AALBORG UNIVERSITY

Social Media/Network Analysis: Introduction & Basic Structural Analysis

Peter Dolog
dolog@cs.aau.dk
http://people.cs.aau.dk/~dolog

Based on the ‘Social Media Mining: An Introduction’ book (Chapters 1-3) by Reza Zafarani, Mohammad Ali Abbasi and Huan Liu, and also slides from Bo Thiesson

Outline

- Introduction to social media analysis
- Basic graph theory
- Connectivity
- Centrality
- Structural similarity

Introduction

Douglas Wray - http://instagr.am/p/nm695/ @ThreeShipsMedia

Do people really talk about donuts? (on Twitter)

1 week of tweets mentioning "donut" or "doughnuts"

- Week of Feb 6-12, 2012.
- Matched ~180k messages

Can be used to find things such as

- In which locations do people eat donuts
- Preferred restaurants for eating donuts
- Preferred kind of donuts
- What people drink with donuts
- What is the mood, when eating donuts
- Etc.

Beyond donuts…

Drugs, diseases, and contagions

- Paul and Dredze 2011; Sadilek, Kautz and Silenzio 2012, **Denecke et al. 2013**

Crises, disasters, and wars

- Starbird et al. 2010; Al-Ani, Mark & Semaan 2010; Monroy-Hernandez et al. 2012

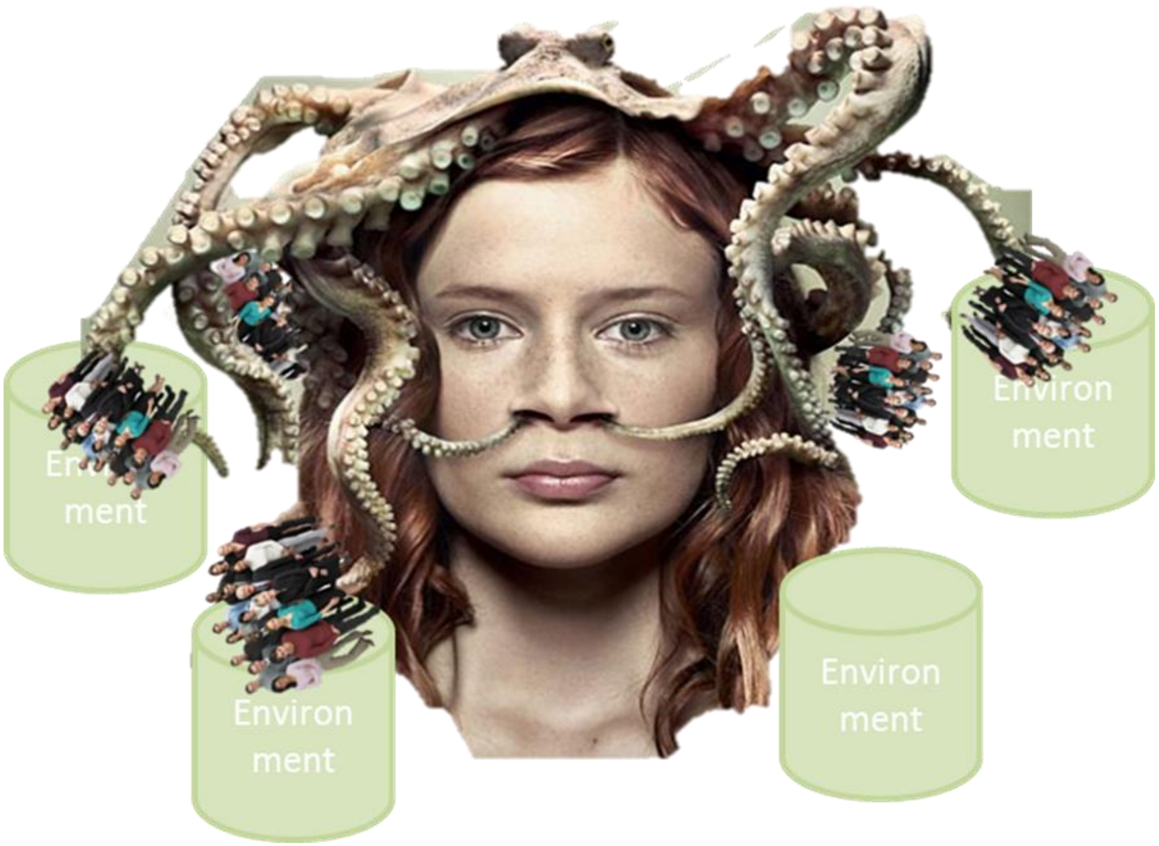
Public Sentiment

- Political and election indices, market insights

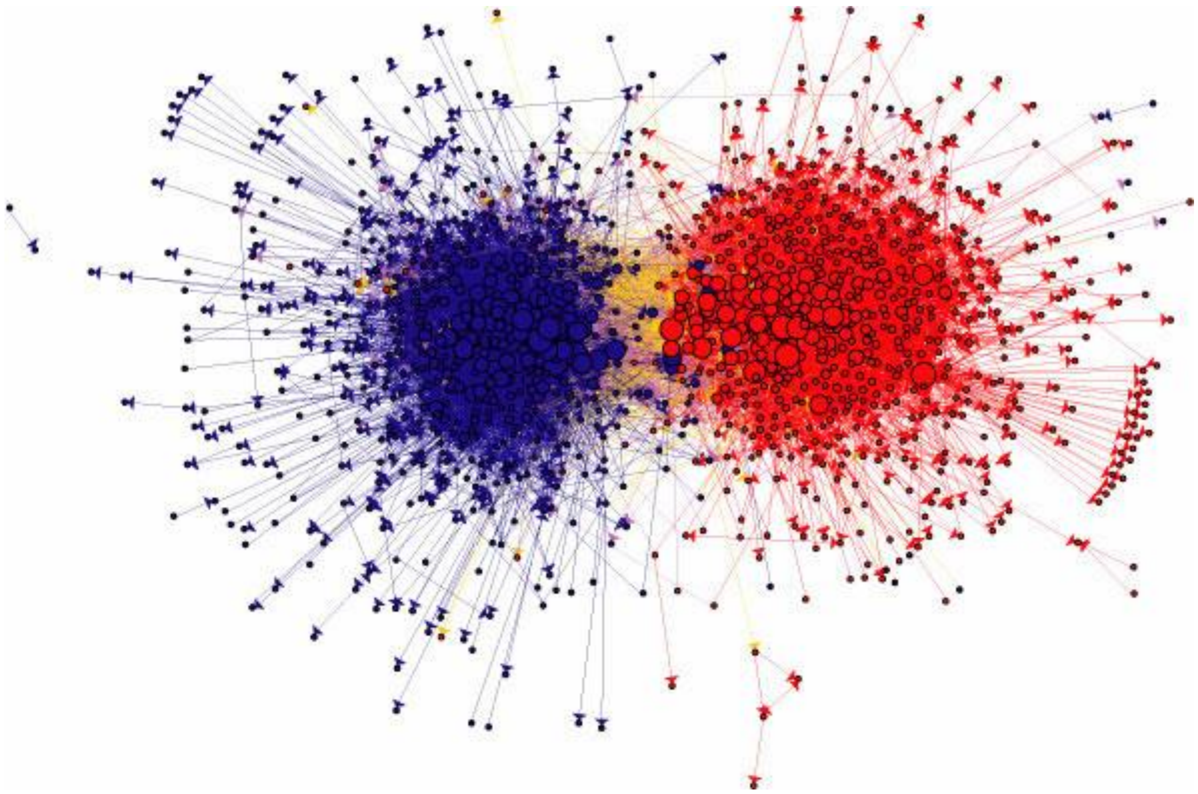
Everyday life

Social Media Landscape (2011) (FredCavazza.net)

Information gathering:
"The human sensor network"



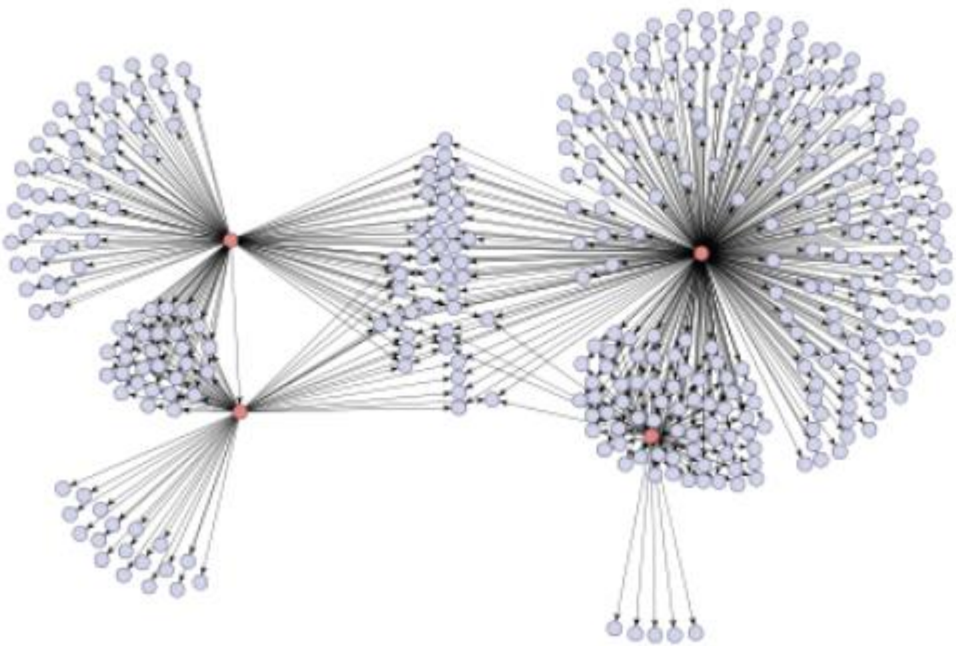
Political blogs

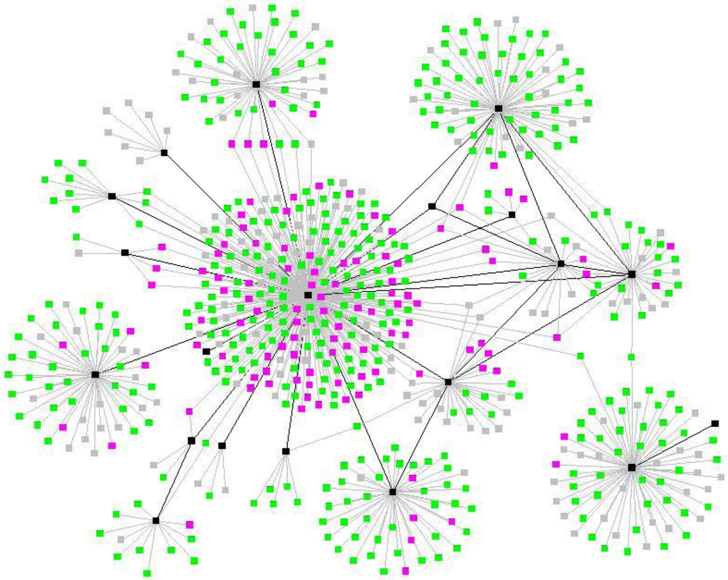


Adamic and Glance, 2005: The political blogosphere and the 2004 U.S. election

Social contagion cascade (viral marketing, Japanese graphic novel)

Biological contagion (tuberculosis outbreak)





Leskovec, Adamic & Huberman, 2007. The dynamics of viral marketing.

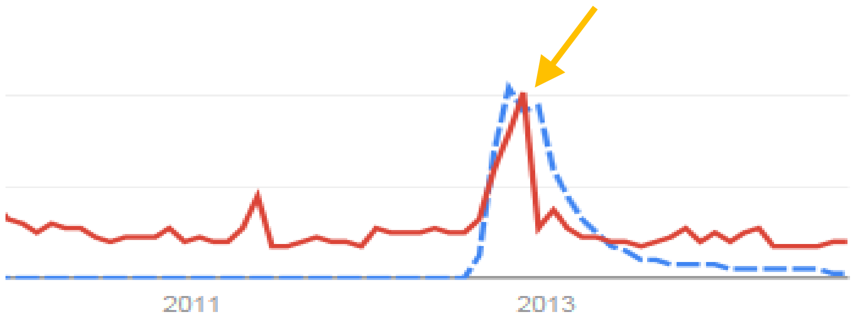
Andre *et al.*, 2007. Transmission network analysis to complement routine tuberculosis contact investigations.

Google trends

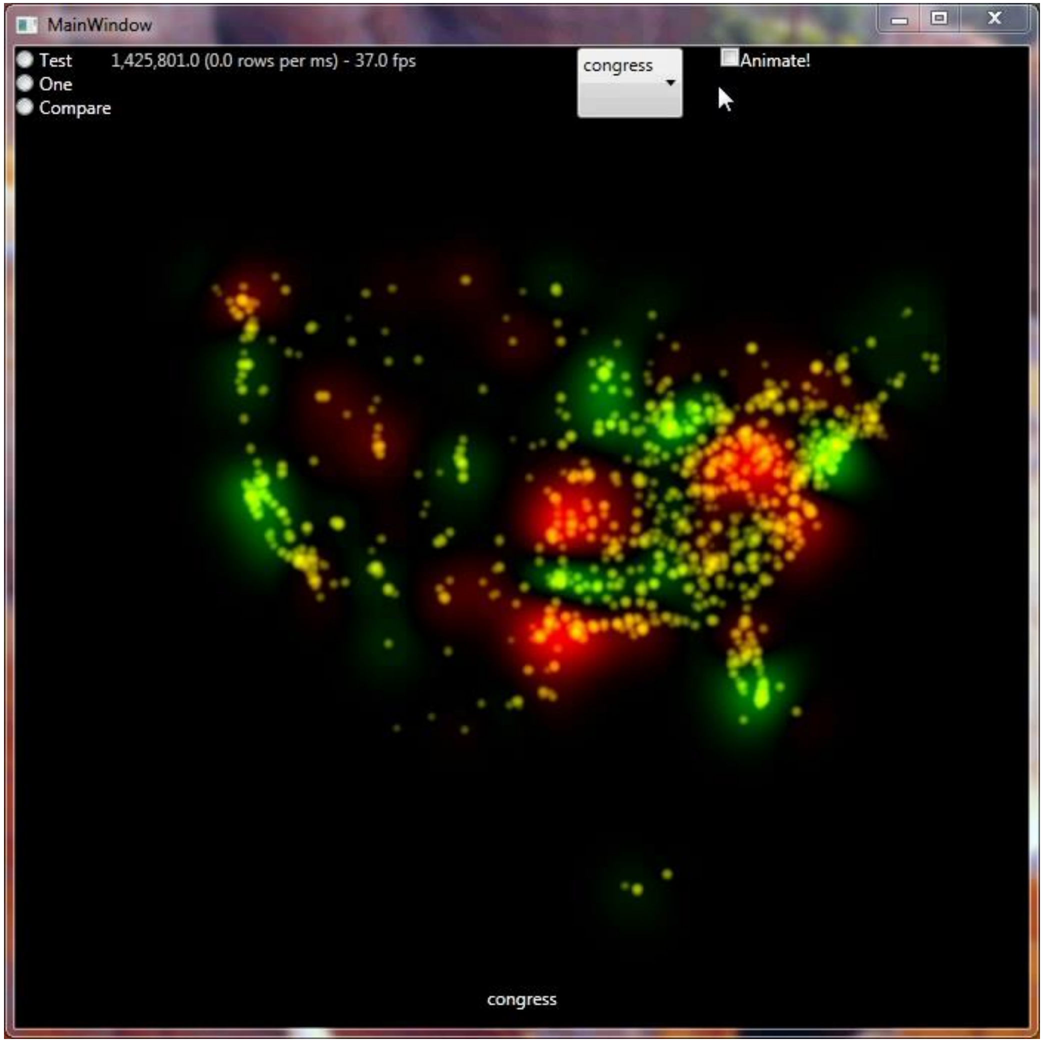
Gangnam style

Obama

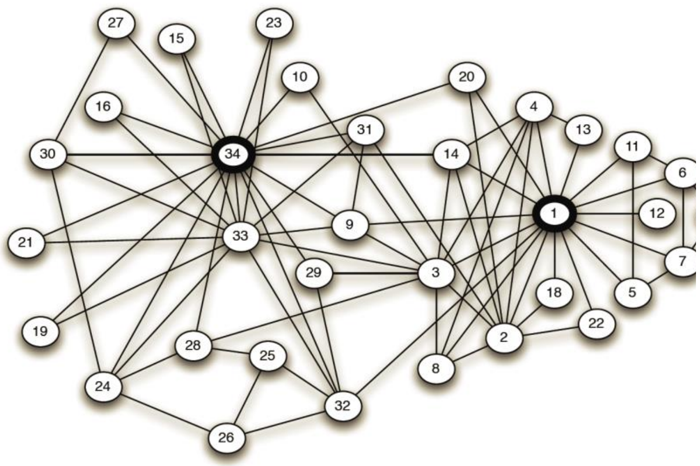
President Obama got re-elected

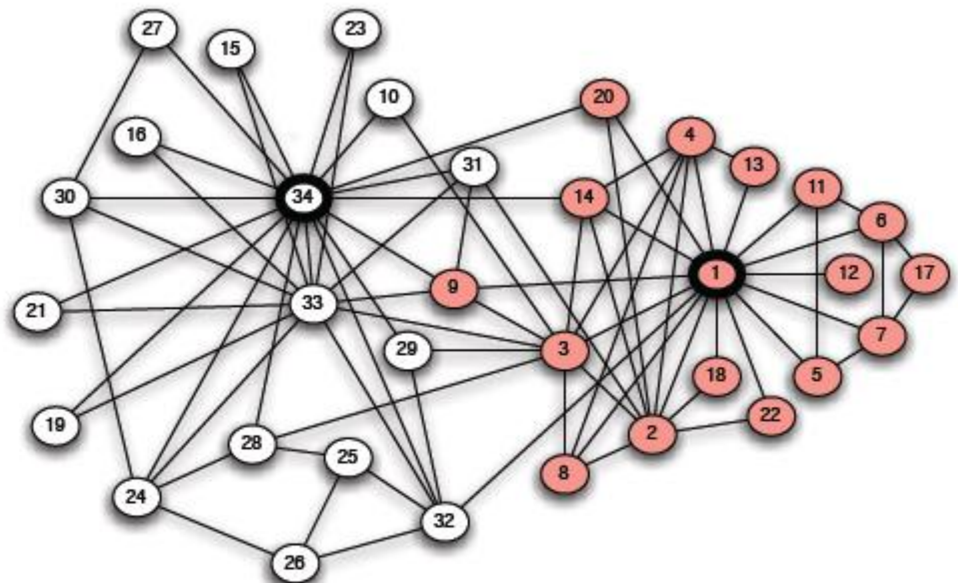


Location based sentiment towards US Congress



Karate club friendships





Zachary, 1977. An information flow model for conflict and fission in small groups

The Big $$$ Question

How does Facebook, Twitter, Microsoft, Google, LinkedIn, NSA, PET, etc. use your social media data?

Where/how do you think social media data could be used?

Many possibilities → new start-ups!

Social Networks & Graph Theory

A social network is made up of actors (people, organizations, groups,...) and ties (social relationships, similarity,...)

A graph consists of nodes and edges between nodes.

$\Rightarrow$ Social network problems can in many cases be represented as a standard graph theoretic problems

Network *Structure* (Statics)

Emphasize purely structural properties

- size, diameter, connectivity, degree distribution, etc.

Structure can reveal:

- communities
- "important" actors, centrality, etc. Who are the most influential members of a network?
- robustness and vulnerabilities
- anomalies, social awkward behavior
- can also impose constraints on dynamics

Less emphasis on what actually occurs on network

- web pages are linked… but people surf the web
- buyers and sellers exchange goods and cash
- friends are connected… but have specific interactions
- transfer of knowledge, ideas, recommendations, commercial promotions…

Network *Dynamics*

Emphasis on what happens on networks

Examples:

- spread of disease/meme/fad in a social network
- transfer of knowledge, ideas, recommendations, viral marketing…
- spread of wealth in an economic (social) network

Statics and dynamics often closely linked

- rate of disease spread (dynamic) depends critically on network connectivity (static)
- distribution of wealth depends on network topology

Network *Formation*

Why does a particular network structure emerge?

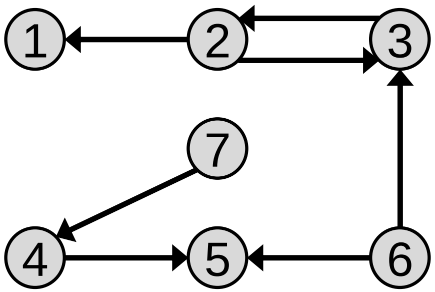
Plausible processes for network formation?

What are the characteristics and why?

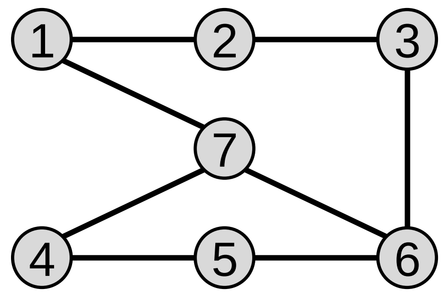
- Power-law degree distribution
- High clustering coefficient
- Small average path length

Basic Graph Theory

Directed and Undirected graphs



Directed graph
Well suited to represent follow/follower relationships seen on e.g. Twitter

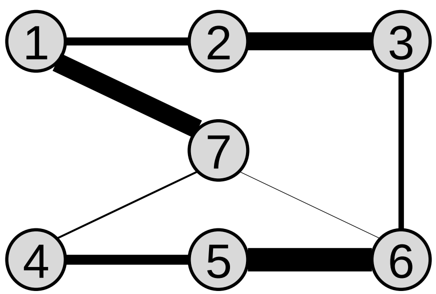


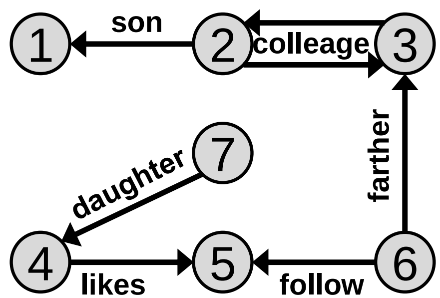
Undirected graph
Well suited to represent undirectional relationships such as Facebook friends

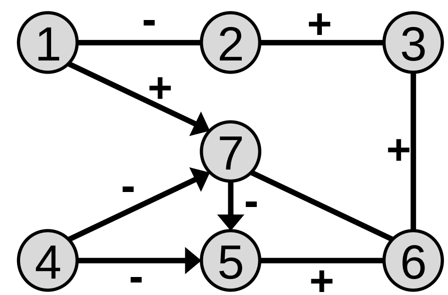
Graph: $G = (V, E)$

Nodes: $V = \{v_1, v_2, \ldots, v_n\}$
Edges: $E = \{e_1, e_2, \ldots, e_m\}$

Weighted, Labeled, and Signed Graphs







Weighted (undirected) graph
Well suited to represent **intensity** of relationships such as number of interactions (e.g. email) or **similarity** (e.g. sentiments towards products)

Labeled (directed) graph
Well suited to represent **types** of relationships

Signed (mixed) graph
Well suited to represent **friends and foes** (undirected) or **social status** (directed)

Graph: $G = (V, E, W)$

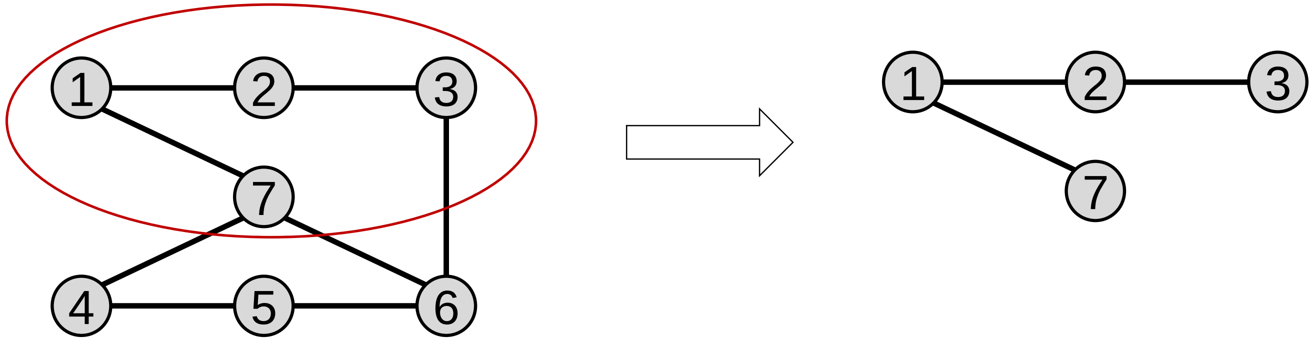
Nodes: $V = \{v_1, v_2, \dots, v_n\}$
Edges: $E = \{e_1, e_2, \dots, e_m\}$
Edge weights/labels/signs: $W = \{w_1, w_2, \dots, w_m\}$

Possible to annotate nodes as well!

Subgraph

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

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



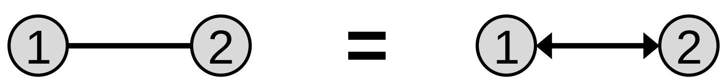
Degree (of a node)

For directed graph

- In-degree of node v_i = number of edges pointing into the node. d_i^{in}
- Out-degree of node v_i = number of edges pointing away from the node. d_i^{out}
- Degree = In-degree + Out-degree. d_i

For undirected graph

- Undirected edge = two opposite directed edges (aka. bi-directional or reciprocal edge)



- In-degree = Out-degree = number of edges connected to node
- Degree = 2 times the number of edges

Degree Distribution

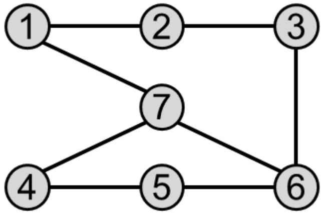
Degree distribution:

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

n_d :number of nodes with degree d

n :number of nodes

(Unrealistic) example:



$p_2 = 0/7, p_4 = 5/7, p_6 = 2/7$

Degree Distribution (for network)

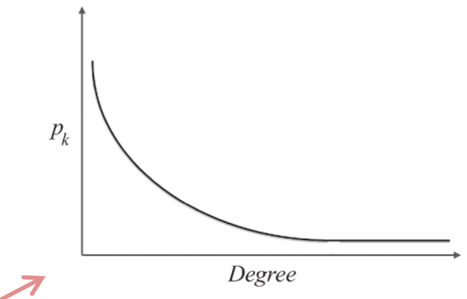
More realistic examples

- Many internet sites are visited <1,000 times a month whereas few are visited more than a million times daily.
- Social media users are often active on a few sites whereas a few individuals are active on hundreds of sites.
- On Facebook there exist many individuals with a few friends and a handful of users with many thousands of friends
- On Twitter many individuals follow (or are followed by) a few other individuals, and a few follows (or are followed) by a huge crowd

Me and you(?) ⟷ Obama or Justin Bieber

Power-Law Degree Distribution

- Many real-world (social) networks exhibit a *power-law* distribution.
- Power laws seem to dominate in cases where the quantity being measured can be viewed as a type of *popularity*. (e.g., node degree)
- A power-law distribution implies that small occurrences are common, whereas large instances are extremely rare.

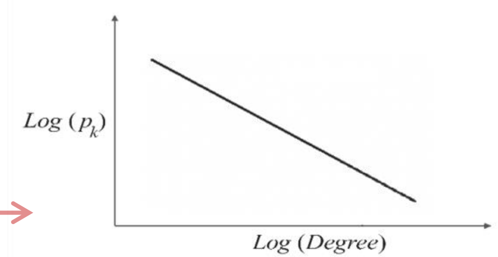


(a) Power-Law Degree Distribution

Power-Law degree distribution:

$$p_d = \beta d^{-\alpha}$$

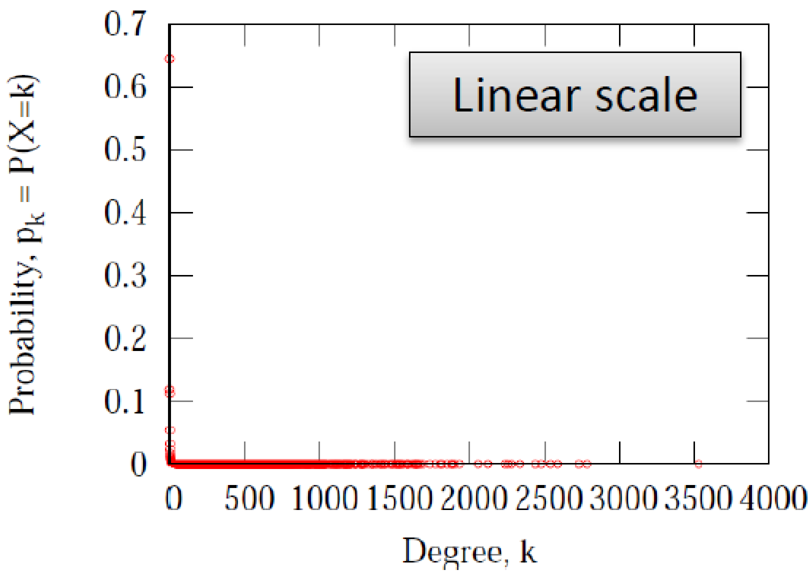
$$\log(p_d) = \log(\beta) - \alpha \log(d)$$

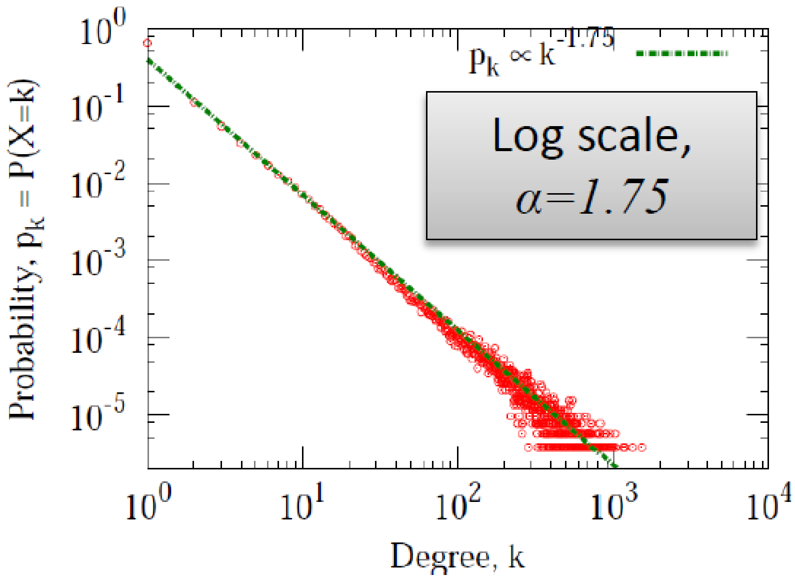


(b) Log-Log Plot of Power-Law Degree Distribution

α: the power-law exponent and its value is typically in the range of [2, 3]
β: power-law intercept

Flickr: Power-Law Degree Distribution





From: Jure Leskovec, ICML '09 tutorial

Graph representation

Visual graph good for intuition, but not for computation!

For computations we instead use

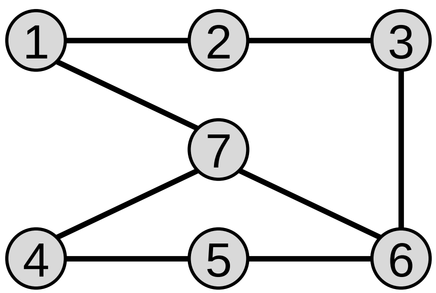
- Adjacency Matrix
- Adjacency List
- Edge list

Adjacency Matrix

Properties

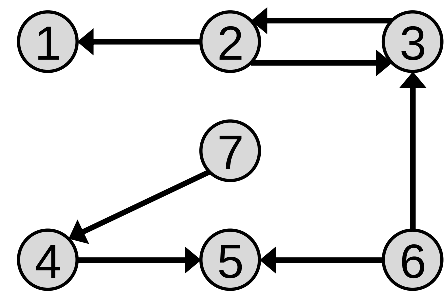
Symmetric

Asymmetric



A	1	2	3	4	5	6	7
1	0	1	0	0	0	0	1
2	1	0	1	0	0	0	0
3	0	1	0	0	0	1	0
4	0	0	0	0	1	0	1
5	0	0	0	1	0	1	0
6	0	0	1	0	1	0	1
7	1	0	0	1	0	1	0

Weights/Labels/Signs instead of 0/1

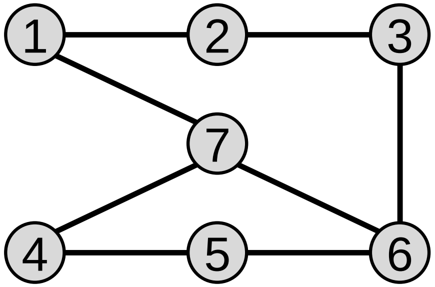


A	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
2	1	0	1	0	0	0	0
3	0	1	0	0	0	0	0
4	0	0	0	0	1	0	0
5	0	0	0	0	0	0	0
6	0	0	1	0	1	0	0
7	0	0	0	1	0	0	0

Social media networks have very sparse adjacency matrices

Adjacency List (aka. sparse representation)

- Every node maintains a list of all the nodes that it connects to (in direction of arrow)
- Recall: undirected edges are bi-directional
- The list is usually sorted based on the node order or other preferences

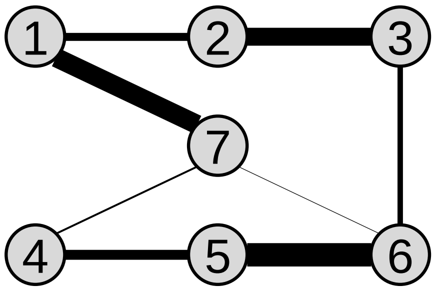


Node	Connects to
1	2, 7
2	1, 3
3	2, 6
4	5, 7
5	4, 6
6	3, 5, 7
7	1, 4, 6

Space efficient for sparse network

Adjacency List

- …and with Weights/Labels/Signs



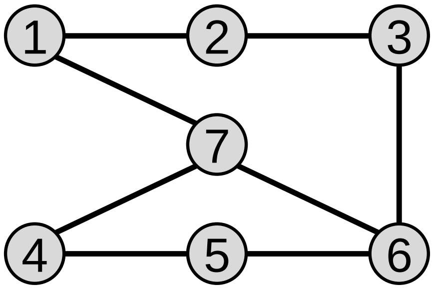
Node	Connects to
1	2:4, 7:11
2	1:4, 3:8
3	2:8, 6:2
4	5:6, 7:1
5	4:6, 6:14
6	3:2, 5:16, 7:1
7	1:11, 4:1, 6:1

Q: Does it look like something you have seen before?

A: Inverted index with terms replaced by nodes and documents replaced by "connects to" nodes

Edge List

- Each element is an edge and is usually represented as *(u, v)*, denoting that node *u* is connects to node *v* via a directed or undirected edge (semantic must be specified)



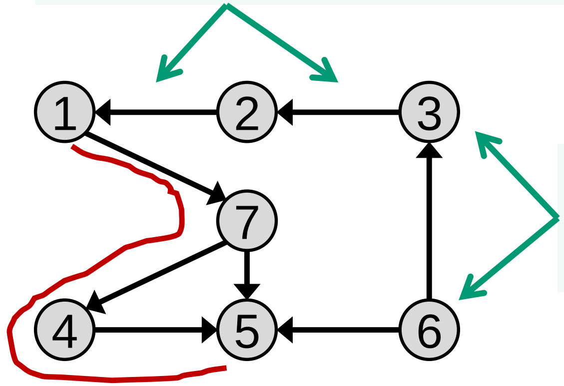
Edge list
(1,2)
(1,7)
(2,3)
(3,6)
(4,5)
(4,7)
(5,6)
(6,7)

- Not very useful for our tasks (personal opinion)

Connectivity

Path

Two edges are incident, if sharing one node

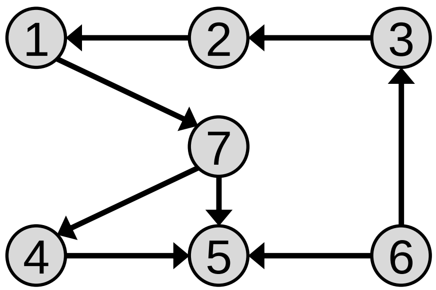


Two nodes are adjacent, if connected with edge

Length of path $1 \rightarrow 7 \rightarrow 4 \rightarrow 5$ is 3

- A walk (following edges) in the graph, where all nodes and edges are distinct is called a path (obeys directions in directed graph)
- A closed path (first node = end node) is called a cycle
- The length of a path (or cycle) is the number of traversed edges

Shortest path



$l_{1,5} = 2$

$diameter_G = 5$
(from v_6 to v_4)

Shortest Path is the path between two nodes that has the shortest length.

- $l_{i,j}$ denotes the shortest path between nodes v_i and v_j

An n-hop neighborhood of a node is the set of nodes that are reachable by shortest paths of length $\leq n$ from the node.

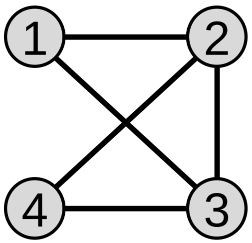
The diameter of graph is the length of the longest shortest path between any nodes in the graph

- $diameter_G = \max_{i,j} l_{i,j}$

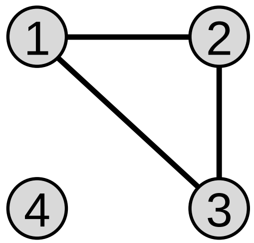
Connectivity

- 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 is not connected.

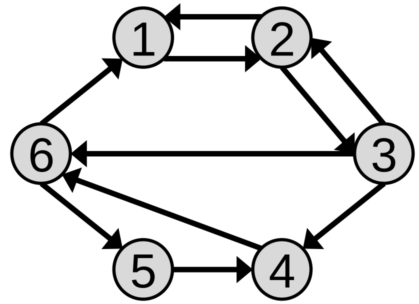
Connectivity (cont.)



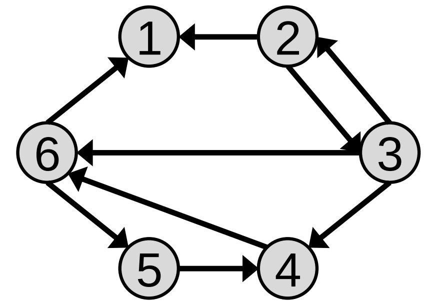
Connected



Disconnected



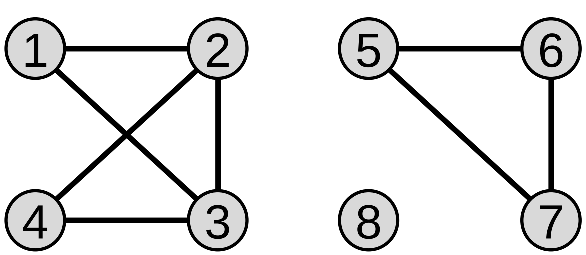
Strongly connected



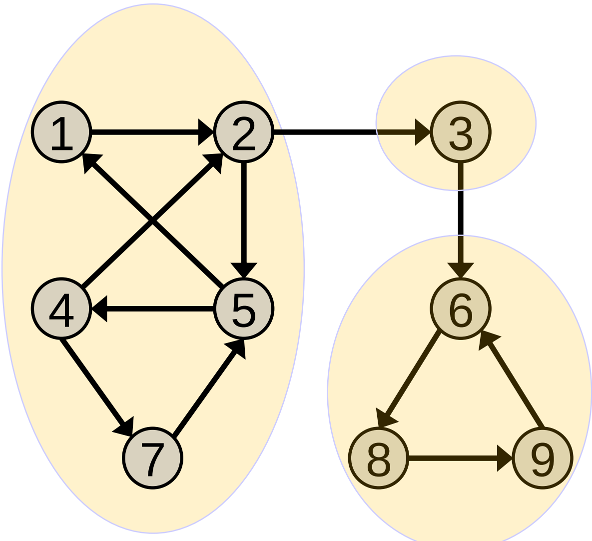
Weakly connected

Component

- A (connected) 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, component is strongly connected if there is a path from u to v and one from v to u for every pair (u,v).
- The component is weakly connected if replacing directed edges with undirected edges results in a connected component.



3 components



1 weakly connected components

3 strongly connected components

Graph Traversal Algorithms

Graph/Tree Traversal Algorithms

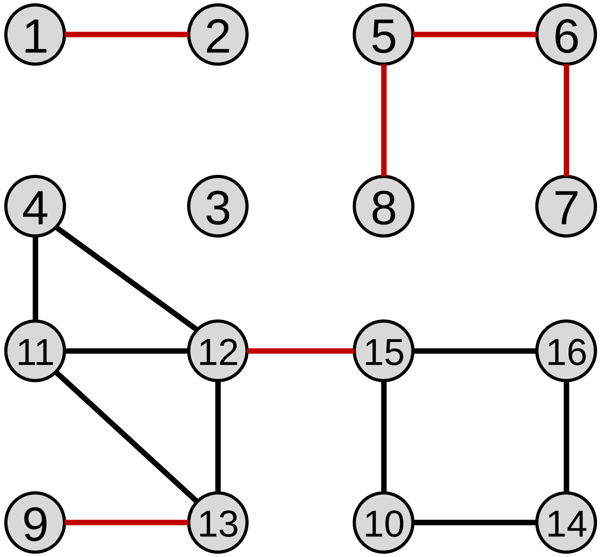
- Depth-First Search (DFS)
- Breadth-First Search (BFS)

Shortest Path Algorithms

- Dijktra’sAlgorithm (1 to many)
- Bellman-Ford Algorithm (1 to many)
- Floyd-Warshall Algorithm (many to many)

Bridges (cut-edges)

Bridges are edges whose removal will increase the number of connected components



(Simple) Bridge Detection Algorithm

(Zafarani, Abbasi & Liu (2014): Social Media Mining: An Introduction)

Algorithm 2.7 Bridge Detection Algorithm

Require: Connected graph $G(V, E)$

return Bridge Edges
$bridgeSet = \{\}$
for $e(u, v) \in E$ **do**
    G' = Remove e from G
    Disconnected = False;
    if BFS in G' starting at u does not visit v **then**
        Disconnected = True;
    end if
    if Disconnected **then**
        $bridgeSet = bridgeSet \cup \{e\}$
    end if
end for
Return $bridgeSet$

Network measures

Network measures – Why?

1. Who are the central figures (influential individuals) in the network? – Centrality

2. Who are the “gate-keepers” (influential individuals) in the network? – Centrality (another type)

3. Who are the like-minded users and how can we find these similar individuals? – Similarity

Centrality

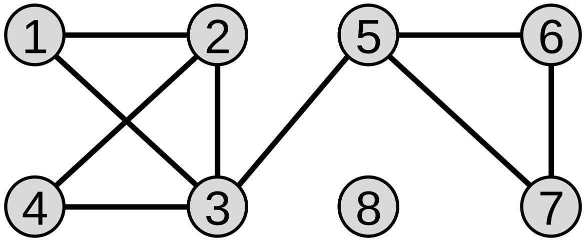
- **Defines how important a node is in a network**
- **Many different centrality measures**

Degree Centrality

- Ranks nodes with more connections higher
- Intuition: Important/influential people have more friends/connections.
- Degree Centrality:

$$C_d(v_i) = d_i$$

- Is a local measure of centrality



$C_d(v_3) = 4$

$C_d(v_8) = 0$

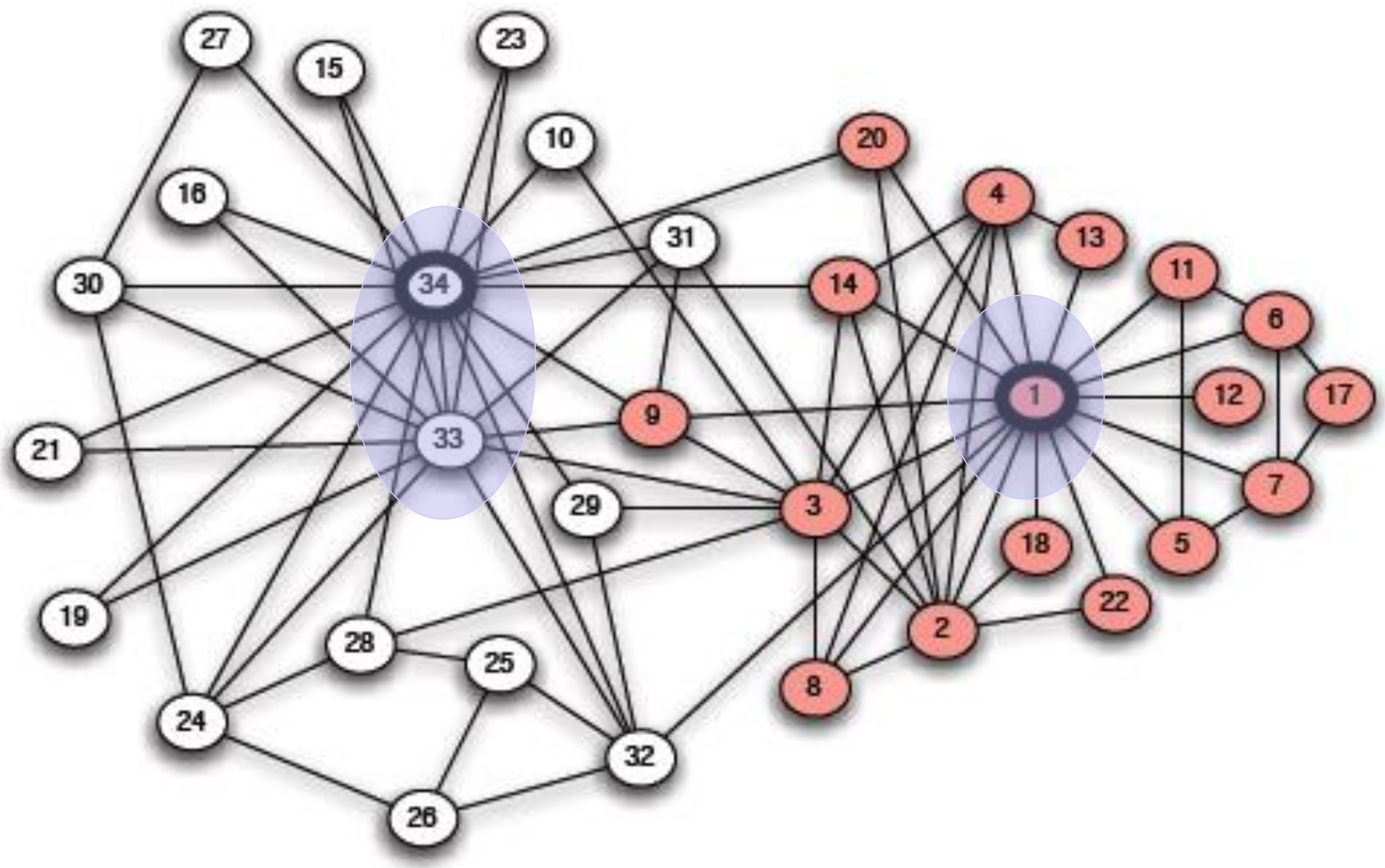
For directed graph

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

-or- $= d_i^{out}$ (gregariousness)

-or- $= d_i^{in} + d_i^{out}$

Karate club – degree centrality



Eigenvector Centrality

Q: Look similar to something you have seen before?

- Is a global measure of centrality
- Ranks nodes with more <u>connected</u> connections higher
- Intuition: Important/influential people have more <u>influential</u> friends/connections.
- Eigenvector Centrality:

$\boldsymbol{A}_{i,j} = 0 \text{ when } i \not\to j$

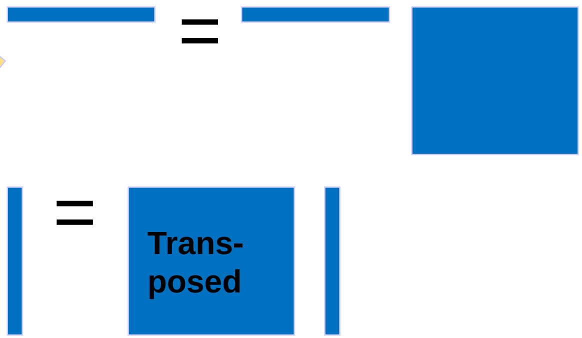
$$C_e(v_j) = \sum_{i=1}^{n} C_e(v_i)\boldsymbol{A}_{i,j} = \sum_{i:i\to j}^{n} C_e(v_i)\boldsymbol{A}_{i,j}$$

Matrix notation, the way we talk about it:

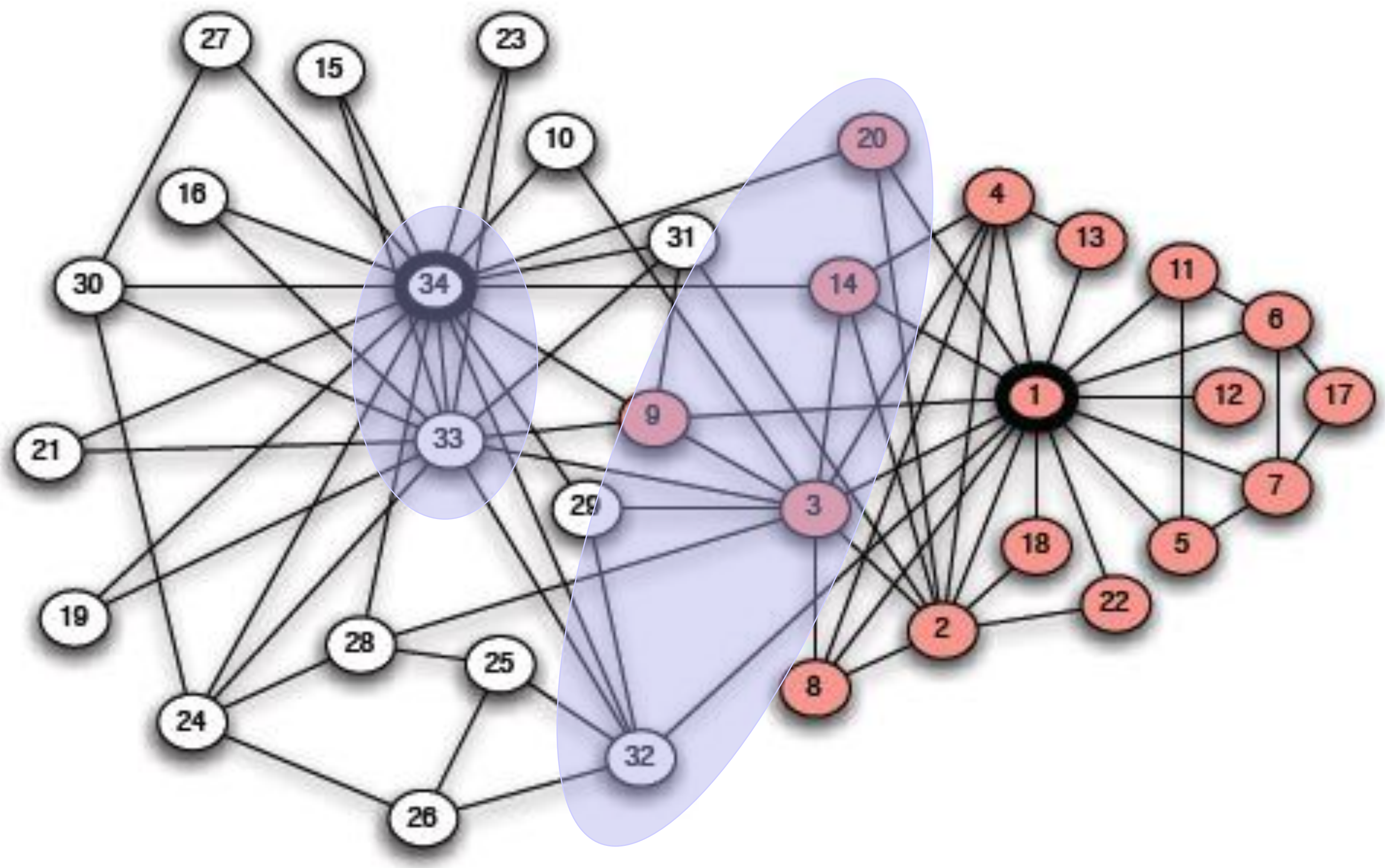
$$C_e = C_e\boldsymbol{A}$$

-or-

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



Karate club – eigenvector centrality



Katz Centrality

- Similar to Eigenvector centrality with "teleporting"
- Katz Centrality:

Vector of all 1's

$$C_K = aC_K\boldsymbol{A} + b\bar{1}$$

a and b are chosen positive constants that determines tradeoff between graph structure relations and teleporting (a must be smaller than $1/\lambda$, where λ is the principal eigenvalue for $\boldsymbol{A}$)

-or-

Matrix of all 1's

$$C_K = C_K((1-\alpha)\boldsymbol{A} + \alpha\mathbf{1})$$

Method:

1. Compute principal eigen-value λ from A
2. Use λ to select a, b (or α)
3. Compute principal eigen-vector (the centrality measure)

PageRank Centrality

- In Katz an influential node distributes all its influence along all of its out-links. – no matter how many out-links (e.g. friends)
- PageRank centrality differentiates between nodes with few and nodes with many out-links
- PageRank Centrality:

$$C_P = C_p\big((1-\alpha)\boldsymbol{P} + \alpha\boldsymbol{U}\big)$$
$$= C_p\boldsymbol{P}_{Pagerank}$$

$$\boldsymbol{P} = \begin{pmatrix} P_{1,1} & \cdots & P_{1,j} & \cdots & P_{1,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ P_{i,1} & \cdots & P_{i,j} & \cdots & P_{i,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ P_{n,1} & \cdots & P_{n,j} & \cdots & P_{n,n} \end{pmatrix}$$

$P_{i,j}$ = probability of moving from page $\boldsymbol{i}$ to page $\boldsymbol{j}$

$$= \frac{1}{out\text{-}degree(i)}$$

Worth noticing:

- Each row sums to 1

Eigenvector, Katz, PageRank Centrality *computation*

Use Power Iteration Method – as defined by the measure.

Book suggests inversion – May be prohibitively expensive for larger networks; Matrix inversion complexity is between $O(n^{2.373})$ and $O(n^3)$

Betweenness Centrality

- Ranks "gate keeping" nodes higher
- Intuition: Important/influential people are those who brokers information in the flow between groups.
- Betweenness Centrality:

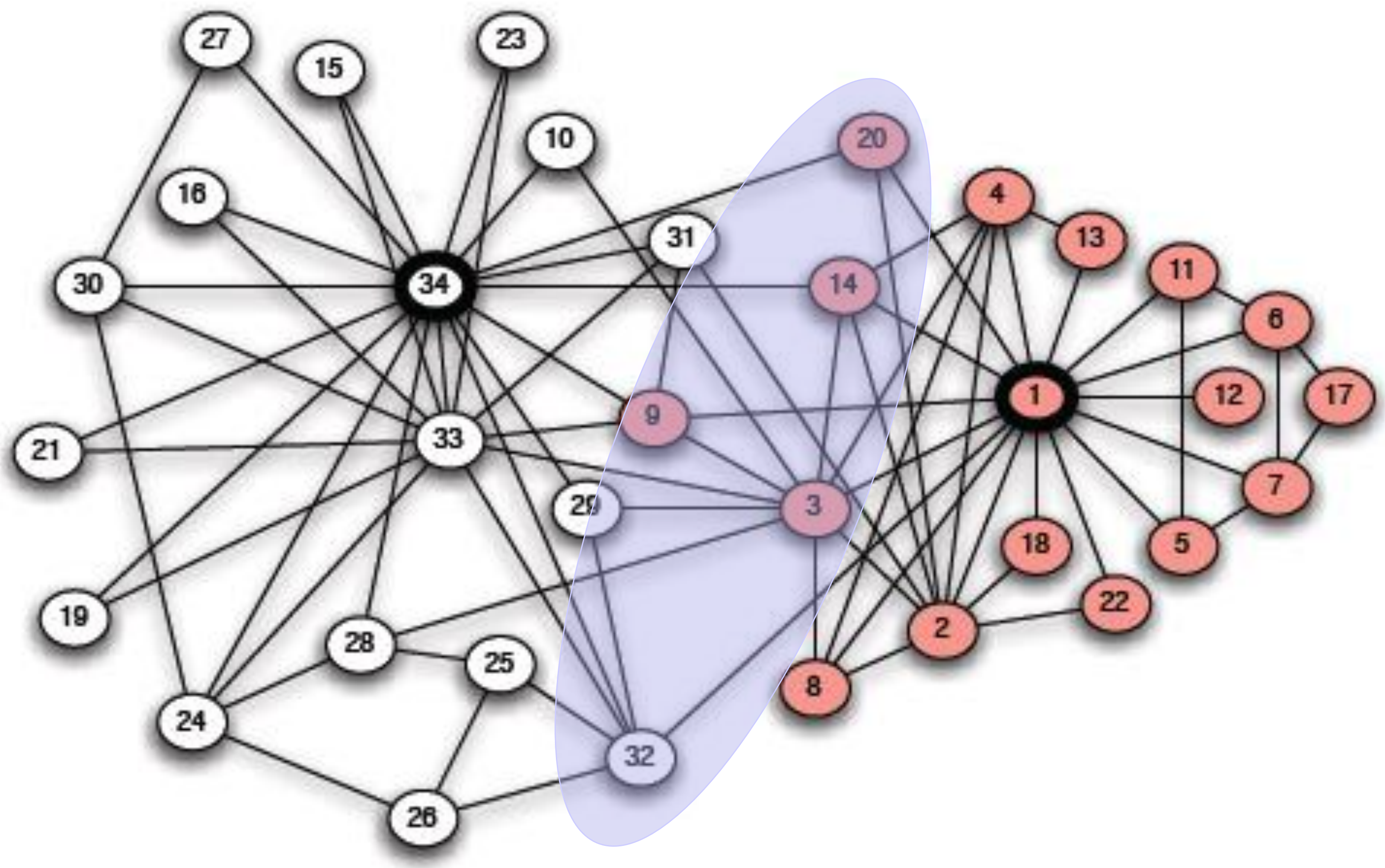
$$C_b(v_i) = \sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

σ_{jk} :the number of shortests paths from node v_j to node v_k

$\sigma_{jk}(i)$:the number of shortests paths from node v_j to node v_k that pass through node v_i

- Compute using n times Dijkstra's algo – once for each node (better methods exist)

Karate club – betweenness centrality

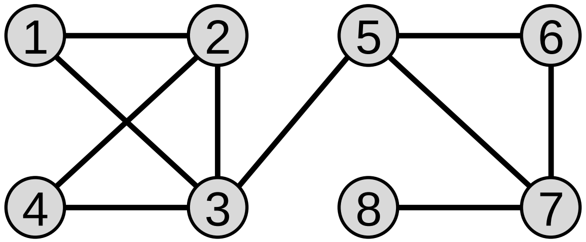


Closeness Centrality

- Ranks nodes with smaller average distance to other nodes higher
- Intuition: Important/influential people can quickly reach other people
- Closeness Centrality:

$$C_c(v_i) = \frac{1}{\overline{l_i}}$$

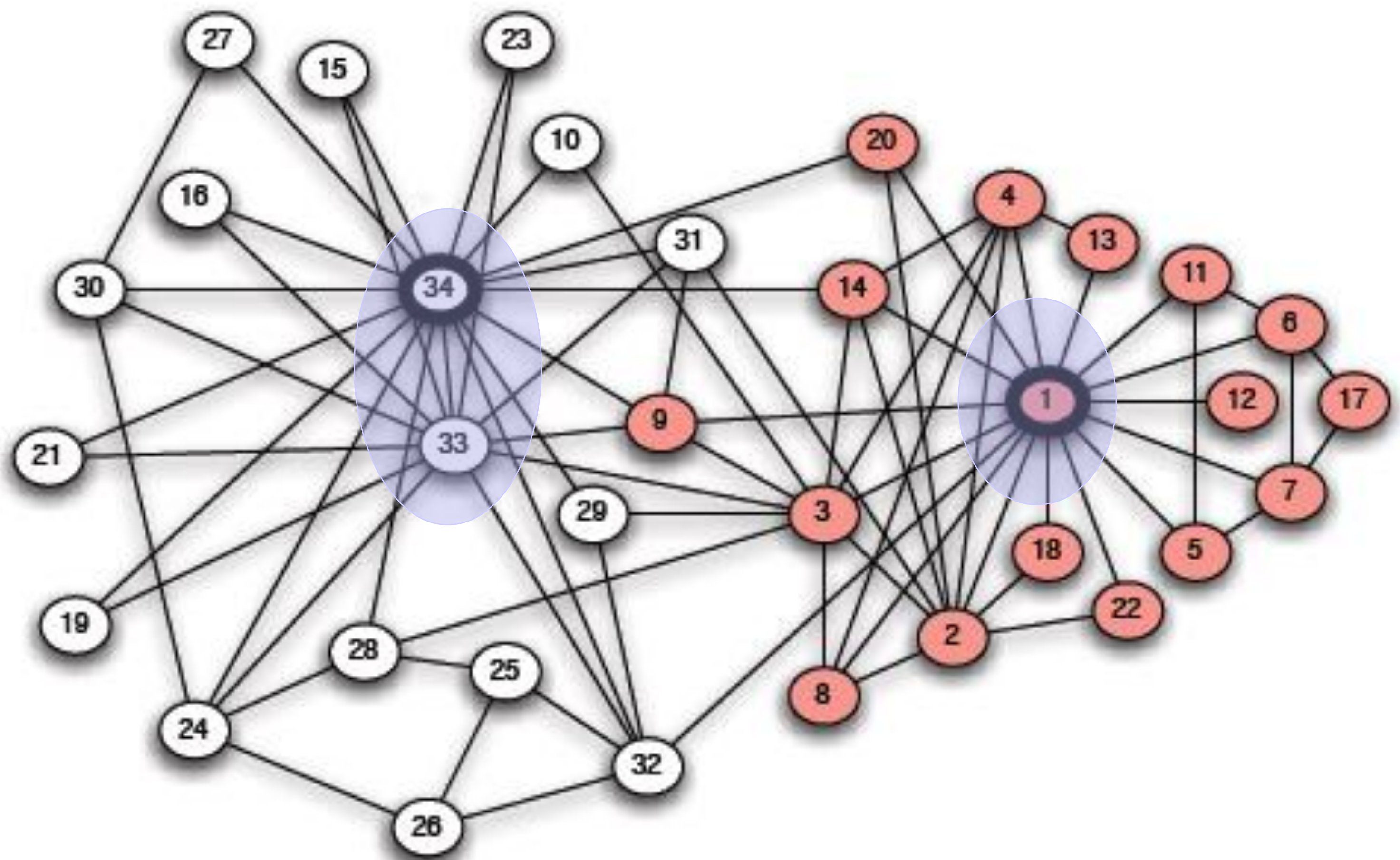
$$\overline{l_i} = \frac{1}{n-1} \sum_{j \neq i} l_{i,j}$$



$$\overline{l_3} = \frac{1+1+1+1+2+2+3}{8-1} = 11/7$$

$$C_c(v_3) = 7/11$$

Karate club – closeness centrality

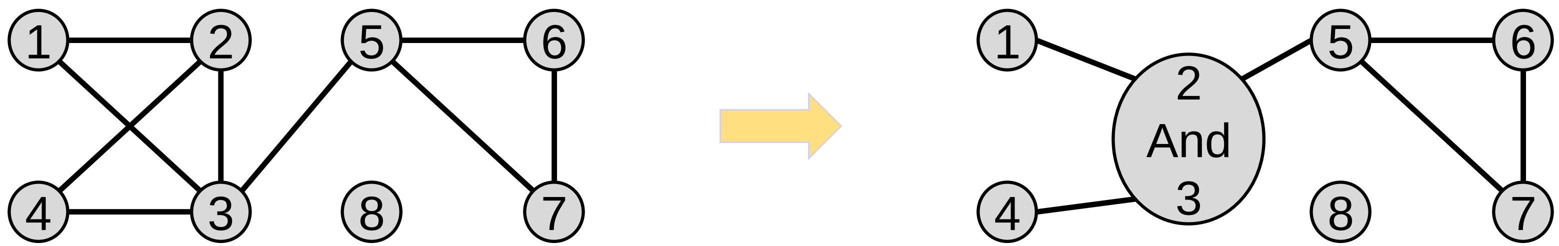


Group Centrality

All centrality measures defined so far measure centrality for a single node. These measures can be generalized for a group of nodes.

A simple approach is to replace all nodes in a group with a super node

- The group structure is disregarded.



Structural Similarity Measures
(between nodes)

Structural Equivalence: Definitions

Vertex similarity:

$$Sim_{Vertex}(v_i, v_j) = |N(v_i) \cap N(v_j)|$$

Jaccard similarity

$$Sim_{Jaccard}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j)|}$$

Cosine similarity

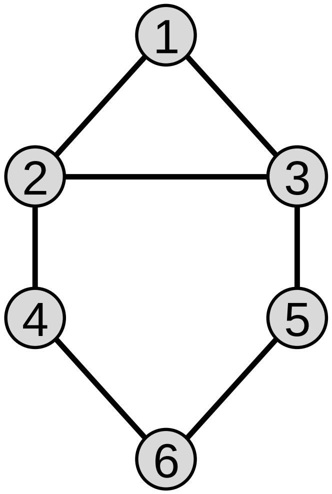
$$Sim_{Cosine}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{\sqrt{|N(v_i)||N(v_j)|}}$$

Q: What should we do if a node is without neighbors

Pearson similarity (aka. Pearson's correlation coefficient)

$Sim_{Pearson}(v_i, v_j)$ similar to standardized vector representation of Cosine similarity

Structural Equivalence: Example



$$Sim_{Vertex}(v_2, v_5) = |\{1,3,4\} \cap \{3,6\}| = 1$$

$$Sim_{Jaccard}(v_2, v_5) = \frac{|\{1,3,4\} \cap \{3,6\}|}{|\{1,3,4,6\}|} = \frac{1}{4}$$

$$Sim_{Cosine}(v_2, v_5) = \frac{|\{1,3,4\} \cap \{3,6\}|}{\sqrt{|\{1,3,4\}||\{3,6\}|}} = \frac{1}{\sqrt{6}}$$

$$Sim_{Pearson}(v_2, v_5) = \frac{1}{\sqrt{6}}$$

Similarities are NOT comparable across different similarity measures
Similarities are only comparable if computed with same measure

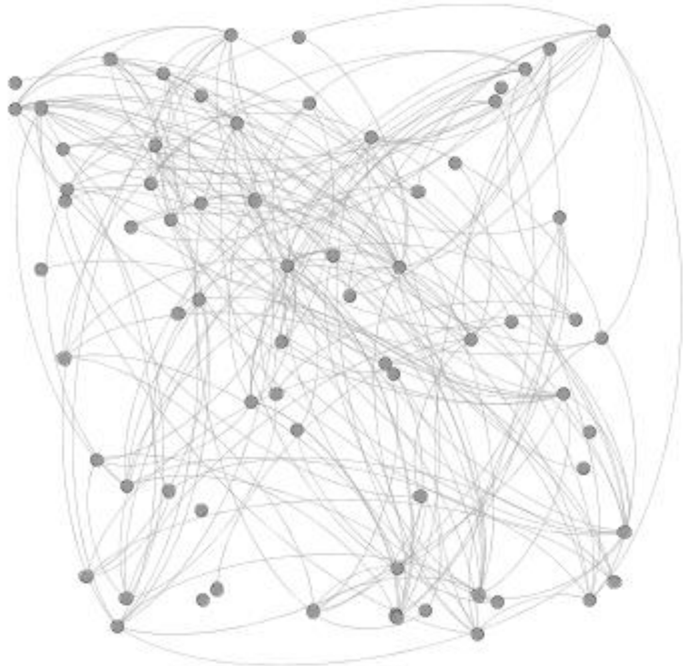
Some Fun ☺

Analyze your ~~Facebook~~ network

1. Download data from e.g.:
   - https://github.com/gephi/gephi/wiki/Datasets
   - (Everything from jazz musicians network to networks of super heroes to twitter mentions & retweets to …)
2. Download and install graph visualization tool Gephi
   - http://www.gephi.org
3. Open network(s) in Gephi and start exploring

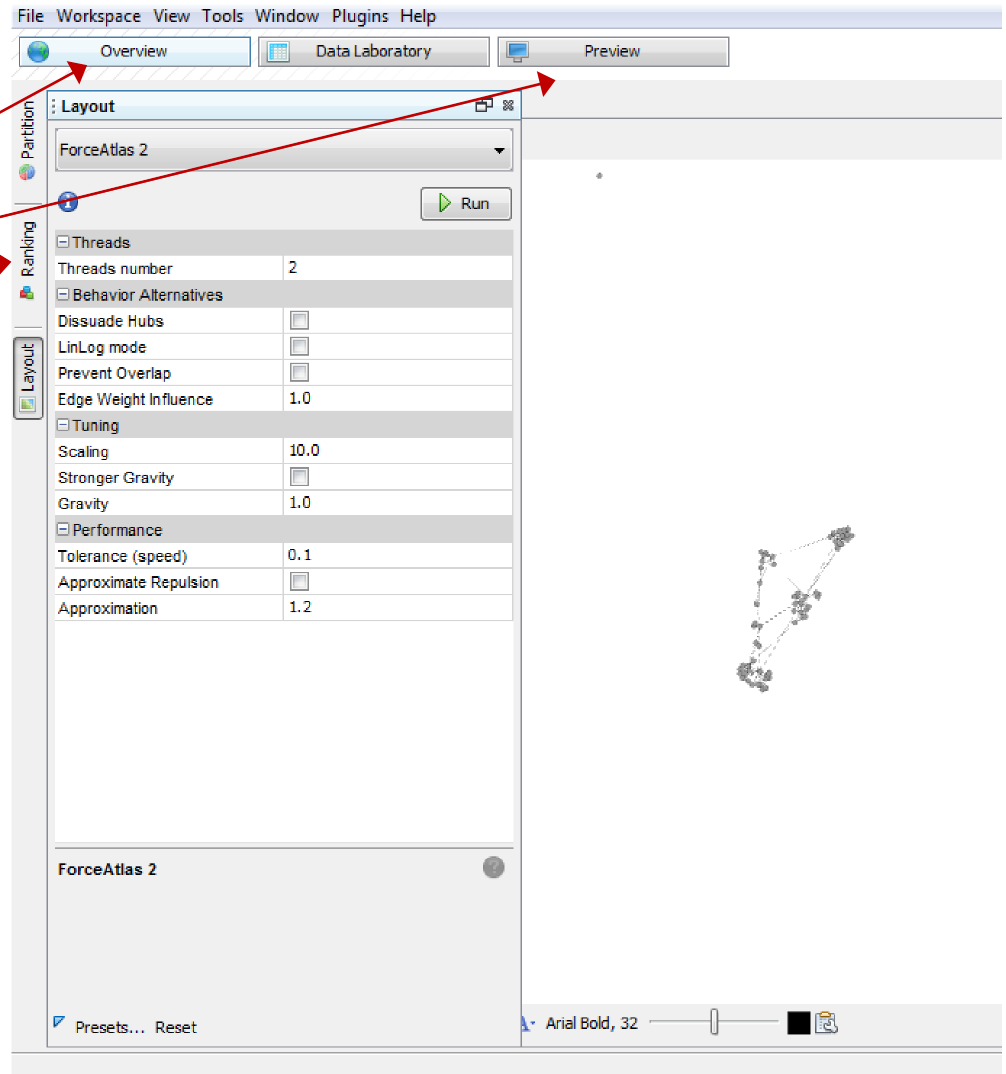
Gephi – Open your network

- From File menu select Open and then select the network file
- At first it looks like a big hairball, so we'll change the layout to make some sense of the connections



Gephi

Notice



Goto

'Overview->

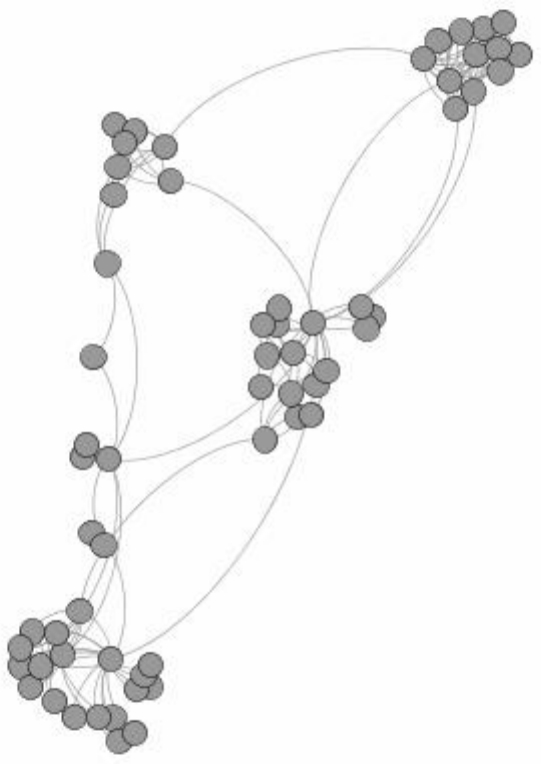
Layout

Gephi - Layout

- From the Layout module on the left side chose Force Atlas2 from the Dropdown Menu, then click run Force atlas makes connected nodes attract each other, while unconnected nodes are pushed towards the periphery

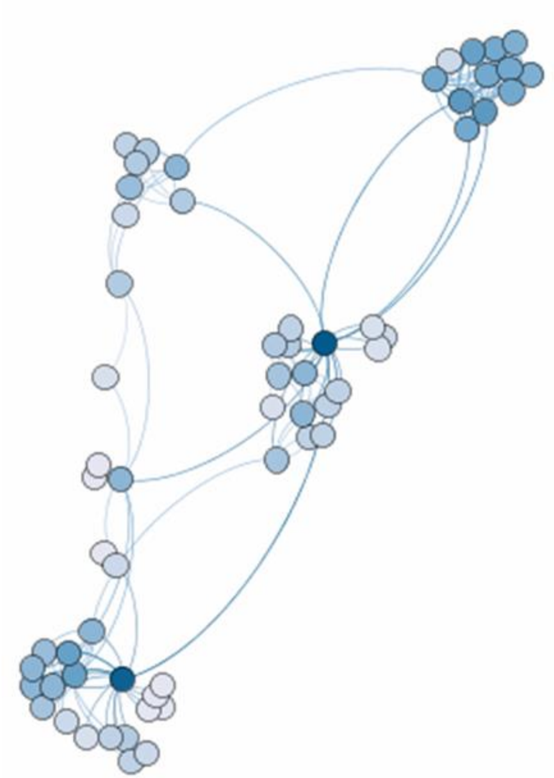
- Click stop when it seems that the layout has converged towards a stable state

- Go back to preview and hit the refresh button



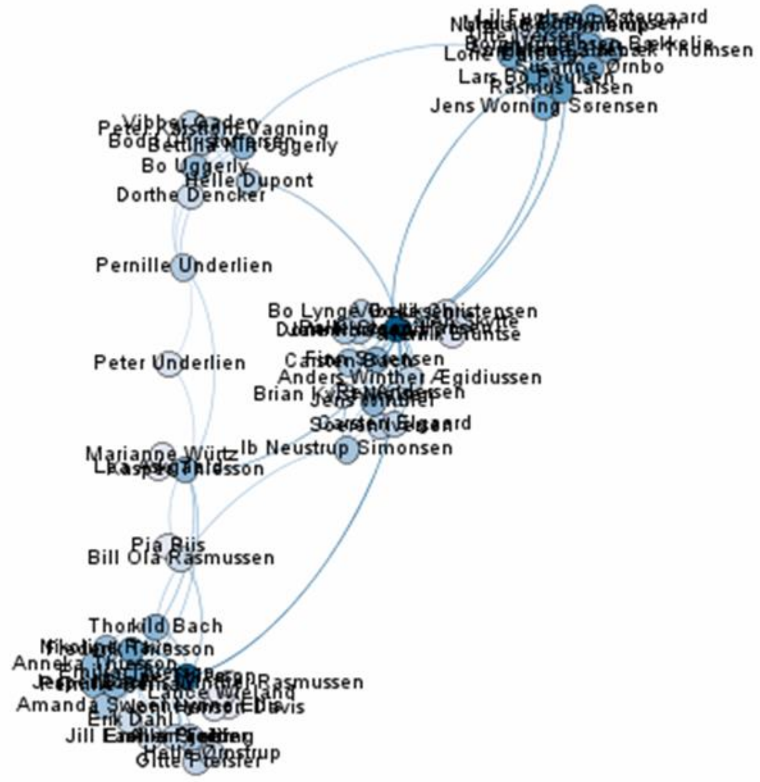
Gephi – Centrality Ranking

- Go back to ‘Overview’ and go to attribute tab
- Under nodes and Ranking, choose e.g. the ‘Degree’ option or any centrality option
- Run one of the ranking module to the right. E.g., Eigenvector centrality.
- Go back to preview and hit the refresh button



Gephi - Labels

- Make some sense out of the graph by adding labels (in the Layout tab)
- Remember to hit ‘Refresh’ every time you make a change.



Gephi - Explore

- Play around and figure out all the interesting little secrets in your personal network.

- Explore some of the different centrality measures that we have discussed today. Do the measures make sense or are you surprised?