Network analysis

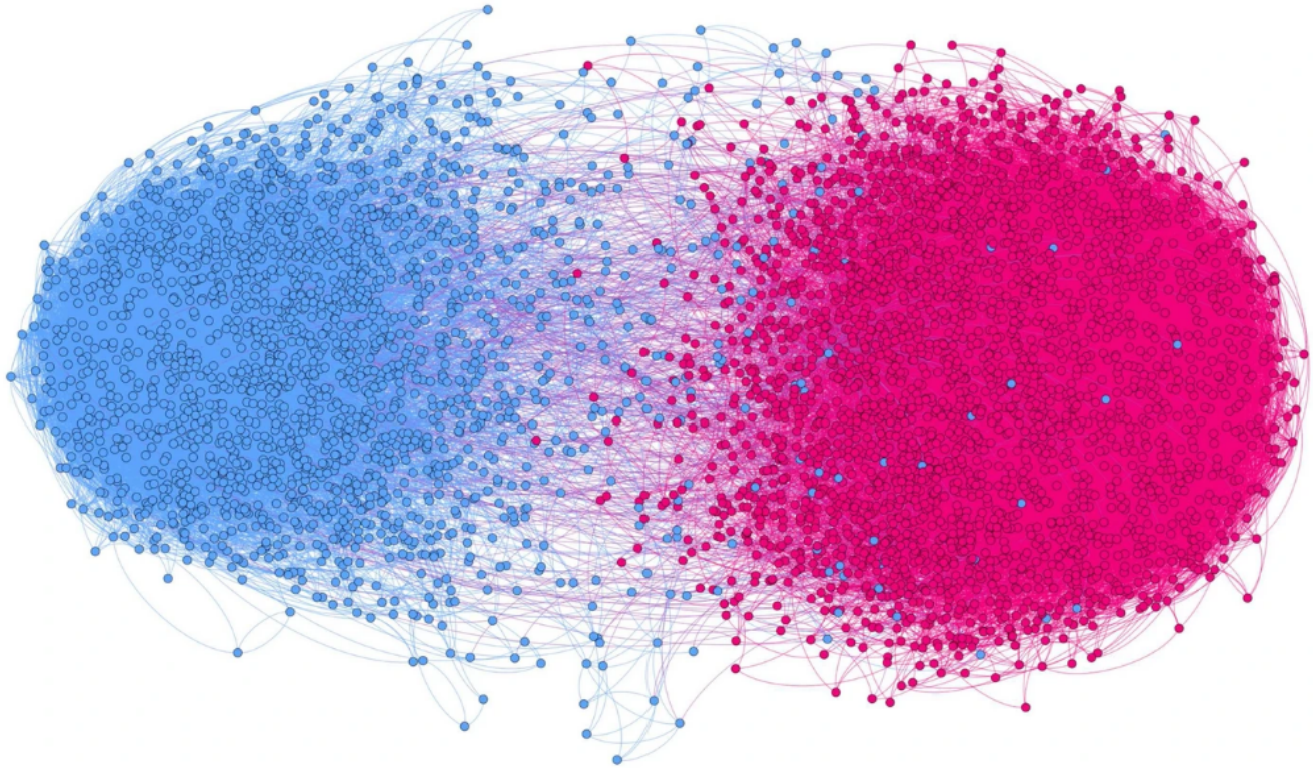
GOV 2018 TF – Yuning Liu, Ruofan Ma

Apr 2024

Content

Online social network

From: Social influence and unfollowing accelerate the emergence of echo chambers



Example of a polarized and segregated network on Twitter. The network visualizes retweets of political hashtags from the 2010 US midterm elections. The nodes represent Twitter users and there is a directed edge from node i to node j if user j retweeted user i. Colors represent political preference: red for conservatives and blue for progressives [20]. For illustration purposes, only the nodes in the $k = 3$ core are visualized. See Methods for more details

Social network for romantic relationships in a high school

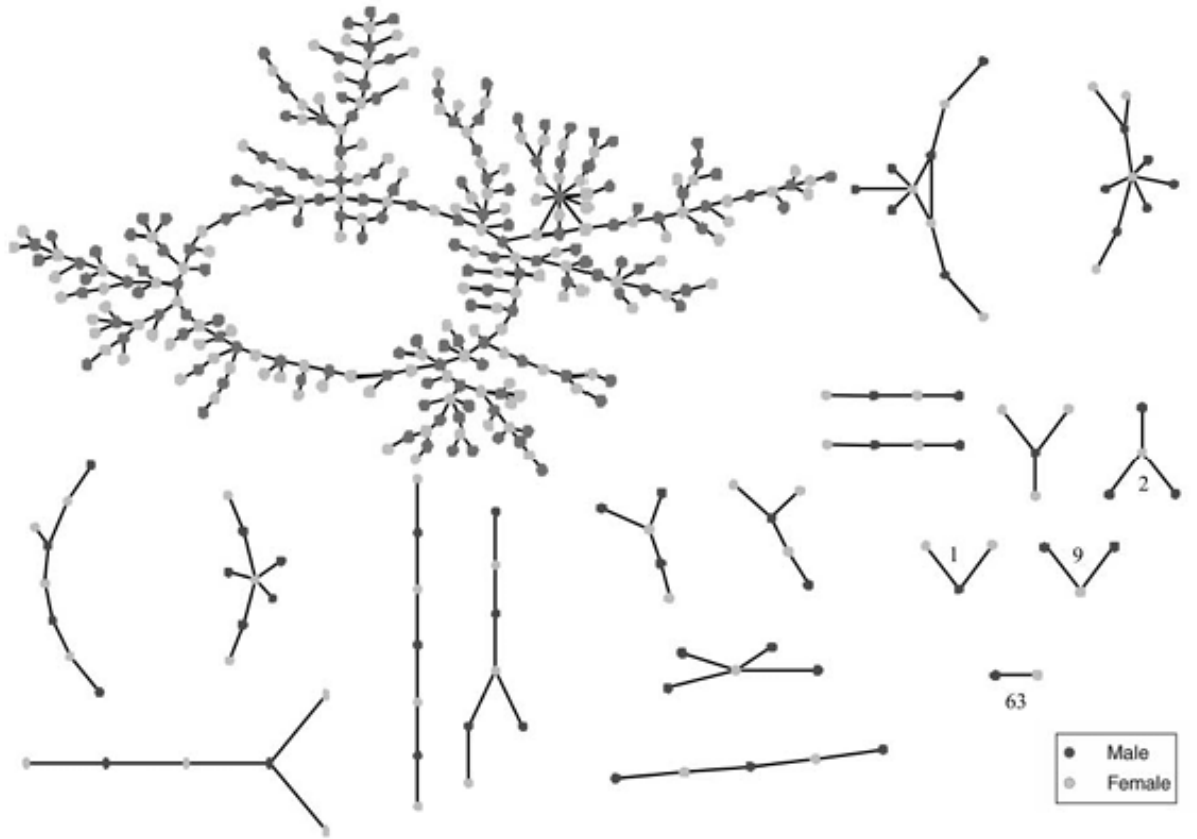


FIG. 2.—The direct relationship structure at Jefferson High

Networks and their representation

Basic elements:

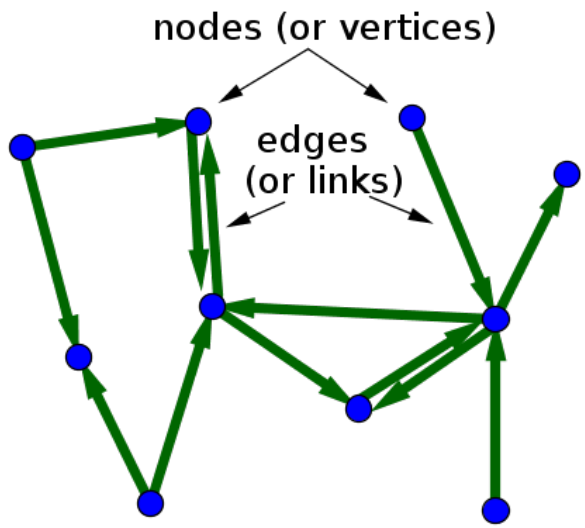
- Network or graph: collection of vertices joined by edges
- Vertices might also be called: nodes, sites, actors
- Edges might also be called: links, bonds, ties

Network	**Vertex**	**Edge**
Internet	Computer or router	Cable or wireless data connection
WWW	Web page	Hyperlink
Citation network	Article, patent, legal casee	Citation
Cosponsorship	Legislators	Cosponsored bills
Friendship network	Person	Friendship
Neural network	Neuron	Synapse

Further basic definitions

Graph G defined as $G(N, L)$

- Set of nodes (vertices) N: nodes can have attributes
- Set of links (edges) L:
  - Directed (arcs) or undirected (edges)
  - Unweighted or weighted (distance, traveling time, number of cosponsored bills, etc.)
  - Links of different types can exist (multiplex networks)



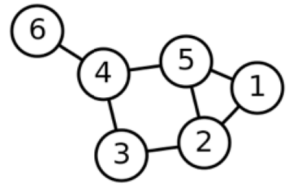
Edge list and adjacency matrix

Representing a network mathematically

- Edge list: n vertices, labeled 1, ..., n; with edge between vertices i and j referred to by (i ; j). A complete network can be specified by giving the value of n and a list of all the edges.
- Adjacency matrix: A of a simple graph is the matrix with elements A_ij such that

$$A_{ij} = \begin{cases} 1, & \text{if there is an edge between vertices i and j} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Edge list and adjacency matrix



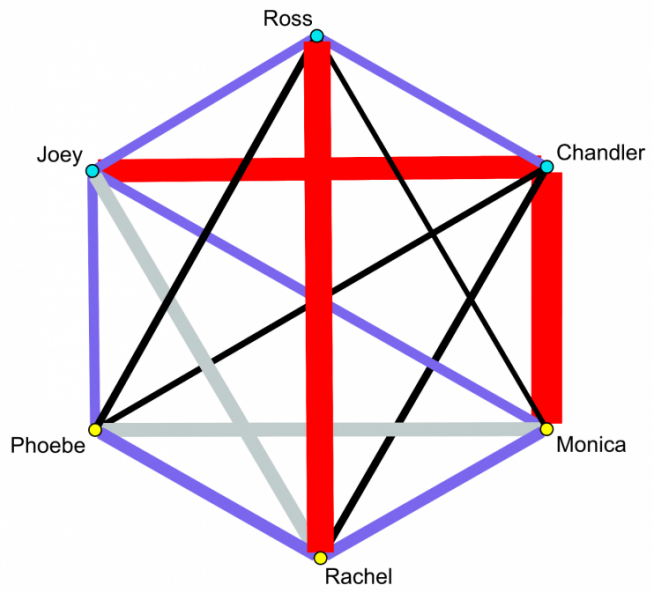
Edge list

1	2
1	5
2	1
2	5
3	2
3	4
4	3
4	5
4	6
5	1
5	2
5	4
6	4

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

Weighted network

- E.g. Internet edges have weights representing amount of data flowing; cosponsorship network has edges representing number of times two legislators cosponsor a bill.
- These are weighted or valued networks, and can be represented by simply giving elements of the adjacency matrix values equal to the weights of the corresponding connections.



In this diagram, the size of the chord for each character's section represents how many times they said the name of the connecting character.

Directed networks

Directed network/directed graph

- Network where each edge has a direction, pointing from one vertex to another
- Edges are then called directed edges, and can be represented by lines with arrows on them.
- Adjacency matrix of a directed network has matrix elements:

$$A_{ij} = \begin{cases} 1, & \text{if there is an edge from vertices j to i} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

  Note direction of edge is from second index to first.
- Adjacency matrix is generally asymmetric

Degrees in networks

- Degree of a vertex in a graph is the number of edges connected to it
- Denote degree of vertex i by k_i.
- For an undirected graph of n vertices, the degree can be written in terms of the adjacency matrix as

$$K_i = \sum_{j=i}^{n} A_{ij} \tag{3}$$

- Each edge in undirected graph has two ends: if m edges total, then 2m ends of edges
- Number of ends of edges is also equal to the sum of degrees of all vertices!
- We can calculate the mean degree of a vertex in a network

Density of a network

$$\rho = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)} = \frac{c}{n-1}$$

- Density ρ of a graph, is the fraction of max number of edges that are actually present.
- Network for which density ρ tends to a constant as n goes to ∞ is said to be dense
- Most real world networks for social scientists considered sparse

Paths in a network

- Any sequence of vertices such that every consecutive pair of vertices in the sequence is connected by an edge in the network

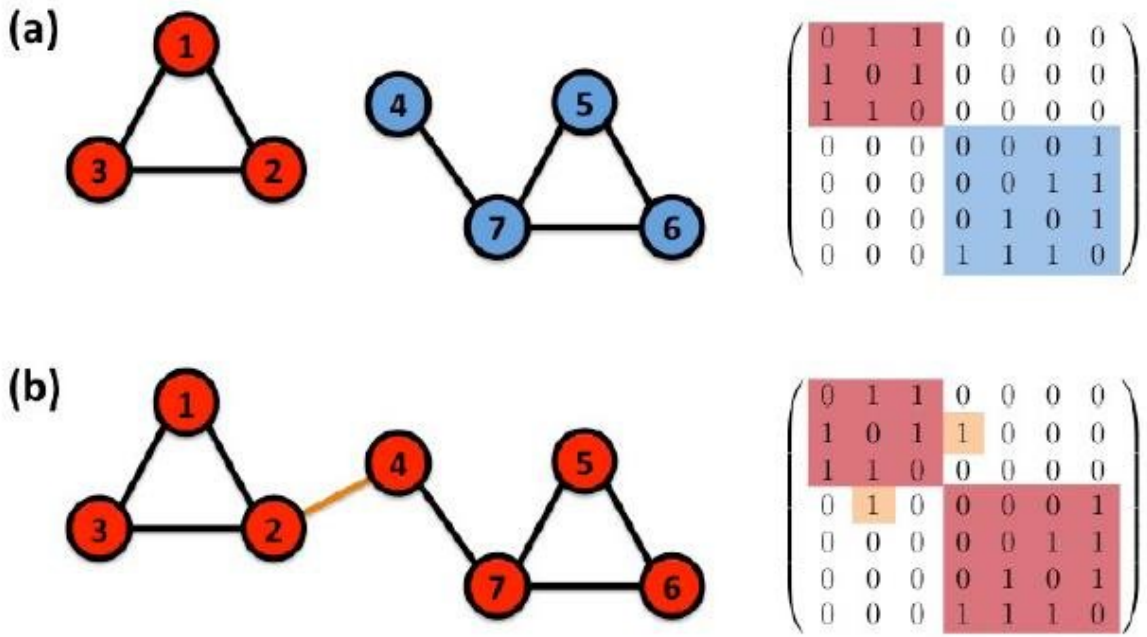
To calculate number of paths of a given length r on a network: for either a directed/undirected simple graph, element $A_{ij} = 1$ if there is an edge from vertex j to vertex i, 0 otherwise. The product $A_{ik}A_{kj}$ is 1 if there is a path of length 2 from j to i via k and 0 otherwise. The total number N_{ij} of paths of length 2 from j to i via any other vertex is:

$$N_{ij}^{(2)} = \sum_{k=1}^{n} A_{ik}A_{kj} = [\mathbf{A}^2]_{ij}$$

where $[\cdots]_{ij}$ denotes the ijth element of a matrix

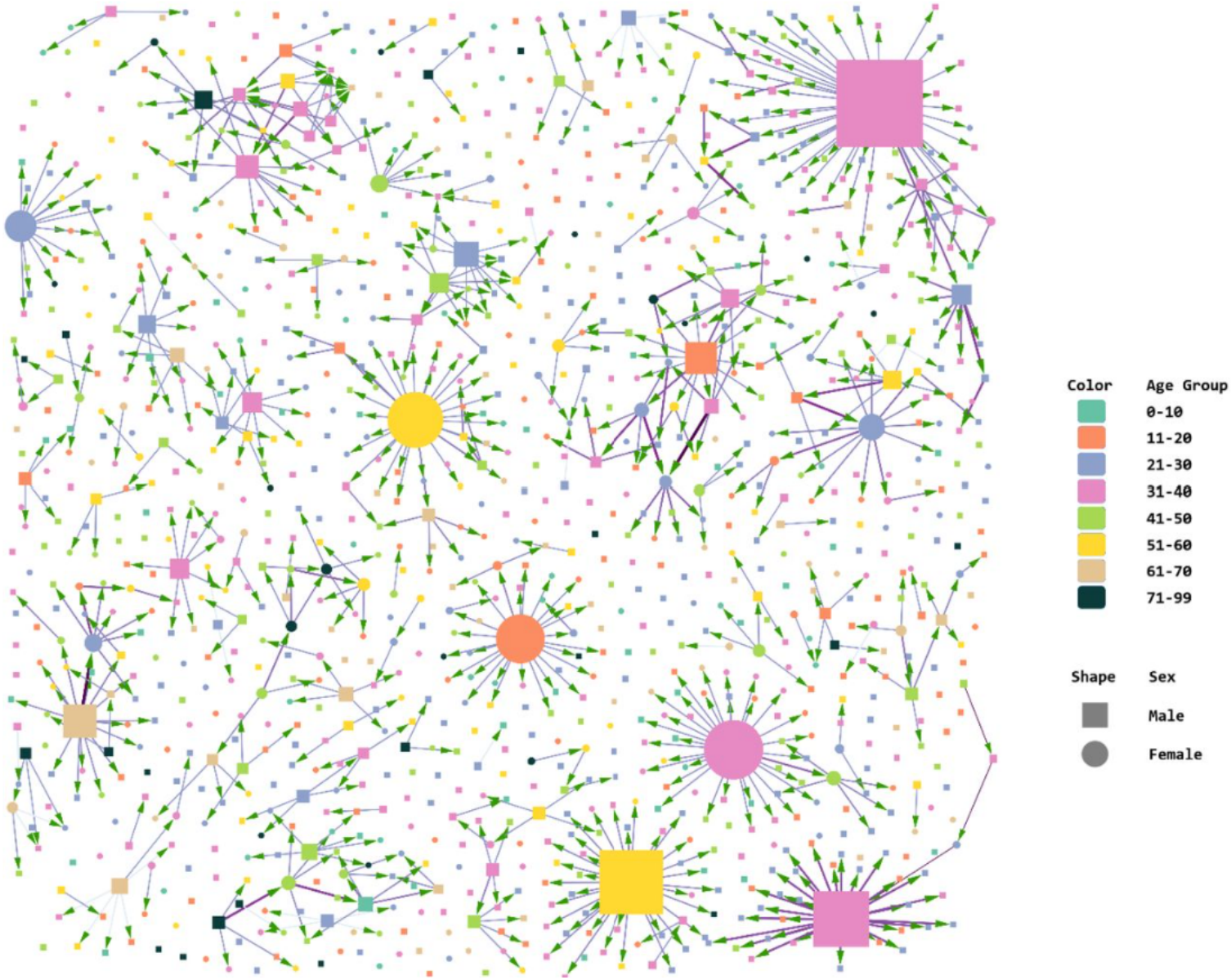
Components in a network

- Subset of vertices of a network such that there exists at least one path from each member of that subset to each other member, such that no other vertex in the network can be added to the subset while preserving this property (also called maximal subsets).
- Adjacency matrix of a network with more than one component can be written in block diagonal form.



Centrality

Which are the most important or central vertices in a network?



Degree Centrality

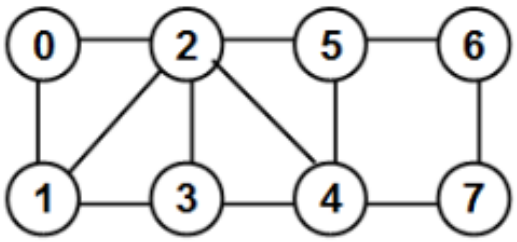
- Number of edges connected to a vertex.
- Important because: reasonable to assume individuals who have connections to many others might have more influences, or number of citations a paper receives from others may be a crude measure for whether a paper is influential
- Awards one ‘centrality point’ for every network neighbor a vertex has

Eigenvector centrality

- Imagine now that network neighbor's are not equally important – but that a vertex's importance in a network is increased by having connections to other vertices that are themselves important.
- Now we award vertices points proportional to the sum of the points of its neighbors.
- Relationships originating from high-scoring nodes contribute more to the score of a node than connections from low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores.

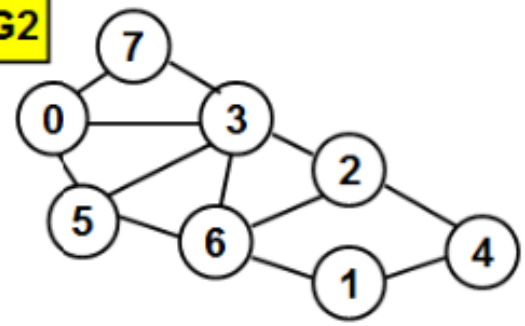
Eigenvector centrality

G1



Degree-based Ordering		EVC-based Ordering	
Vertex	Degree	Vertex	EVC
2	5	2	0.5364
4	4	4	0.4321
1	3	3	0.3974
3	3	1	0.3596
5	3	5	0.3355
0	2	0	0.2681
6	2	7	0.1749
7	2	6	0.1527

G2



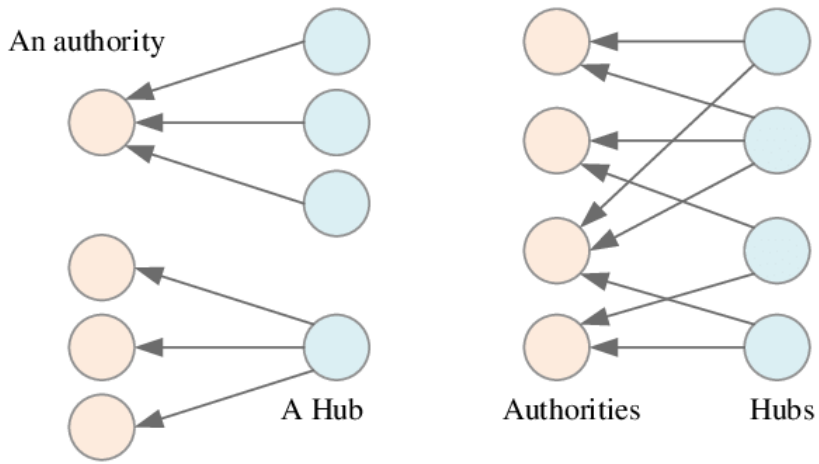
Degree-based Ordering		EVC-based Ordering	
Vertex	Degree	Vertex	EVC
3	5	3	0.5364
6	4	6	0.4321
0	3	5	0.3974
2	3	0	0.3596
5	3	2	0.3355
1	2	7	0.2681
4	2	1	0.1749
7	2	4	0.1527

Mapping of Vertices	
G1	G2
2	3
4	6
3	5
1	0
5	2
0	7
7	1
6	4

Hubs and Authorities

What about when vertices point to others with high centrality?

- E.g. citation network where paper such as review article cites other articles that are canons on the subject, but itself has relatively little information on it.
- Two types of important vertices here: authorities are vertices that contain useful information on a topic of interest and hubs are the vertices that tell us where the best authorities are; an authority may also be a hub and vice versa.



Closeness centrality

What about measuring mean distance between nodes?

- Suppose d_{ij} is the length of a geodesic path from i to j, ie the number of edges along the path. The mean geodesic distance from i to j averaged over all j vertices in the network:

$$l_i = \frac{1}{n} \sum_j d_{ij} \tag{4}$$

- Statistic above takes low values for vertices separated from others by only a short geodesic distance on average – these vertices might have better access to info at other vertices or more direct influence on other vertices
- E.g. in a social network, a person with a lower average distance to others might find their opinions reach others in the community more quickly than someone with a higher average distance

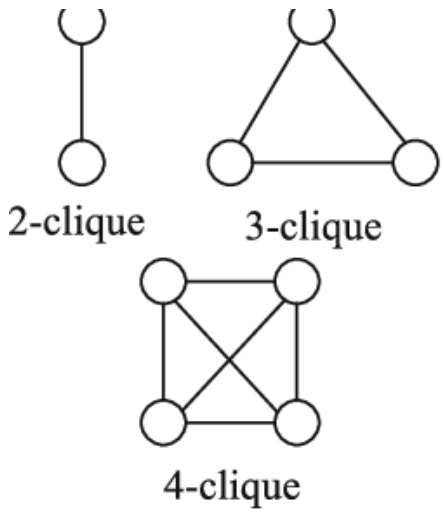
Betweenness centrality

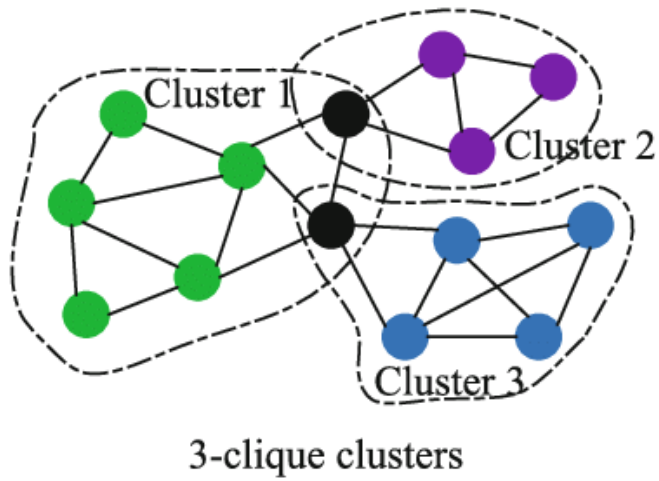
What about measuring how much a node lies in in paths between other nodes?

- Betweenness centrality captures: the extent to which a certain vertex lies on the shortest paths between other vertices. In other words, it helps identify individuals who play a "bridge spanning" role in a network.
- Since messages passing down each geodesic path at same rate (if we assume exchange of message with equal probas per unit time) the number passing through each vertex is proportion of geodesic paths the vertex lies on
- Number of geodesic paths = betweenness centrality
- Intuition: high betweenness nodes might have a lot of influence in a network by virtue of their control over info passing to others

Cliques

- Clique is a maximal subset of vertices in an undirected network such that every member of the set is connected by an edge to every other (maximal such that no other vertex in the network can be added that would still retain the property detailed)
- The existence of cliques in a sparse network indicates a highly cohesive subgroup





3-clique clusters

Component

A component in an undirected network is a maximal subset of vertices such that each is reachable by some path from each of the others, none of which are connected to any other vertex in the graph.

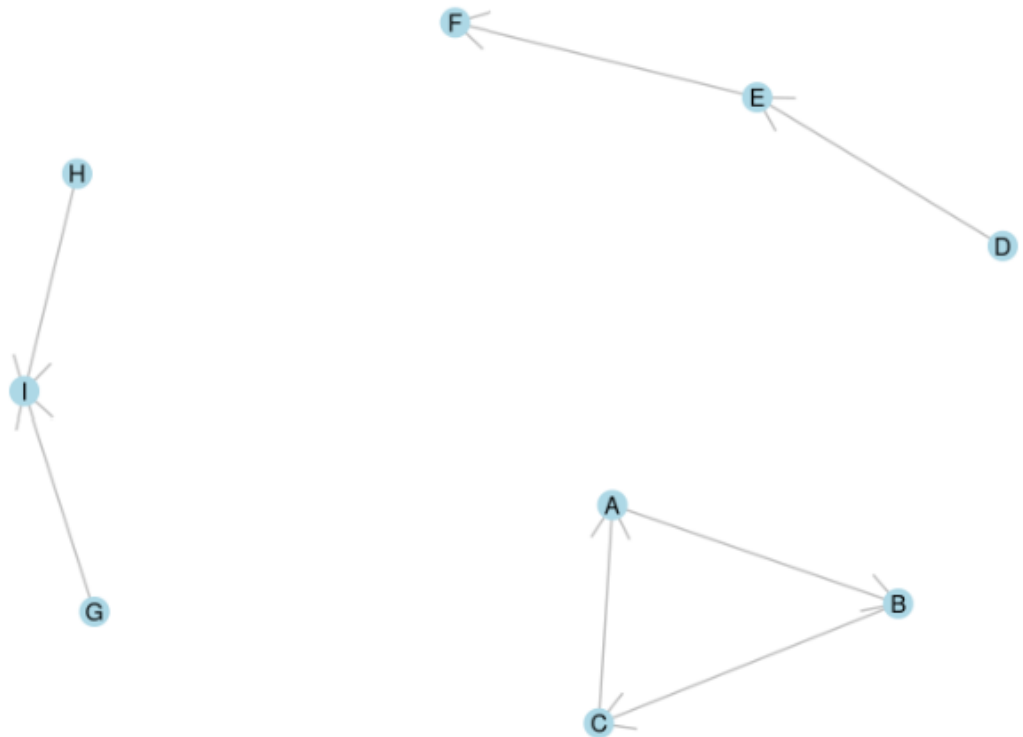
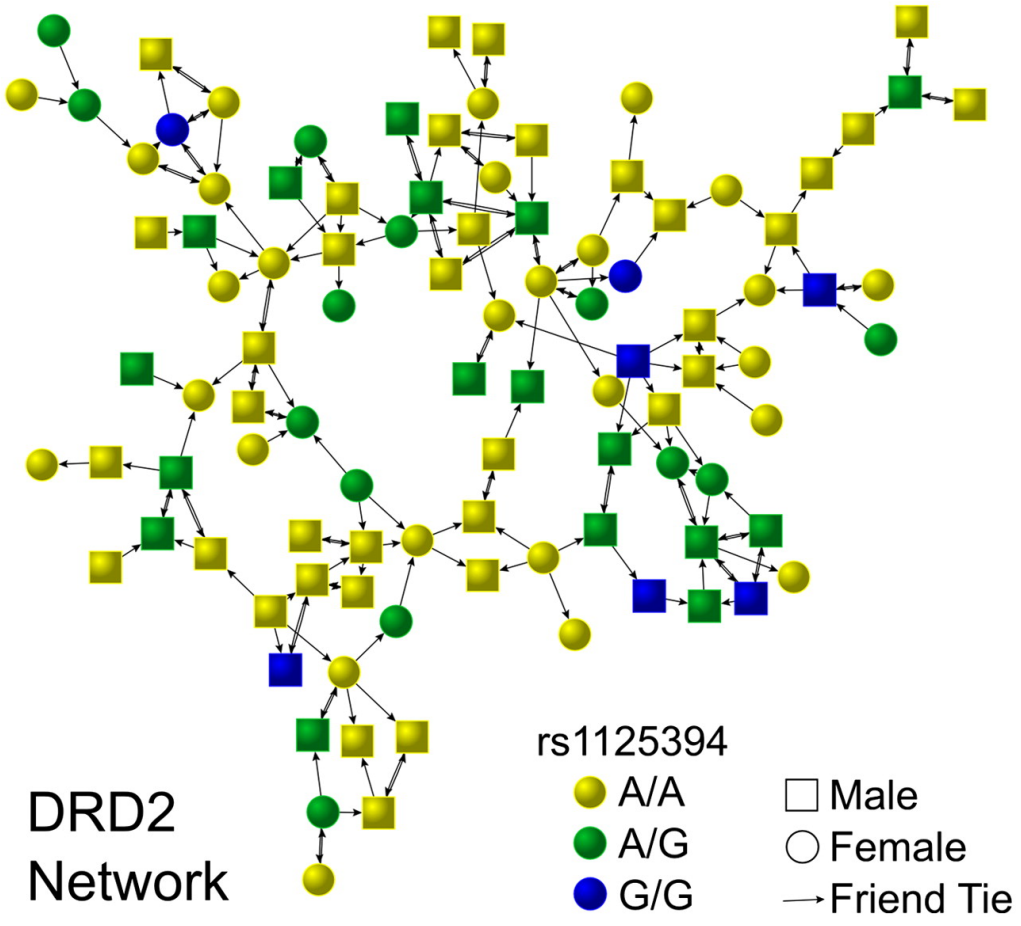


Figure 7.2: A directed graph with three connected components, one strongly connected, one weakly connected and one unilaterally connected

Homophily

Birds of a feather...

- Tendency for nodes to associate with other nodes that are similar – homophily, or assortative mixing
- Fowler, Settle, Christakis. 2011. Correlated genotypes in friendship networks". PNAS



Community structures – cluster

- Discrete clusters of nodese that are densely connected, and loosely connected to other clusters often occur in real networks
- How do we find them, and quantify the degree of community structure?

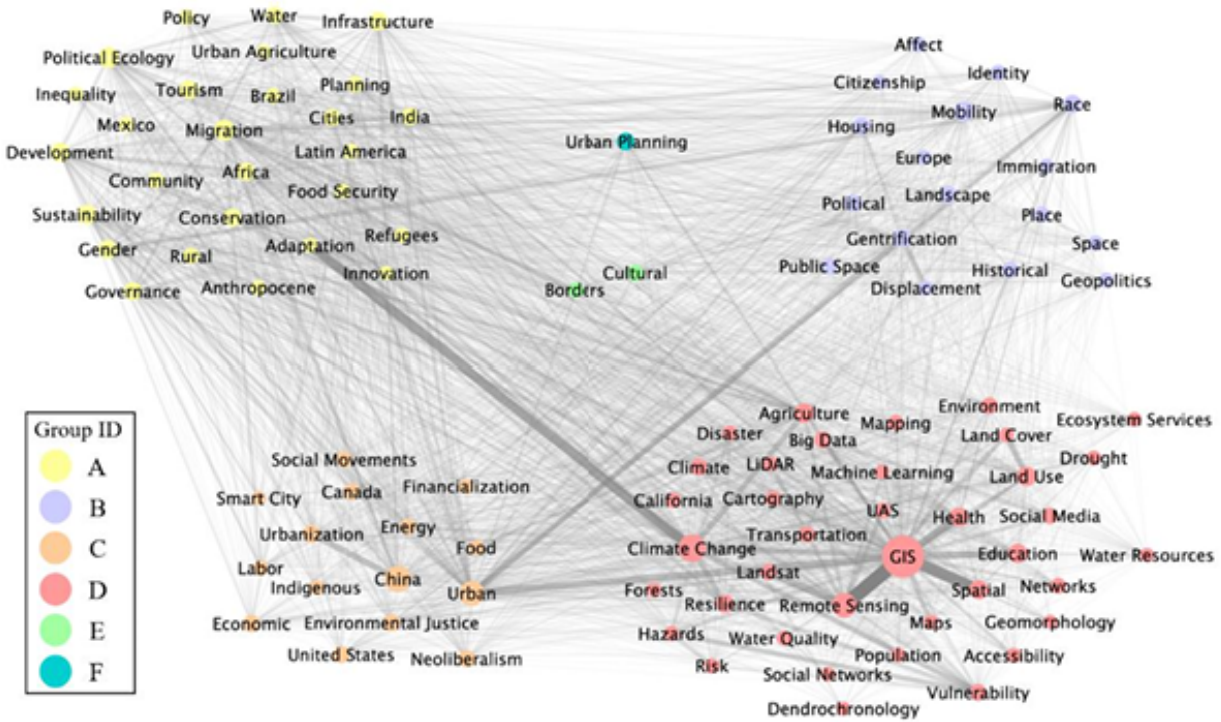


Figure 3. Network clustering with Leading eigenvector algorithm

Modularity

How much clusters are discrete

- Modularity index Q is the proportion of edges that occur within communities, relative to the expected proportion if all edges were placed randomly

$$Q = \frac{1}{2m}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j)$$

m is total # edges in a network, A_{ij} adjacency matrix element, k_i and k_j are degrees of nodes i and j, c_i and c_j refer to communities to which i and j belong, and $\delta(c_i, c_j)$ is the Kroenecker delta function, which equals 1 when $c_i = c_j$ and 0 otherwise.

- There exist a class of modularity-based methods of community detection that find partitions in the network so that Q is maximized
- Since many possible partitions – exhaustive searches aren't usually feasible. Instead modularity-optimization techniques rely on search algorithms that use different approaches (e.g. agglomerative versus divisive) with different strengths/weaknesses
- igraph offers community detection functions