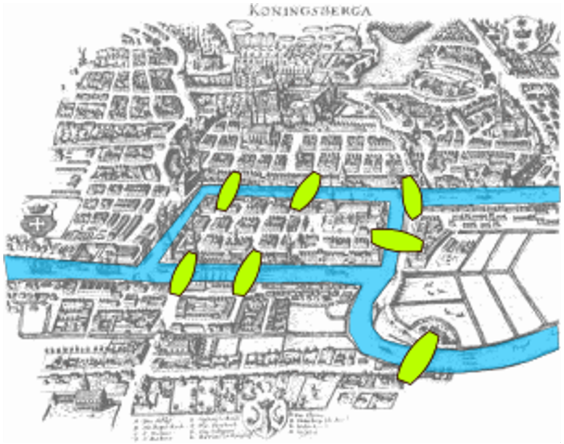
Social Media Mining

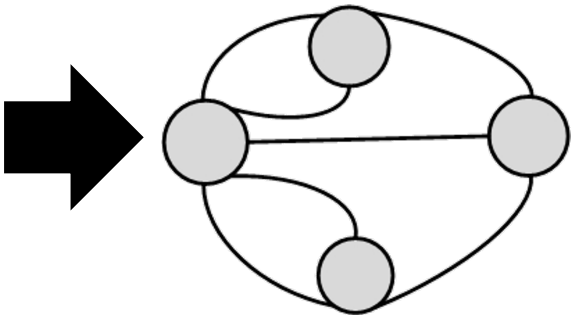
Graph Essentials

Bridges of Konigsberg

- There are 2 islands and 7 bridges that connect the islands and the mainland
- Find a path that crosses each bridge exactly once



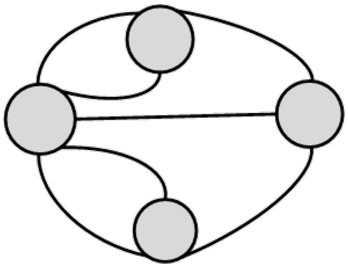
City Map (From Wikipedia)



Graph Representation

Modeling the Problem by Graph Theory

- The key to solve this problem is an ingenious graph representation
- Euler proved that since except for the starting and ending point of a walk, one has to enter and leave all other nodes, thus these nodes should have an even number of bridges connected to them
- This property does not hold in this problem

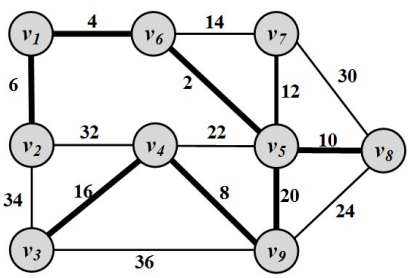


Networks

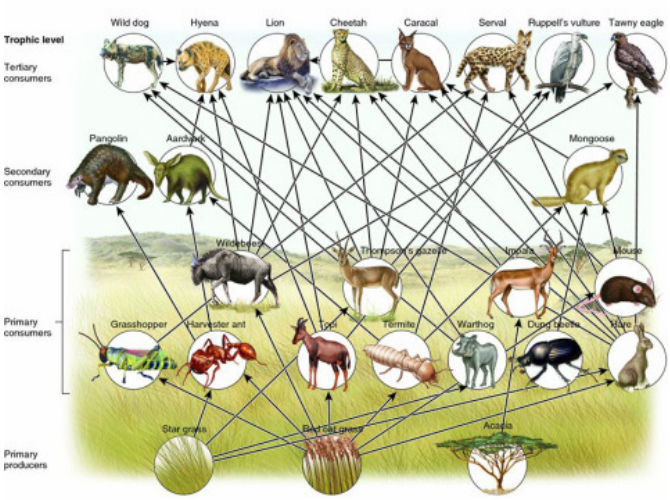
- A network is a graph.
  - Elements of the network have meanings
- Network problems can usually be represented in terms of graph theory

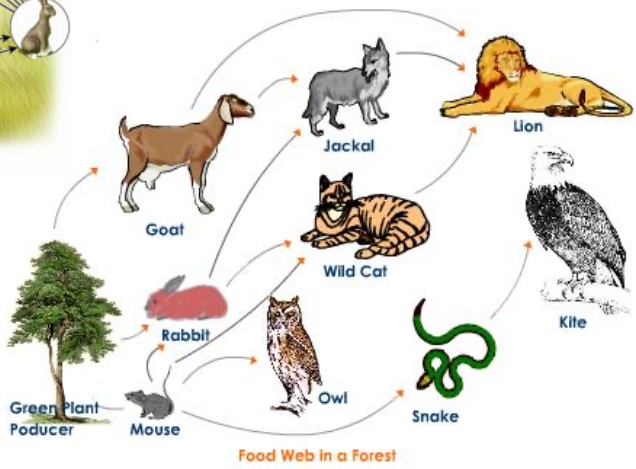
Twitter example:

- Given a piece of information, a network of individuals, and the cost to propagate information among any connected pair, find the minimum cost to disseminate the information to all individuals.



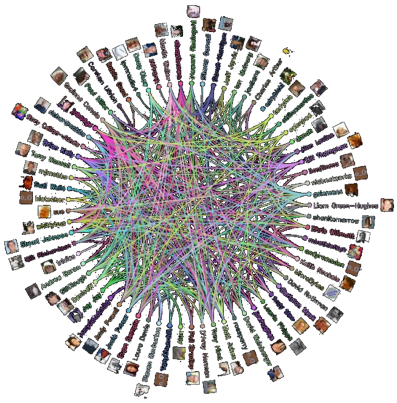
Food Web

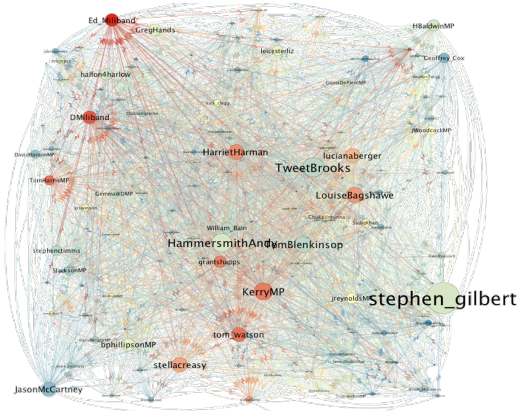




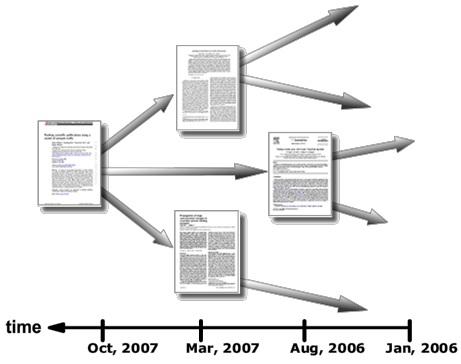
Network are Pervasive

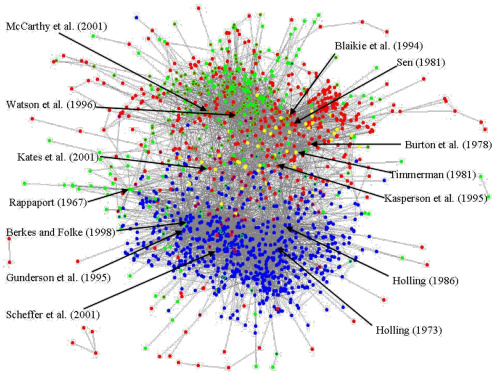
Twitter Networks



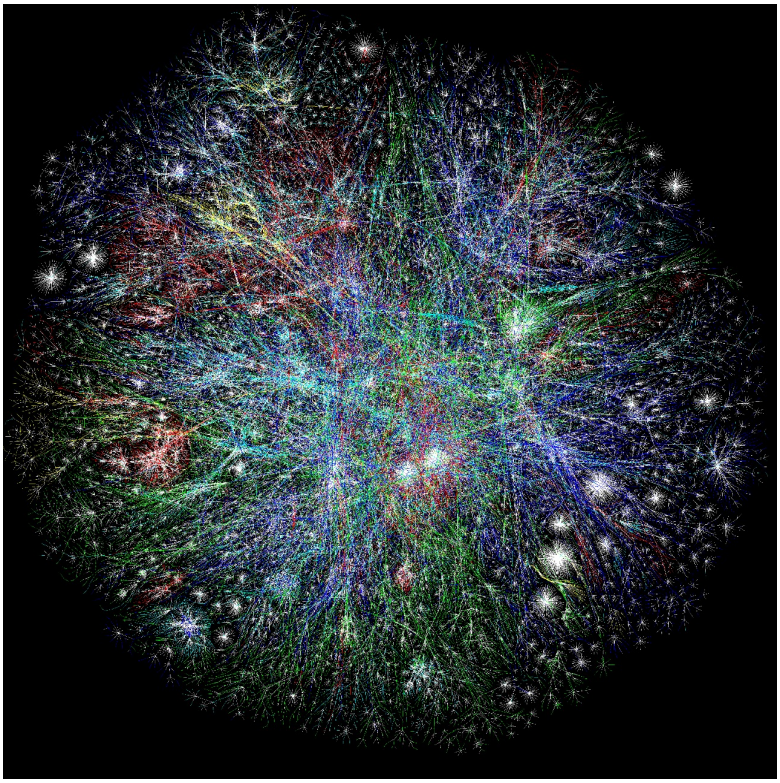


Citation Networks





Internet



Network of the US Interstate Highways

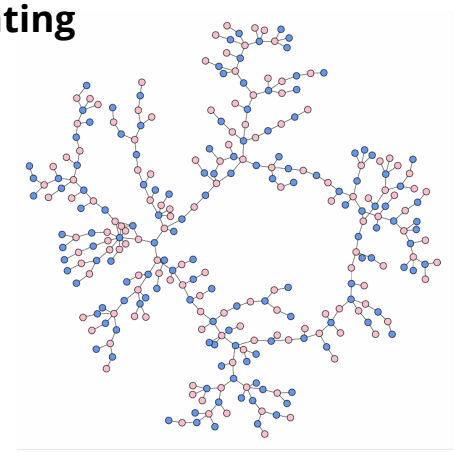
NY State Road Network

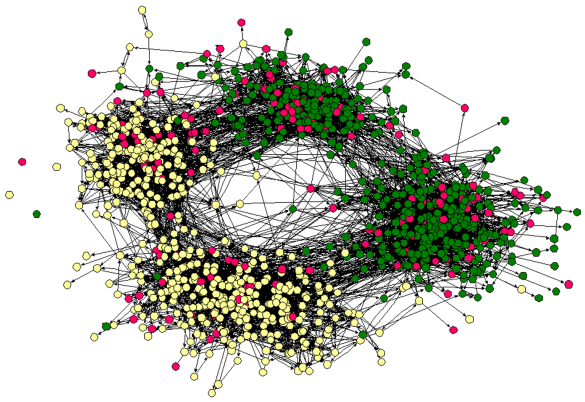
Social Networks and Social Network Analysis

- A social network
  - A network where elements have a social structure
    - A set of actors (such as individuals or organizations)
    - A set of ties (connections between individuals)
- Social networks examples:
  - your family network, your friend network, your colleagues ,etc.
- To analyze these networks we can use Social Network Analysis (SNA)
- Social Network Analysis is an interdisciplinary field from social sciences, statistics, graph theory, complex networks, and now computer science

Social Networks: Examples

High school dating





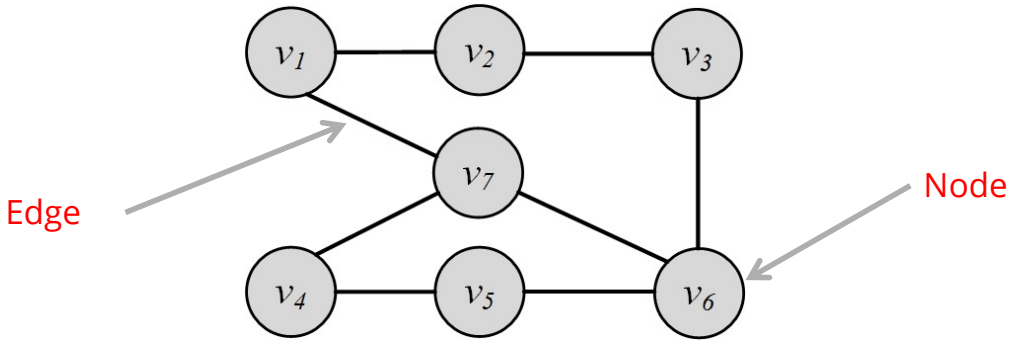
High school friendship

Graph Basics

Nodes and Edges

A network is a graph, or a collection of points connected by lines

- Points are referred to as **nodes, actors,** or **vertices** (plural of **vertex**)
- Connections are referred to as **edges** or **ties**



Nodes or Actors

- In a friendship social graph, nodes are people and any pair of people connected denotes the friendship between them
- Depending on the context, these nodes are called nodes, or actors
  - 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 size of the graph is $|V| = \mathbf{n}$

Edges

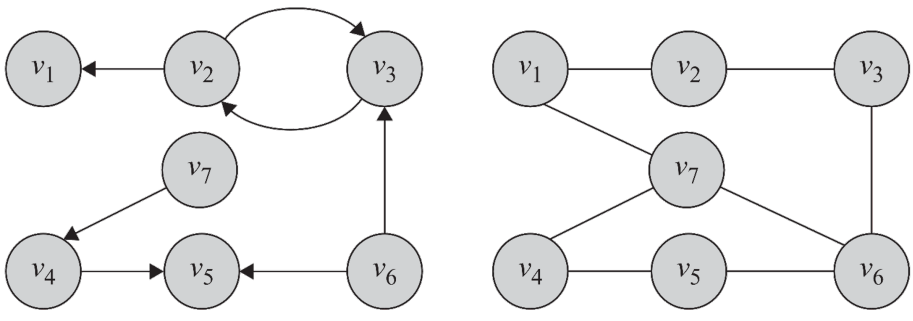
- Edges connect nodes and are also known as **ties** or **relationships**
- In a social setting, where nodes represent social entities such as people, edges indicate internode relationships and are therefore known as relationships or (social) ties

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

- Number is 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

- Edges are represented using their end-points $e(v_2, v_1)$.
- In undirected graphs both representations are the same

Neighborhood and Degree (In-degree, out-degree)

For any node v, in an undirected graph, the set of nodes it is connected to via an edge is called its neighborhood and is represented as $N(v)$

- *In directed graphs we have incoming neighbors $N_{in}(v)$ (nodes that connect to v) and outgoing neighbors $N_{out}(v)$.*

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 degrees in an undirected graph is twice the number of edges

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

- **Lemma 1.** The number of nodes with odd degree is even
- **Lemma 2.** In any directed graph, the summation of in-degrees is equal to the summation of out-degrees,

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

Degree Distribution

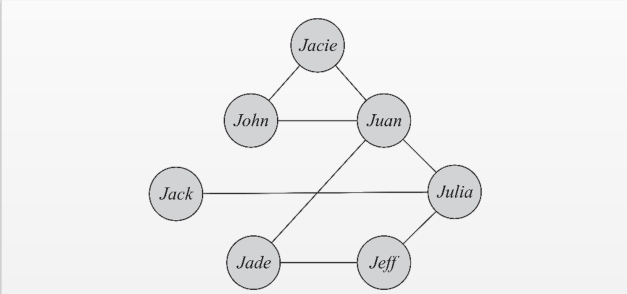
When dealing with very large graphs, how nodes' degrees are distributed is an important concept to analyze and is called ***Degree Distribution***

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

$$p_d = \frac{n_d}{n}$$

n_d is the number of nodes with degree d

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



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

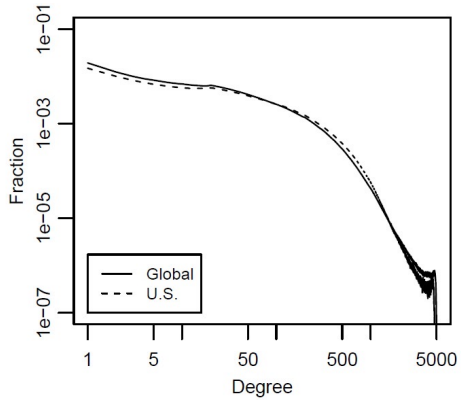
Degree Distribution Plot

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**)



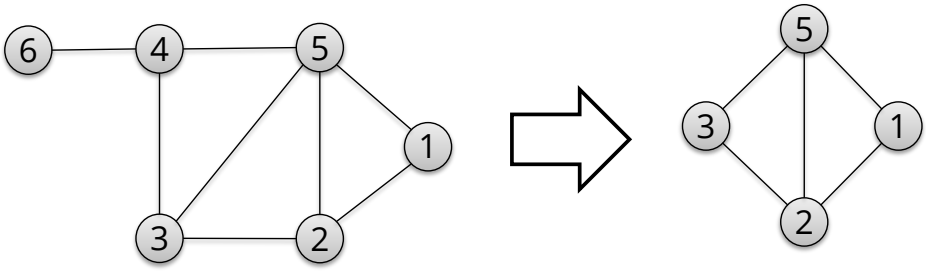
Facebook Degree Distribution

Subgraph

- Graph G can be represented as a pair $G(V, E)$ where V is the node set and E is the edge set

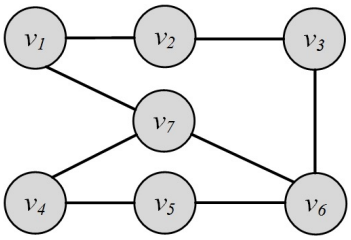
- $G'(V', E')$ is a subgraph of $G(V, E)$

$$V' \subseteq V$$
$$E' \subseteq (V' \times V') \cap E$$



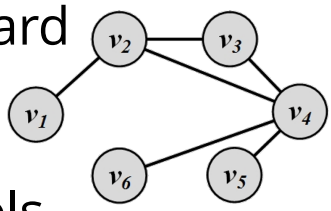
Graph Representation

- **Adjacency Matrix**
- **Adjacency List**
- **Edge List**



Graph Representation

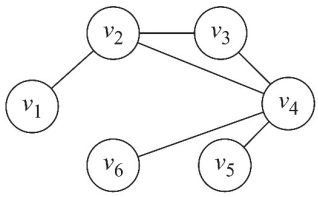
- Graph representation is straightforward and intuitive, but it cannot be effectively manipulated using mathematical and computational tools



- We are seeking representations that can store these two sets in a way such that
  - Does not lose information
  - Can be manipulated easily by computers
  - Can have mathematical methods applied easily

Adjacency Matrix (a.k.a. sociomatrix)

$$A_{ij} = \begin{cases} 1, \text{ if there is an edge between nodes } v_i \text{ and } v_j \\ 0, \text{ otherwise} \end{cases}$$



(a) Graph

	v_1	v_2	v_3	v_4	v_5	v_6
v_1	0	1	0	0	0	0
v_2	1	0	1	1	0	0
v_3	0	1	0	1	0	0
v_4	0	1	1	0	1	1
v_5	0	0	0	1	0	0
v_6	0	0	0	1	0	0

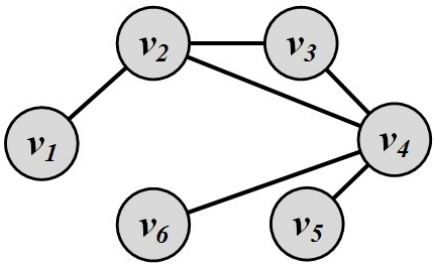
(b) Adjacency Matrix

Diagonal Entries are self-links or loops

Social media networks have very sparse Adjacency matrices

Adjacency List

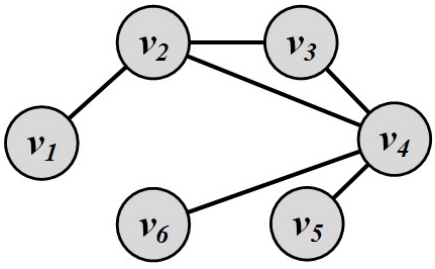
- In an adjacency list for every node, we maintain a list of all the nodes that it is connected to
- The list is usually sorted based on the node order or other preferences



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

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 via an edge



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

Types of Graphs

- **Null, Empty, Directed/Undirected/Mixed, Simple/Multigraph, Weighted, Signed Graph, Webgraph**

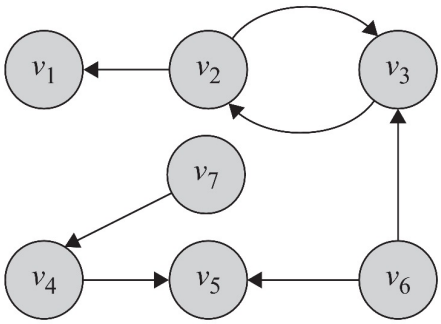
Null Graph and Empty Graph

- A **null graph** is one where the node set is empty (there are no nodes)
  - Since there are no nodes, there are also no edges

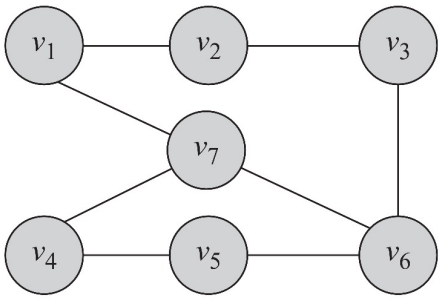
$$G(V, E), V = E = \emptyset$$

- An **empty graph** or **edge-less graph** is one where the edge set is empty, $E = \emptyset$
- The node set can be non-empty.
  - A null-graph is an empty graph.

Directed/Undirected/Mixed Graphs



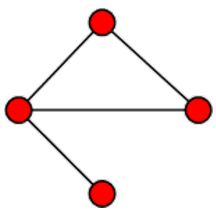
- The adjacency matrix for directed graphs is often not symmetric ($\boldsymbol{A \neq A^T}$)
  - $\boldsymbol{A_{ij} \neq Aji}$
  - We can have equality though



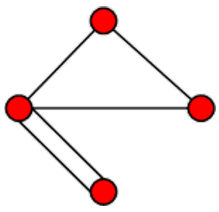
The adjacency matrix for undirected graphs is symmetric ($\boldsymbol{A = A^T}$)

Simple Graphs and Multigraphs

- Simple graphs are graphs where only a single edge can be between any pair of nodes
- Multigraphs are graphs where you can have multiple edges between two nodes and loops



Simple graph

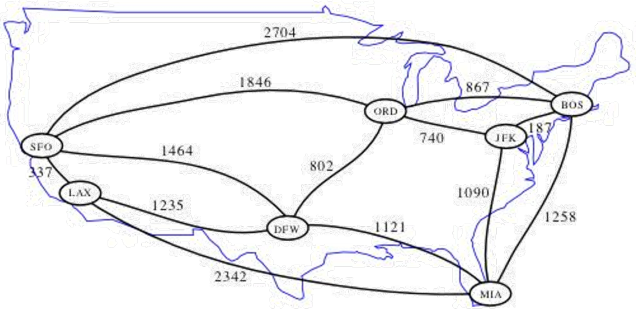


Multigraph

- The adjacency matrix for multigraphs can include numbers larger than one, indicating multiple edges between nodes

Weighted Graph

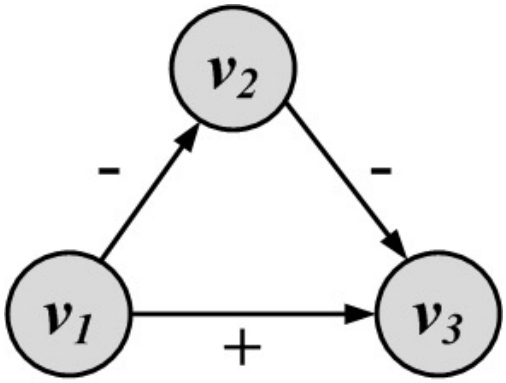
- A weighted graph $\boldsymbol{G(V, E, W)}$ is one where edges are associated with weights
  - For example, a graph could represent a map where nodes are airports and edges are routes between them
    - The weight associated with each edge could represent the distance between the corresponding cities



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

Signed Graph

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

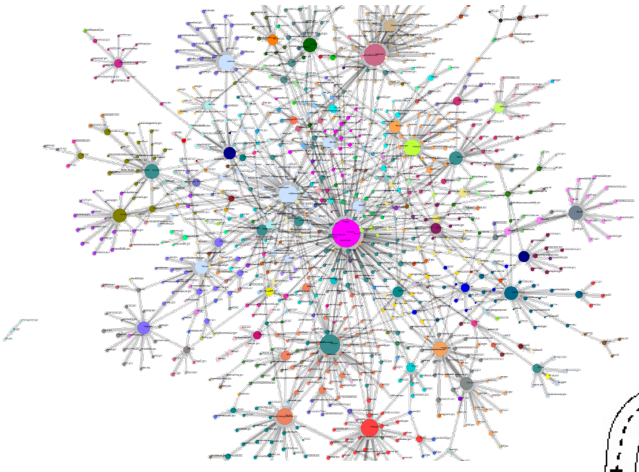


- It is used to represent friends or foes
- It is also used to represent social status

Webgraph

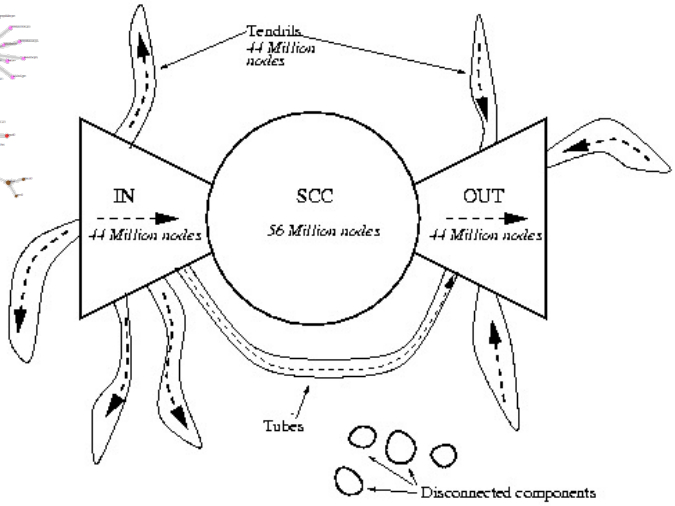
- A webgraph is a way of representing how internet sites are connected on the web
- In general, a web graph is a directed multigraph
- Nodes represent sites and edges represent links between sites.
- Two sites can have multiple links pointing to each other and can have loops (links pointing to themselves)

Webgraph



Government Agencies

Bow-tie structure



Broder et al –
200 million pages, 1.5 billion links

Connectivity in Graphs

- **Adjacent nodes/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 on end-point

When the graph is directed, edge directions must match for edges to be incident

An edge in a graph can be traversed when one starts at one of its end-nodes, moves along the edge, and stops at its other end-node.

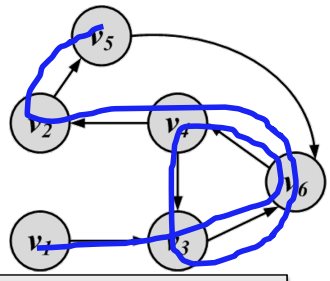
Walk, Path, Trail, Tour, and Cycle

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: $e_1, e_2, \ldots, en$
  - A sequence of nodes: $v_1, v_2, \ldots, vn$

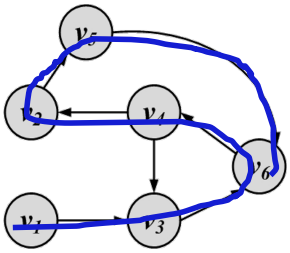
- Length of walk:
  the number of visited edges

Length of walk= 8



Trail

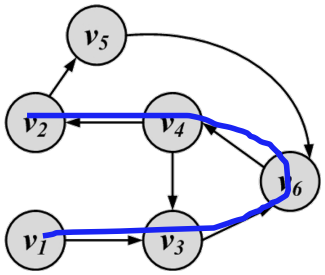
- A trail is a walk where **no edge is visited more than once** and all walk edges are distinct



- 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** and a closed path is called a **cycle**
- The length of a path or cycle is the number of edges visited in the path or cycle

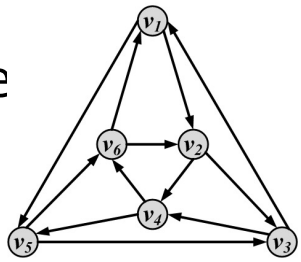


Length of path= 4

Examples

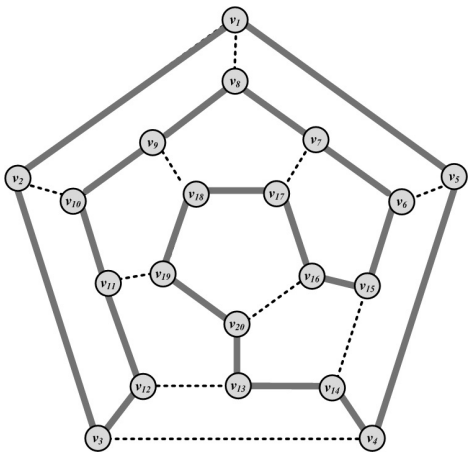
Eulerian Tour

- All edges are traversed only once
  - Konigsberg bridges



Hamiltonian Cycle

- A cycle that visits all nodes



Random walk

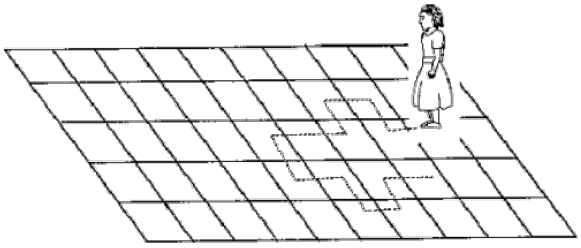
- A walk that in each step the next node is selected randomly among the neighbors
  - The weight of an edge can be used to define the probability of visiting it
  - For all edges that start at v_i the following equation holds

$$\sum_x w_{i,x} = 1, \forall i, j \quad w_{i,j} \geq 0$$

Random Walk: Example

Mark a spot on the ground

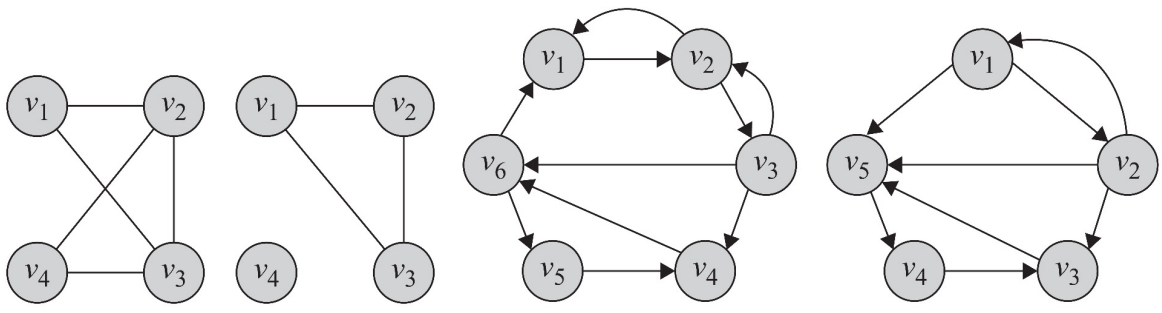
- Stand on the spot and flip the coin (or more than one coin depending on the number of choices such as left, right, forward, and backward)
- If the coin comes up heads, turn to the right and take a step
- If the coin comes up tails, turn to the left and take a step
- Keep doing this many times and see where you end up



Connectivity

- **A node v_i is connected to node v_j** (or reachable from v_j) if it is adjacent to it 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 following the edge directions
- A graph is **disconnected,** if it not connected.

Connectivity: Example

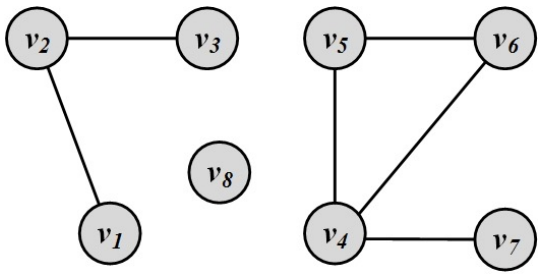


(a) Connected (b) Disconnected (c) Strongly connected (d) Weakly connected

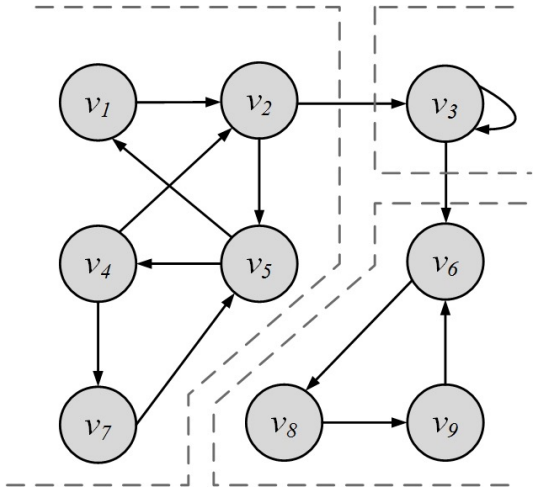
Component

- A **component** in an undirected graph is a connected **subgraph**, i.e., there is a path between every pair of nodes inside the component
- In directed graphs, we have a **strongly connected** components when there is a path from u to v and one from v to u for every pair of nodes u and v.
- The component is **weakly connected** if replacing directed edges with undirected edges results in a connected component

Component Examples:



3 components



3 Strongly-connected components

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 concept of the neighborhood of a node can be generalized using shortest paths. An **n-hop neighborhood** of a node is the set of nodes that are within n hops distance from the node.

Diameter

The diameter of a graph is the length of the longest shortest path between any pair of nodes between any pairs 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 web?
  - Average distance of roughly **19 links** between two randomly selected Web pages
  - If the size of the Web mushrooms grows to 10 times its current size, the degrees of separation would only rise slightly, from 19 to 21
  - The "19-click" finding could also provide a ballpark estimate for how deeply a Web crawler needed to dig

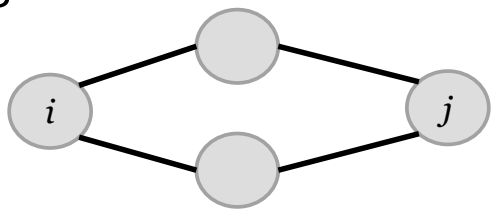
Adjacency Matrix and Connectivity

- Consider the following adjacency matrix

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \dots & A_{1n} \\ A_{21} & A_{22} & A_{23} & \dots & A_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ A_{d1} & A_{d2} & A_{d3} & \dots & A_{dn} \end{bmatrix}$$

- Number of Common neighbors between node i and node j

$$\sum_k A_{ik} A_{jk} = A_i \cdot A_j$$



- That's element of [*ij*] of matrix $A \times A^T = A^2$
- Common neighbors are paths of length 2
- Similarly, what is A^3?

Special Graphs