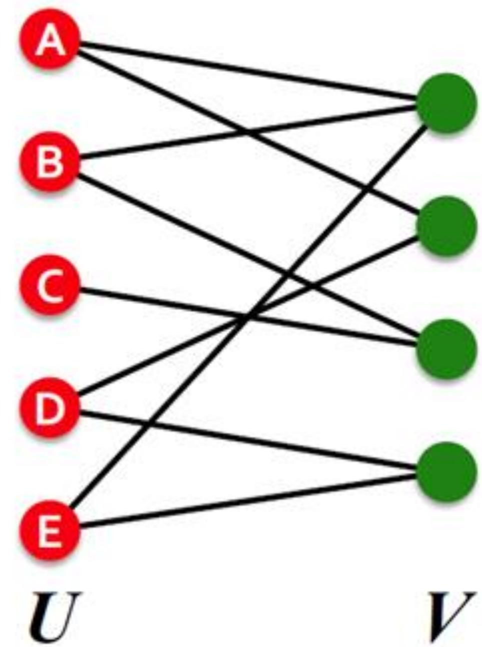
- A set of disconnected trees is called a **forest**



A forest containing 3 trees

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 (they authored)
  - Actors-to-Movies (they appeared in)
  - Users-to-Movies (they rated)
  - Recipes-to-Ingredients (they contain)

Line graph

- The **line graph** of an undirected graph *G* is another graph L(*G*) that represents the adjacencies between edges of *G.*
- L(G) is constructed in the following way:
  - for each edge in G, make a vertex in L(G);
  - for every two edges in G that have a vertex in common, make an edge between their corresponding vertices in L(G).

Each edge in the graph corresponds to a node in the line graph.

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