NETWORK BASICS

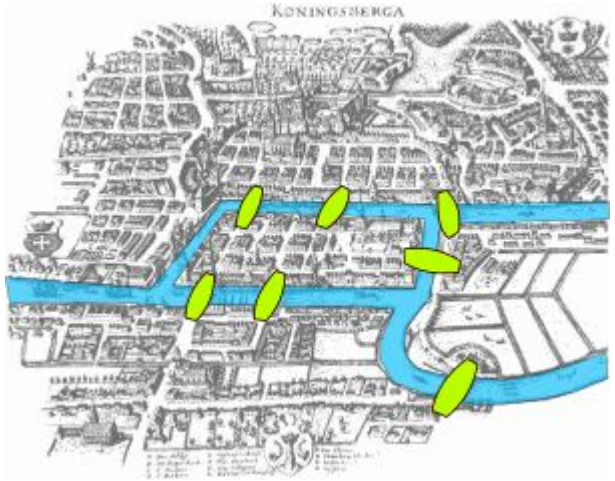
Kristina Lerman

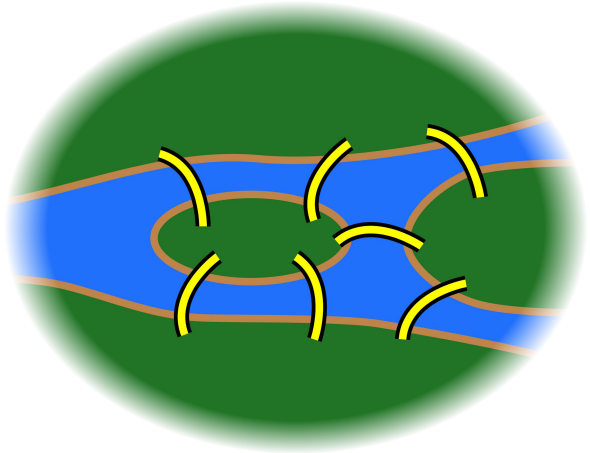
DSCI 531

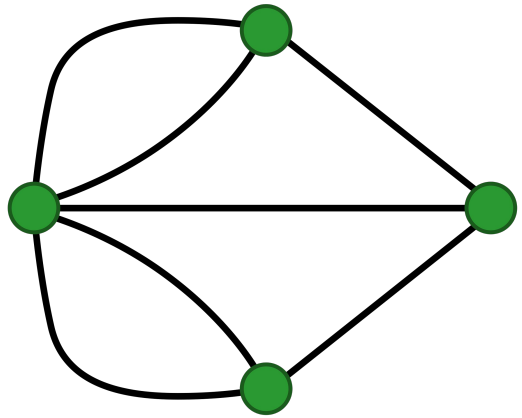
Spring 2025

Networks and graphs

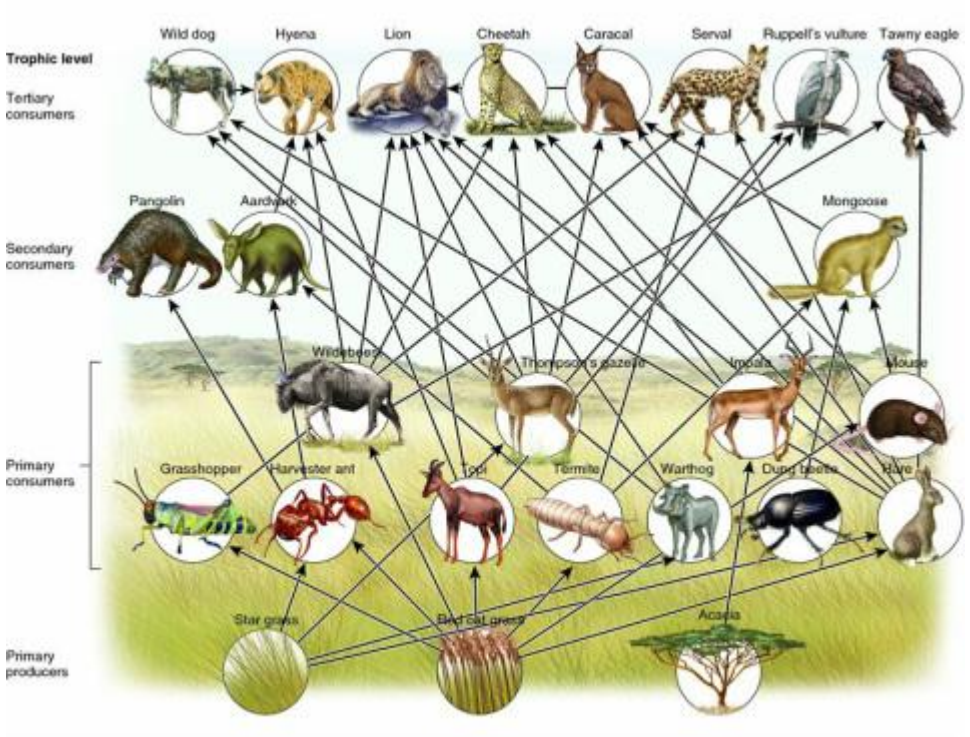
- The seven bridges of Konigsberg: can one cross the city, traversing each bridge only once?
- Euler (1736) proved it impossible
- His insight gave birth to graph theory
  - Physical layout does not matter; only the pattern of connections matters

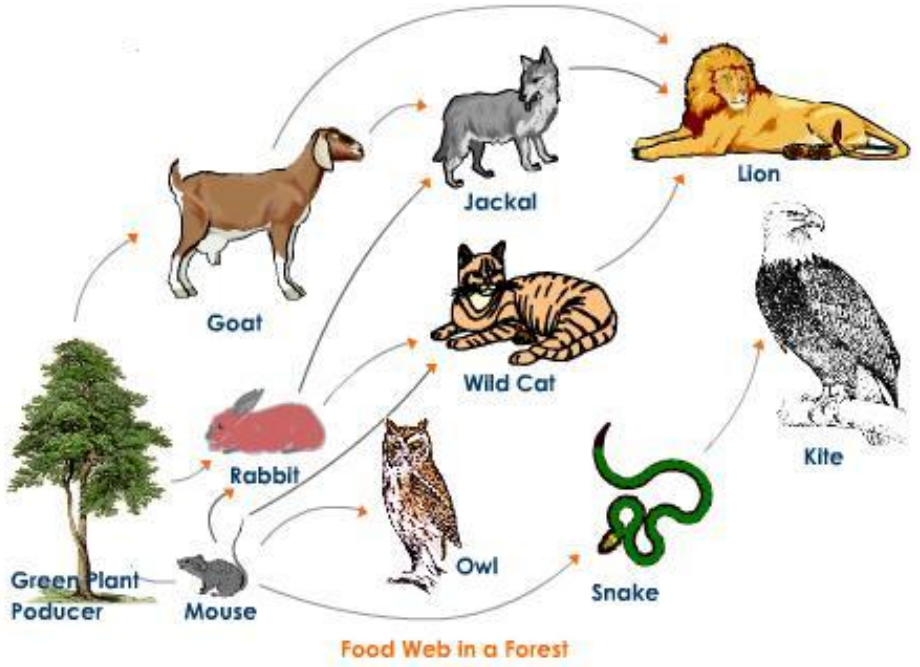






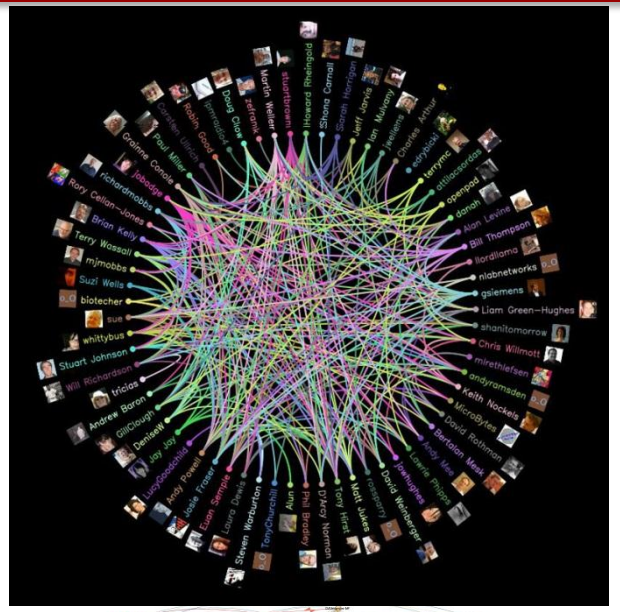
Food Web

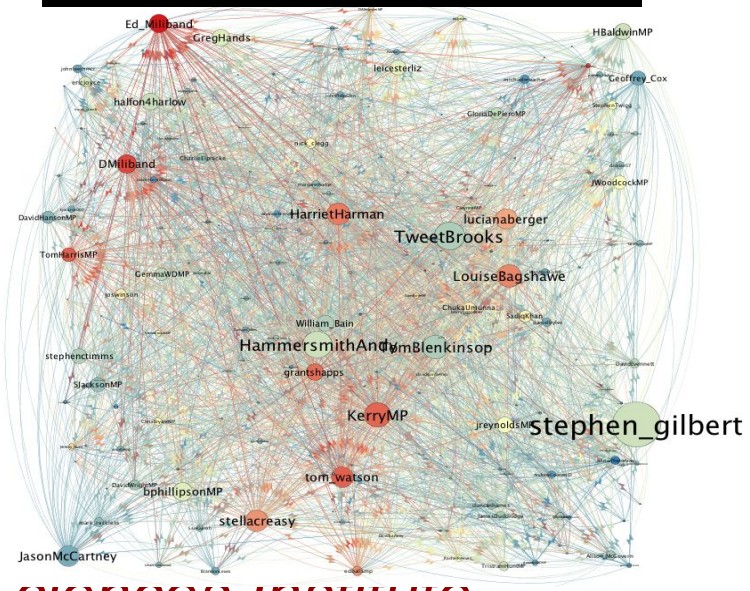

Trophic level
Tertiary consumers
Secondary consumers
Primary consumers
Primary producers
Wild dog
Hyena
Lion
Cheetah
Caracal
Serval
Ruppell's vulture
Tawny eagle
Pangolin
Aardvark
Mongoose
Wildebeest
Thomson's gazelle
Impala
Mouse
Grasshopper
Harvester ant
Topi
Termite
Warthog
Dung beetle
Hare
Star grass
Elliot cat grass
Acacia


Goat
Jackal
Lion
Wild Cat
Kite
Rabbit
Owl
Snake
Green Plant
Poducer
Mouse
Food Web in a Forest

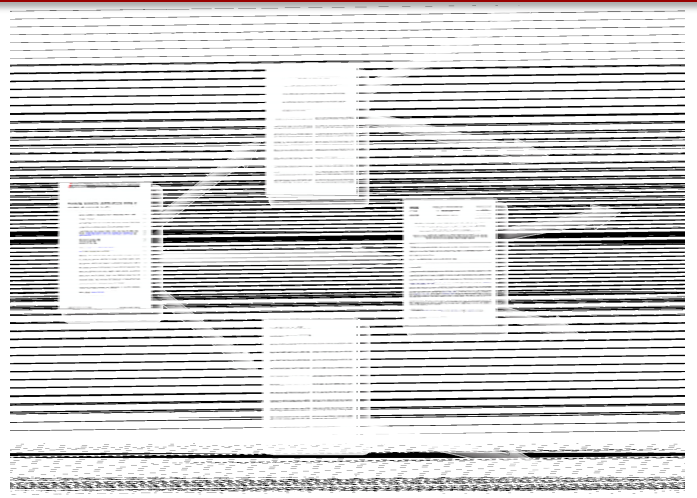
Networks are Pervasive

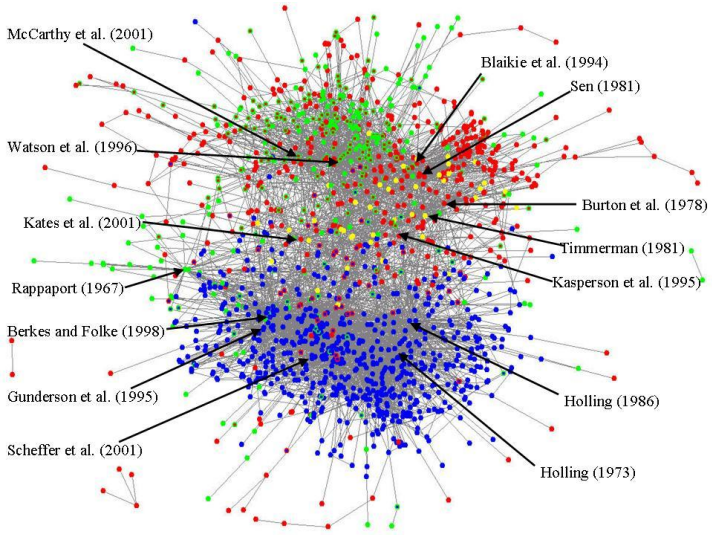
Twitter Networks



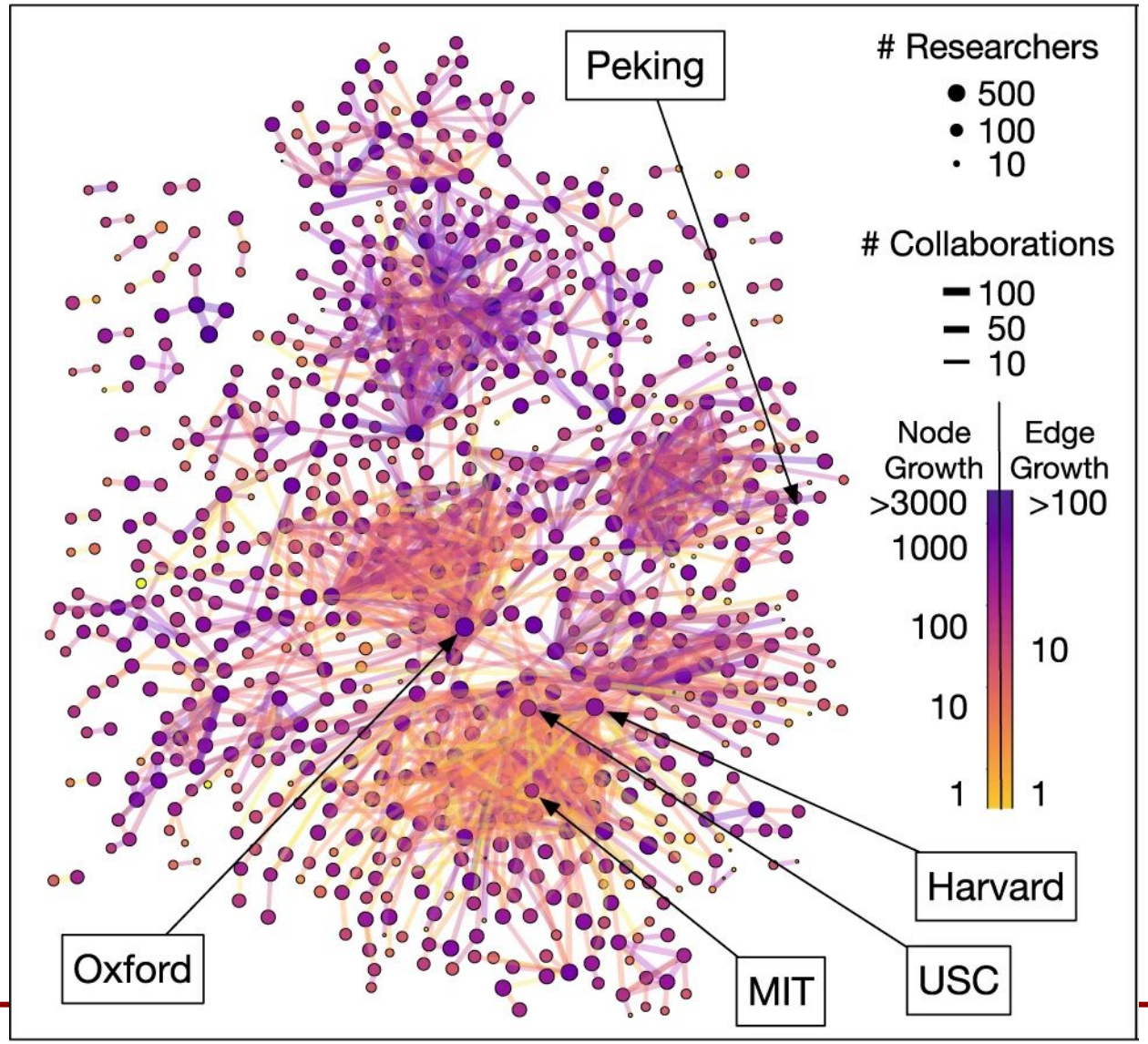


Citation Networks

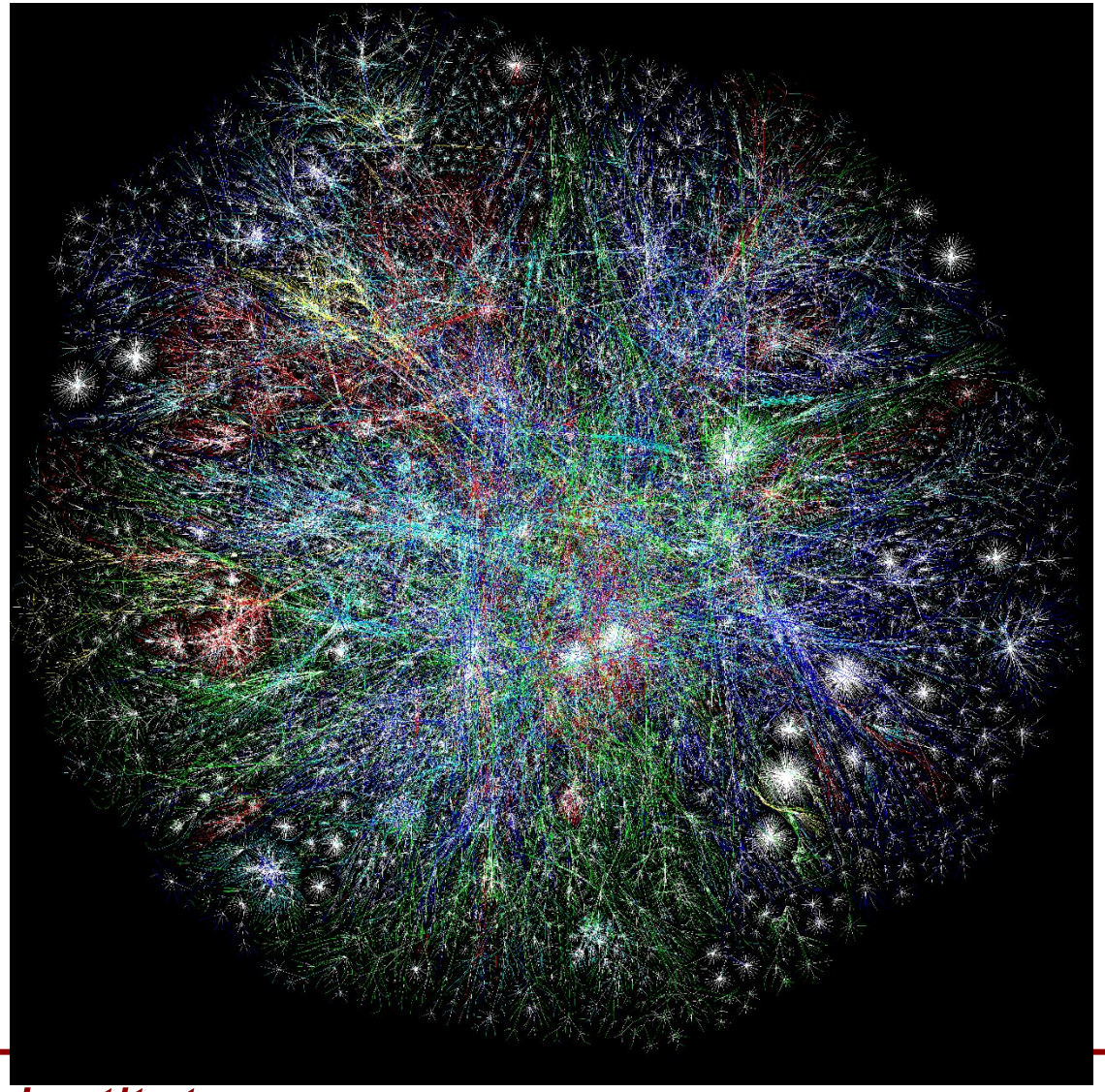




Collaborations between institutions



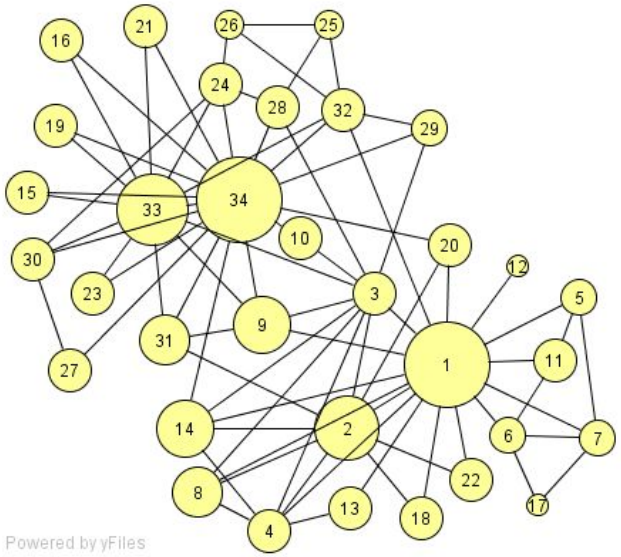
Internet



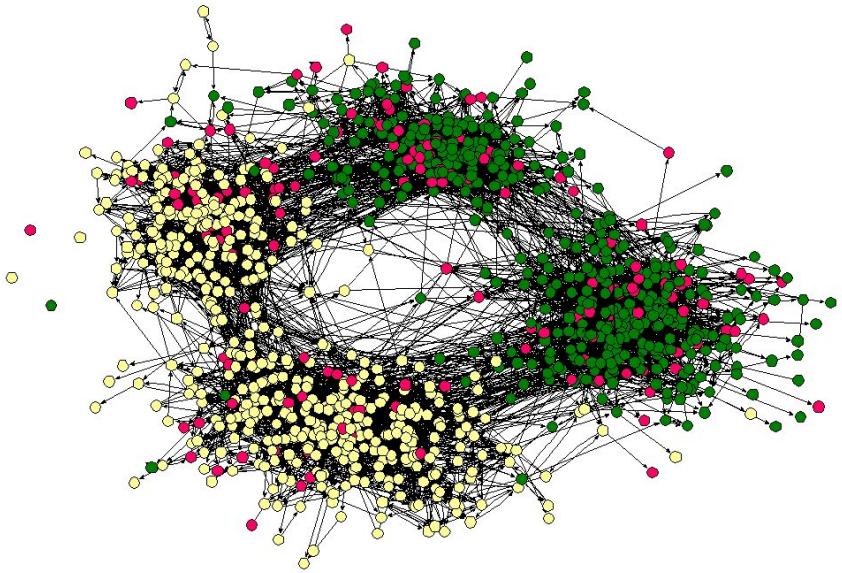
Social Networks and Social Network Analysis

- 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

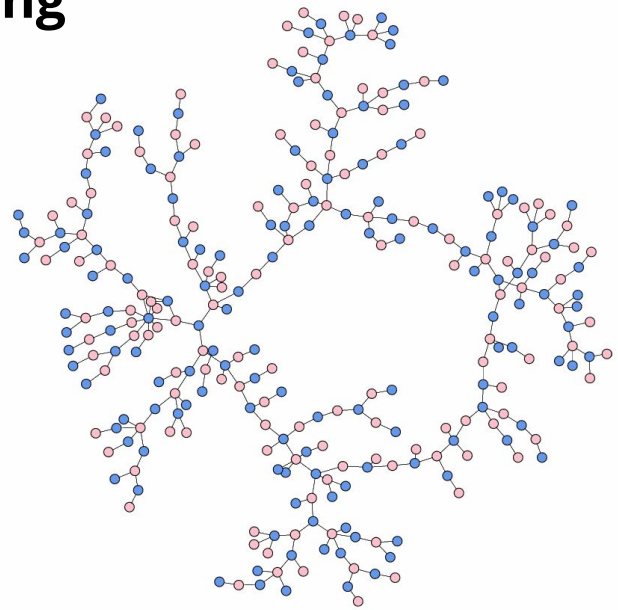


Friendship network of members of a college karate club
(Zachary 1972)



High school friendship

High school dating



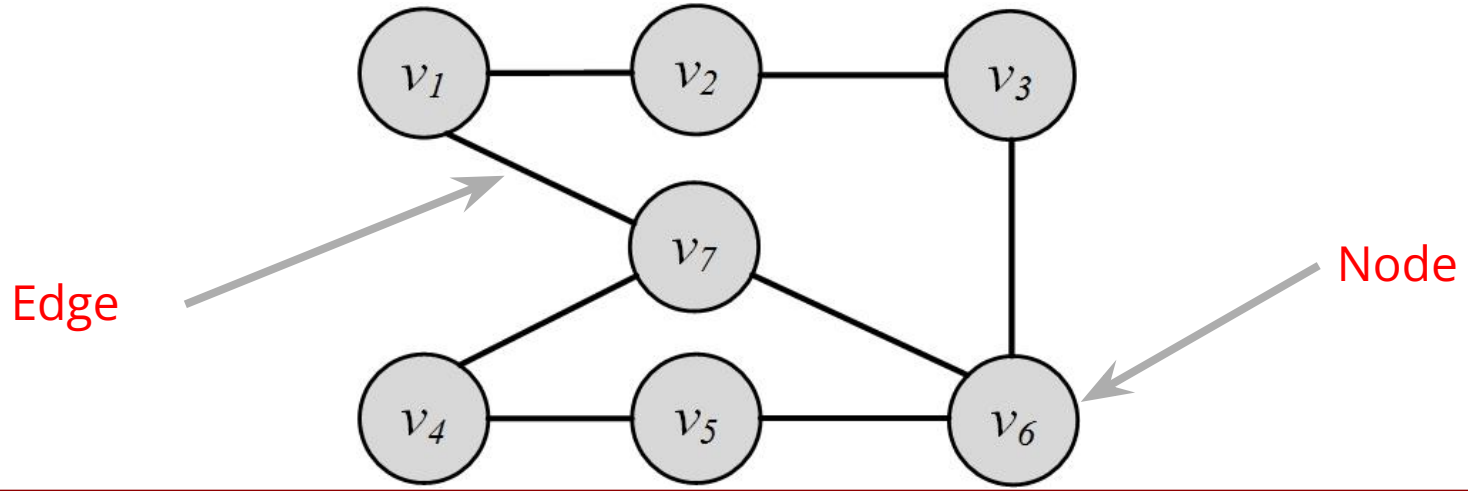
Online social networks

Graph Basics

Nodes and Edges

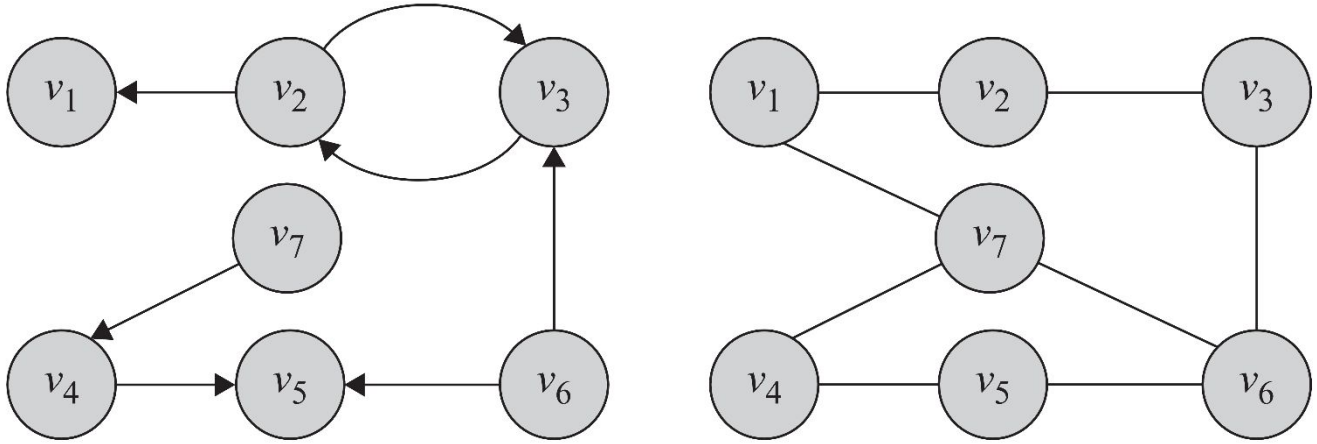
A network is a graph, or a collection of points connected by lines

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



Directed Edges and Directed Graphs

- Edges can have directions.



(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-degree is the number of edges pointing towards a node

d_i^{out} – Out-degree is the number of edges pointing away from a node

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

Degree Distribution

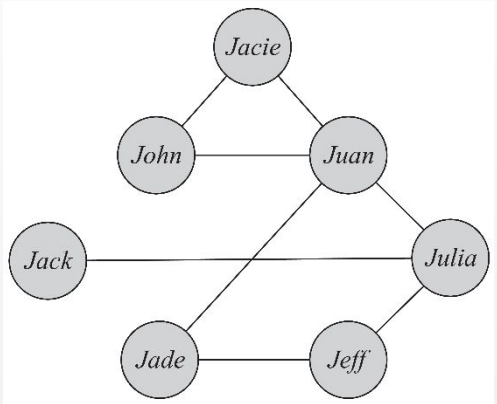
When dealing with very large graphs, how nodes' degrees are distributed is an important concept to analyze and is called ***Degree Distribution***

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

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

n_d is the number of nodes with degree d

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



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

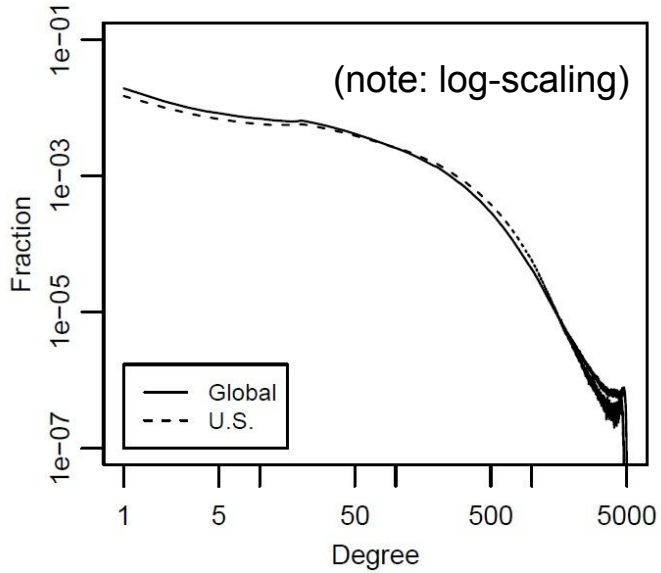
Degree Distribution Plot

- 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.

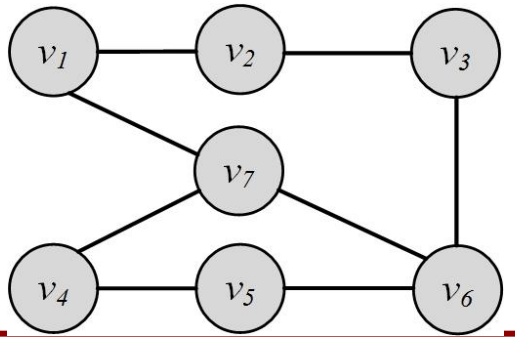
(**Power-law degree distribution**)



Facebook Degree Distribution

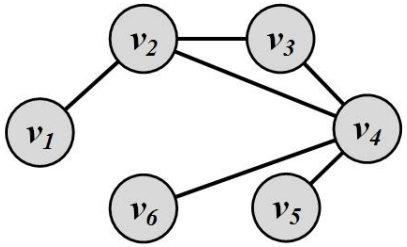
Graph Representation

- Adjacency Matrix
- Adjacency List
- Edge List



Graph Representation

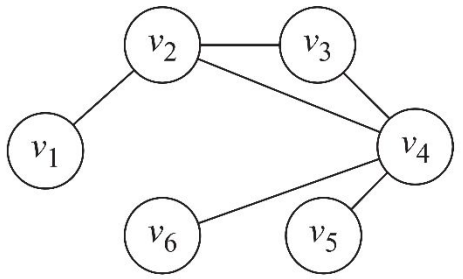
- Graph representation is straightforward and intuitive, but it cannot be effectively manipulated using mathematical and computational tools



- 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

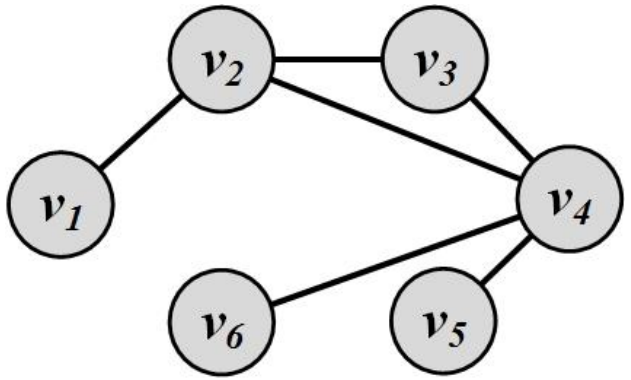
(b) Adjacency Matrix

Diagonal Entries are self-links or loops

Social media networks have very sparse Adjacency matrices

Adjacency List

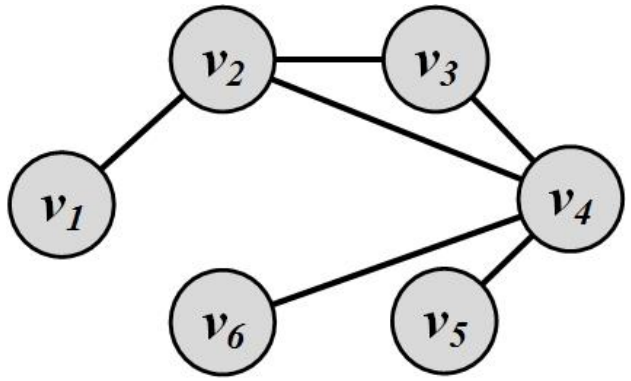
- In an adjacency list for every node, we maintain a list of all the nodes that it is connected to
- The list is usually sorted based on the node order or other preferences



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

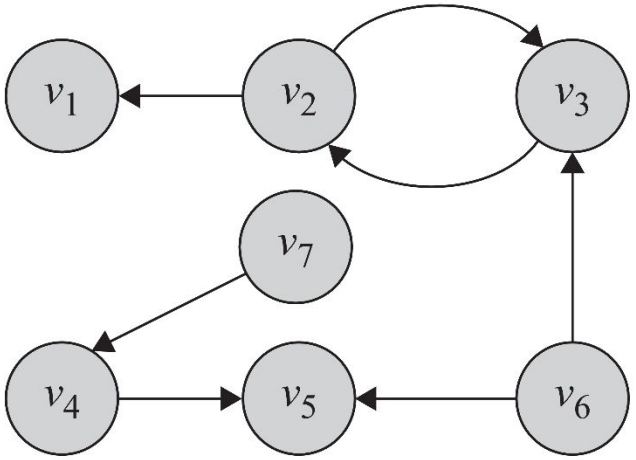


(v_1, v_2)
(v_2, v_3)
(v_2, v_4)
(v_3, v_4)
(v_4, v_5)
(v_4, v_6)

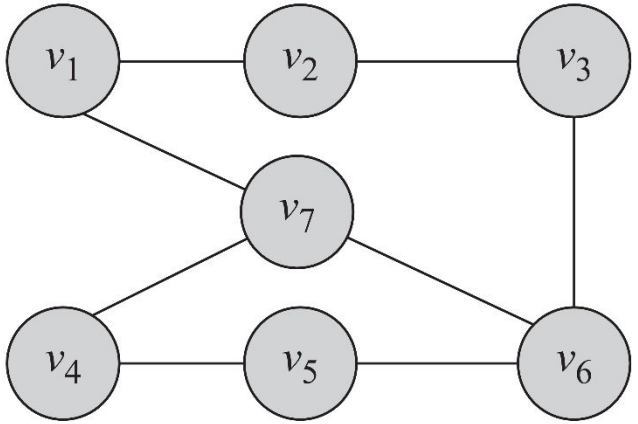
Types of Graphs

Directed/Undirected/Mixed, Simple/Multigraph, Weighted, Signed Graph, Ego-network

Directed/Undirected/Mixed Graphs



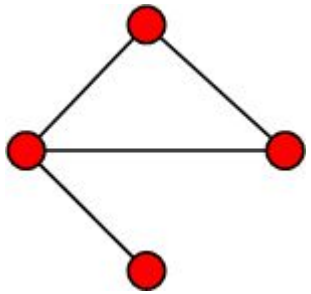
- The adjacency matrix for directed graphs is often not symmetric ($\boldsymbol{A} \neq \boldsymbol{A}^{\boldsymbol{T}}$)
  - $\boldsymbol{A}_{\boldsymbol{ij}} \neq \boldsymbol{A}_{\boldsymbol{ji}}$
  - We can have equality though



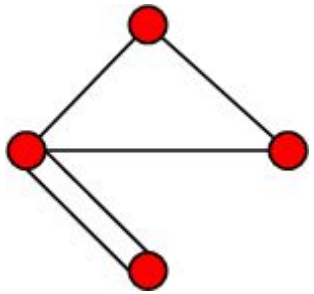
- The adjacency matrix for undirected graphs is symmetric ($\boldsymbol{A} = \boldsymbol{A}^{\boldsymbol{T}}$)

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

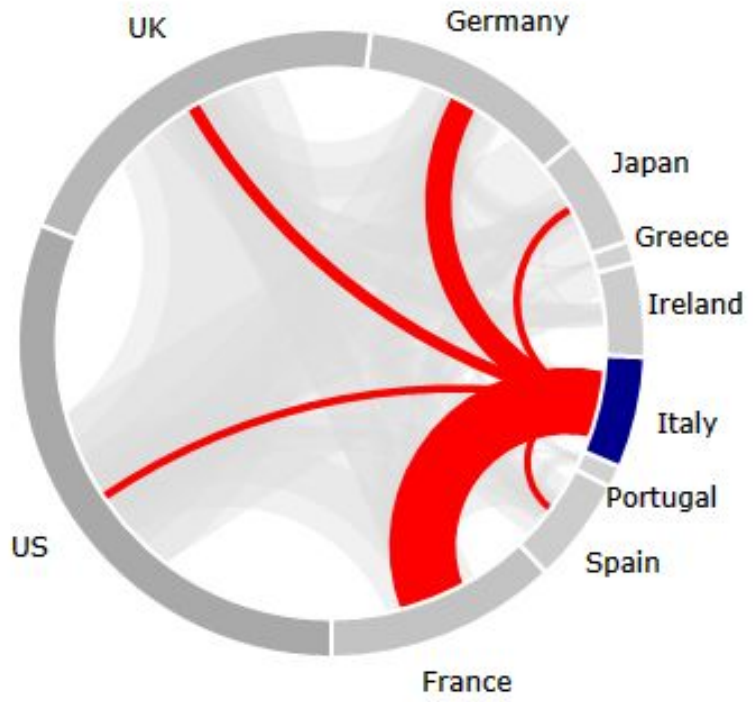


Multigraph

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

Weighted Graph

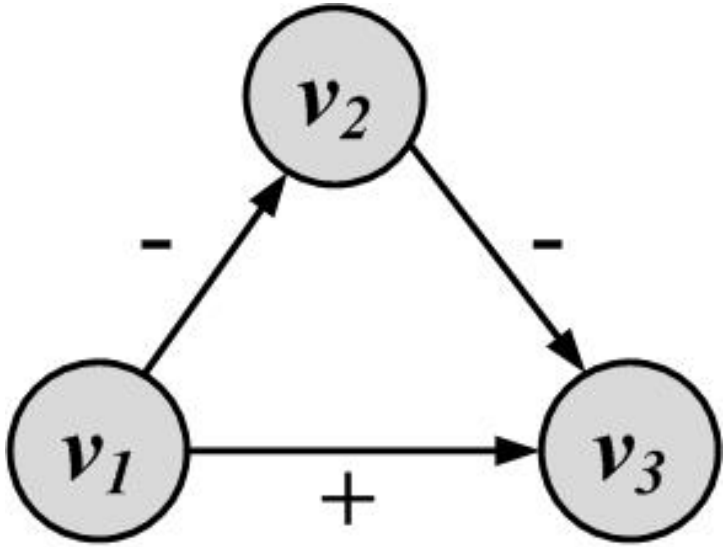
- A weighted graph $\boldsymbol{G(V, E, W)}$ is one where edges are associated with weights



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

Signed Graph

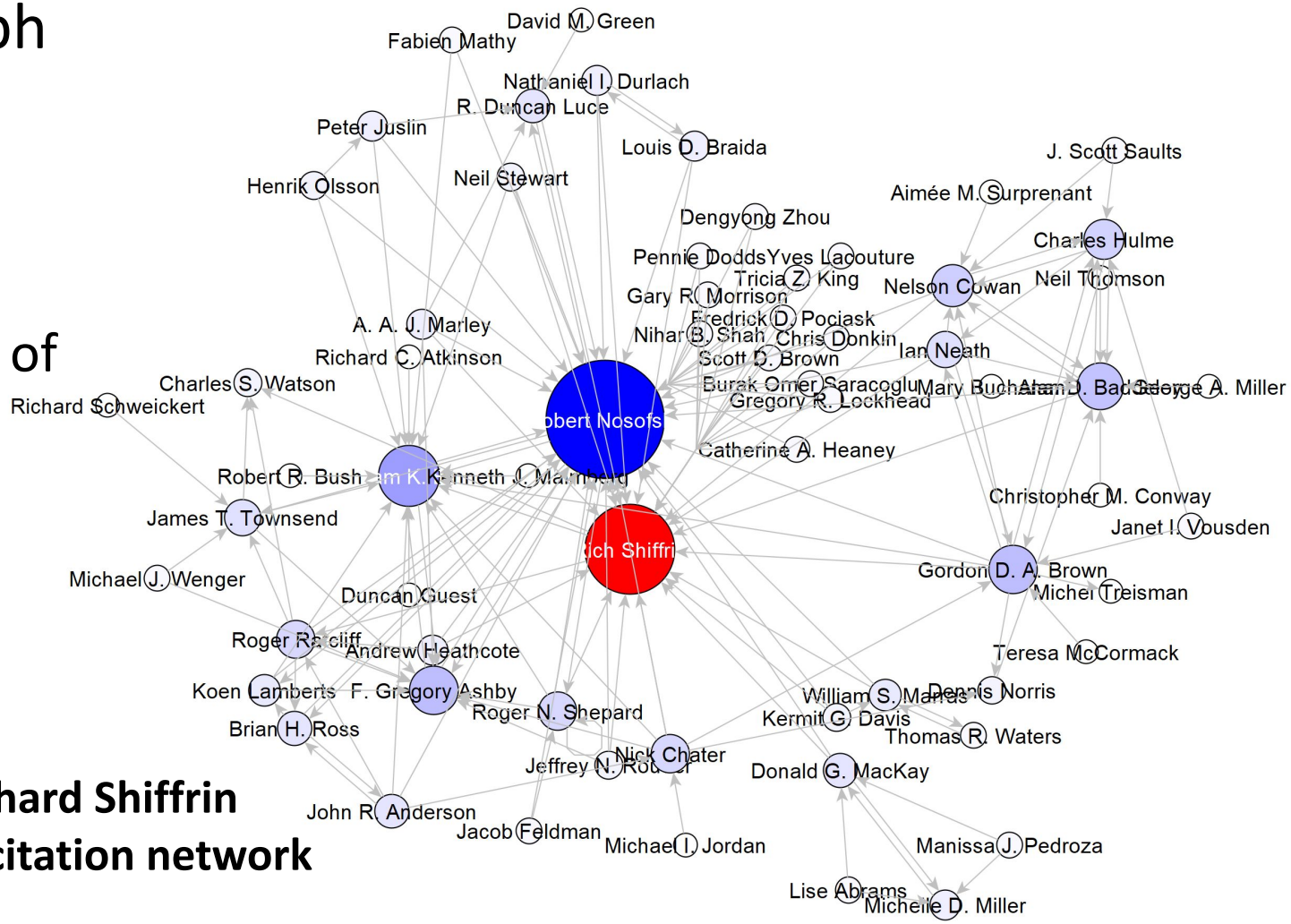
- Weights are binary (0/1, -1/1, +/-)



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

Ego-networks

- An **ego-network** is a subgraph that contains the **focal node** (ego), its **neighbors** and the **connections** between them
  - Represents a node-level view of a network



Ego network of Richard Shiffrin within the author-citation network

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 an 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.

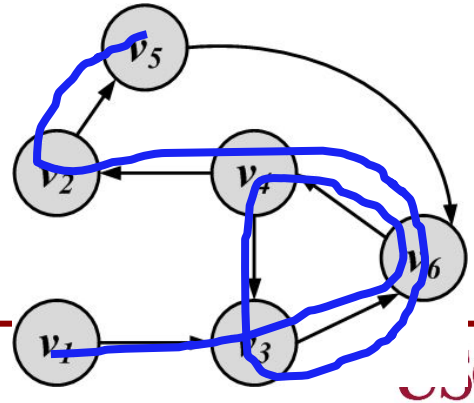
Walk, Path, Trail, Tour, and Cycle

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

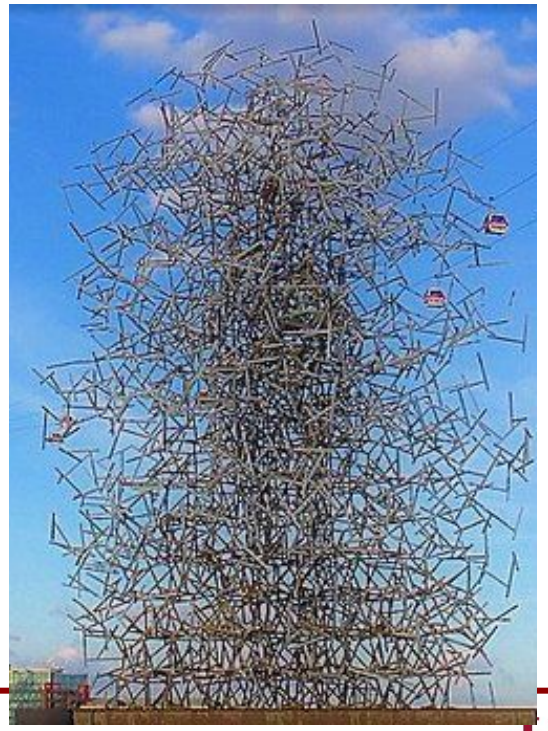
Trail

- 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**
- 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

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

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



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

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

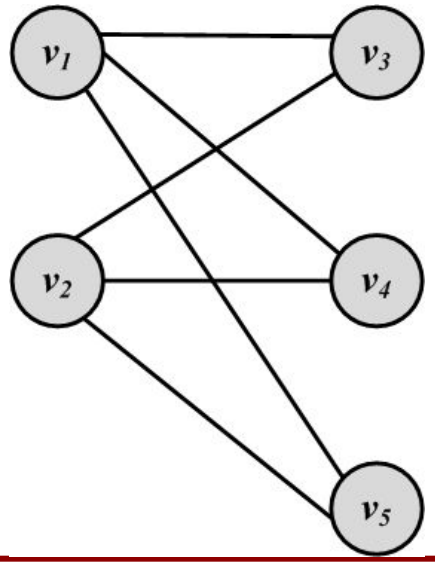
- How big is the diameter of Facebook?

Special Subgraphs

Bipartite Graphs

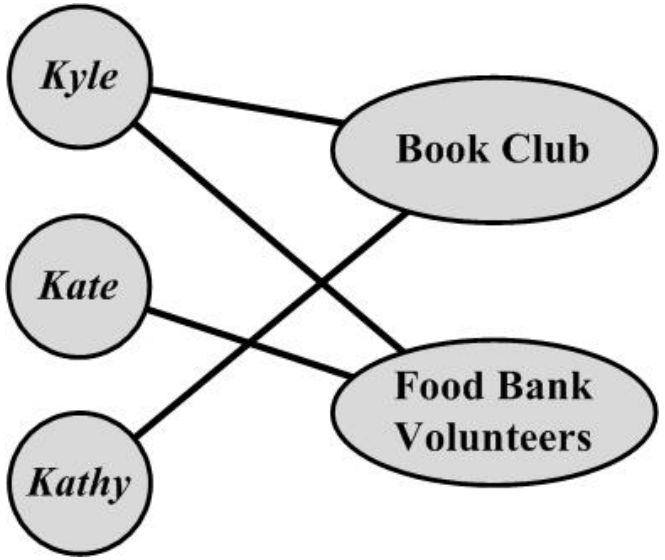
A bipartite graph $G(V, E)$ is a graph where the node set can be partitioned into two sets such that, for all edges, one end-point is in one set and the other end-point is in the other set.

$$\begin{cases} V = V_L \cup V_R, \\ V_L \cap V_R = \emptyset, \\ E \subset V_L \times V_R \end{cases}$$



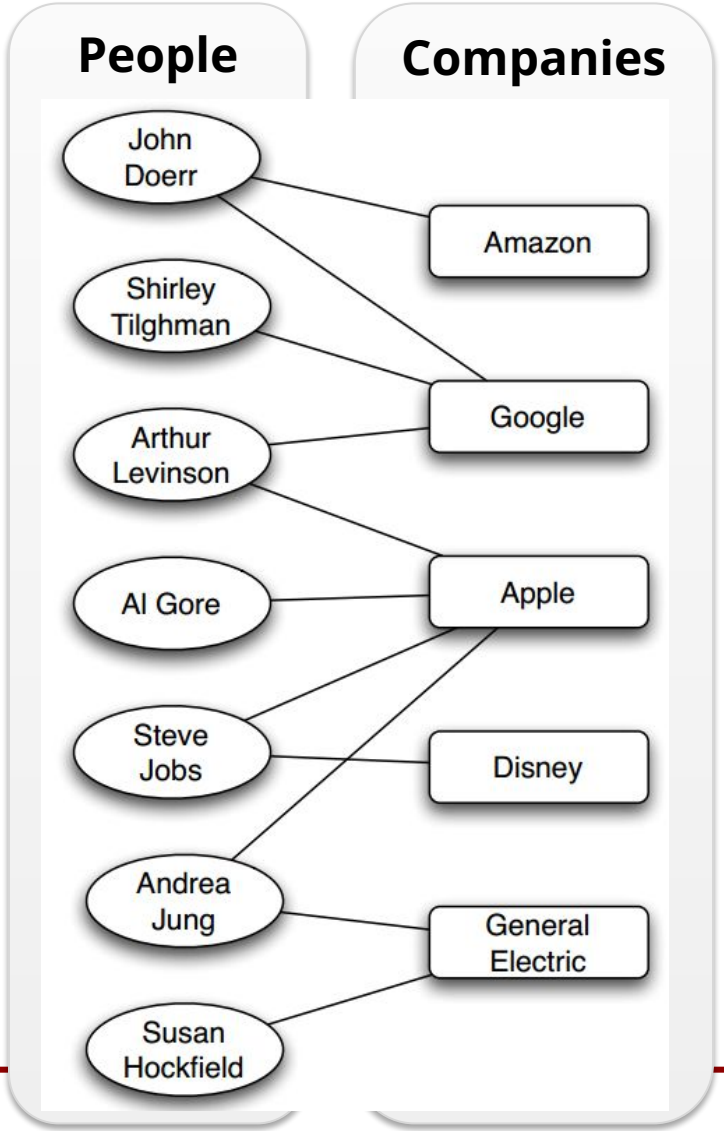
Affiliation Networks

An affiliation network is a bipartite graph. If an individual is associated with an affiliation, an edge connects the corresponding nodes.



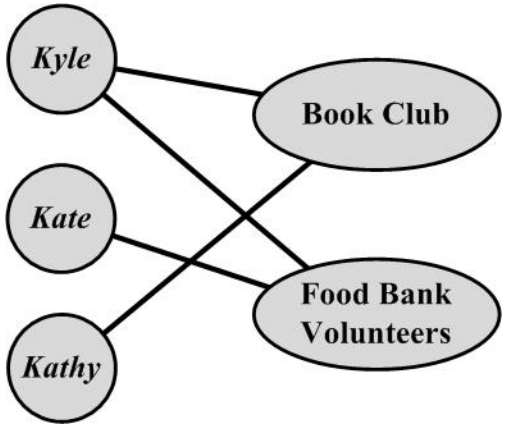
Affiliation Networks: Membership

Affiliation of people on corporate boards of directors



Bipartite Representation / one-mode Projections

- We can save some space by keeping membership matrix *X*



$$X = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

- What is XX^T? *Similarity between users - [Bibliographic Coupling]*
- What is X^TX? *Similarity between groups - [Co-citation]*

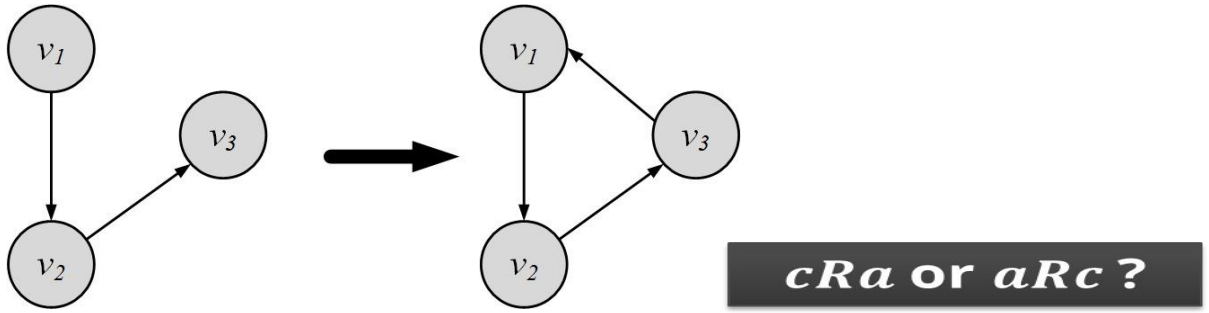
Elements on the diagonal are number of groups
the user is a member of
OR
number of users in the group

Friendship Patterns

- Transitivity/Reciprocity
- Status/Balance

Transitivity

- Mathematic representation:
  - For a transitive relation R: $aRb \wedge bRc \rightarrow aRc$



- In a social network:
  - ***Transitivity is when a friend of my friend is my friend***
  - Transitivity in a social network leads to a denser graph, which in turn is closer to a complete graph
  - We can determine how close graphs are to the complete graph by measuring transitivity

[Global] Clustering Coefficient

- **Clustering coefficient** measures transitivity in undirected graphs
  - Count paths of length two and check whether the third edge exists

$$C = \frac{|\text{Closed Paths of Length 2}|}{|\text{Paths of Length 2}|}$$

When counting triangles, since every triangle has 6 closed paths of length 2

$$C = \frac{(\text{Number of Triangles}) \times 6}{|\text{Paths of Length 2}|}$$

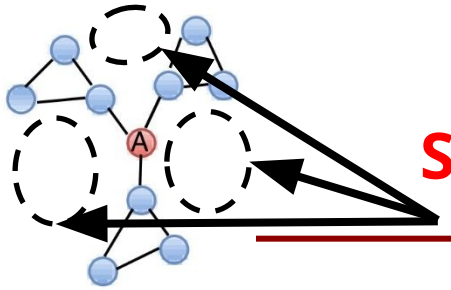
[Local] Clustering Coefficient

- Local clustering coefficient measures transitivity at the node level
  - Commonly employed for undirected graphs
  - Computes how strongly neighbors of a node v (nodes adjacent to v) are themselves connected

$$C(v_i) = \frac{\text{Number of Pairs of Neighbors of } v_i \text{ That Are Connected}}{\text{Number of Pairs of Neighbors of } v_i}.$$

In an undirected graph, the denominator can be rewritten as: $\binom{d_i}{2} = d_i(d_i - 1)/2$

Provides a way to determine **structural holes**

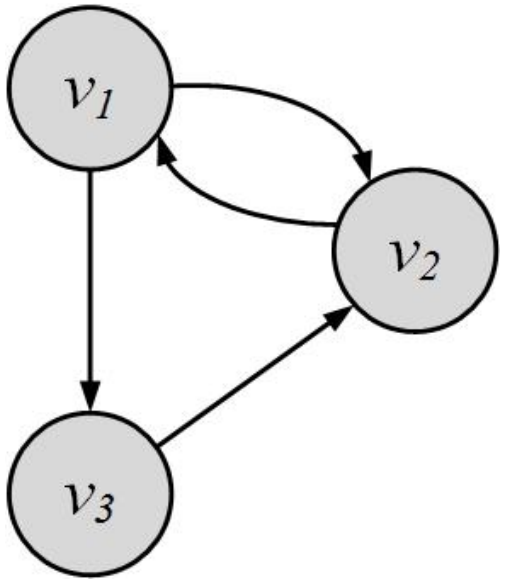


Structural Holes

Reciprocity

If you become my friend, I'll be yours

- Reciprocity is simplified version of transitivity
  - It considers closed loops of length 2
- If node v is connected to node u,
  - u by connecting to v, exhibits reciprocity



Social Balance Theory

Social balance theory

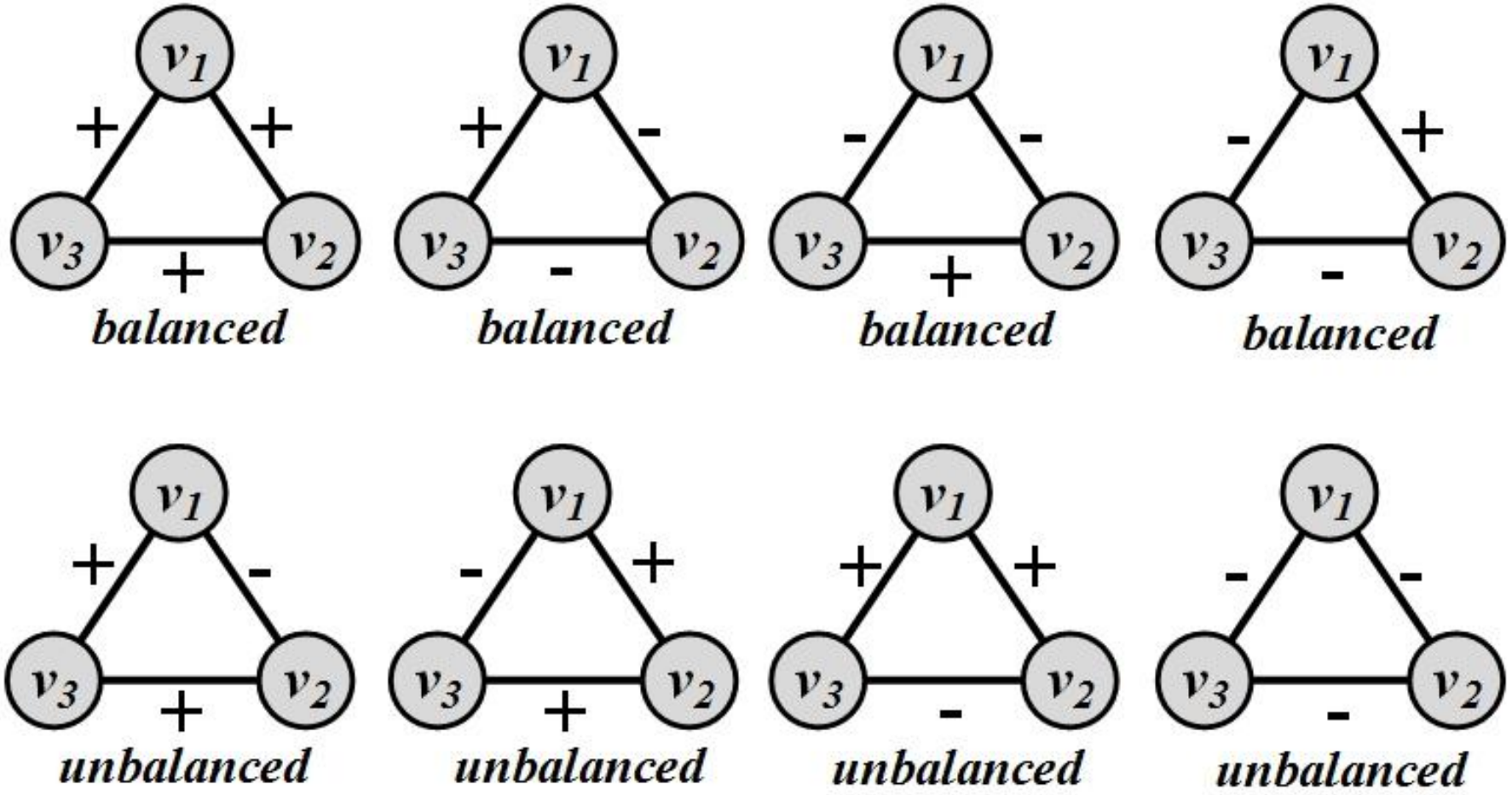
- Consistency in friend/foe relationships among individuals
- Informally, friend/foe relationships are consistent when

$$\begin{aligned}&\textit{The friend of my friend is my friend,}\\&\textit{The friend of my enemy is my enemy,}\\&\textit{The enemy of my enemy is my friend,}\\&\textit{The enemy of my friend is my enemy.}\end{aligned}$$

- In the network
  - Positive edges demonstrate friendships ($w_{ij} = 1$)
  - Negative edges demonstrate being enemies ($w_{ij} = -1$)
- Triangle of nodes i, j, and k, is balanced, if and only if
  - w_{ij} denotes the value of the edge between nodes i and j

$$w_{ij}w_{jk}w_{ki} \geq 0.$$

Social Balance Theory: Possible Combinations



For any cycle, if the multiplication of edge values become positive, then the cycle is socially balanced

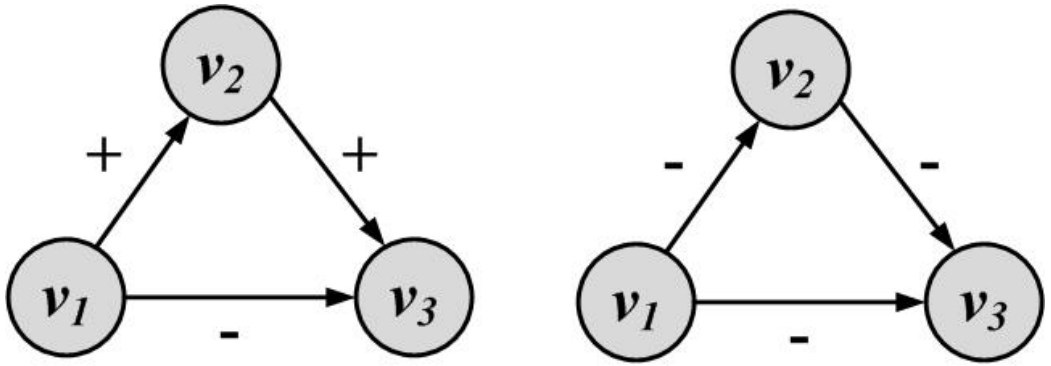
Social Status Theory

- **Status:** how prestigious an individual is ranked within a society

- **Social status theory:**
  - How consistent individuals are in assigning status to their neighbors
  - Informally,

If X has a higher status than Y and Y has a higher status than Z, then X should have a higher status than Z.

Social Status Theory: Example



Unstable configuration Stable configuration

- A directed **'+'** edge from node X to node Y shows that Y has a higher status than X and a **'-'** one shows vice versa

Measuring Centrality

Why Do We Need Measures?

- Who are the important actors (influential individuals) in the network?
  - **Centrality**
- What interaction patterns are common in friends?
  - **Reciprocity and Transitivity**
  - **Balance and Status**
- Who are the like-minded users and how can we find these similar individuals?
  - **Similarity**
- To answer these and similar questions, one first needs to define measures for quantifying **centrality**, **level of interactions**, and **similarity**, among others.

Centrality in terms of those who you are connected to

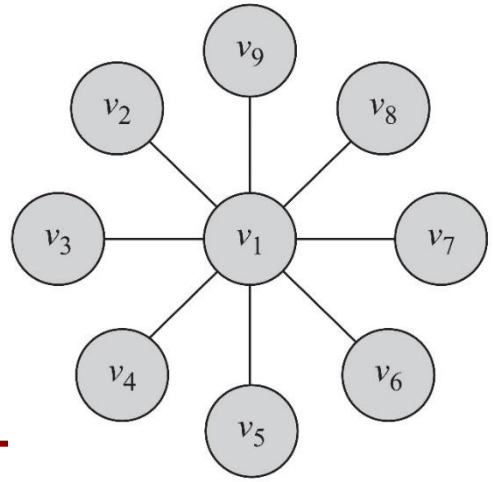
Degree Centrality

- **Degree centrality**: ranks nodes with more connections higher in terms of centrality

$$C_d(v_i) = d_i$$

- d_i is the degree (number of friends) for node v_i
  - i.e., the number of length-1 paths (can be generalized)

In this graph, degree centrality for node v_1 is d_1=8 and for all others is $d_j = 1, j \neq 1$



Degree Centrality in Directed Graphs

- In directed graphs, we can either use the in-degree, the out-degree, or the combination as the degree centrality value:
- In practice, mostly in-degree is used.

$$C_d(v_i) = d_i^{\text{in}} \quad (prestige)$$

$$C_d(v_i) = d_i^{\text{out}} \quad (gregariousness)$$

$$C_d(v_i) = d_i^{\text{in}} + d_i^{\text{out}}$$

d_i^{out} is the number of outgoing links for node v_i

Normalized Degree Centrality

- Normalized by the maximum possible degree

$$C_d^{\text{norm}}(v_i) = \frac{d_i}{n-1}$$

- Normalized by the maximum degree

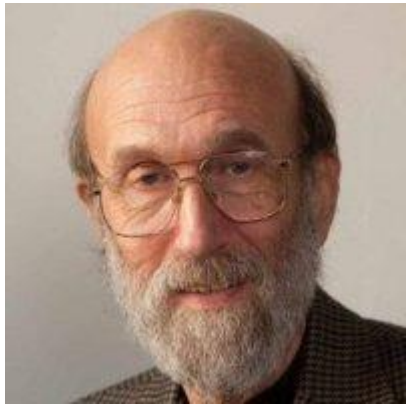
$$C_d^{\max}(v_i) = \frac{d_i}{\max_j d_j}$$

- Normalized by the degree sum

$$C_d^{\text{sum}}(v_i) = \frac{d_i}{\sum_j d_j} = \frac{d_i}{2|E|} = \frac{d_i}{2m}$$

Eigenvector Centrality

- Having more friends does not by itself guarantee that someone is more important
  - Having more **important friends** provides a stronger signal



Phillip Bonacich

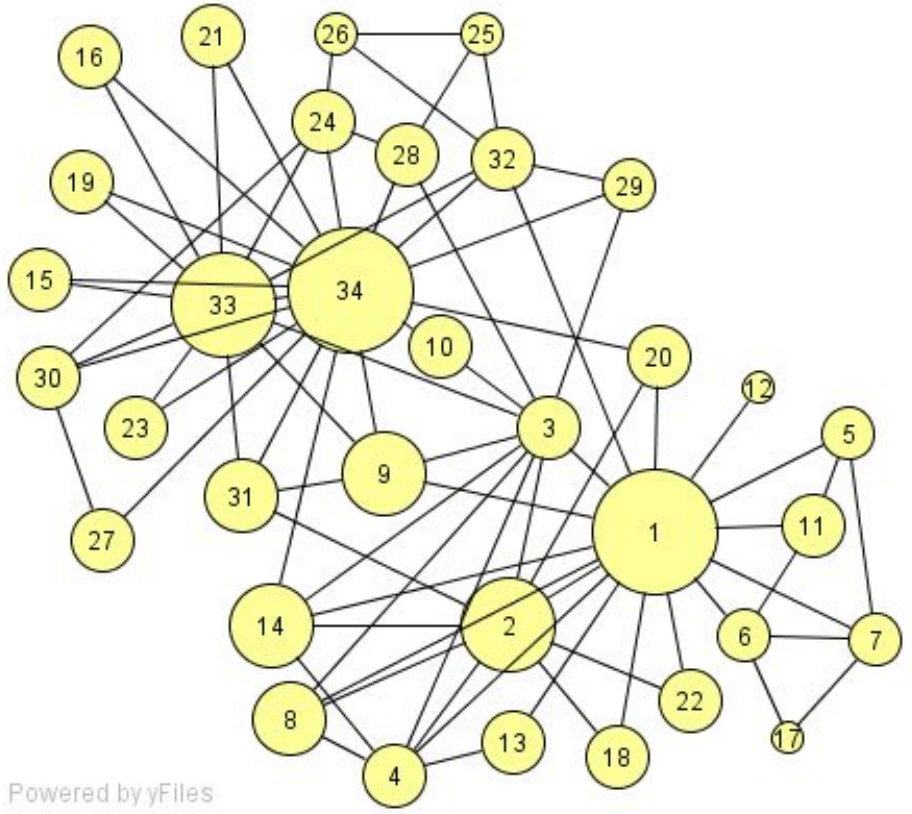
- Eigenvector centrality generalizes degree centrality by incorporating the importance of the neighbors (undirected)

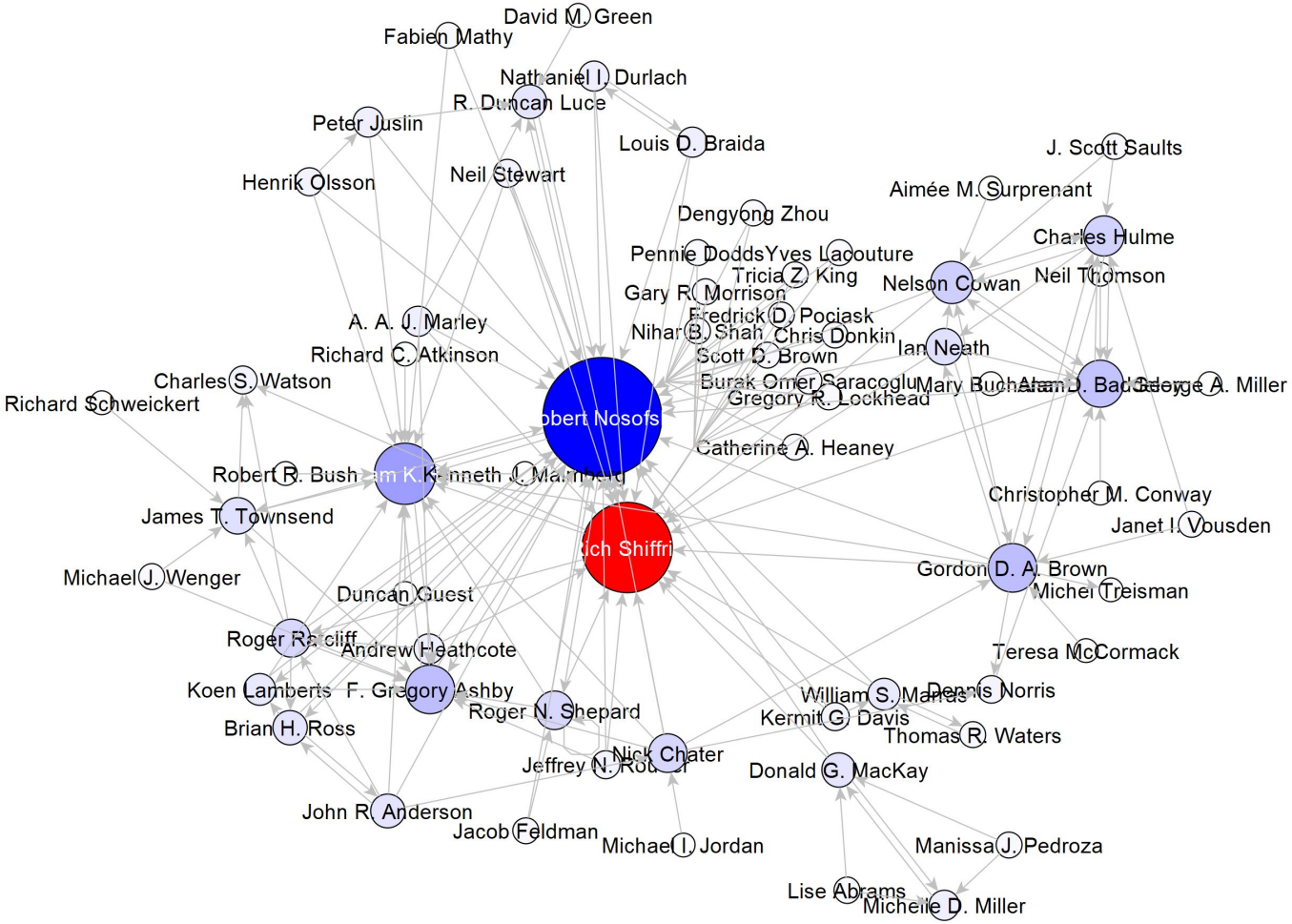
- For directed graphs, we can use incoming or outgoing edges

PageRank

- Problem with Eigenvector Centrality:
  - In directed graphs, once a node becomes an authority (high centrality), it passes **all** its centrality along **all** of its out-links
- This is less desirable since not everyone known by a well-known person is well-known
- **Solution?**
  - We can divide the value of passed centrality by the number of outgoing links, i.e., out-degree of that node
  - Each connected neighbor gets a fraction of the source node's centrality

Examples: Size related to Pagerank Centrality

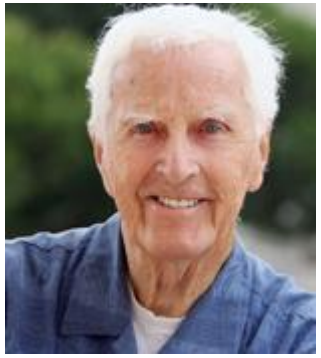




Centrality in terms of how you connect others
(information broker)

Betweenness Centrality

Another way of looking at centrality is by considering how important nodes are in connecting other nodes



Linton Freeman

$$C_b(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

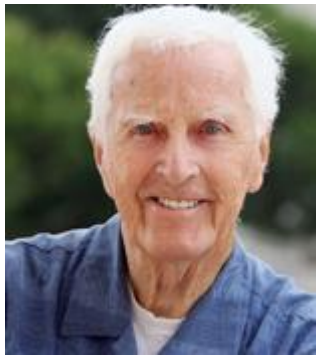
σ_{st} The number of shortest paths from vertex s to t – *a.k.a.* information pathways

$\sigma_{st}(v_i)$ The number of **shortest paths** from s to t that pass through v_i

Centrality in terms of how fast you can reach others

Closeness Centrality

- The intuition is that influential/central nodes can quickly reach other nodes
- These nodes should have a smaller average shortest path length to others



Linton Freeman

Closeness centrality: $C_c(v_i) = \frac{1}{\bar{l}_{v_i}}$

$$\bar{l}_{v_i} = \frac{1}{n-1} \sum_{v_j \neq v_i} l_{i,j}$$

An Interesting Comparison!

Comparing three centrality values

- Generally, the 3 centrality types will be positively correlated
- When they are not (or low correlation), it usually reveals interesting information

	Low Degree	Low Closeness	Low Betweenness
High Degree		*Node is embedded in a community that is far from the rest of the network*	*Ego's connections are redundant - communication bypasses the node*
High Closeness	*Key node connected to important/active alters*		*Probably multiple paths in the network, ego is near many people, but so are many others*
High Betweenness	*Ego's few ties are crucial for network flow*	*Very rare! Ego monopolizes the ties from a small number of people to many others.*	

Assortativity

Social Forces

- **Social Forces** connect individuals in different ways
- When individuals get connected, we observe distinguishable patterns in their connectivity networks.
  - **Assortativity**, also known as *social similarity or homophily*
- In networks with assortativity:
  - Similar nodes are connected to one another more often than dissimilar nodes.
- Social networks are assortative
  - A high similarity between friends is observed
  - We observe similar behavior, interests, activities, or shared attributes such as language among friends

Why are connected people similar?

Influence

- The process by which a user (i.e., influential) affects another user
- The influenced user becomes more similar to the influential figure.
  - **Example:** If most of our friends/family members switch to a cellphone company, we might switch [i.e., become influenced] too.

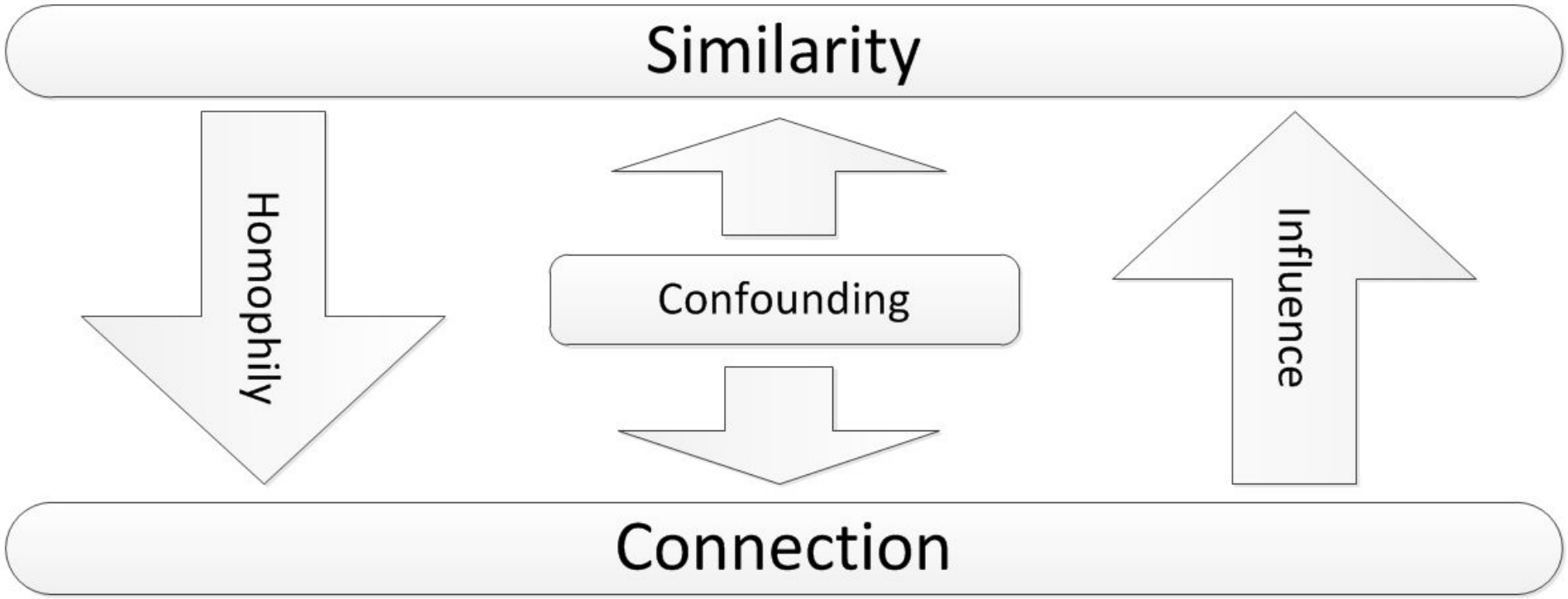
Homophily

- Similar individuals becoming friends due to their high similarity
  - **Example:** Two musicians are more likely to become friends.

Confounding

- The environment's effect on making individuals similar
  - **Example:** Two individuals living in the same city are more likely to become friends than two random individuals

Influence, Homophily, and Confounding

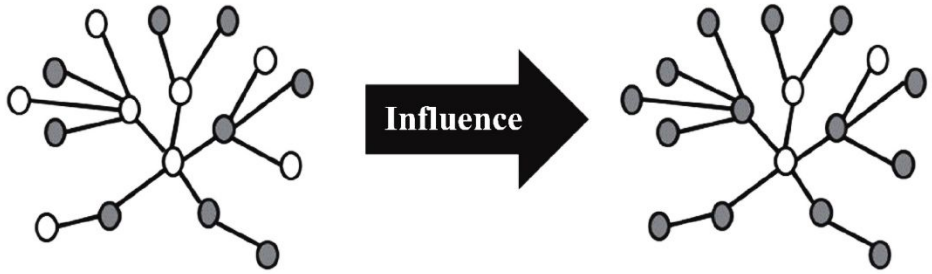


Source of Assortativity in Networks

Both influence and Homophily generate similarity in social networks

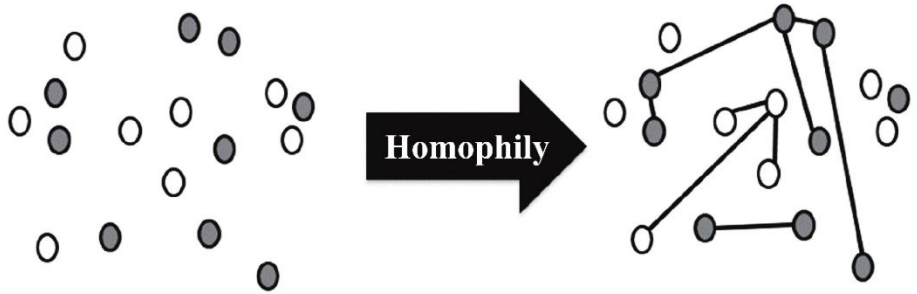
Influence

Makes connected nodes similar to each other



Homophily

Selects similar nodes and links them together



Assortativity Example

The city's draft tobacco control strategy says more than 60% of under-16s in Plymouth smoke regularly

BBC | News | Sport | Weather | Travel | TV | Radio | More... | Search

DEVON

BBC RADIO DEVON
Listen Live Listen Again

BBC Local
Devon
Things to do
People & Places
Nature & Outdoors
History
Religion & Ethics
Arts & Culture
BBC Introducing
TV & Radio

Local BBC Sites
News
Sport
Weather
Travel

Neighbouring Sites
Cornwall
Dorset
Somerset

Related BBC Sites
England

Page last updated at 14:58 GMT, Monday, 14 June 2010 15:58 UK

E-mail this to a friend Printable version

Patches for Plymouth's young smokers

By Jo Irving
BBC Devon website

More than 60% of Plymouth's under-16s smoke

MORE FROM DEVON
NEWS
SPORT
WEATHER
TRAVEL

ELSEWHERE ON THE WEB
Plymouth NHS Trust Stop Smoking Service

Why?

- Smoker friends influence their non-smoker friends

Influence

- Smokers become friends
  - Can this explain smoking behavior?

Homophily

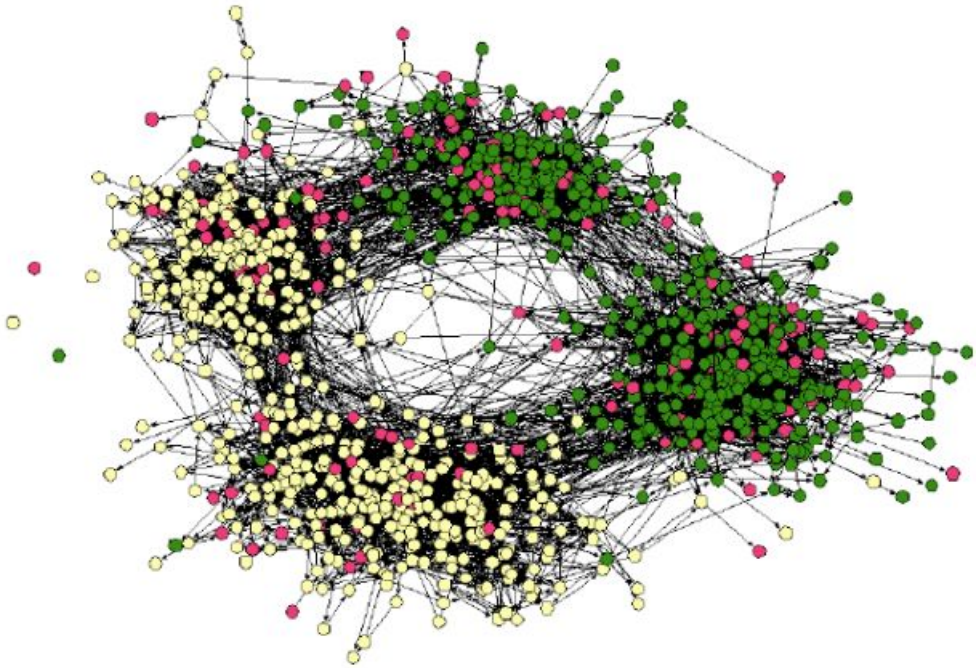
- There are lots of places that people can smoke

Confounding

Measuring Assortativity

Assortativity example

- The friendship network of high school & middle school students,
- Color shows race
- Divisions show high school vs middle school.



Measuring Assortativity for Nominal Attributes

- Assume **nominal** attributes are assigned to nodes
  - Example: race

- Edges between nodes of the same type can be used to measure assortativity of the network
  - Same type = nodes that share an attribute value
  - Node attributes could be nationality, race, sex, etc.

$$\frac{1}{m}\sum_{(v_i,v_j)\in E}\delta(t(v_i),t(v_j)) = \frac{1}{2m}\sum_{ij}A_{ij}\delta(t(v_i),t(v_j))$$

$t(v_i)$ denotes type of vertex v_i

$$\delta(x,y) = \begin{cases} 0, & \text{if } x \neq y \\ 1, & \text{if } x = y \end{cases}$$

Kronecker delta function

Assortativity Significance

- **Assortativity significance**
  - The difference between **measured assortativity** and **expected assortativity**
  - The higher this difference, the more significant the assortativity observed

Example

- In a school, 50% of the population is **white** and the other 50% is **hispanic**.
- We expect 50% of the connections to be between members of different races.
- If all connections are between members of different races, then we have a significant finding

Assortativity Significance

Assortativity

Expected assortativity
(according to configuration model)

$$Q = \frac{1}{2m}\sum_{ij} A_{ij}\delta(t(v_i), t(v_j)) - \frac{1}{2m}\sum_{ij} \frac{d_i d_j}{2m}\delta(t(v_i), t(v_j))$$

This is modularity