Social Networks Measures

- **Single-node Measures**: Based on some properties of specific nodes
- **Graph-based measures**: Based on the graph-structure of the network

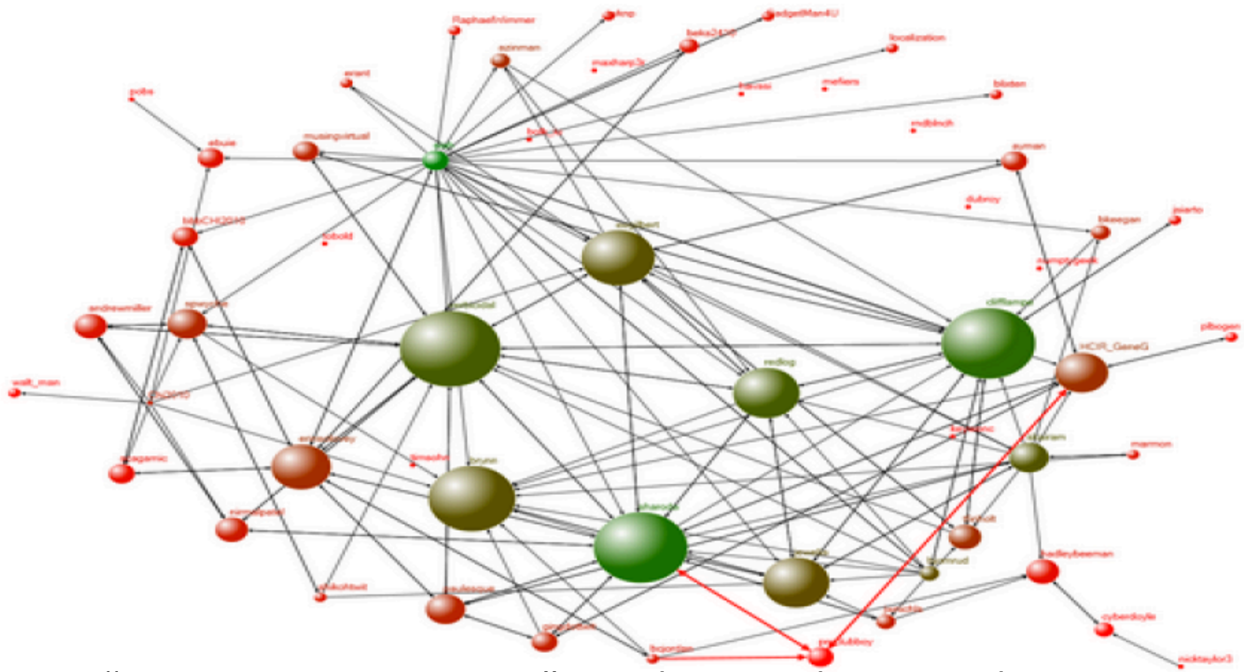
Graph-based measures of social influence

- Previously surveyed measures of influence, such as buzz, applause etc. are based on surface metrics (e.g. number of retweets, etc): graph-based measures go more in-depth.
- Objective here: model the social network as a graph
- Use graph-based methods/algorithms to identify "relevant players" in the network
  - Relevant players = more influential, according to some criterion
- Use graph-based methods to identify communities (community detection)
- Use graph-based methods to analyze the "spread" of information

Graph-based measures of social influence

- **Use graph-based methods/algorithms to identify "relevant players" in the network**
  - Relevant players = more influential, according to some criterion
- Use graph-based methods to identify global network properties and communities (community detection)
- Use graph-based methods to analyze the "spread" of information

Modeling a Social Network as a graph



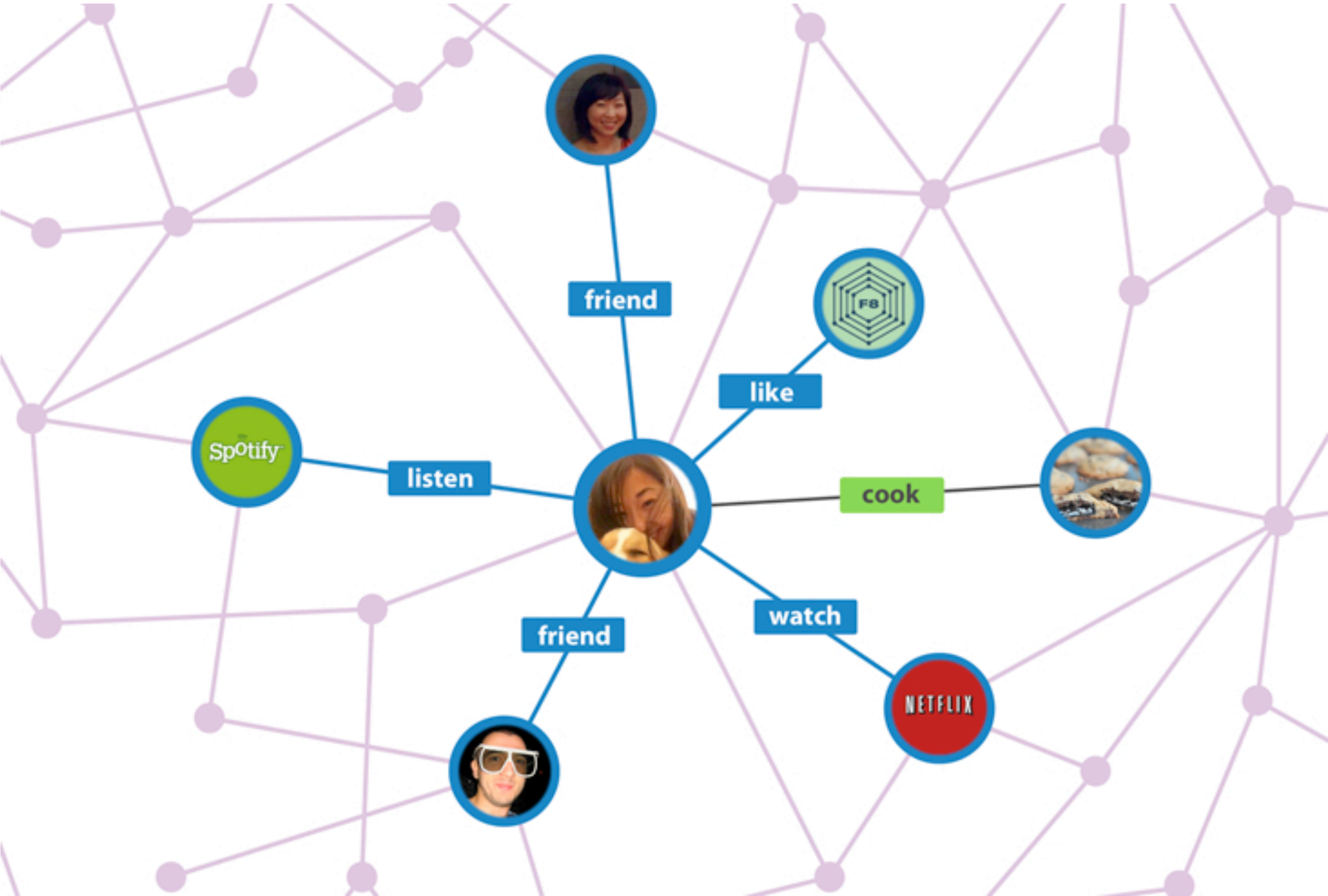
NODE= “actor, vertices, points” i.e. the social entity who participates in a certain network

EDGE= “connection, edges, arcs, lines, ties” is defined by some type of relationship between these actors (e.g. friendship, reply/re-tweet, partnership between connected companies..)

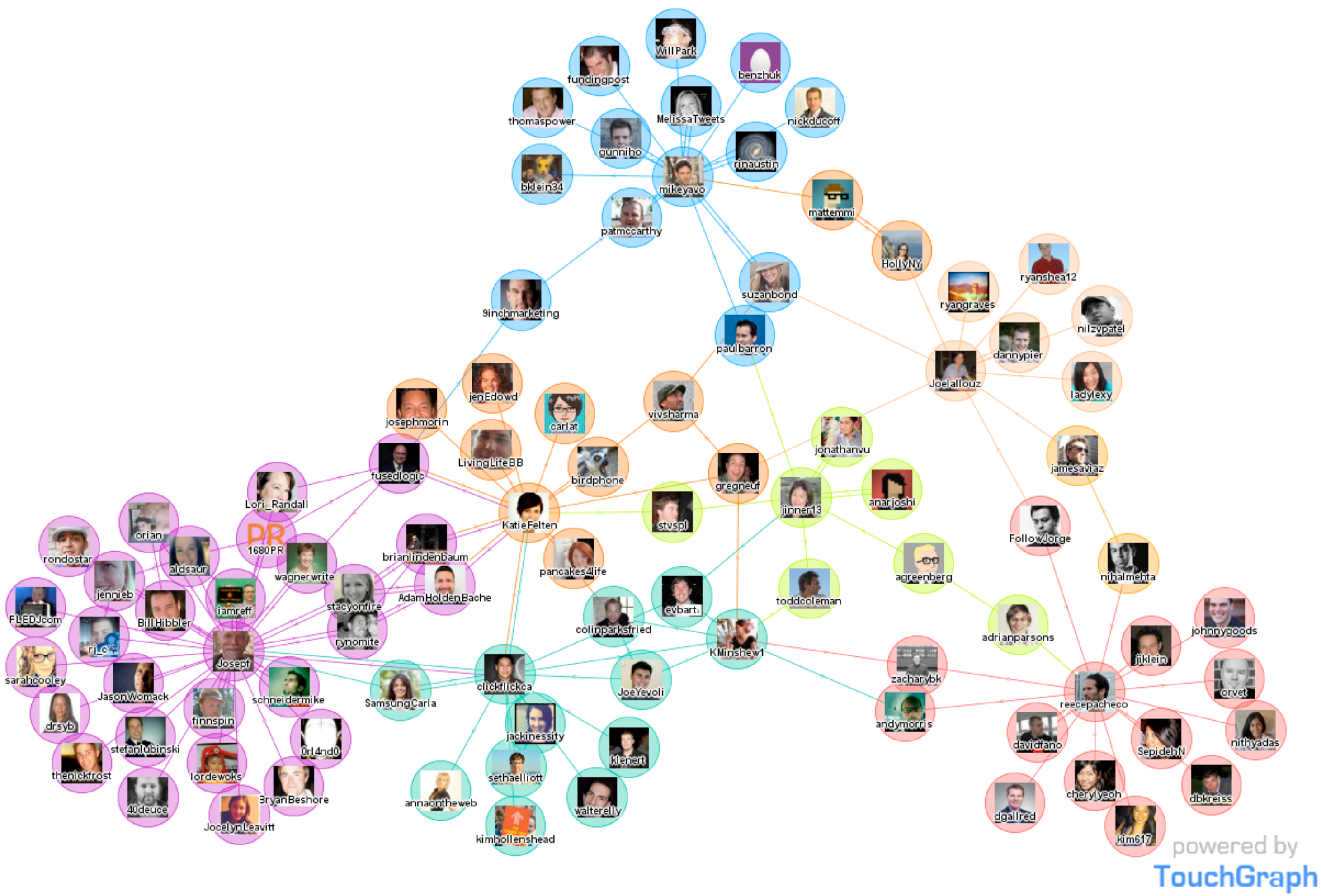
SN = graph

- A network can then be represented as a graph data structure
- We can apply a variety of measures and analysis to the graph representing a given SN
- Edges in a SN can be **directed or undirected** (e.g. friendship, co-authorship are usually undirected, emails are directed)

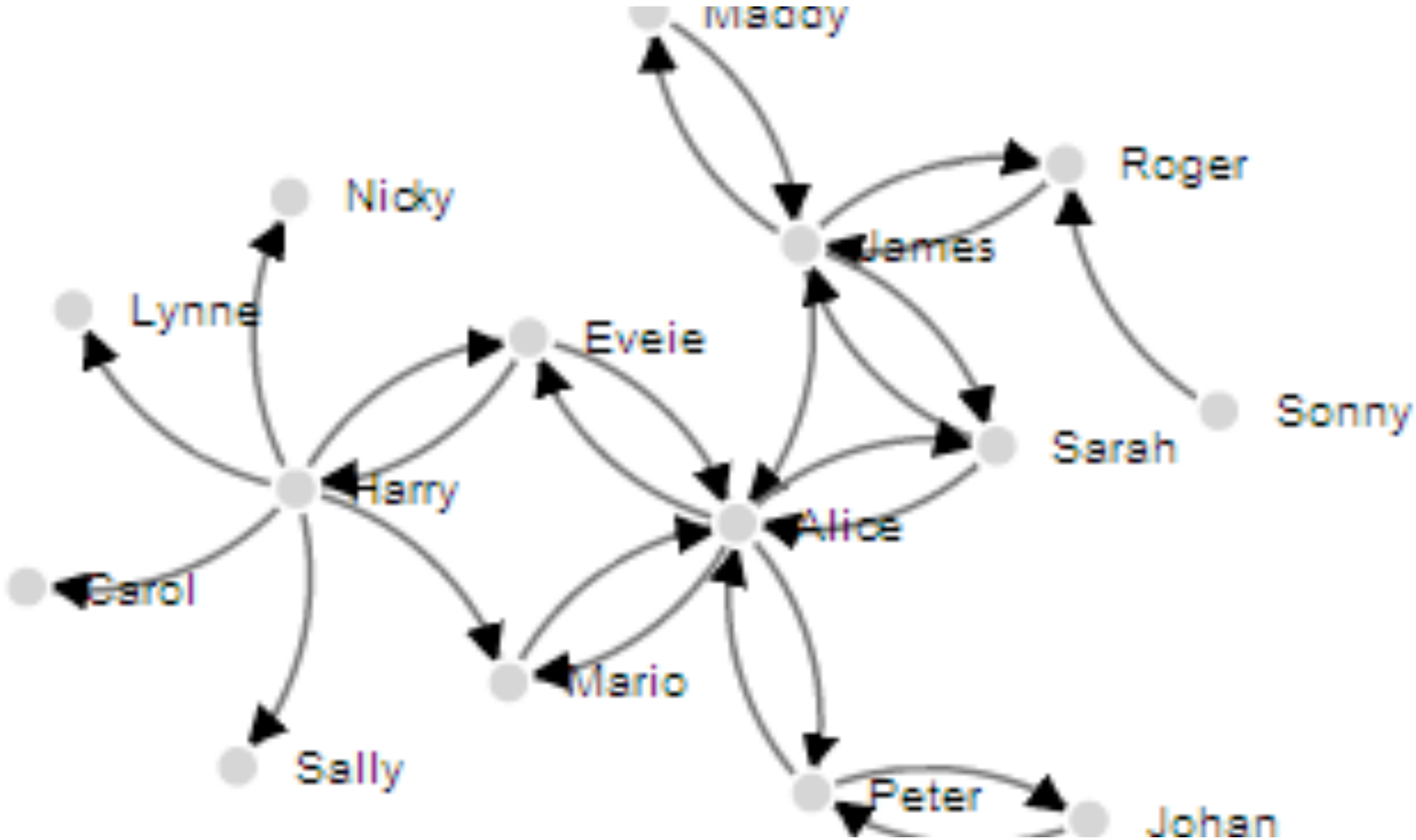
What is the meaning of edges?



Facebook in undirected (friendship is mutual)



Twitter is a directed graph (friendship is not necessarily bidirectional)

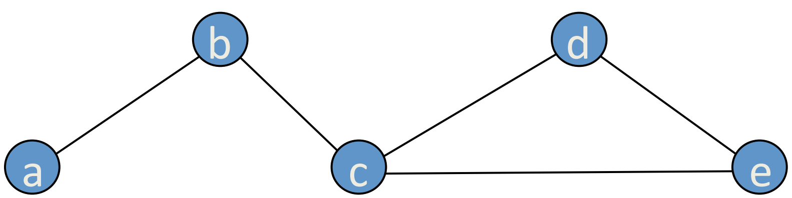


Social Network as a graph

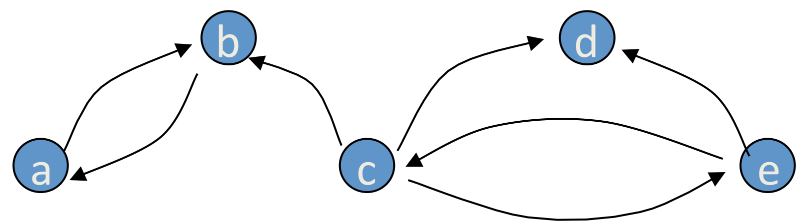
In general, a relation can be:

Binary or Valued

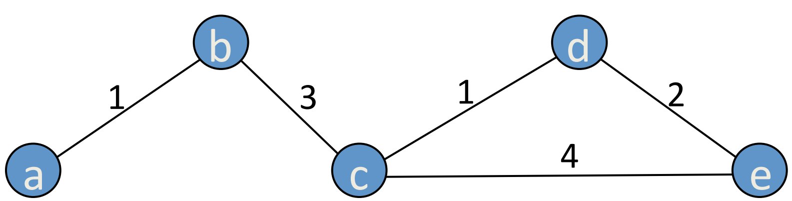
Directed or Undirected



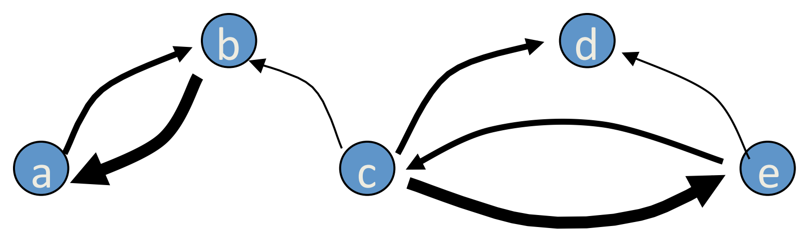
Undirected, binary



Directed, binary



Undirected, Valued

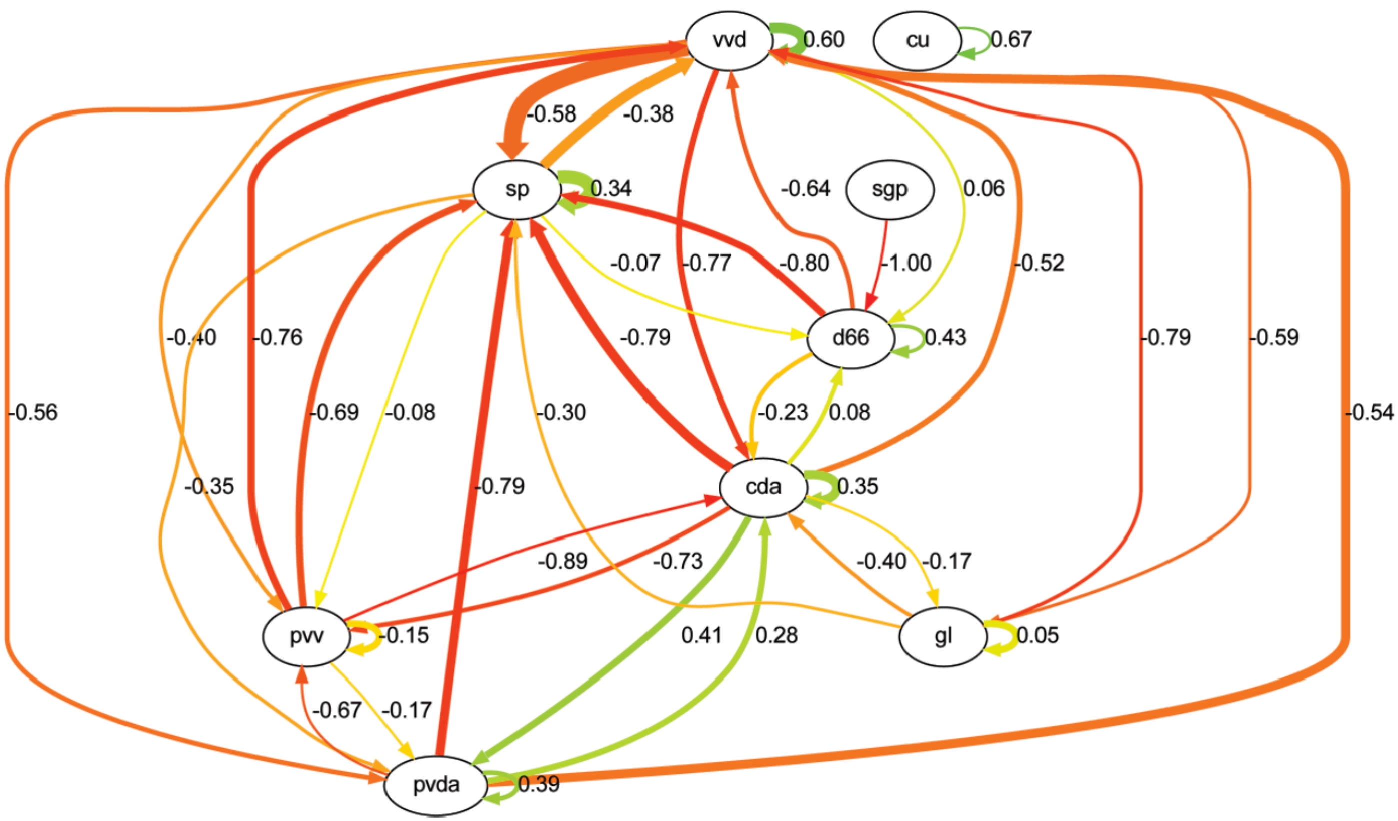


Directed, Valued

Example of directed, valued: Sentiment relations among parties during a political campaign.

Color: positive (green) negative (red).

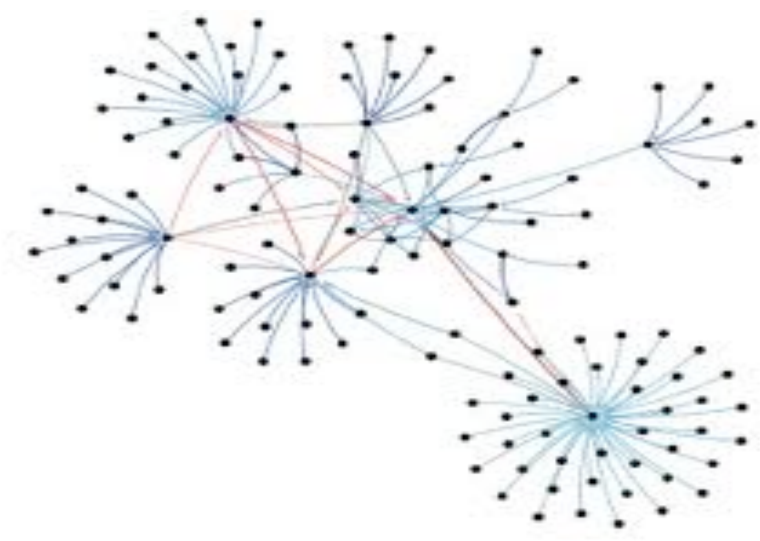
Intensity (thikness of edges): related to number of mutual references



Graph-based measures of social influence: key players

Key players

- Using graph theory, we can identify **key players** in a social network
- Key players are nodes (or actors, or vertexes) with some measurable connectivity property
- Two important concepts in a network are the ideas of **centrality** and **prestige** of an actor.
- Centrality more suited for undirected, prestige for directed



Measuring Networks: Centrality

Centrality refers to (one dimension of) *location,* identifying *where* an actor resides in a network. Mostly used for **undirected** networks.

- For example, we can compare actors at the edge of the network to actors at the center.

- In general, this is a way to formalize intuitive notions about the distinction between *insiders and outsiders*.

Measuring Networks: **Centrality**

Conceptually, centrality is fairly straight forward: we want to identify which nodes are in the 'center' of the network. In practice, identifying exactly what we mean by 'center' is somewhat complicated.

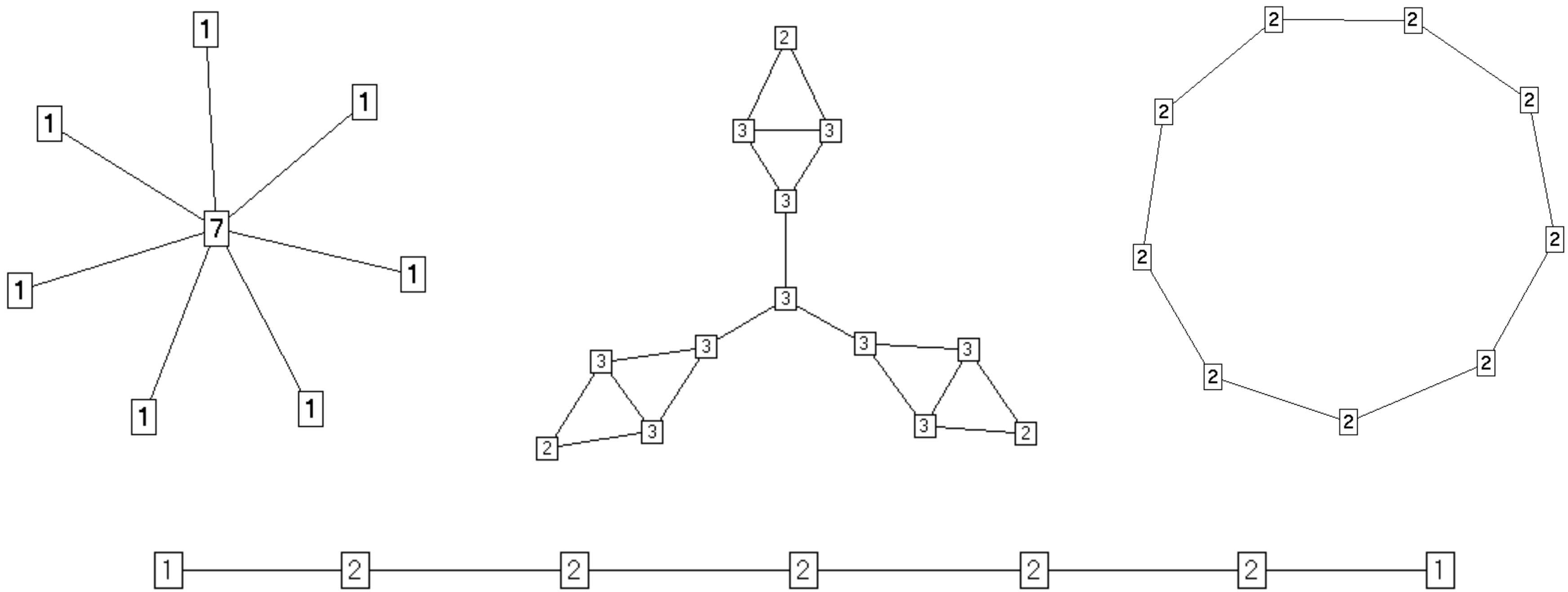
Three standard centrality measures capture a wide range of "importance" in a network:

- Degree
- Closeness
- Betweenness

Measuring Networks: **Centrality**

1.Centrality Degree

The most intuitive notion of centrality focuses on ***degree***. Degree is the number of ties, and the actor with the most ties is the most important:



$$C_D = d(n_i) = X_{i+} = \sum_j X_{ij}$$

Measuring Networks: Closeness **Centrality**

A second measure of centrality is **closeness centrality**. An actor is considered important if he/she is relatively close to all other actors.

Closeness is based on the inverse of the distance of each actor to every other actor in the network.

Closeness Centrality:

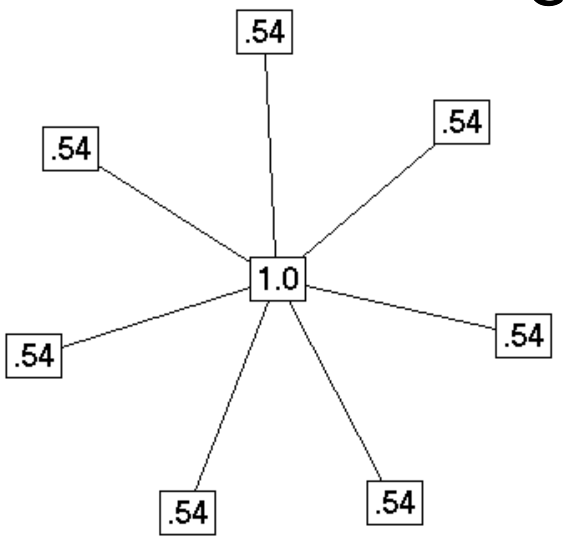
$$C_c(n_i) = \left[\sum_{j=1}^{g} d(n_i, n_j) \right]^{-1}$$

Normalized Closeness Centrality

$$C'_C(n_i) = (C_C(n_i))(g-1)$$

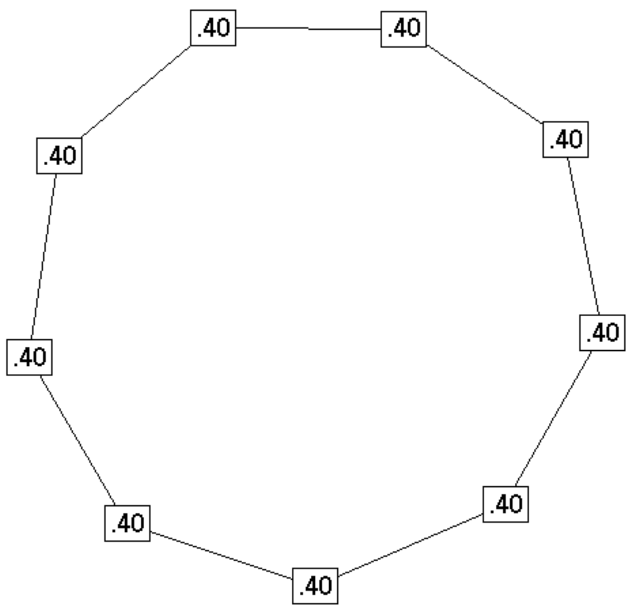
Measuring Networks: Centrality

Closeness Centrality in the examples



Distance	Closeness	normalized
0 1 1 1 1 1 1 1	.143	1.00
1 0 2 2 2 2 2 2	.077	.538
1 2 0 2 2 2 2 2	.077	.538
1 2 2 0 2 2 2 2	.077	.538
1 2 2 2 0 2 2 2	.077	.538
1 2 2 2 2 0 2 2	.077	.538
1 2 2 2 2 2 0 2	.077	.538
1 2 2 2 2 2 2 0	.077	.538

$$C_c(n_i) = \left[\sum_{j=1}^{g} d(n_i, n_j)\right]^{-1}$$



Distance	Closeness	normalized
0 1 2 3 4 4 3 2 1	.050	.400
1 0 1 2 3 4 4 3 2	.050	.400
2 1 0 1 2 3 4 4 3	.050	.400
3 2 1 0 1 2 3 4 4	.050	.400
4 3 2 1 0 1 2 3 4	.050	.400
4 4 3 2 1 0 1 2 3	.050	.400
3 4 4 3 2 1 0 1 2	.050	.400
2 3 4 4 3 2 1 0 1	.050	.400
1 2 3 4 4 3 2 1 0	.050	.400

Measuring Networks: **Centrality**

Closeness Centrality in the examples

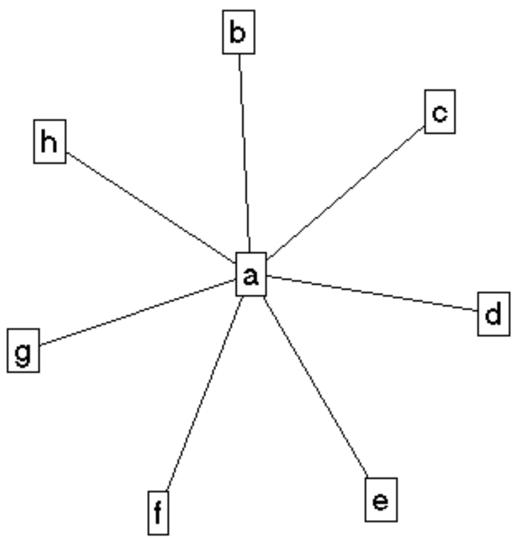
Distance	Closeness	normalized
0 1 2 3 4 5 6	.048	.286
1 0 1 2 3 4 5	.063	.375
2 1 0 1 2 3 4	.077	.462
3 2 1 0 1 2 3	.083	.500
4 3 2 1 0 1 2	.077	.462
5 4 3 2 1 0 1	.063	.375
6 5 4 3 2 1 0	.048	.286

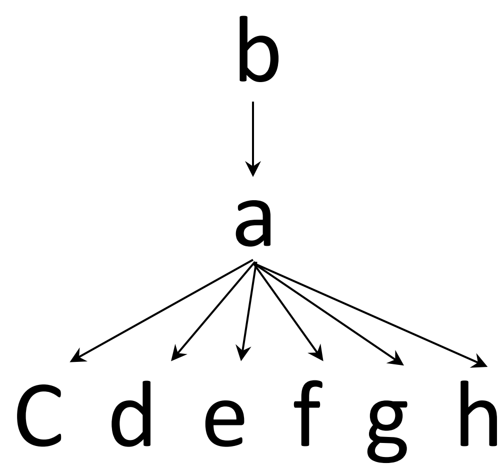
$$C_c(n_i) = \left[\sum_{j=1}^{g} d(n_i, n_j) \right]^{-1}$$

Measuring Networks: **Betweenness Centrality**

Model based on communication flow: A person who lies on communication paths can control communication flow, and is thus important.

Betweenness centrality counts the number of <u>geodesic</u> paths between *i* and *k* **that actor *j* resides on**. Geodesics are defined as the shortest path between points





Measuring Networks: **Betweenness Centrality**

$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i) / g_{jk}$$

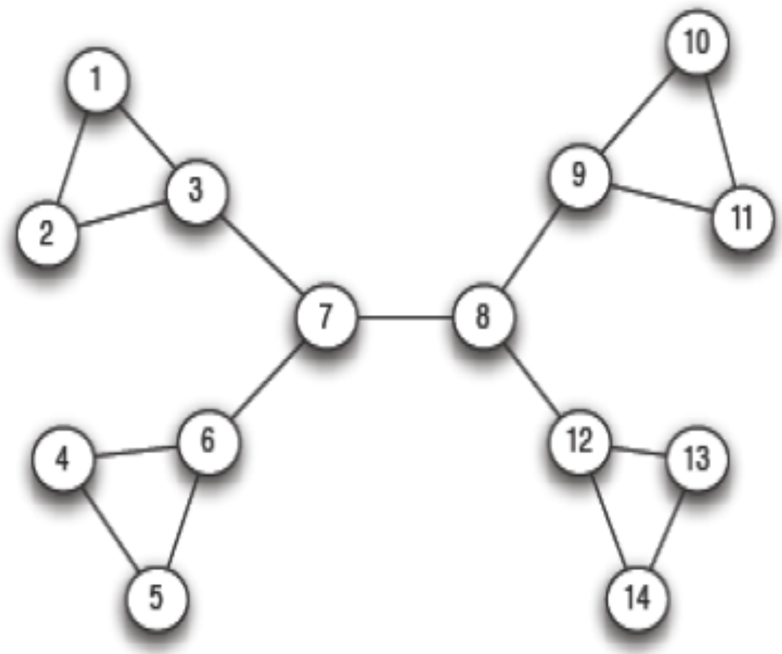
Where g_{jk} = the number of geodesics (shortest) connecting *jk*, and g_{jk} (ni)= the number of such paths that node *i* is on (count also in the start-end nodes of the path).

Can also compute **edge betweenness** in the very same way

How to Compute Betweenness?

Example (edge betweennes)

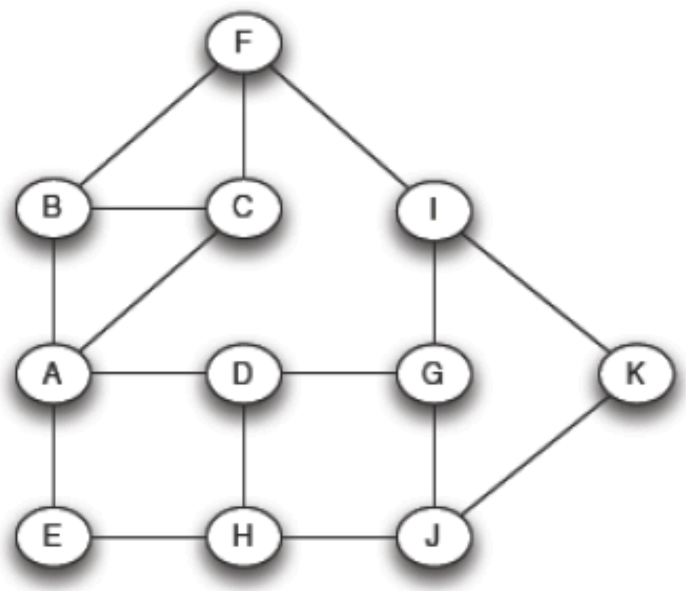
- Edge **7-8**: each pair of nodes between [1-7] and [8-14]; each pair with traffic = 1; total 7 x 7 = 49
- Edge **3-7**: each pair of nodes between [1-3] and [4-14]; each pair with traffic = 1; total 3 x 11 = 33
- Edge **1-3**: each pair of nodes between [1] and [3-14] (not node 2); each pair with traffic = 1; total 1 x 12 = 12 (similar for edges 2-3, 4-6, 5-6, 9-10, 9-11, 12-13, and 12-14)
- Edge **1-2**: each pair of nodes between [1] and [2] (no other); each pair with traffic = 1; total 1 x 1 = 1 (similar for edges 4-5, 10-11, and 13-14)



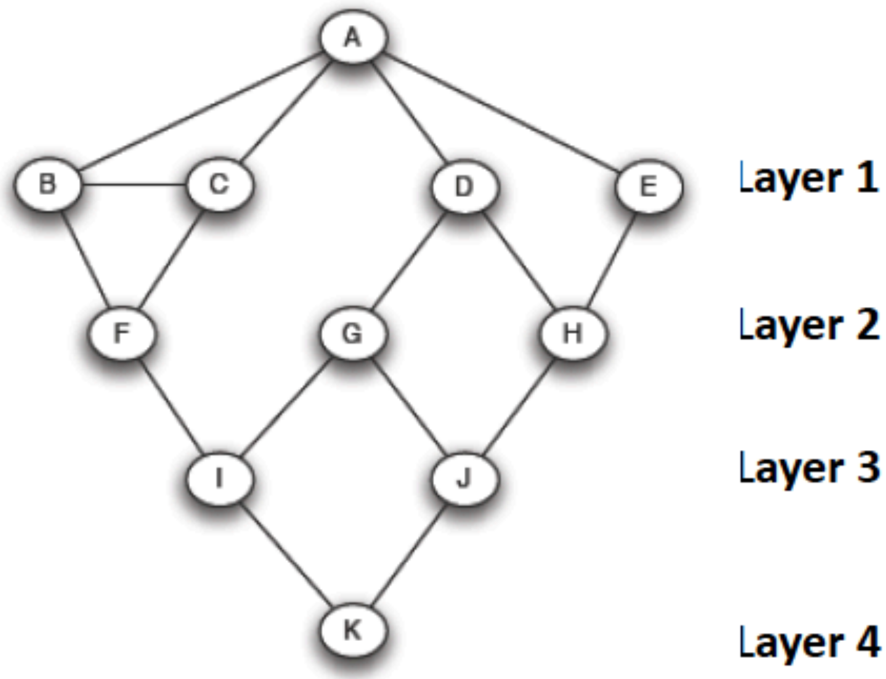
Method (to avoid computing shortest paths for all nodes /edges

- For each node A:
  1. BFS starting at A
  2. Count the number of shortest paths from A to each other node
  3. Based on this number, determine the amount of flow from A to all other nodes

Step 1 (for node A): BFS starting from A



(a) *A sample network*

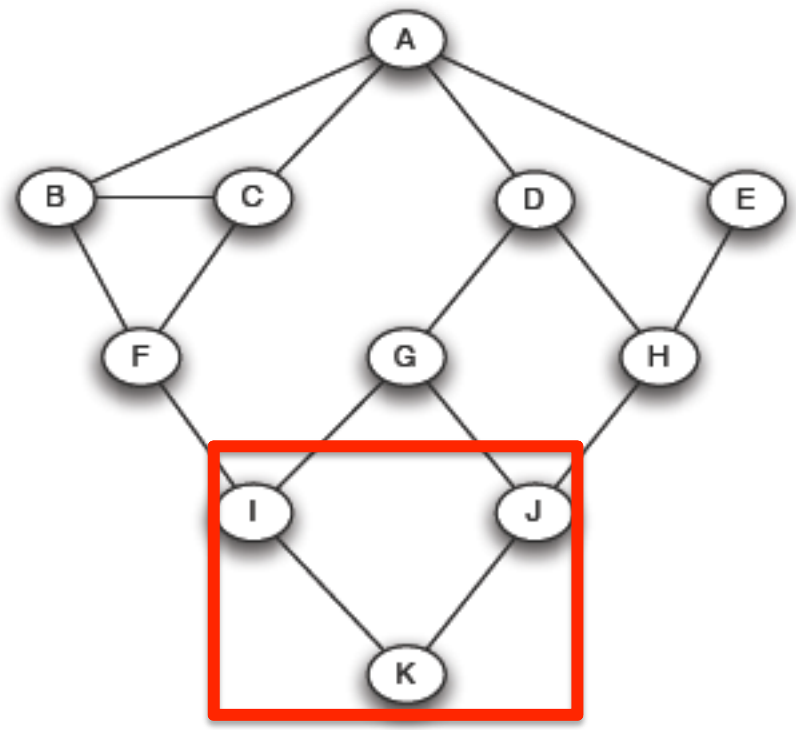


(b) *Breadth-first search starting at node A*

Step 2

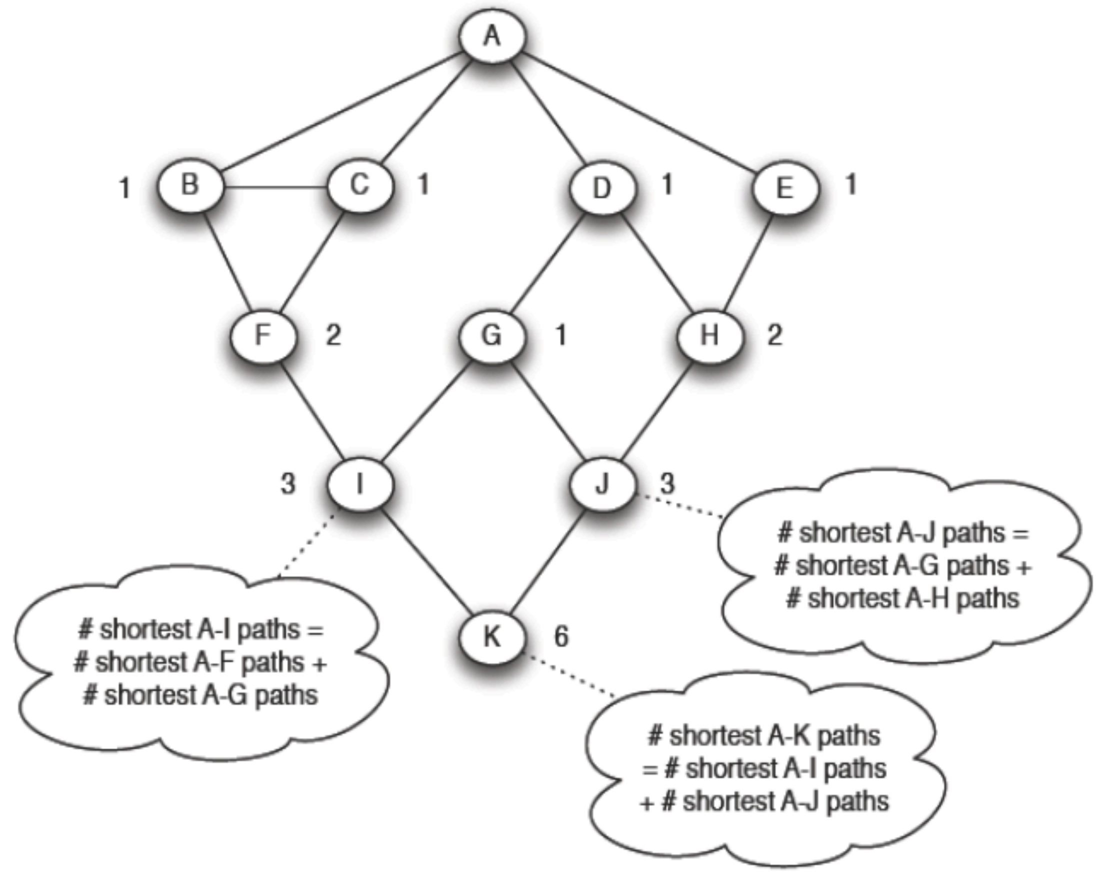
Recursively find paths

- I and J are above K
- All shortest-paths from A to K must take their last step through either I or J
- To be a shortest path to K, a path must first be a shortest path to one of I or J, and then take this last step to K
- The number of shortest paths from A to K is the number of shortest paths from A to I, plus the number of shortest paths from A to J



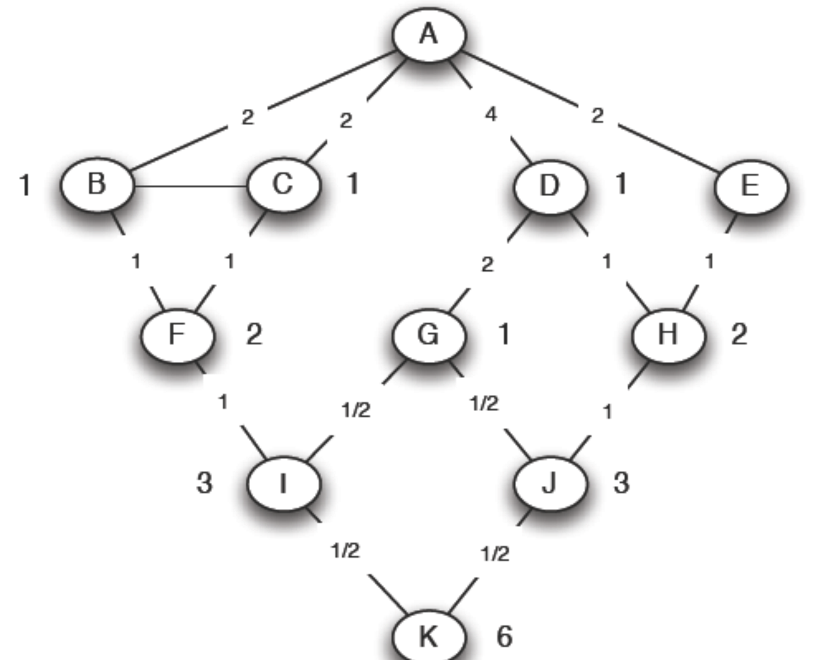
Step 2: tag nodes

- Each node in the first layer has only 1 shortest path from A
- The number of shortest paths to each other node is the sum of the number of shortest paths to all nodes directly above it
- Avoid finding the shortest paths themselves!

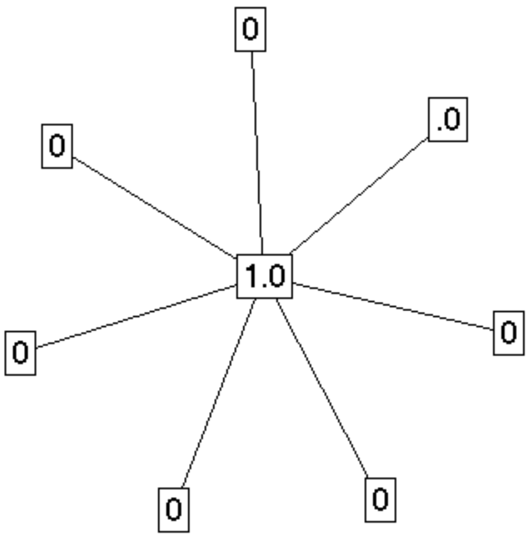


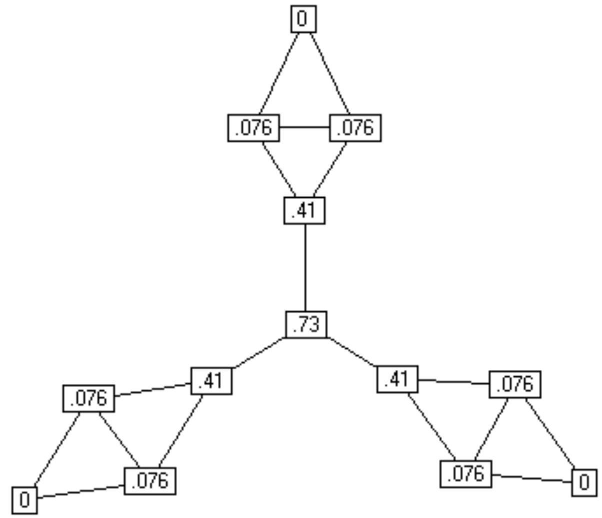
Step 3: compute flow from A

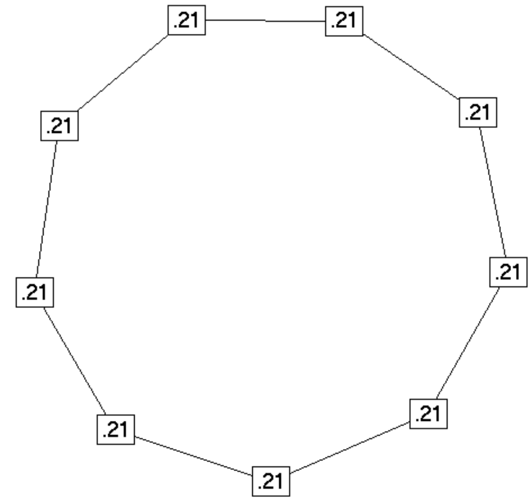
- How the flow from A to all other nodes spreads out across the edges?
- Working up from the lowest layers
  - **1** unit of flow arrives at K and an equal number of the shortest paths from A to K come through nodes I and J => **1/2**-unit of flow on each of these edges
  - **3/2** units of flow arriving at I (**1** unit destined for I plus the **1/2** passing through to K). These **3/2** units are divided in proportion **2** to **1** between F and G => **1** unit to F and **1/2** to G

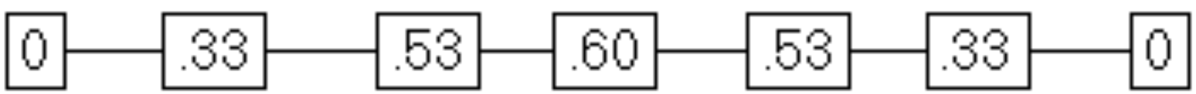


Other examples (node betweenness)



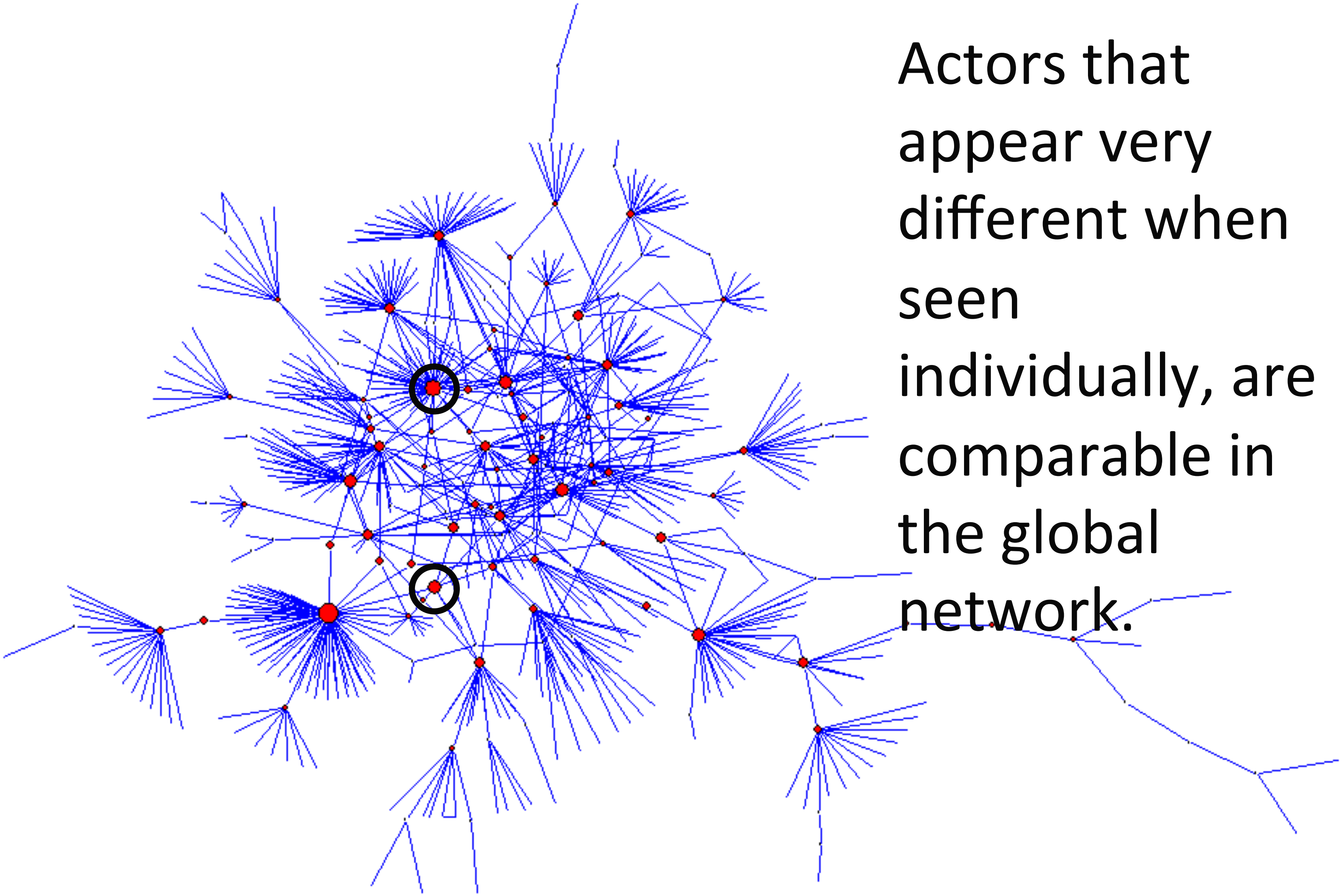






$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i) / g_{jk}$$

Measuring Networks: **Betweenness Centrality**

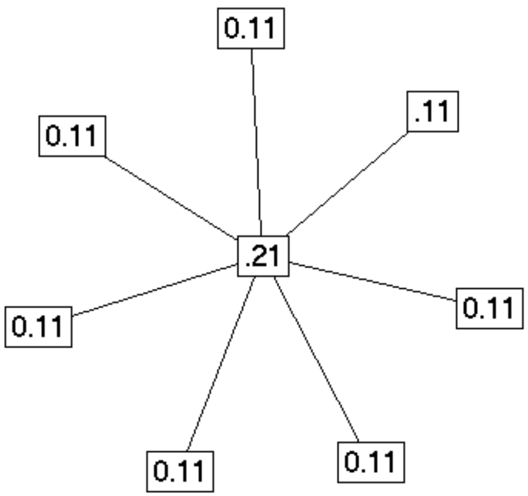


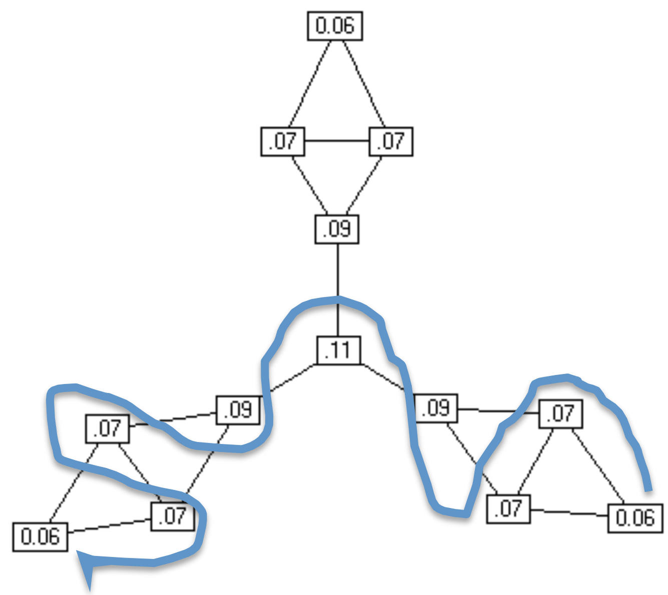
Actors that appear very different when seen individually, are comparable in the global network.

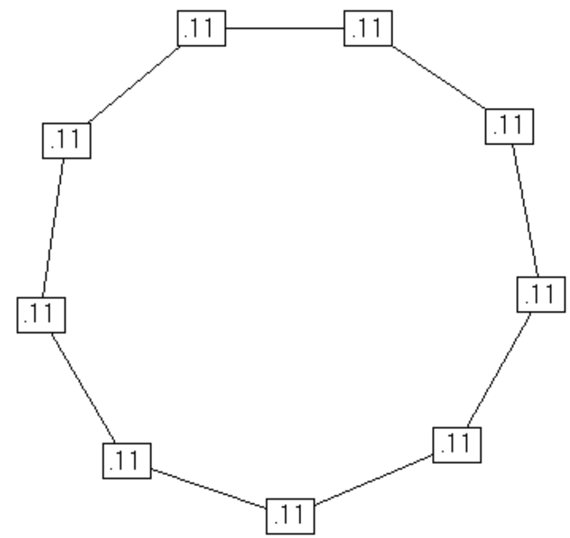
(Node size proportional to betweenness centrality)

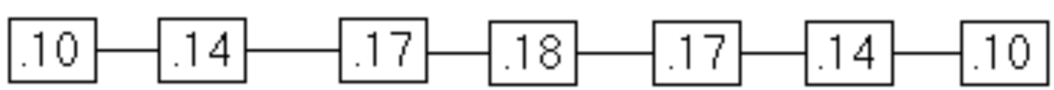
Measuring Networks: **Information Centrality**

It is quite likely that information can flow through paths *other* than the geodesic. The Information Centrality score uses **all paths** in the network, and weights them based on their length.









Computationall very demanding for large graphs!!

Measuring Networks: **Prestige**

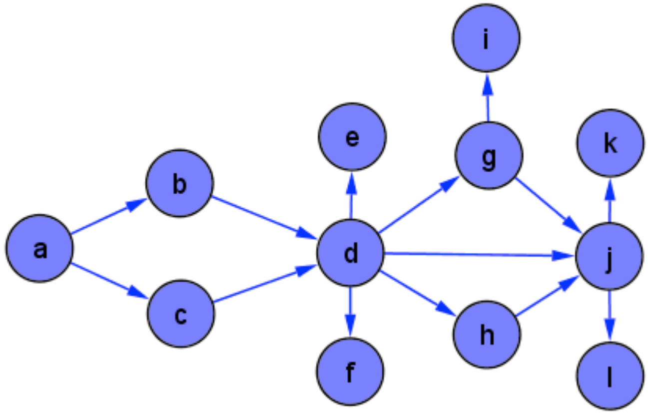
- The term prestige is used for **directed networks** since for this measure the direction is an important property of the relation.
- In this case we can define two different types of prestige:
  – one for outgoing arcs (measures of influence),
  – one for incoming arcs (measures of support).
- Examples:
  – An actor has high influence, if he/she gives hints to several other actors (e.g. in Yahoo! Answers).
  – An actor has high support, if a lot of people vote for him/her (many “likes”)

Measures of prestige

- Influence and support
- Influence domain
- Hubs and authorities
- Brockers

Measuring prestige: influence and support

- **Influence and support**: According to the direction/meaning of a relation, in and outdegree represent support or influence. (e.g., likes, votes for,. . .).

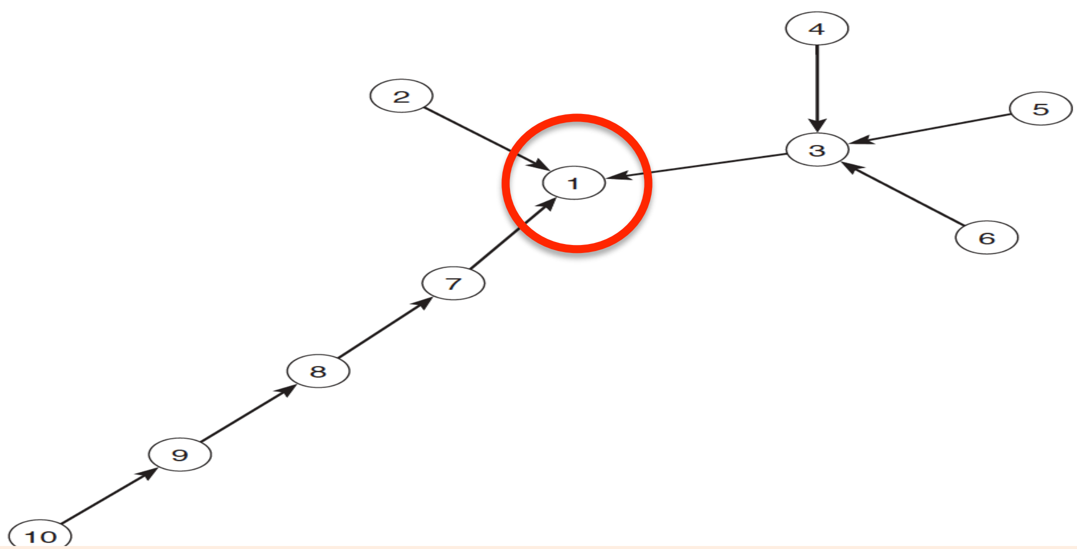


$$InDegree(x) = \#\ \ incoming\ \ edges(x)$$

$$InDegree^{N}(x) = \frac{\#\ \ incoming\ \ edges(x)}{\max\limits_{y \in network}(InDegree^{N}(y))}$$

Measuring prestige: influence domain

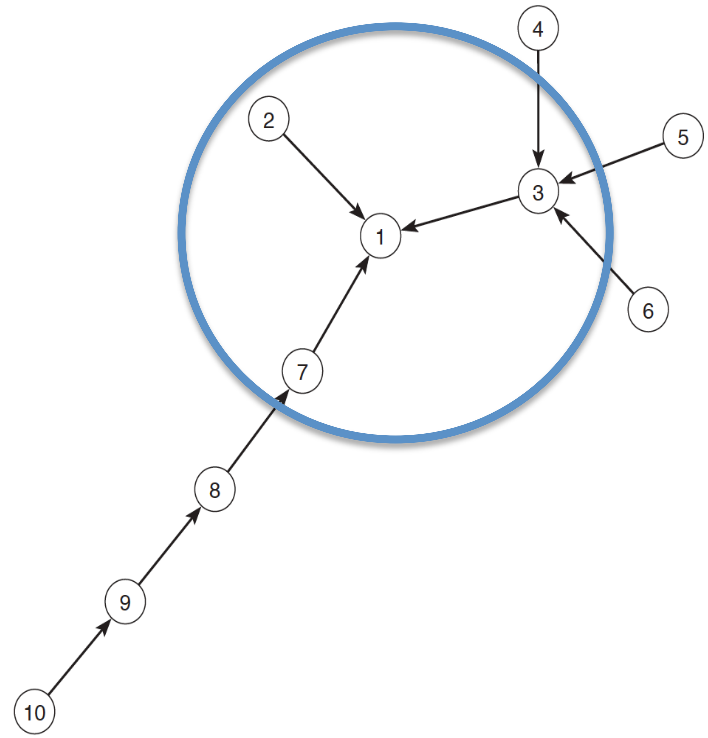
- **Influence domain**: The influence domain of an actor (node) in a directed network is the number (or proportion) of all other nodes *which are connected by a path to this node*.



All other actors are in influence domain of actor 1: Prest(1)=10/10=1.

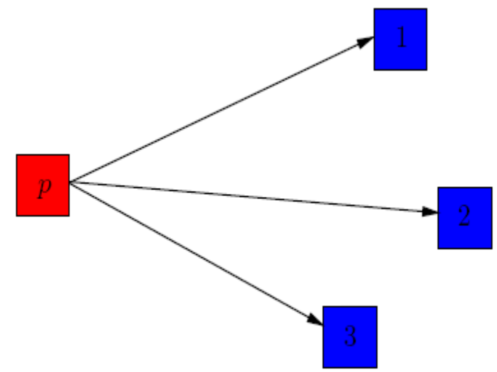
Limits of Influence domain

- Influence domain has an important limitation: *all the nodes contribute equally to influence*.
- Choices by actors 2, 3, and 7 are more important to person 1 than **indirect** choices by 4, 5, 6, and 8. Individuals 9 and 10 contribute even less to the prestige of 1.

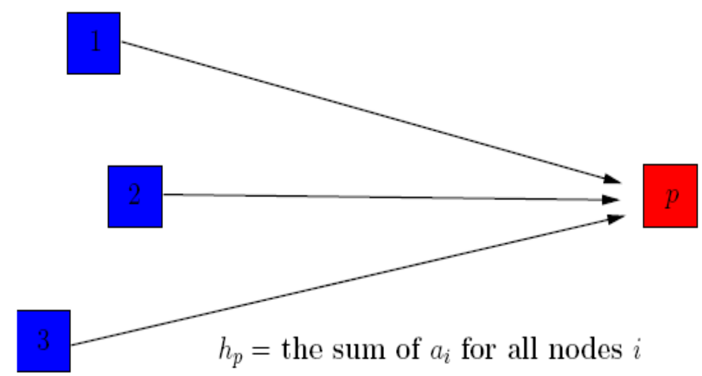


Measuring prestige: Hubs and Authorities, Page Rank

- Hubness is a good measure of influence
- Authority is a good measure of support
- **Kleinberg's algorithm** (HITS) to compute authority and hubness degree of nodes, same as for link analysis
- **Page Rank** is a good measure of support
- HITS, Page Rank: see previous lessons

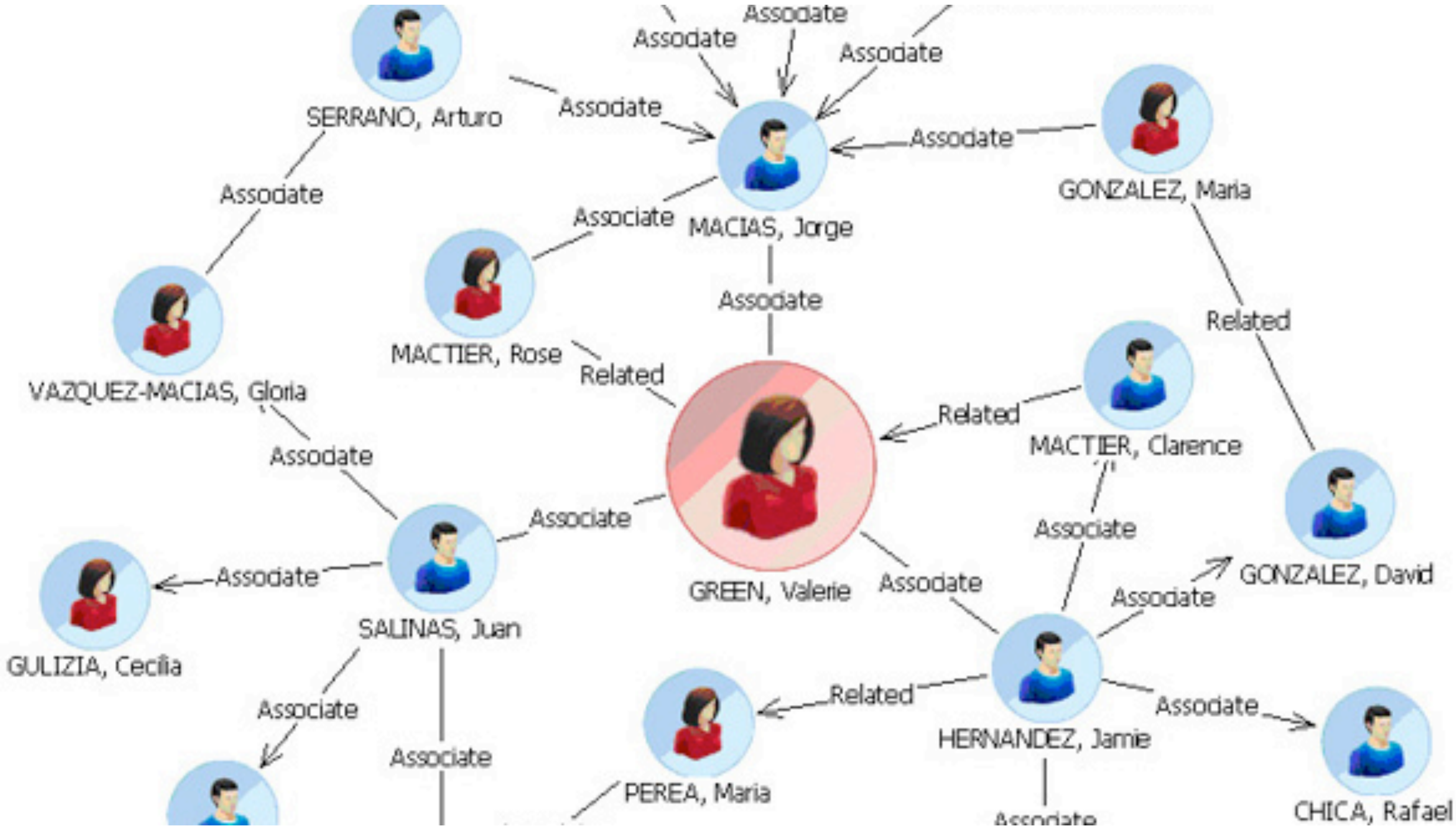


a_p = the sum of h_i for all nodes i pointing to p



h_p = the sum of a_i for all nodes i pointed to by p

Example



If Mrs. Green is the boss, employees referring directly to her are more important

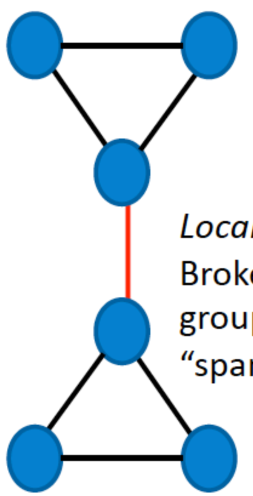
High-level scheme

- Hubs and authorities can be computed in sub-communities, i.e. on parts of a large social network graph, or on the entire graph
- Initial step (create a sub-graph):
  1. Extract from the graph a base set of users that *could* be good hubs or authorities (e.g. with many incoming or outgoing links).
  2. From these, identify a small set of top hub and authority users;

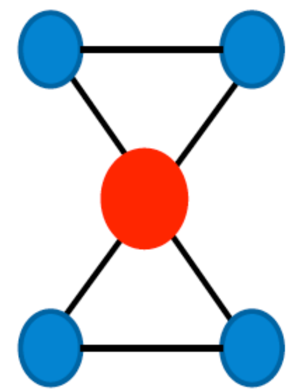
  →using the iterative HITS algorithm.

Measuring prestige: Brockers (bridges)

- Network brokerage: Links between different groups/ communites (very similar to betweenness)



Local bridging ties:
Brokerage of disconnected groups:
“spanning structural holes”

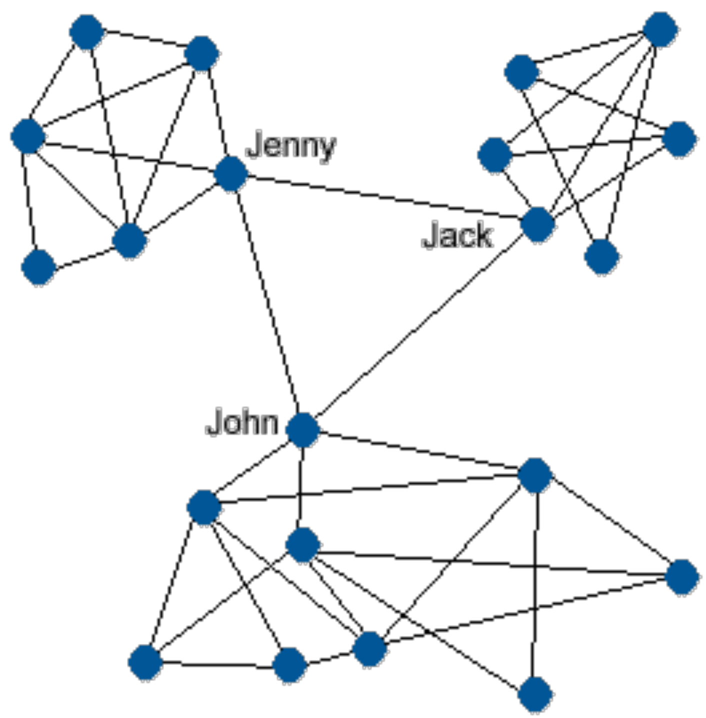


Local cut points:
Brokerage through overlapping group membership

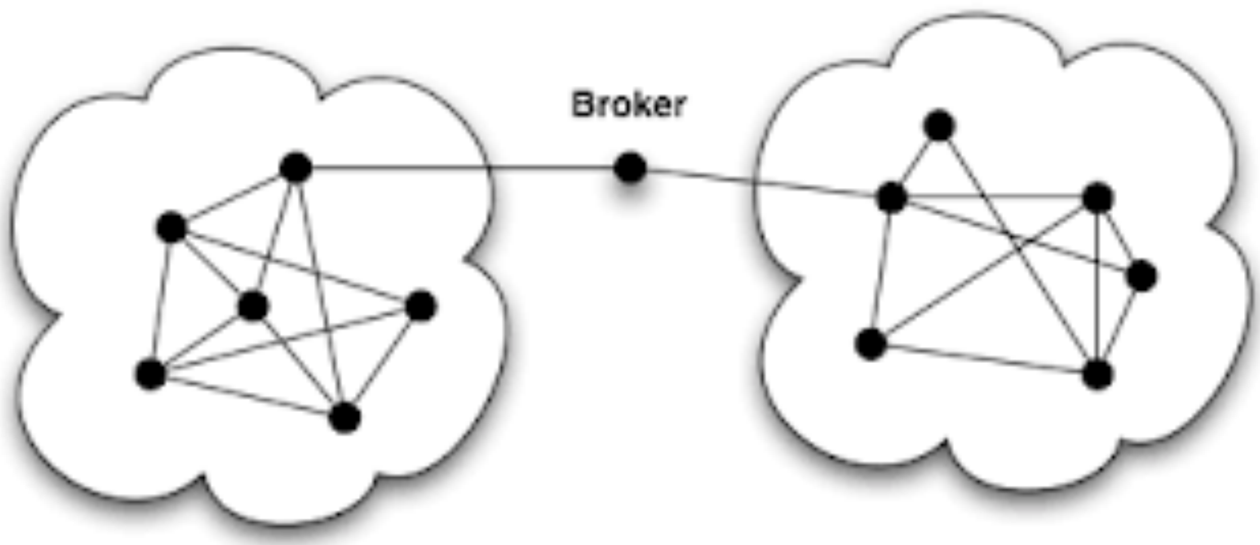
Measuring prestige: Brockers

Finding Brockers

- Brockers are "**intermediaries**", people that create relationships between communities
- As for graph representation, a brocker is a node that, if removed from the graph, **reduces graph connectivity.** For example, it causes the creation of disconnected components (*Jenny*, *Jack* and *John* in the graph)
- Brockers are also called key separators



Example of key separator



Algorithms to identify brockers are all based on some measure of the **graph connectivity**.

Algorithm for KPP_NEG (Keblady 2010)

- Let C_G be a measure of graph connectivity (e.g *reachability,* see later) for a graph G; V is the set of actors in G(nodes, vertexes)
- Algorithm KPP-neg (greedy algorithm)

Compute proposed measure of entire graph, C_G

$\forall\, v_i \in V$, remove v_i from the graph

Compute $C_{G-\{v_i\}}$ for the graph $G - \{v_i\}$.

Rank the nodes based on $|C_G$- $C_{G-\{v_i\}}|$ difference. Larger difference ranks higher.

Top ranked nodes are considered as key separators.

KPP-neg (2)

- A measure of connectivity: *reachability*

Pseudocode 1: *Reach*(v_i) – number of nodes reachable from v_i

---

Go to Source vertex v_i and mark it as *visited* and add to the set $Reach(v_i)$

For each adjacent vertex, A , of v_i,

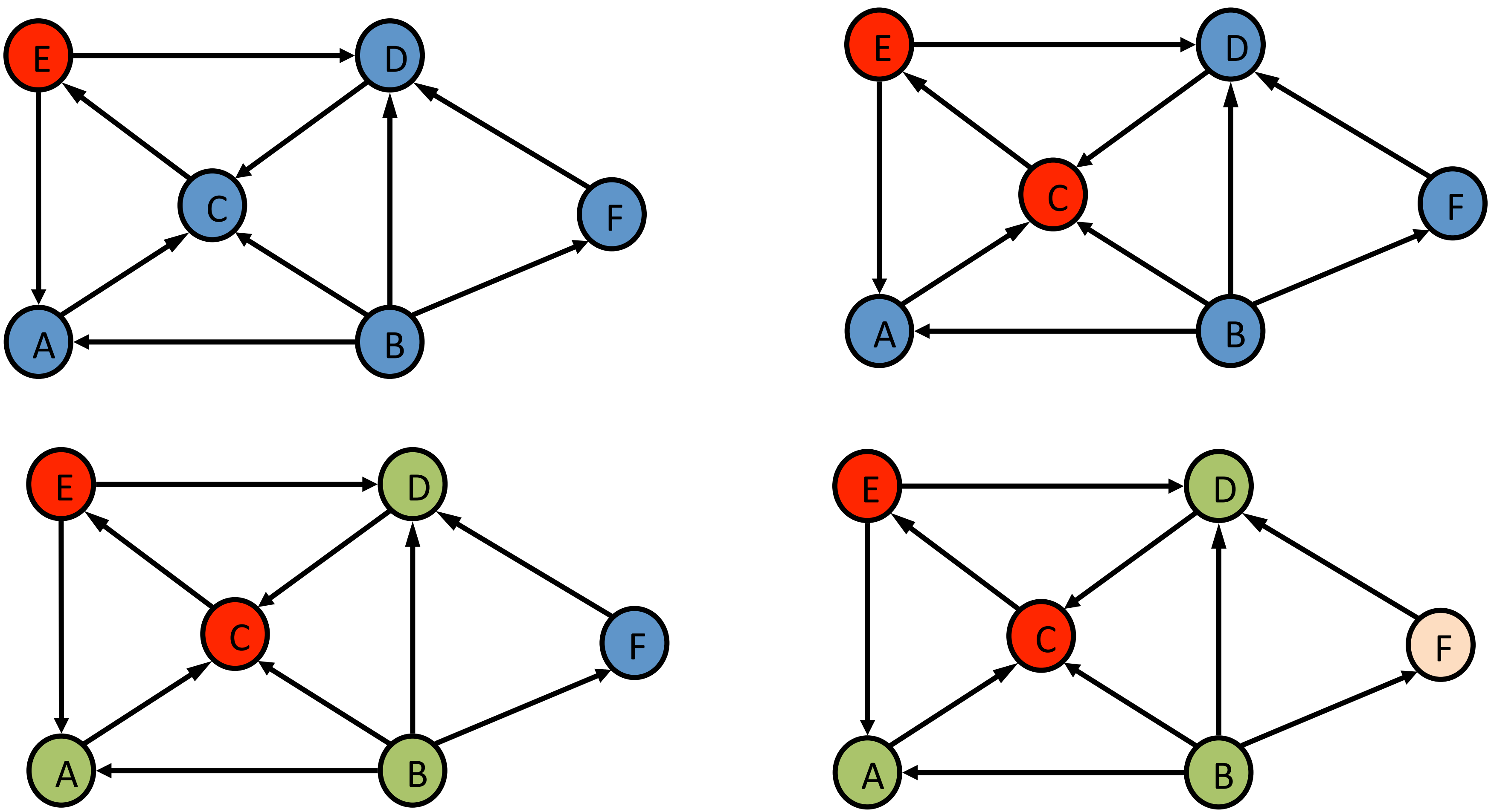
    If A is not already visited,

        Add adjacent vertex A to the set $Reach(v_i)$ and mark A as *visited*

        Call $Reach(A)$

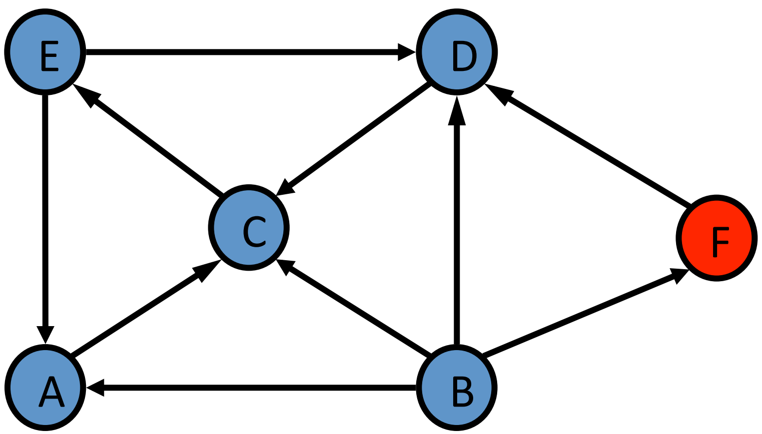
$$C_G = \sum_{i=1}^{n} Reach(v_\mathrm{i})$$

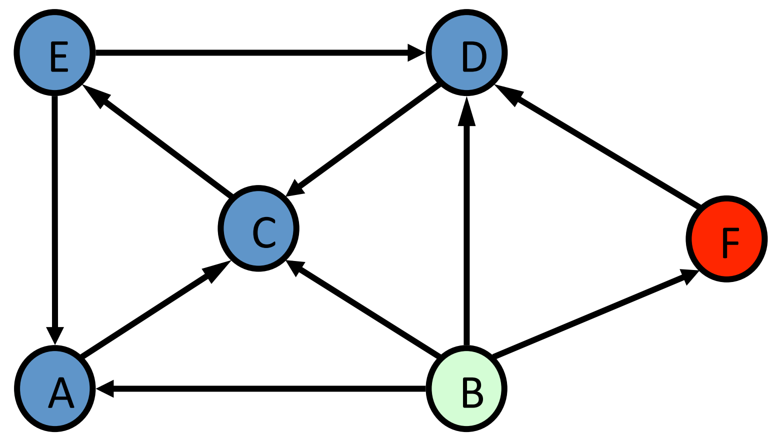
Example



R(E)= E,C,A,B,D,F

Example (2)





R(F)=F,B

NOTE: node reachability is a more accurate measure than previously seen “REACH”

Graph-based measures of social influence

1. Use graph-based methods/algorithms to identify “relevant players” in the network

   Relevant players = more influential, according to some criterion

2. **Use graph-based methods to identify global network properties and communities (community detection)**

3. Use graph-based methods to analyze the “spread” of information

Global Network Analysis

- Global properties of the network
- Community detection
- Spread of influence

Network Centrality

If we want to measure the degree to which the graph **as a whole** is centralized, we look at the **dispersion of centrality**:

Simple!: **variance of the individual centrality scores.**

$$S_D^2 = \left[\sum_{i=1}^{g} (C_D(n_i) - \overline{C}_d)^2 \right] / g$$

Or, using Freeman's general formula for centralization:

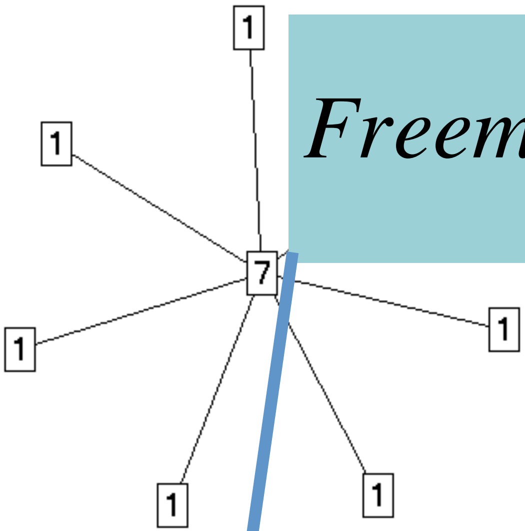
$$C_D = \frac{\sum_{i=1}^{g} \left[C_D(n^*) - C_D(n_i) \right]}{[(g-1)(g-2)]}$$

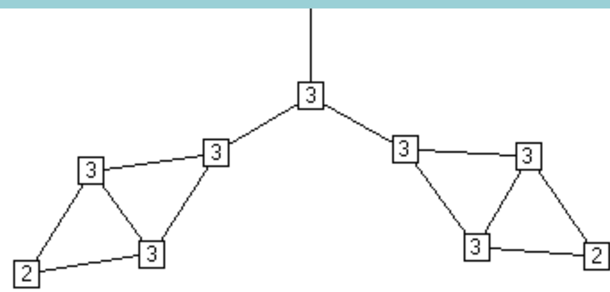
$C_D(n^*)$ is the maximum obtained value , therefore we are measuring the dispersion around that value

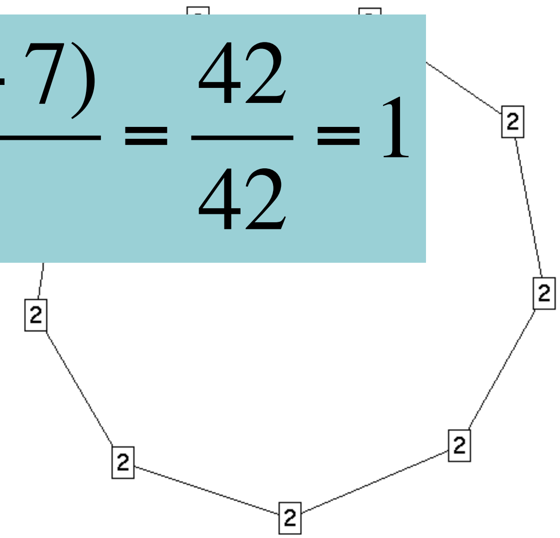
Network Centrality

Degree Centralization Scores

$$Freeman: \frac{(7-1)\times 7+(7-7)}{(8-1)(8-2)} = \frac{42}{42} = 1$$







Freeman: 1.0
Variance: 3.9

Freeman: .02
Variance: .17

Freeman: 0.0
Variance: 0.0

Freeman: .07
Variance: .20

Other Global measures

- Global measures can be defined for each of the node-related measures seen so far (betweeneness, authoritativeness, brokerage..)
- More interesting global analysis refers to **temporal evolution** of the network

Measuring Networks: Time

Time

Two factors that affect network information flow:

Time

- the timing of contacts matters
- simple example: an actor cannot pass information he has not yet received.

Topology - the shape, or form, of the network

- simple example: one actor cannot pass information to another unless they are either directly or indirectly connected (will see later on information spreading)

Measuring Networks: **Time**

Timing in networks

A focus on contact *structure* has often slighted the importance of network *dynamics,* though a number of recent works are addressing this.

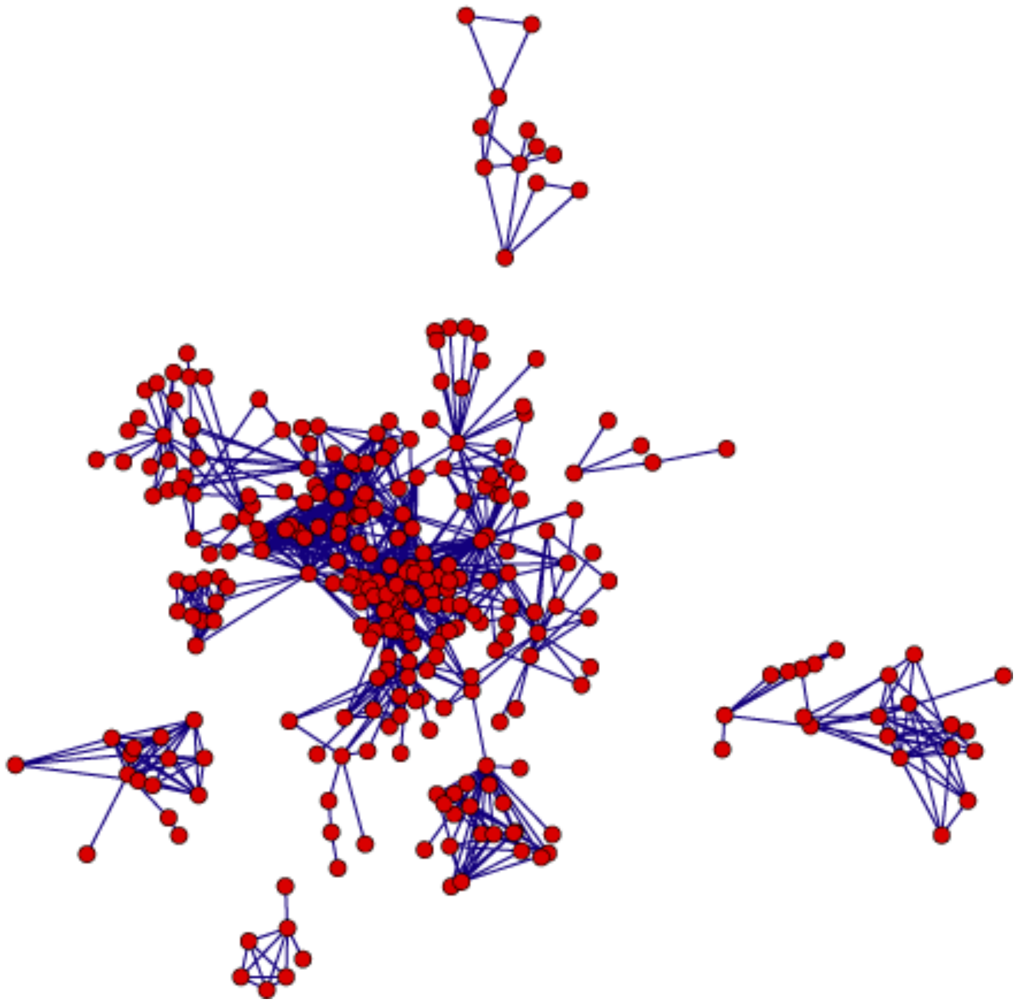
Time affects networks in two important ways:

1) The structure itself evolves, in ways that will affect the topology an thus flow.

2) The timing of contact constrains information flow

Measuring Networks: Time

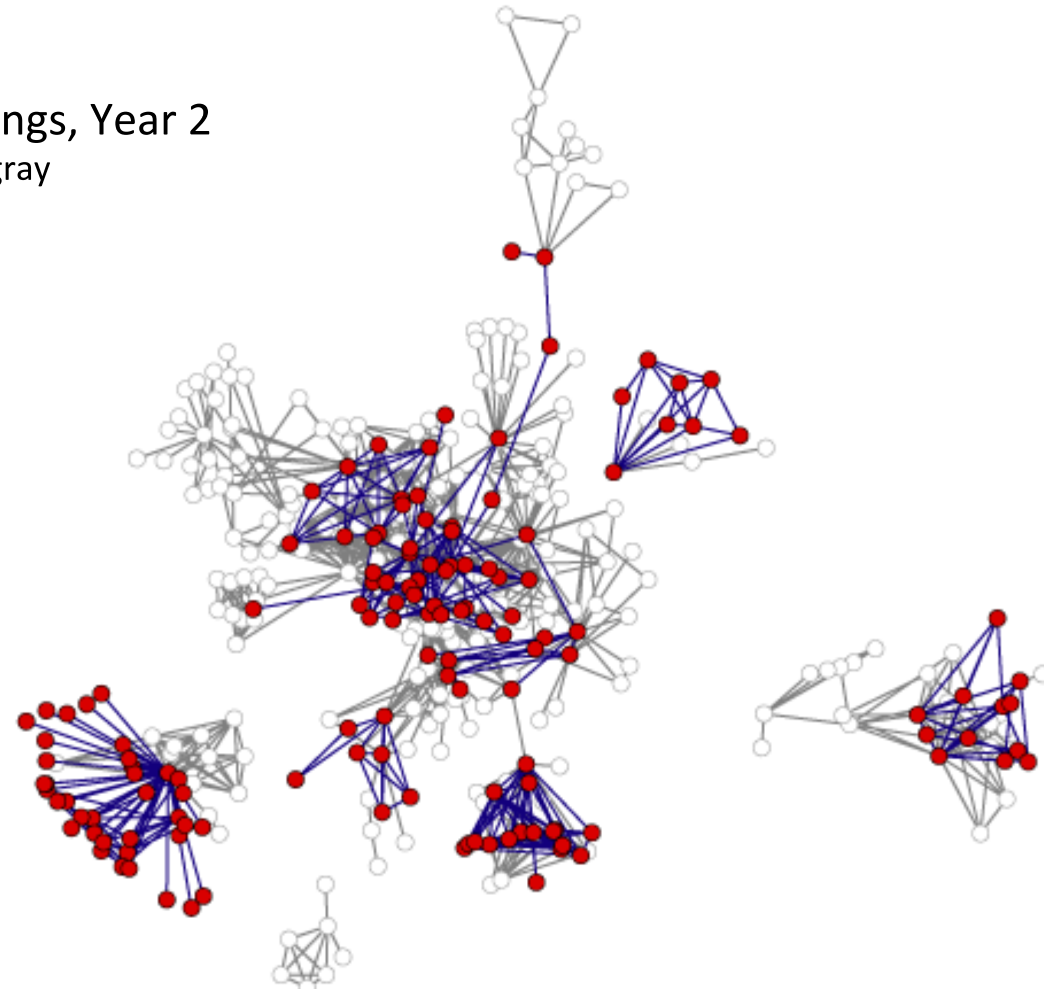
Drug Relations, Colorado Springs, Year 1



Data on drug users in Colorado Springs, over 5 years

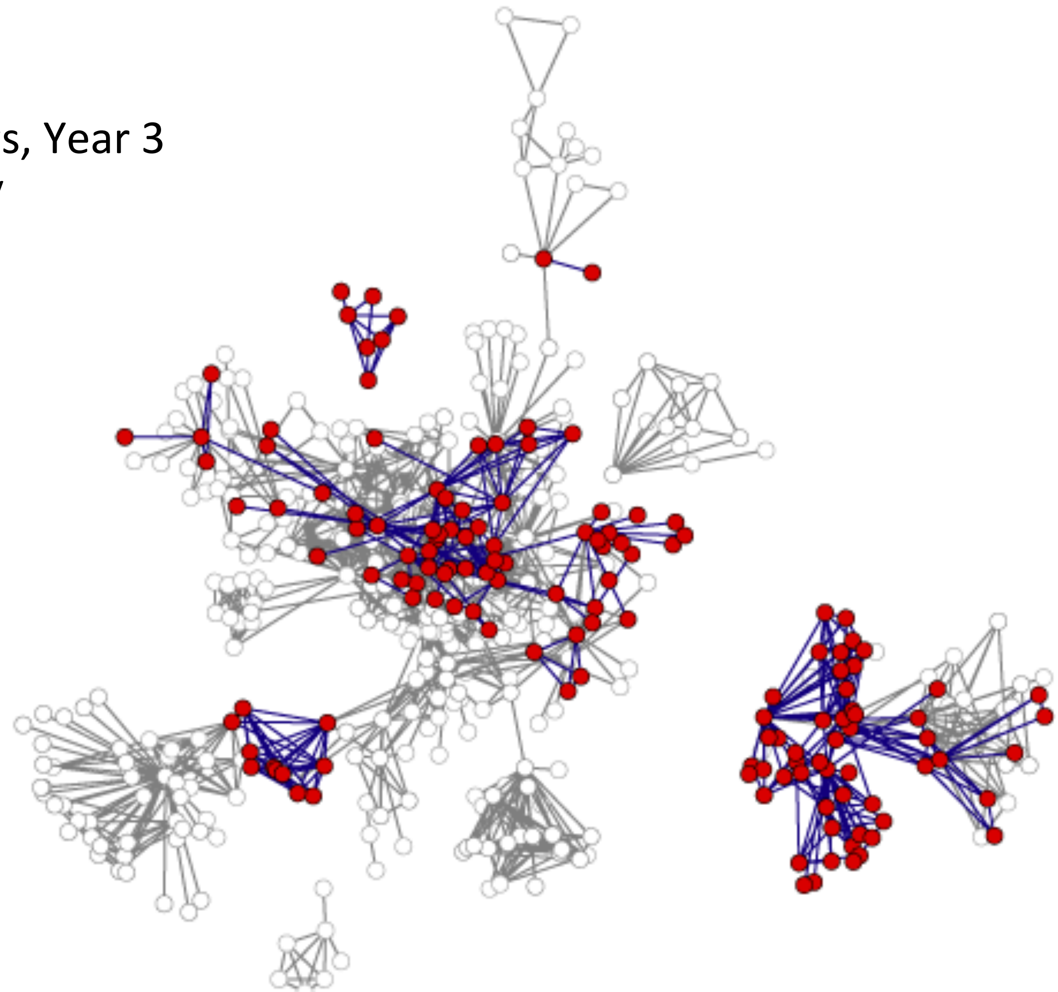
Measuring Networks: **Time**

Drug Relations, Colorado Springs, Year 2
Current year in red, past relations in gray



Measuring Networks: **Time**

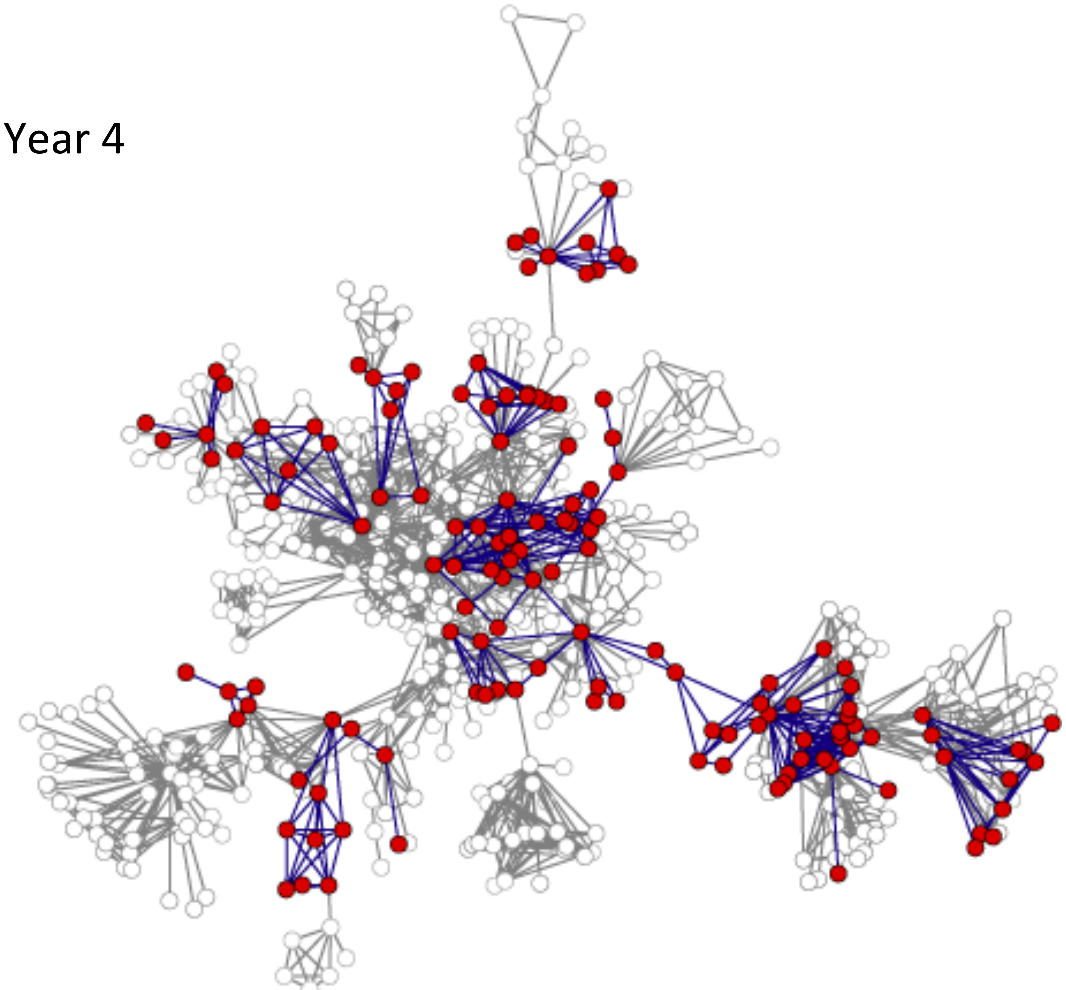
Drug Relations, Colorado Springs, Year 3
Current year in red, past relations in gray



Measuring Networks: Time

Drug Relations, Colorado Springs, Year 4

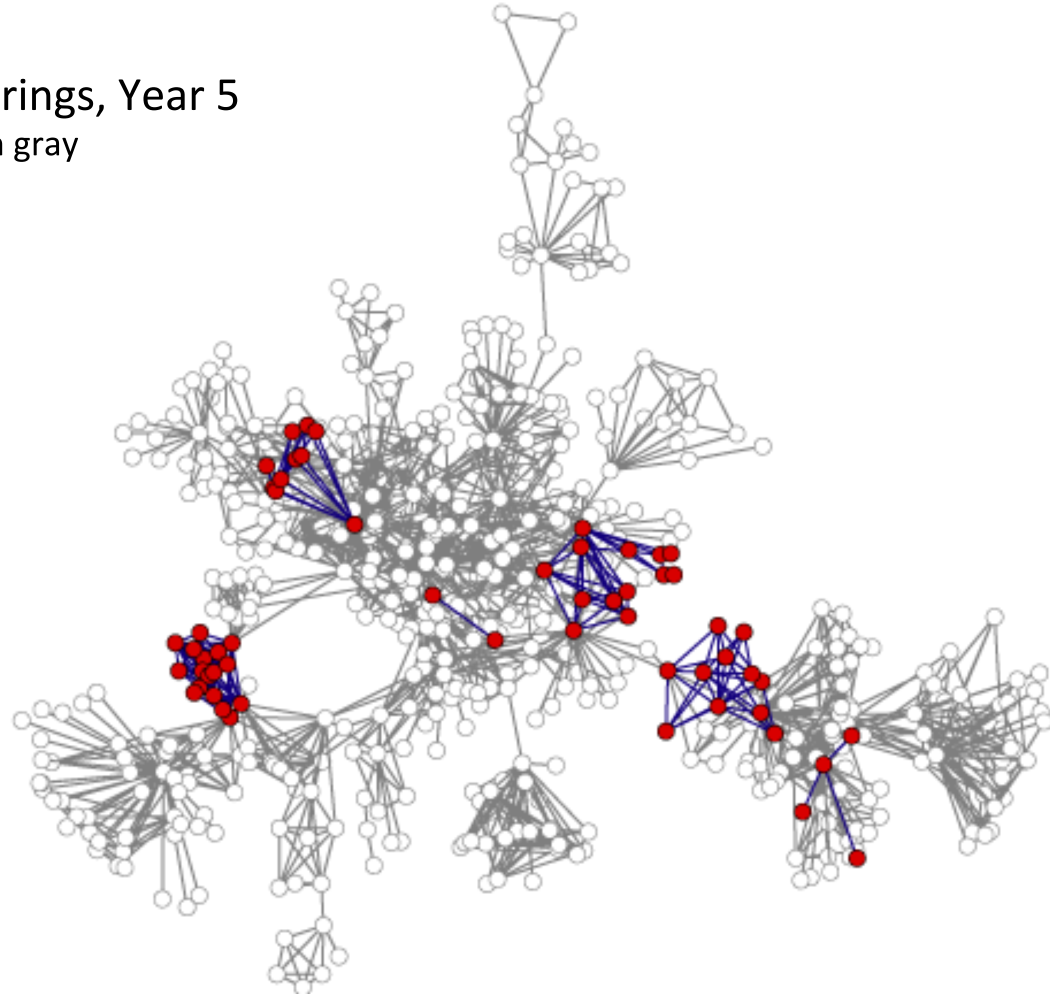
Current year in red, past relations in gray



Measuring Networks: Time

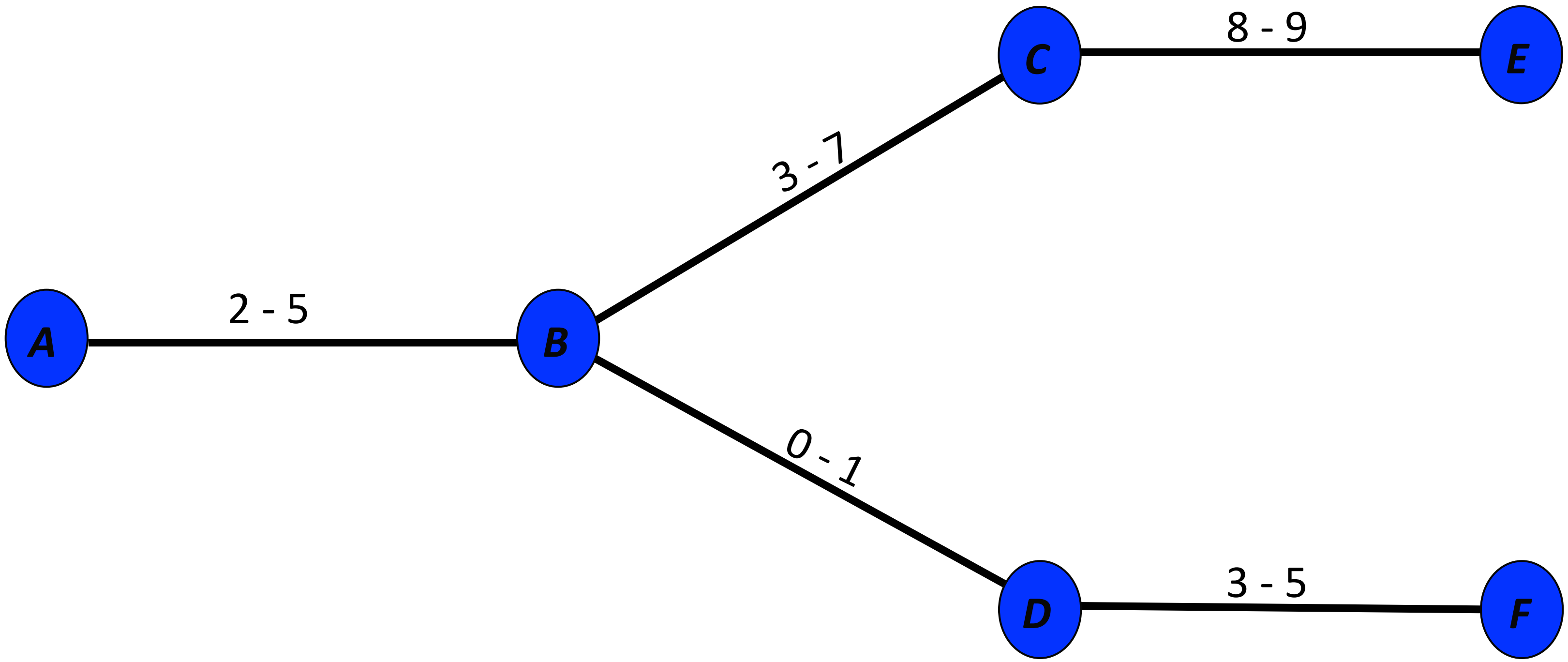
Drug Relations, Colorado Springs, Year 5

Current year in red, past relations in gray



Measuring Networks: Time

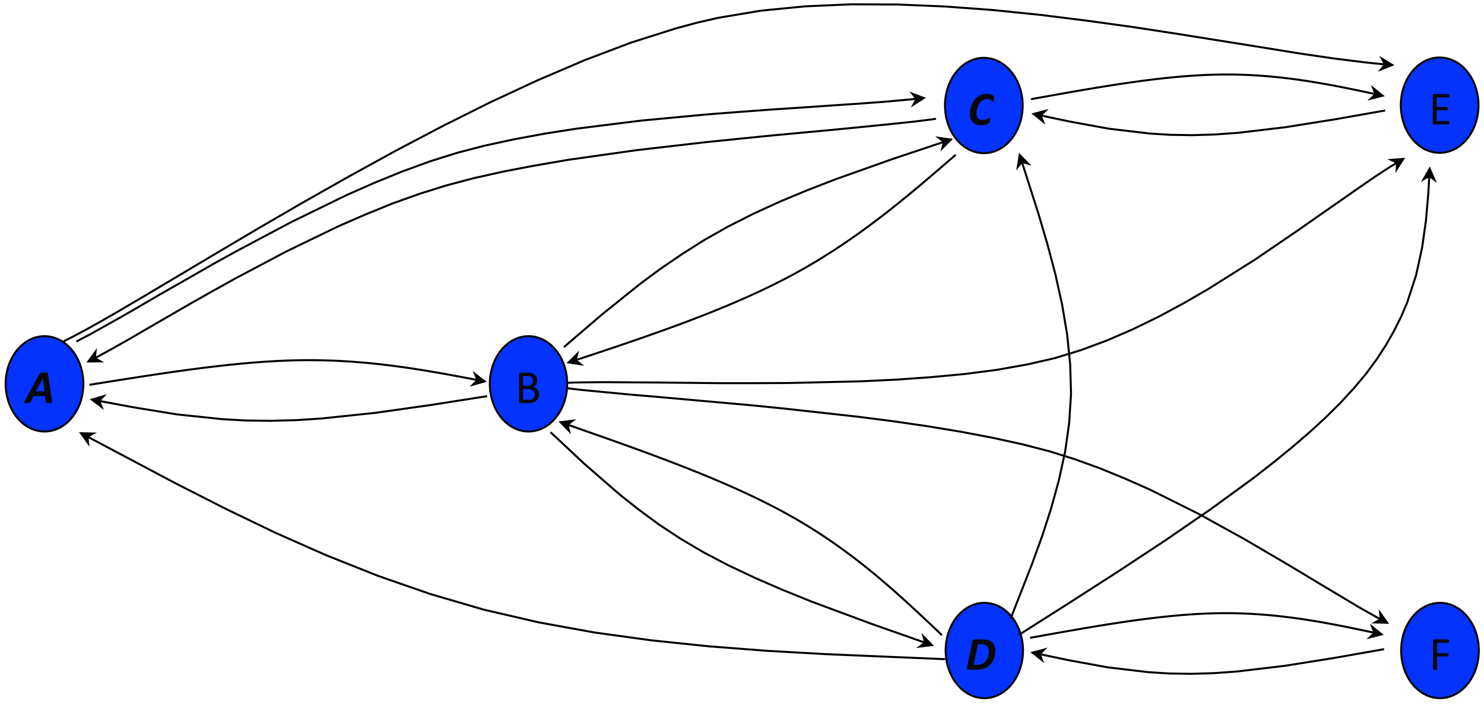
What impact does timing have on flow through the network?



Numbers above lines indicate contact periods

Measuring Networks: **Time**

The path graph for the hypothetical contact network



While clearly important, this is not often handled well by current SNA software.

Global Network Analysis

- Global properties of the network
- **Community detection**
- Spread of influence

Community detection

- Community: It is formed by individuals such that those within a group interact with each other **more frequently than with those outside the group**
  - a.k.a. group, cluster, cohesive subgroup, module in different contexts
- Community detection: discovering groups in a network where individuals' group memberships are not explicitly given