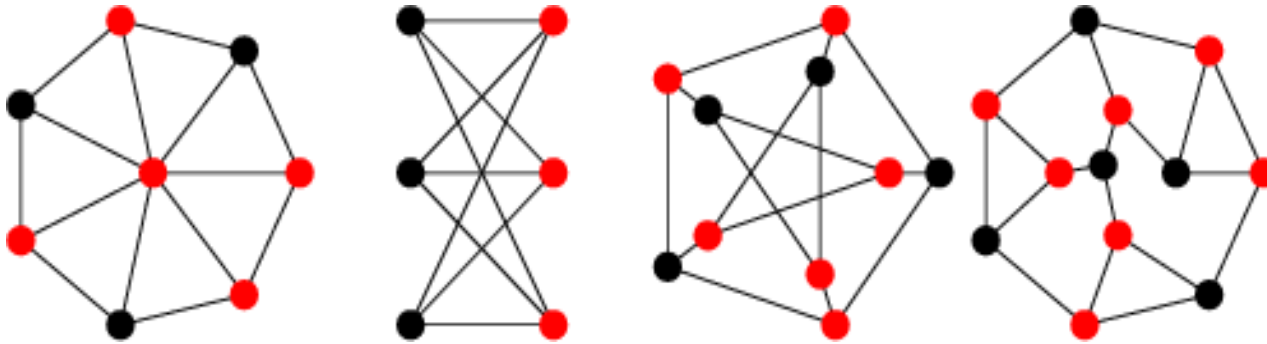


# Social Network Analysis

## **Day 1**

Graph theory, Centrality, Clustering,  
Transitivity



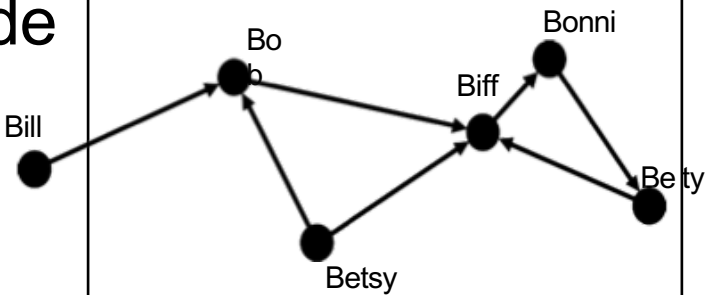
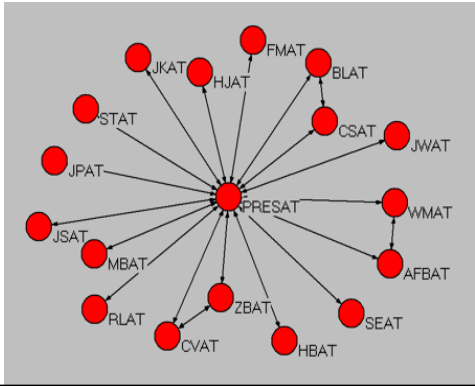
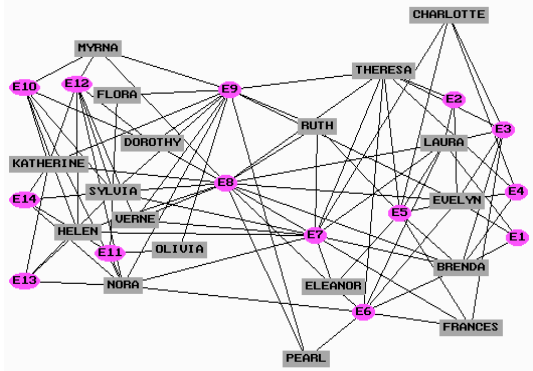
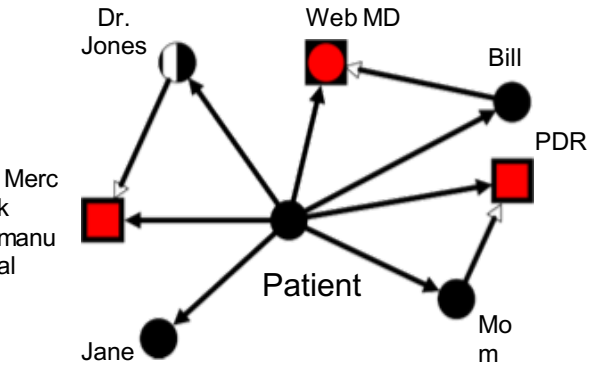
# Graph Theoretic Concepts

- In this section we will cover:
  - Definitions
  - Terminology
  - Adjacency
  - Density concepts
    - E.g, Completeness
  - Walks, trails, paths
  - Cycles, Trees
  - Reachability/Connectedness
    - Connectivity, flows
  - Isolates, Pendants, Centers
  - Components, bi-components
  - Walk Lengths, distance
    - Geodesic distance
  - Independent paths
  - Cutpoints, bridges

# What is a Graph?

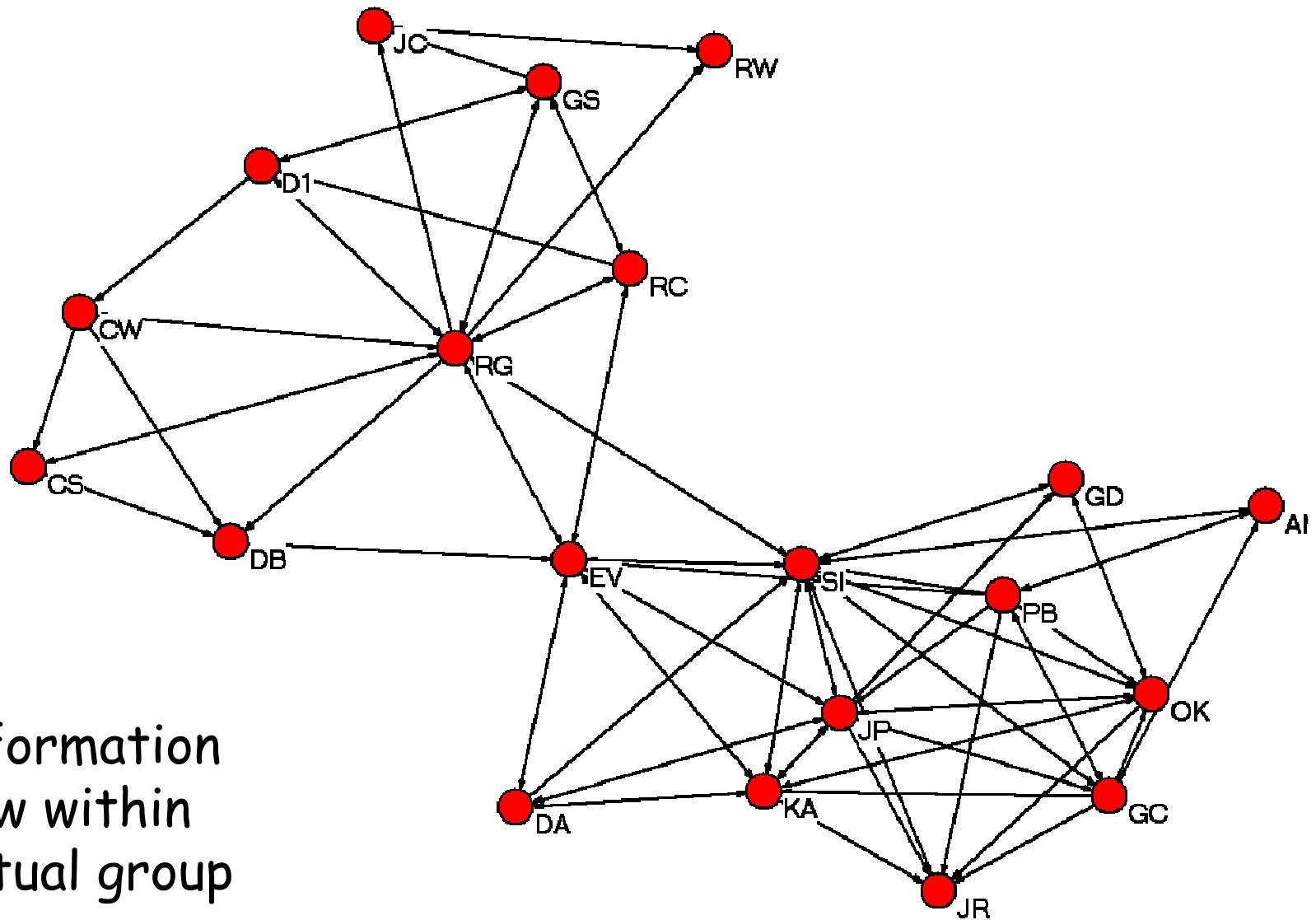
- $G = (V, E)$ 
  - A graph is a set of vertices and edges
    - Vertices, sometimes called nodes, are the actors or entities between which relationships exist
      - People
      - Organizations
    - Edges, sometimes called relations or lines, are the behavior/interaction/relationship of interest
      - Communicates with
      - Trusts
      - Uses

# Kinds of Network Data

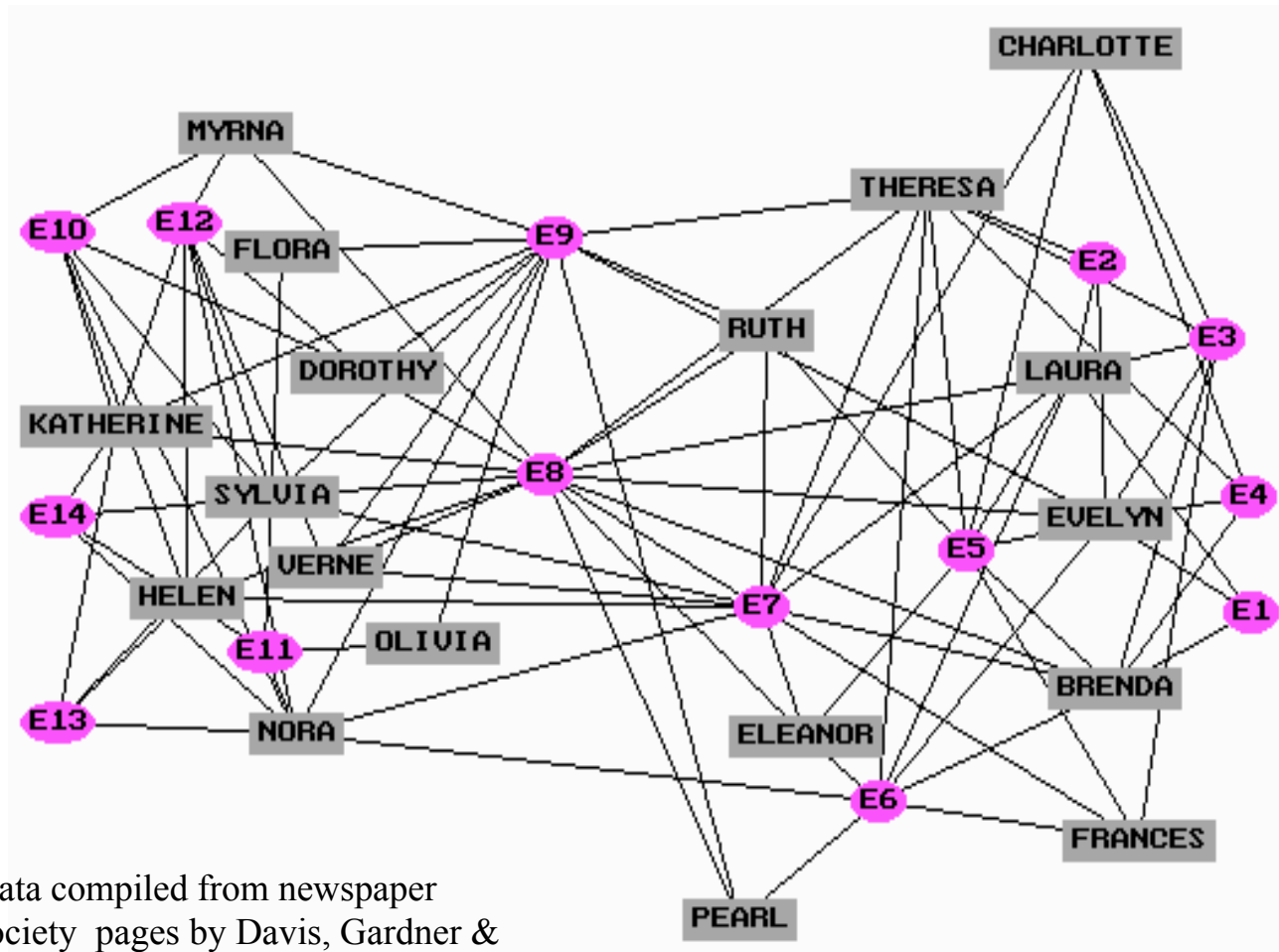
	Complete	Ego
1-mode		
2-mode		



# 1-mode Complete Network



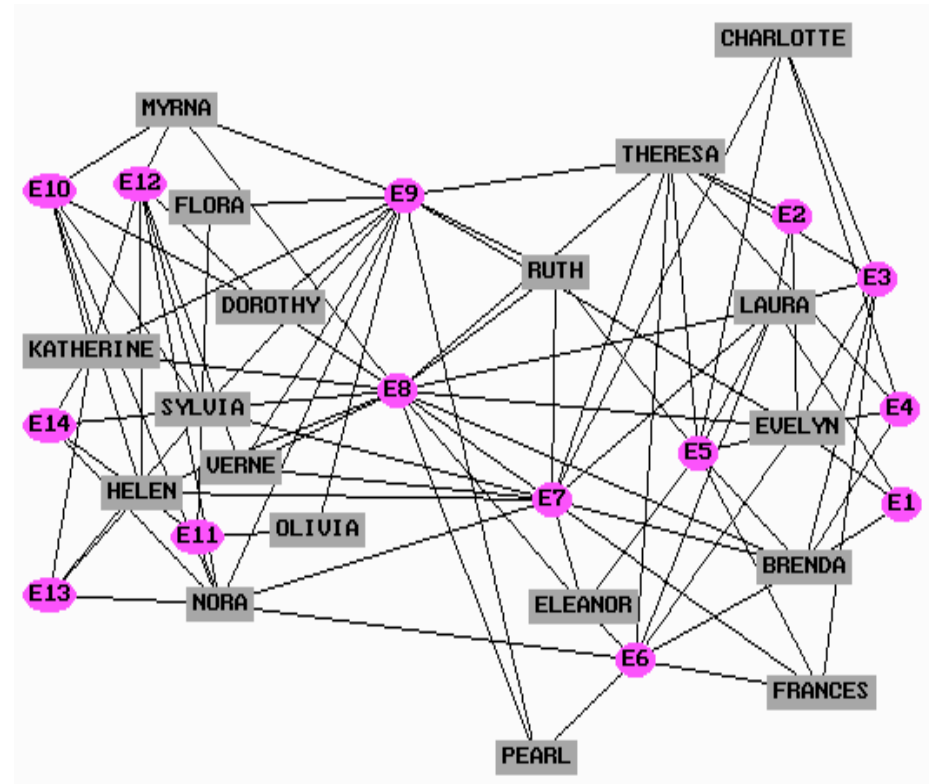
# 2-mode Complete Network



Data compiled from newspaper  
society pages by Davis, Gardner &  
Gardner

# Bipartite graphs

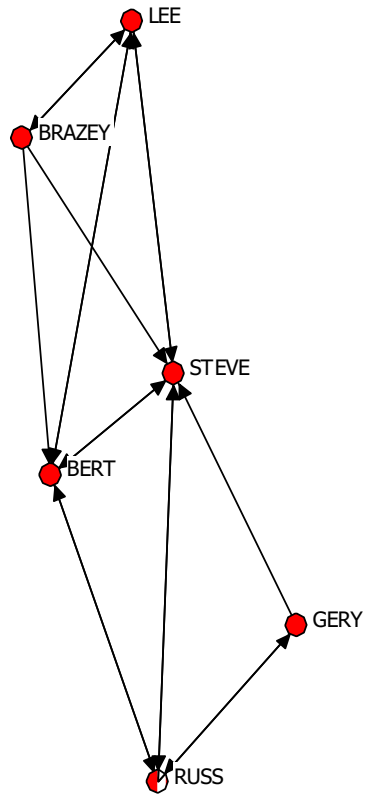
- Used to represent 2-mode data
- Nodes can be partitioned into two sets (corresponding to modes)
- Ties occur only between sets, not within



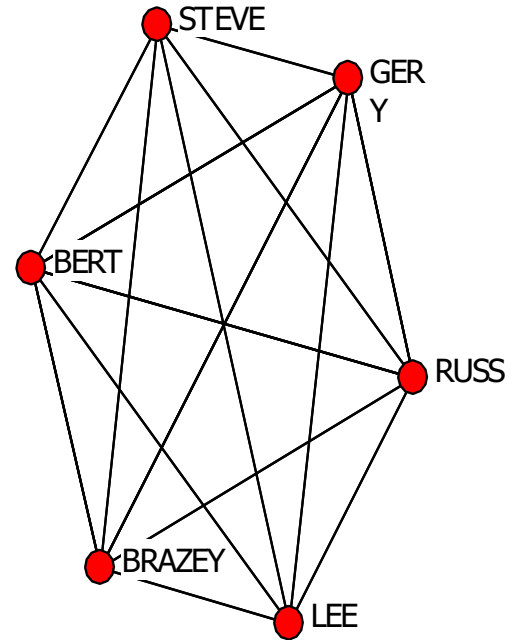
# Complete Network Data vs. Complete Graph

- The term “Complete Network Data” refers to collecting data for/from all actors (vertices) on the graph
  - The opposite if Ego-Network or Ego-Centric Network data, in which data is collected only from the perspective an individual (the ego)
- The term “Complete Graph” refers to a graph where every edge that could exist in the graph, does:
  - For all  $i, j$  ( $j > i$ ),  $v(i, j) = 1$

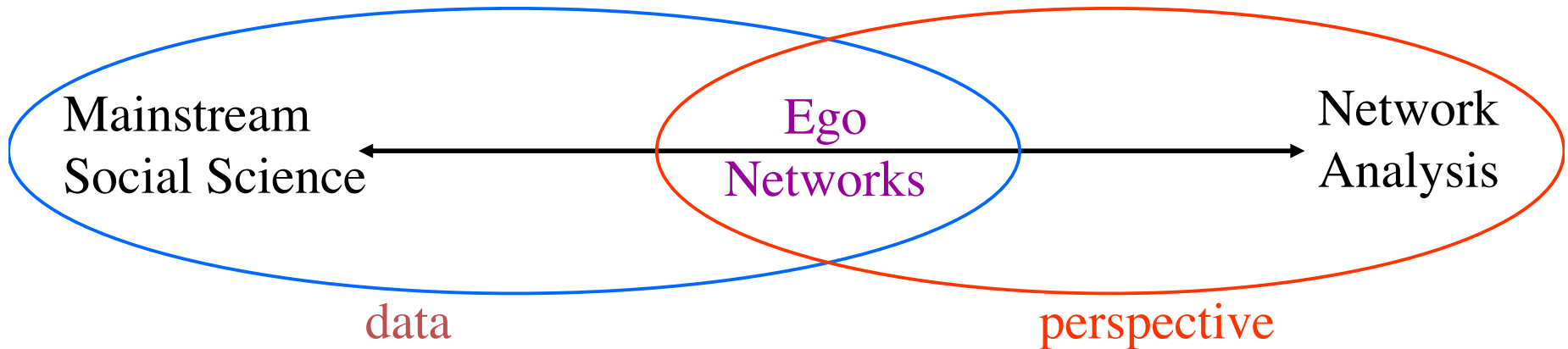
# Complete Network Data



# Complete Graph



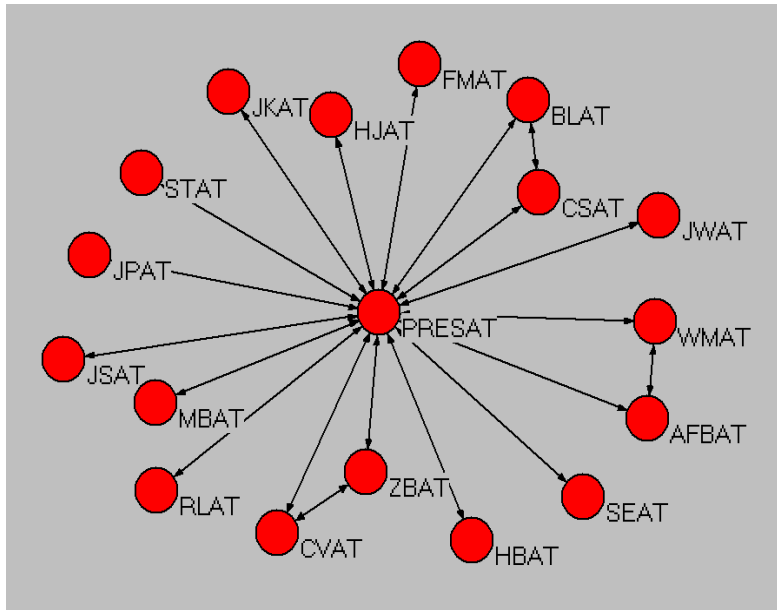
# Ego Network Analysis



- Combine the perspective of network analysis with the data of mainstream social science

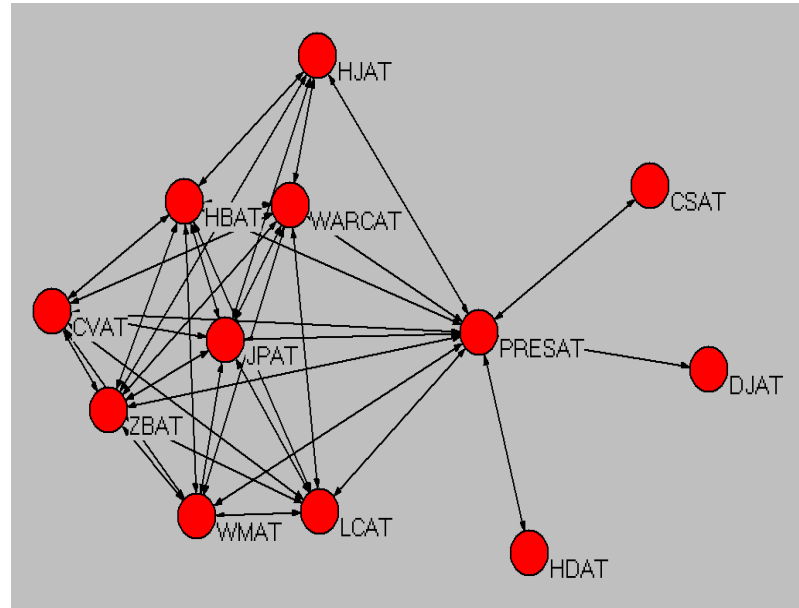
# 1-mode Ego Network

Carter Administration  
meetings



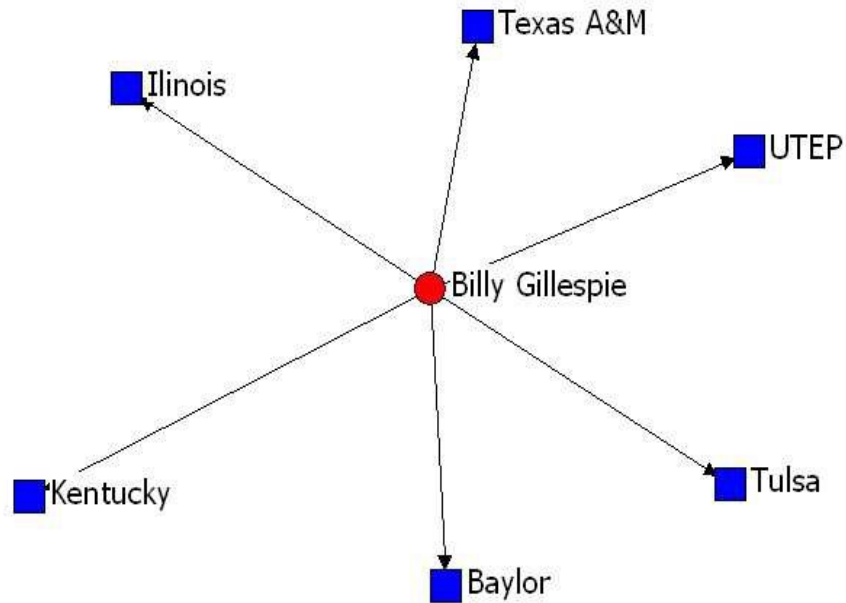
Year 1

Data courtesy of Michael  
Link



Year 4

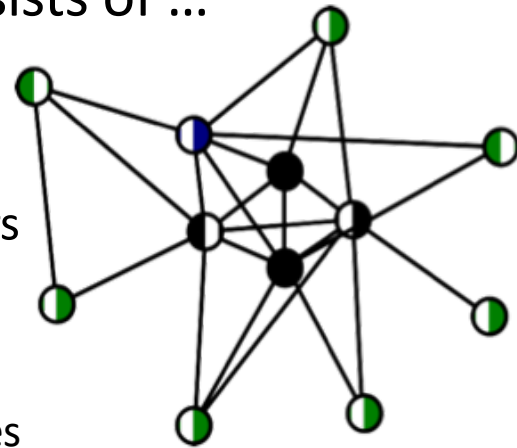
# 2-mode Ego Network





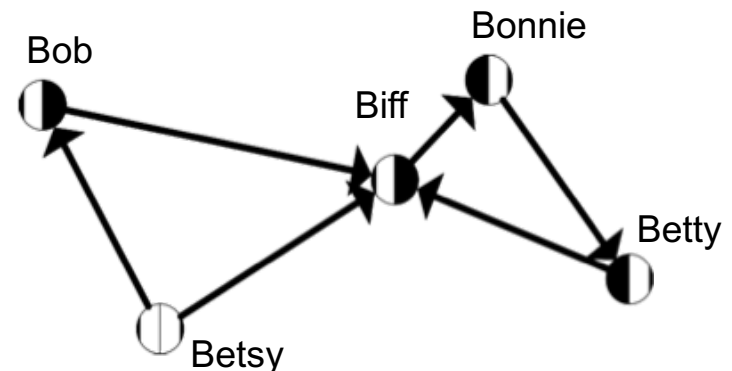
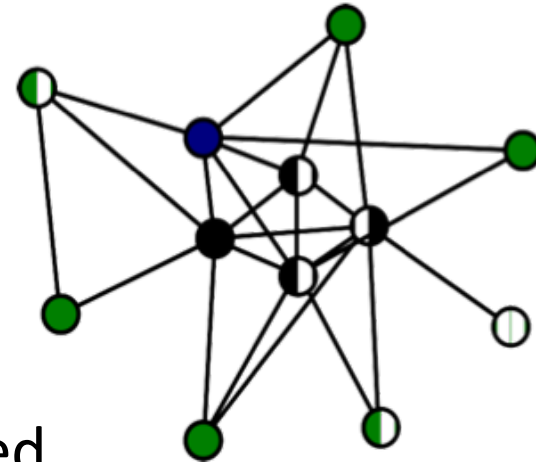
# Undirected Graphs

- An undirected graph  $G(V,E)$  (often referred to simply as a graph or a simple graph) consists of ...
  - Set of nodes|vertices  $V$  representing actors
  - Set of lines|links|edges  $E$  representing ties among pairs of actors
    - An edge is an unordered pair of nodes  $(u,v)$
    - Nodes  $u$  and  $v$  adjacent if  $(u,v) \in E$
    - So  $E$  is subset of set of all pairs of nodes
- Drawn without arrow heads
  - Sometimes with dual arrow heads
- Used to represent logically symmetric social relations
  - In communication with; attending same meeting as



# Directed vs. Undirected Ties

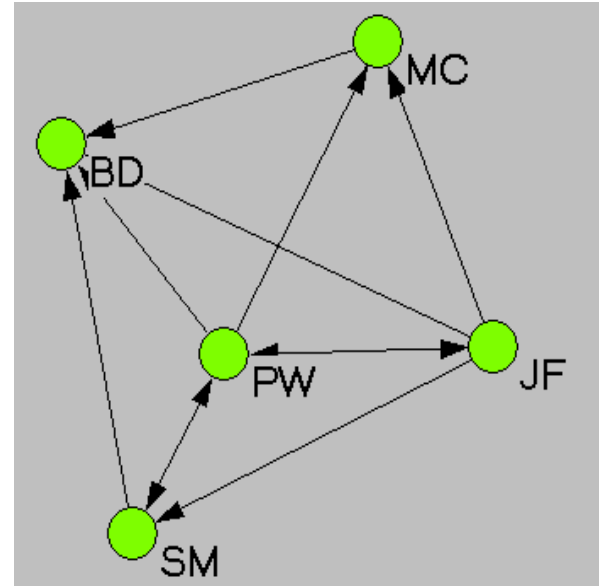
- Undirected relations
  - Attended meeting with
  - Communicates daily with
- Directed relations
  - Lent money to
- Logically vs empirically directed ties
  - Empirically, even undirected relations can be non-symmetric due to measurement error



# Directed Graphs (Digraphs)

- Digraph  $G(V,E)$  consists of ...

- Set of nodes  $V$
- Set of directed arcs  $E$ 
  - An arc is an ordered pair of nodes  $(u,v)$
  - $(u,v) \in E$  indicates  $u$  sends arc to  $v$
  - $(u,v) \in E$  does not imply that  $(v,u) \in E$



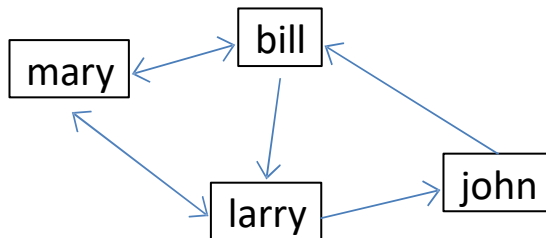
- Ties drawn with arrow heads, which can be in both directions
- Represent logically non-symmetric or anti-symmetric social relations
  - Lends money to

# Transpose Adjacency matrix

- In directed graphs, interchanging rows/columns of adjacency matrix effectively reverses the direction & meaning of ties

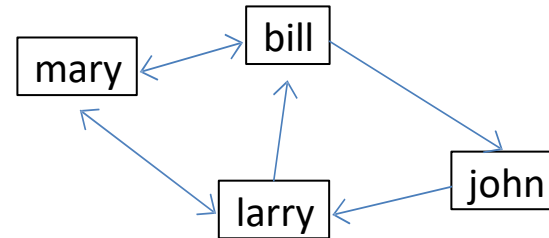
	Mary	Bill	John	Larry
Mary	0	1	0	1
Bill	1	0	0	1
John	0	1	0	0
Larry	1	0	1	0

Gives money to



	Mary	Bill	John	Larry
Mary	0	1	0	1
Bill	1	0	1	0
John	0	0	0	1
Larry	1	1	0	0

Gets money from



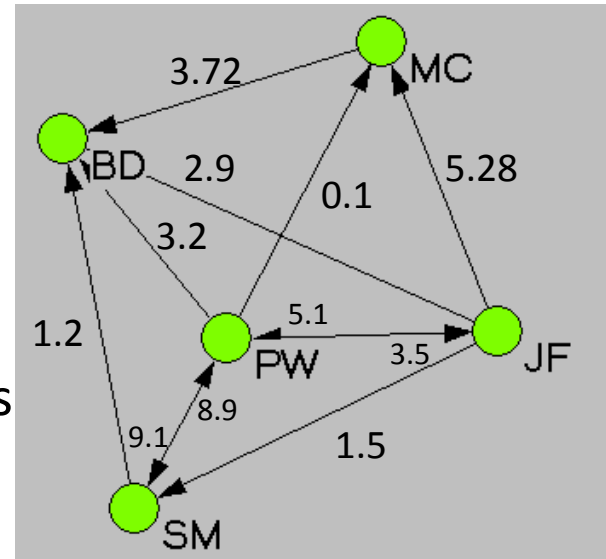
# Valued Digraphs (vigraphs)

- A valued digraph  $G(V,E,W)$  consists of ...

- Set of nodes  $V$
- Set of directed arcs  $E$ 
  - An arc is an ordered pair of nodes  $(u,v)$
  - $(u,v) \in E$  indicates  $u$  sends arc to  $v$
  - $(u,v) \in E$  does not imply that  $(v,u) \in E$
- Mapping  $W$  of arcs to real values

- Values can represent such things as

- Strength of relationship
- Information capacity of tie
- Rates of flow or traffic across tie
- Distances between nodes
- Probabilities of passing on information
- Frequency of interaction



# Valued Adjacency Matrix

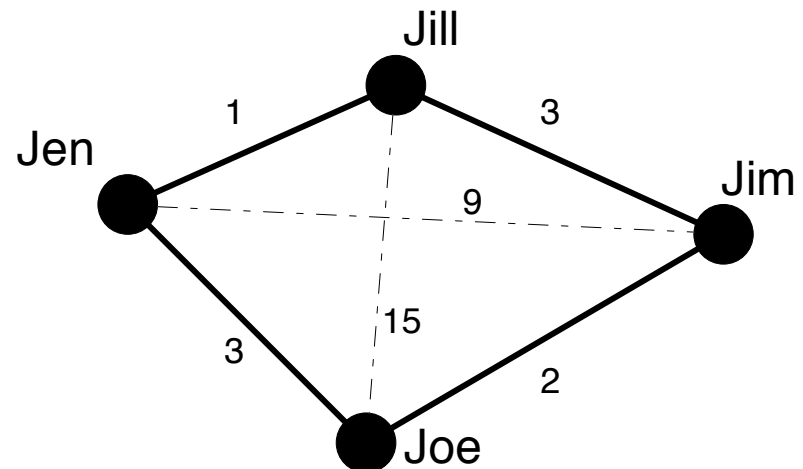
## Dichotomized

	Jim	Jill	Jen	Joe
Jim	-	1	0	1
Jill	1	-	1	0
Jen	0	1	-	1
Joe	1	0	1	-

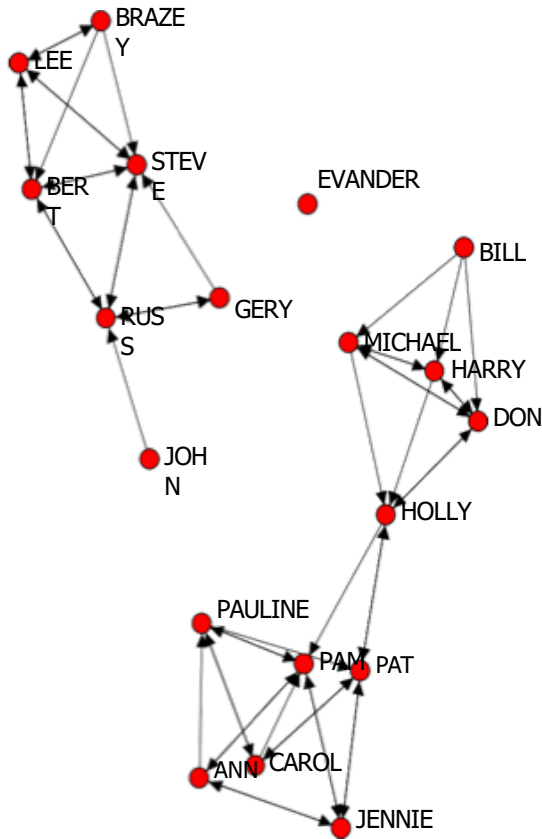
- The diagram below uses solid lines to represent the adjacency matrix, while the numbers along the solid line (and dotted lines where necessary) represent the proximity matrix.
- In this particular case, one can derive the adjacency matrix by dichotomizing the proximity matrix on a condition of  $p_{ij} \leq 3$ .

## Distances btw offices

	Jim	Jill	Jen	Joe
Jim	-	3	9	2
Jill	3	-	1	15
Jen	9	1	-	3
Joe	2	15	3	-



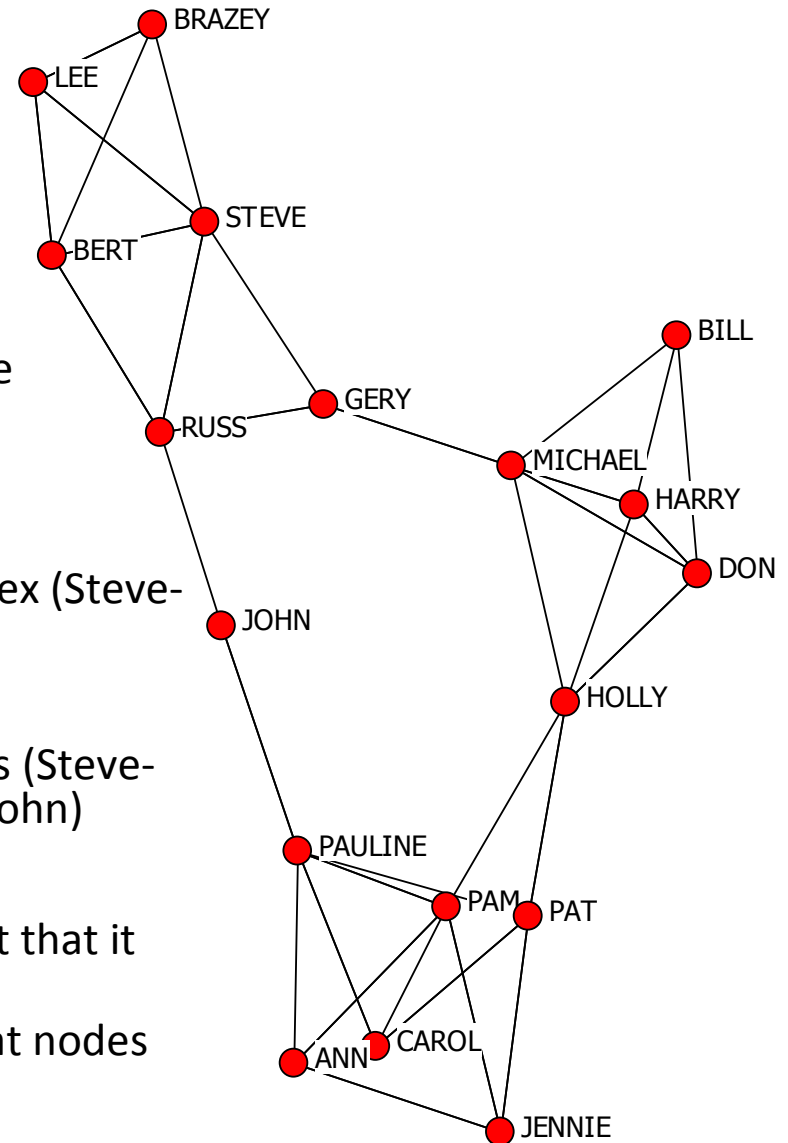
# Node-related concepts



- **Degree**
  - The number of ties incident upon a node
  - In a digraph, we have indegree (number of arcs to a node) and outdegree (number of arcs from a node)
- **Pendant**
  - A node connected to a component through only one edge or arc
    - A node with degree 1
    - Example: John
- **Isolate**
  - A node which is a component on its own
    - E.g., Evander

# Graph traversals

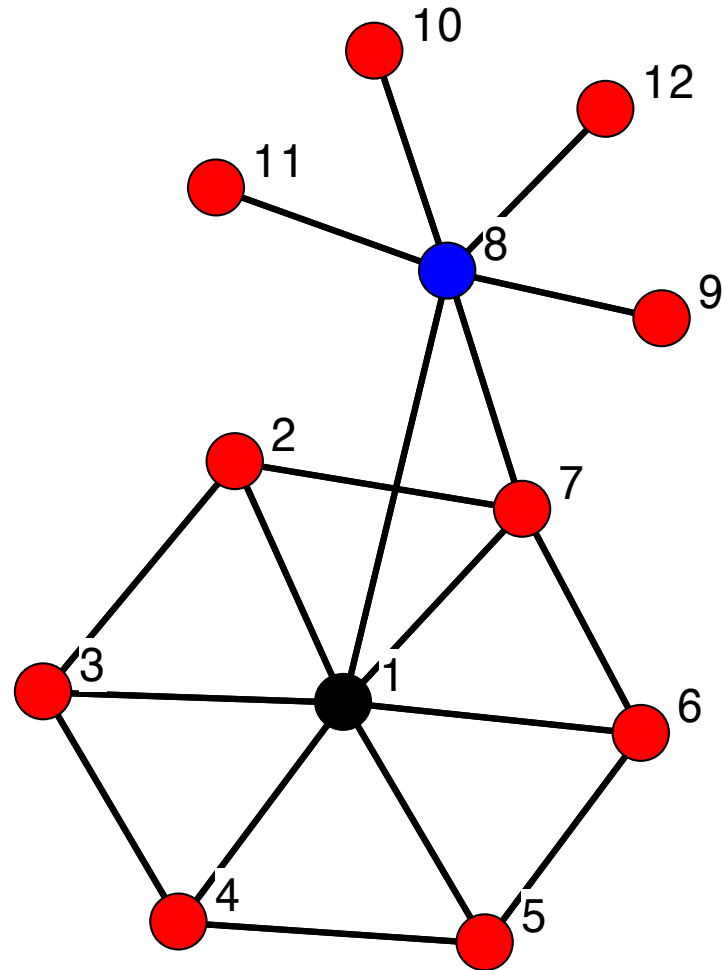
- **Walk**
  - Any unrestricted traversing of vertices across edges (Russ-Steve-Bert-Lee-Steve)
- **Trail**
  - A walk restricted by not repeating an edge or arc, although vertices can be revisited (Steve-Bert-Lee-Steve-Russ)
- **Path**
  - A trail restricted by not revisiting any vertex (Steve-Lee-Bert-Russ)
- **Geodesic Path**
  - The shortest path(s) between two vertices (Steve-Russ-John is shortest path from Steve to John)
- **Cycle**
  - A cycle is in all ways just like a path except that it ends where it begins
  - Aside from endpoints, cycles do not repeat nodes
  - E.g. Brazey-Lee-Bert-Steve-Brazey





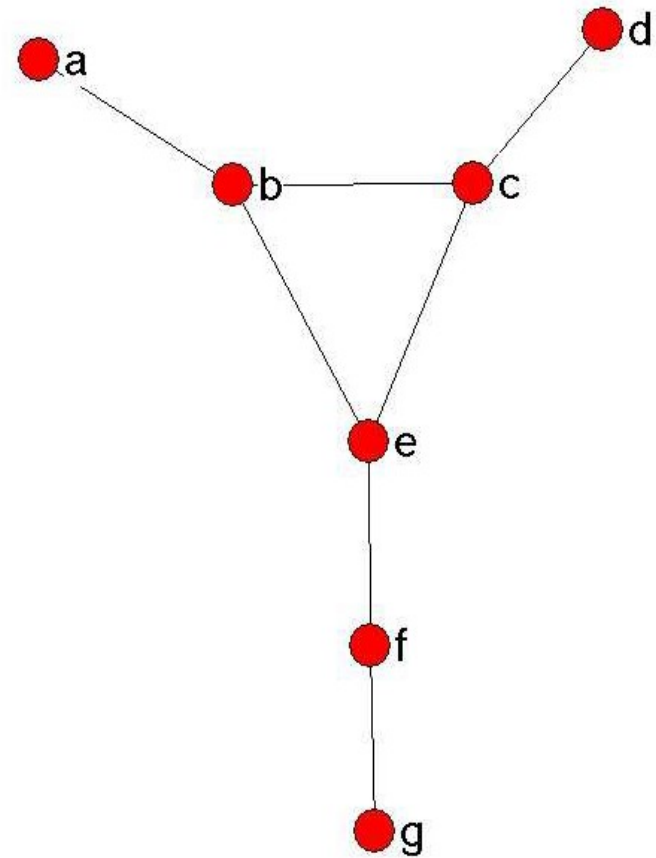
# Length & Distance

- Length of a path (or any walk) is the number of links it has
- The **Geodesic Distance** (aka graph-theoretic distance) between two nodes is the length of the shortest path
  - Distance from 5 to 8 is 2, because the shortest path (5-1-8) has two links



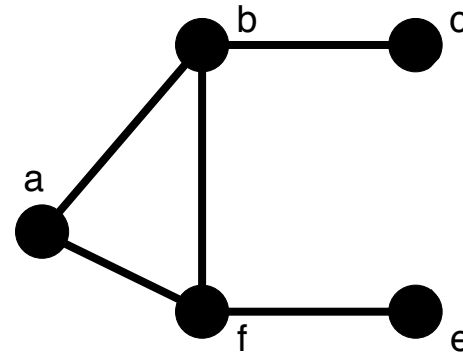
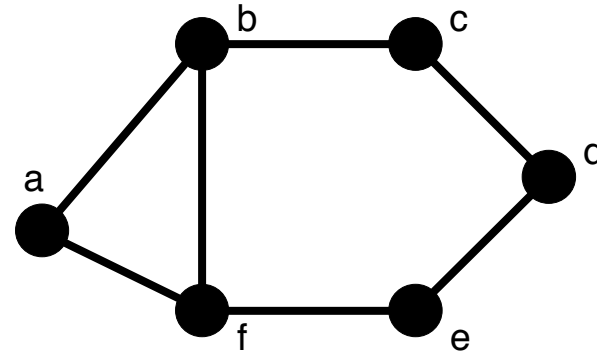
# Geodesic Distance Matrix

	a	b	c	d	e	f	g
a	0	1	2	3	2	3	4
b	1	0	1	2	1	2	3
c	2	1	0	1	1	2	3
d	3	2	1	0	2	3	4
e	2	1	1	2	0	1	2
f	3	2	2	3	1	0	1
g	4	3	3	4	2	1	0



# Subgraphs

- Set of nodes
  - Is just a set of nodes
- A subgraph
  - Is set of nodes together with ties among them
- An induced subgraph
  - Subgraph defined by a set of nodes
  - Like pulling the nodes and ties out of the original graph



Subgraph induced by considering the set  $\{a,b,c,f,e\}$

# Components

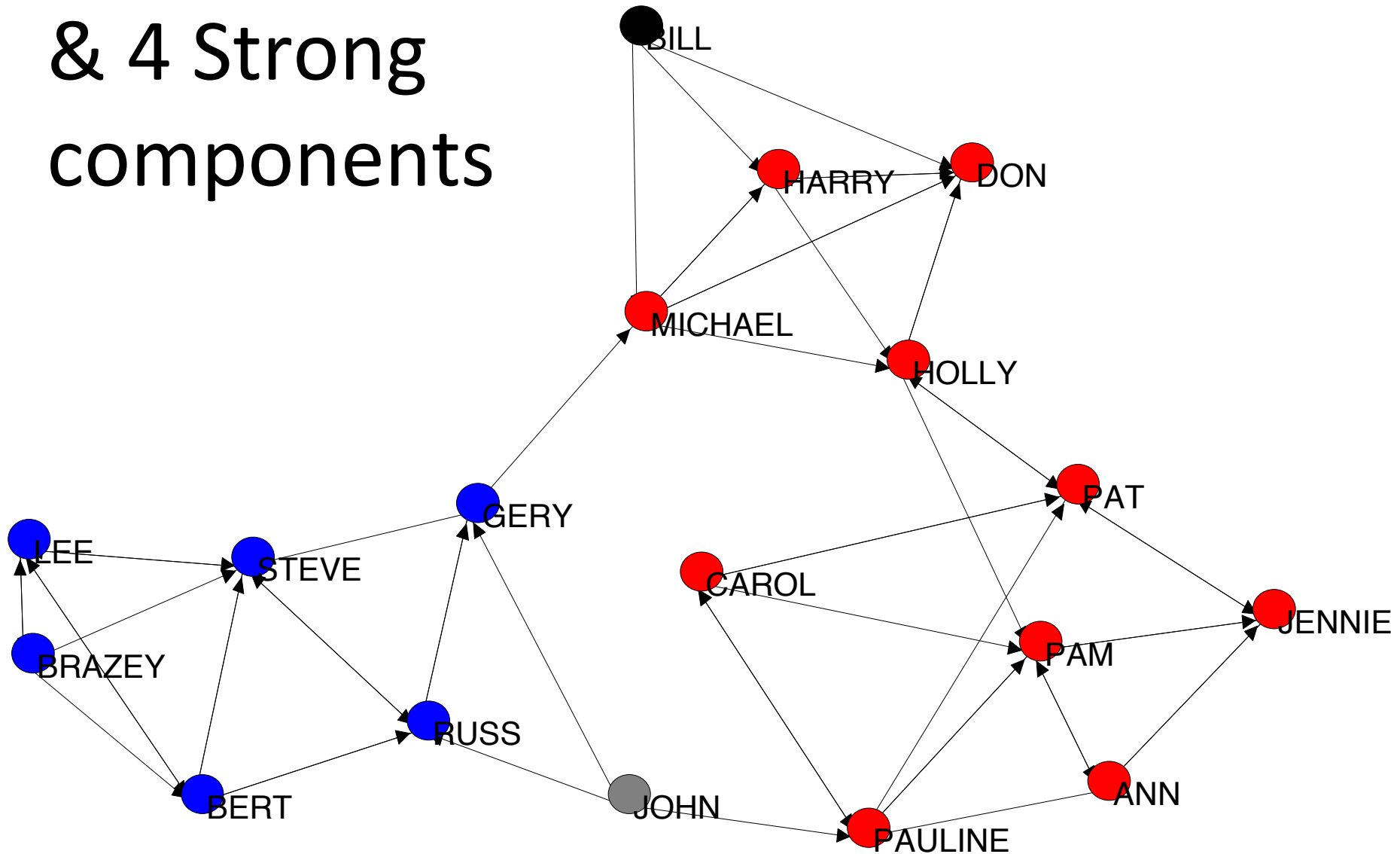
- Maximal sets of nodes in which every node can reach every other by some path (no matter how long)
- A graph is *connected* if it has just one component

It is relations (types of tie) that define different networks, not components. A network that has two components remains one (disconnected) network.

# Components in Directed Graphs

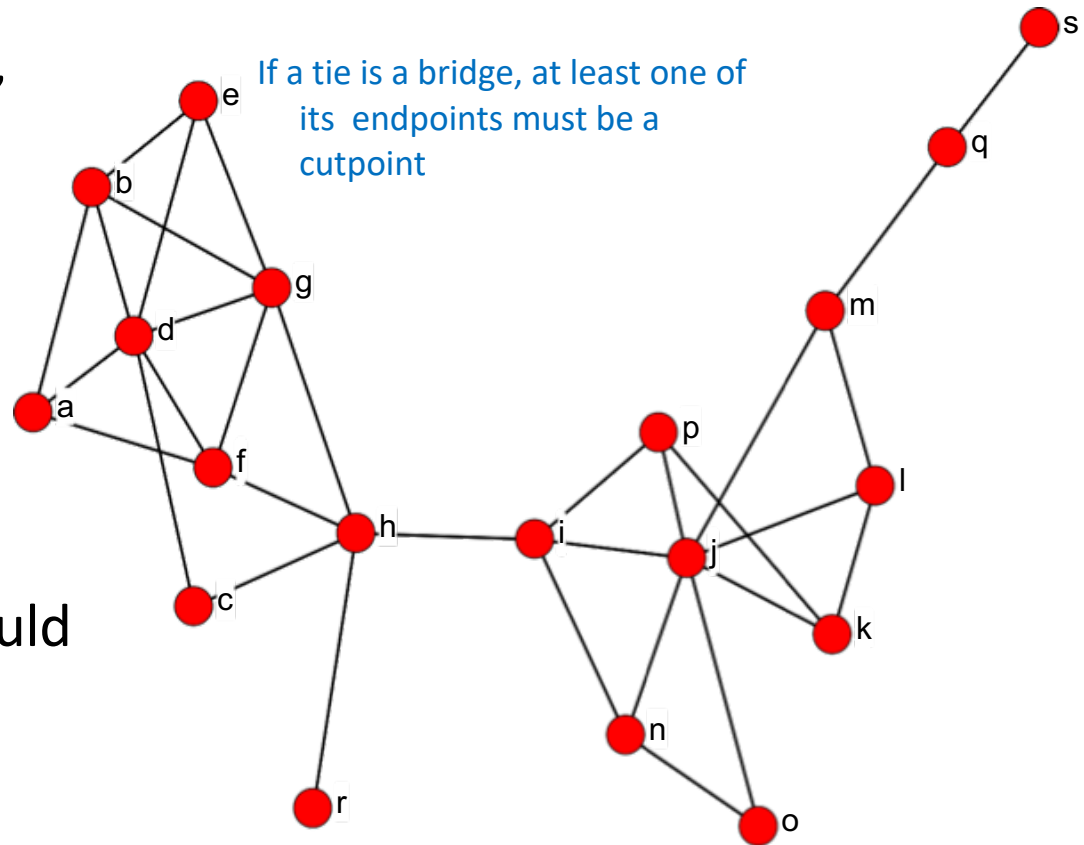
- Strong component
  - There is a directed path from each member of the component to every other
- Weak component
  - There is an undirected path (a weak path) from every member of the component to every other
  - Is like ignoring the direction of ties – driving the wrong way if you have to

# A graph with 1 Weak Component & 4 Strong components



# Cutpoints and Bridges

- **Cutpoint**
  - A node which, if deleted, would increase the number of components
- **Bridge**
  - A tie that, if removed, would increase the number of components



# Data Representation

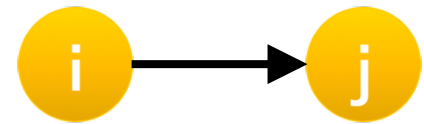
- Adjacency matrix
- Edgelist
- Adjacency/node list



# data representation – adjacency matrix

- Representing edges (who is adjacent to whom) as a matrix

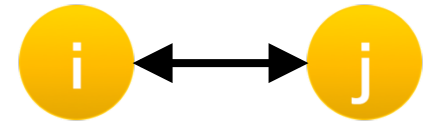
- $A_{ij} = 1$  if node  $i$  has an edge to node  $j$   
= 0 if node  $i$  does not have an edge to  $j$



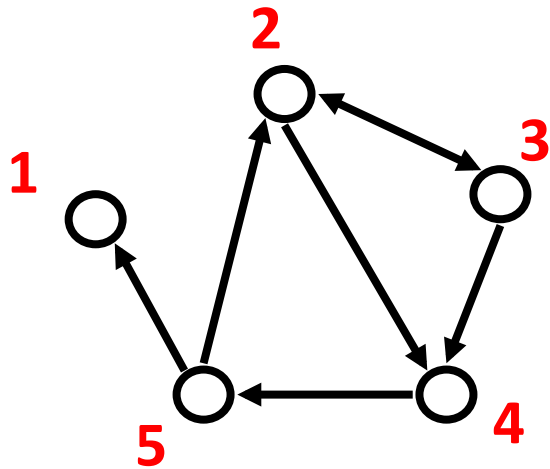
- $A_{ii} = 0$  unless the network has self-loops



- $A_{ij} = A_{ji}$  if the network is undirected, or if  $i$  and  $j$  share a reciprocated edge



# data representation – adjacency matrix



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

Issues:

1. Your dataset will likely contain network data in a non-matrix format;
2. Large, sparse networks take way too much space if kept in a matrix format

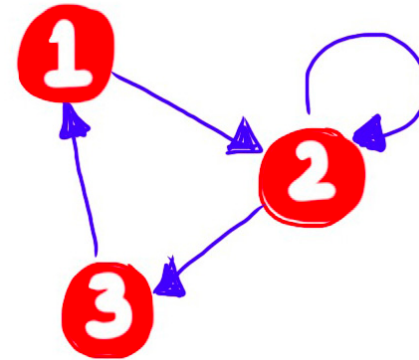
# data representation – adjacency matrix

Which adjacency matrix represents this network?

A) 
$$\begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

B) 
$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

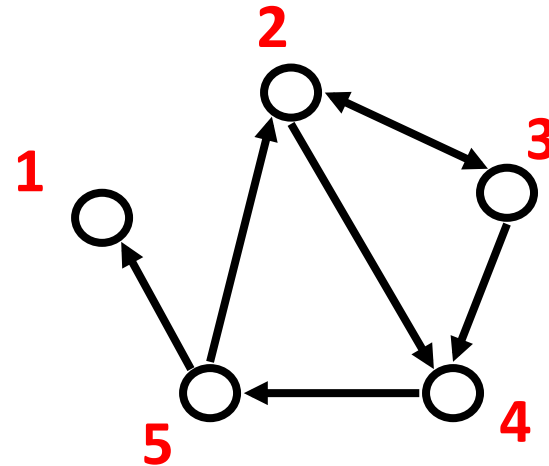
C) 
$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$



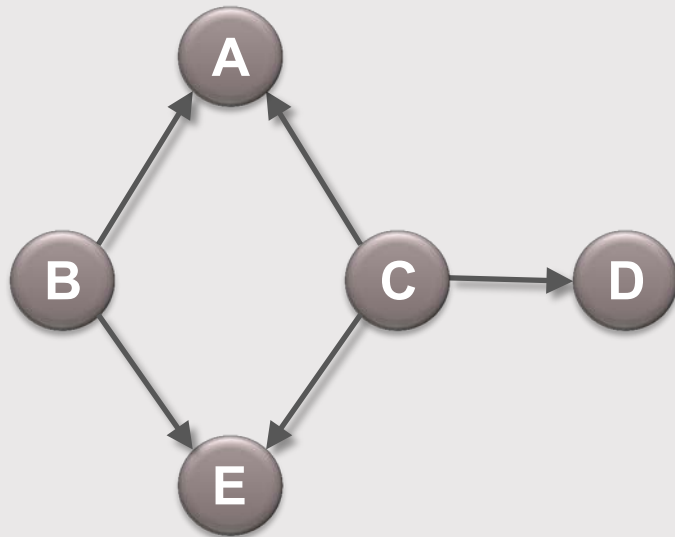
# data representation – edgelist

## ■ Edge list

- 2, 3
- 2, 4
- 3, 2
- 3, 4
- 4, 5
- 5, 2
- 5, 1



# data representation – edgelist with weights



**Source Destination Weight**

**B    A    1**

**B    E    1**

**C    A    1**

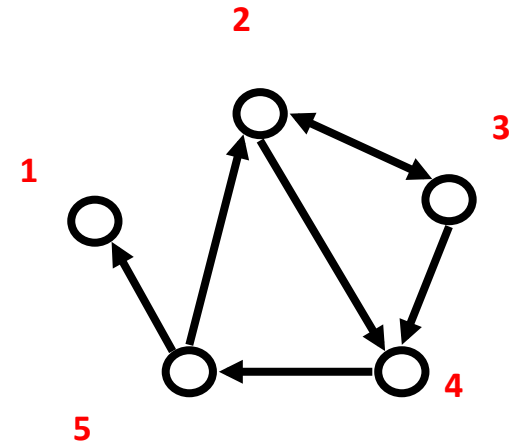
**C    E    1**

**C    D    1**

**Note: Weights are optional.**

# data representation – nodelist

- Adjacency list
  - is easier to work with if network is
    - large
    - sparse
  - quickly retrieve all neighbors for a node
    - 1:
    - 2: 3 4
    - 3: 2 4
    - 4: 5
    - 5: 1 2



# Network Data Collection

## Ego Networks

- Can use standard sampling techniques (e.g. random sample)
- Each respondent describes their own relationships (name generators).

## Complete Networks

- Boundary specification?
- Each respondent reports their own relationships within the network.
- Could use a roster that people use to identify contacts.

## Cognitive Social Structures

- Ask not only for a person's own relationships, but also for perceived relationships between other people in your population.

## Snowball Sampling

- Individuals included in the sample identify contacts (friends, sexual partners, etc.) who are added to the study at the next step.
- Often used in preventive medicine.

## Secondary Data

- Digital traces, social media, hyperlink networks and many more.

# **Centrality, Clustering &Transitivity**



# networks are complex

Can we understand them better without a “ridiculogram”?



## simplifying networks – undirected graph

Consider a classroom with 30 students. How many different possible networks could exist to represent the friendships in that classroom?

# simplifying networks – undirected graph

Consider a classroom with 30 students. How many different possible networks could exist to represent the friendships in that classroom?

(1) How many possible ties ( $N = 30$ )?

$$n^C k = \frac{n!}{(n-k)!k!}$$
$$30^C 2 = \frac{30!}{(30-2)!2!} = \frac{870}{2} = 435 \quad \text{possible dyads/edges}$$

Equivalent to binomial coefficient:

$$\frac{n!}{k!(n-k)!} \equiv \frac{n(n-1)}{2} \quad \text{for } k = 2$$

(2) How many possible graphs (not digraphs)?

Therefore, a set with  $\binom{n}{2}$  pairs of distinct points, if loops/multiple edges are not allowed, each pair determines one possible edge (2 for directed graphs).

This set has  $2^{\binom{n}{2}}$  possible graphs.

That's equivalent to:

$$2^{\frac{n(n-1)}{2}} = 2^{435} > [2^{158}, \dots, 2^{246}]$$

# simplifying networks – undirected graph

Consider a classroom with 30 students. How many different possible networks could exist to represent the friendships in that classroom?

(1) How many possible ties ( $N = 30$ )?

$${}^{30}C_2 = \frac{30!}{(30-2)!2!} = \frac{870}{2} = 435 \text{ possible dyads/edges}$$

Equivalent to:

$$\frac{n!}{k!(n-k)!} \equiv \frac{n(n-1)}{2} \quad \text{for } k = 2$$

(2) How many possible graphs (not digraphs)?

$$2^{n(n-1)} = 2^{870}$$

**For undirected networks:** combinatorics;

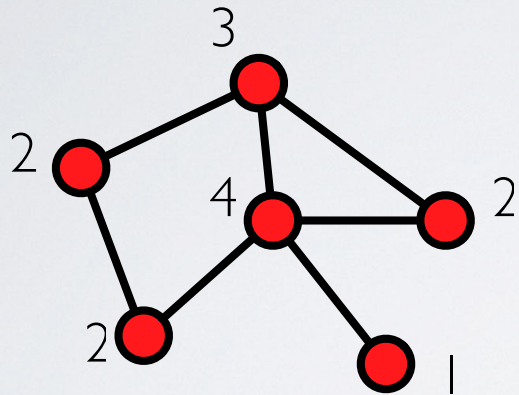
**For directed networks:** permutations;

# How do we determine who is “important” in a Network?

How to describe an individual position in the network?

- **Degree (number of connections)**
- Clustering
- Distance to other nodes
- Centrality, influence, power

# describing networks



**degree:**

number of connections  $k$

$$k_i = \sum_j A_{ij}$$

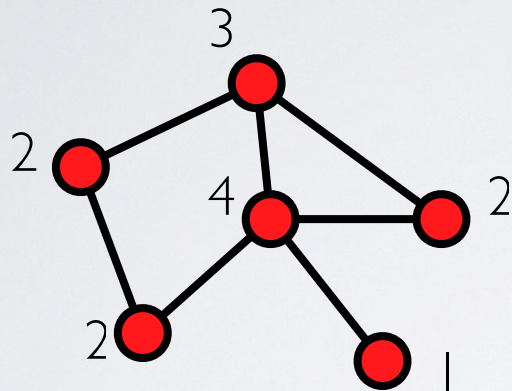
number of edges

$$m = \frac{1}{2} \sum_{i=1}^n k_i = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n A_{ij} = \frac{1}{2} \sum_{j=1}^n \sum_{i=1}^n A_{ji}$$

mean degree

$$\langle k \rangle = \frac{1}{n} \sum_{i=1}^n k_i = \frac{2m}{n}$$

# describing networks



**degree:**

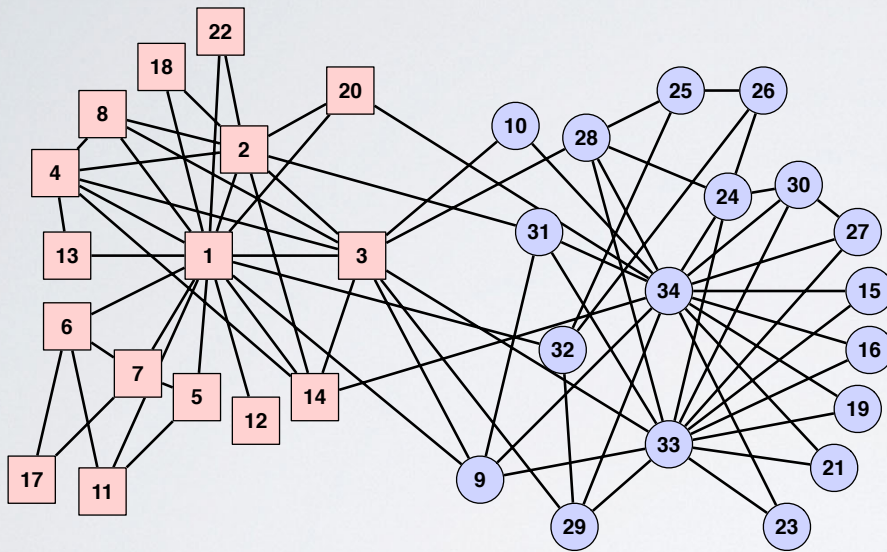
number of connections  $k$

$$k_i = \sum_j A_{ij}$$

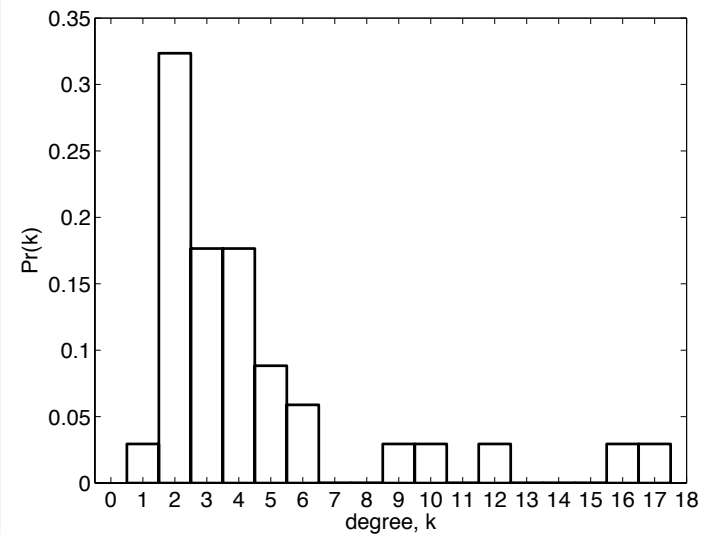
degree sequence  $\{1, 2, 2, 2, 3, 4\}$

degree distribution  $\Pr(k) = \left[ \left(1, \frac{1}{6}\right), \left(2, \frac{3}{6}\right), \left(3, \frac{1}{6}\right), \left(4, \frac{1}{6}\right) \right]$

# describing networks

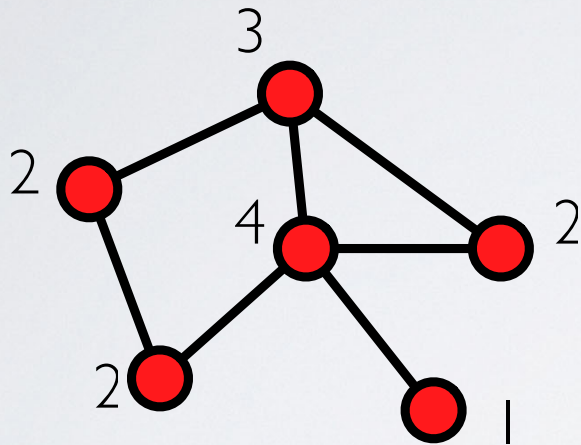


Zachary karate club\*





## describing networks



**degree:**

number of connections  $k$

$$k_i = \sum_j A_{ij}$$

**when does node  
degree matter?**

# describing networks

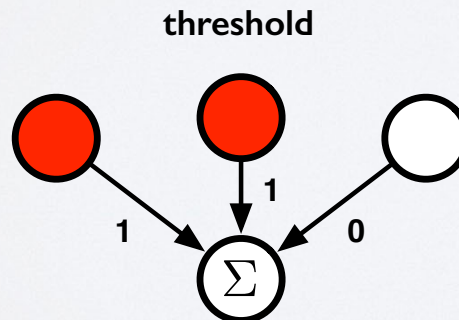
## spreading processes on networks

biological (diseases)

- SIS and SIR models

social (information)

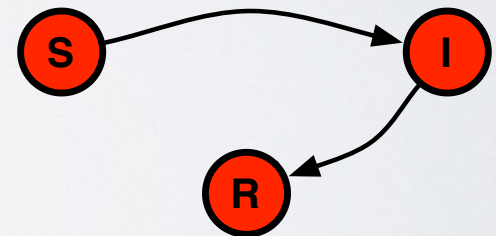
- SIS, SIR models
- threshold models



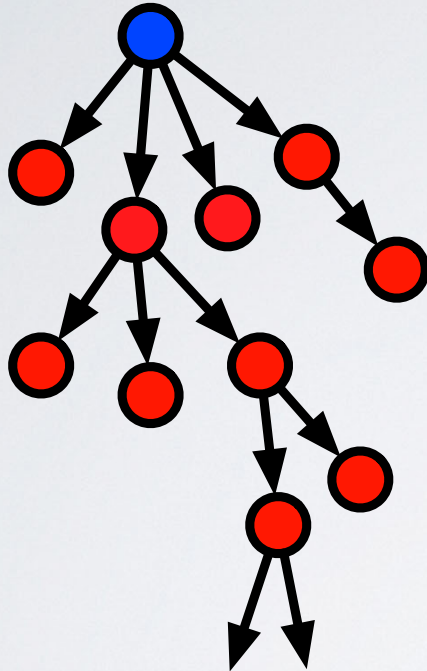
susceptible-infected-susceptible



susceptible-infected-recovered



## describing networks



$$R_0 = 0.923 \dots$$

$R_0$  is the basic reproduction number: the number of infected people an infected person can *reproduce*.

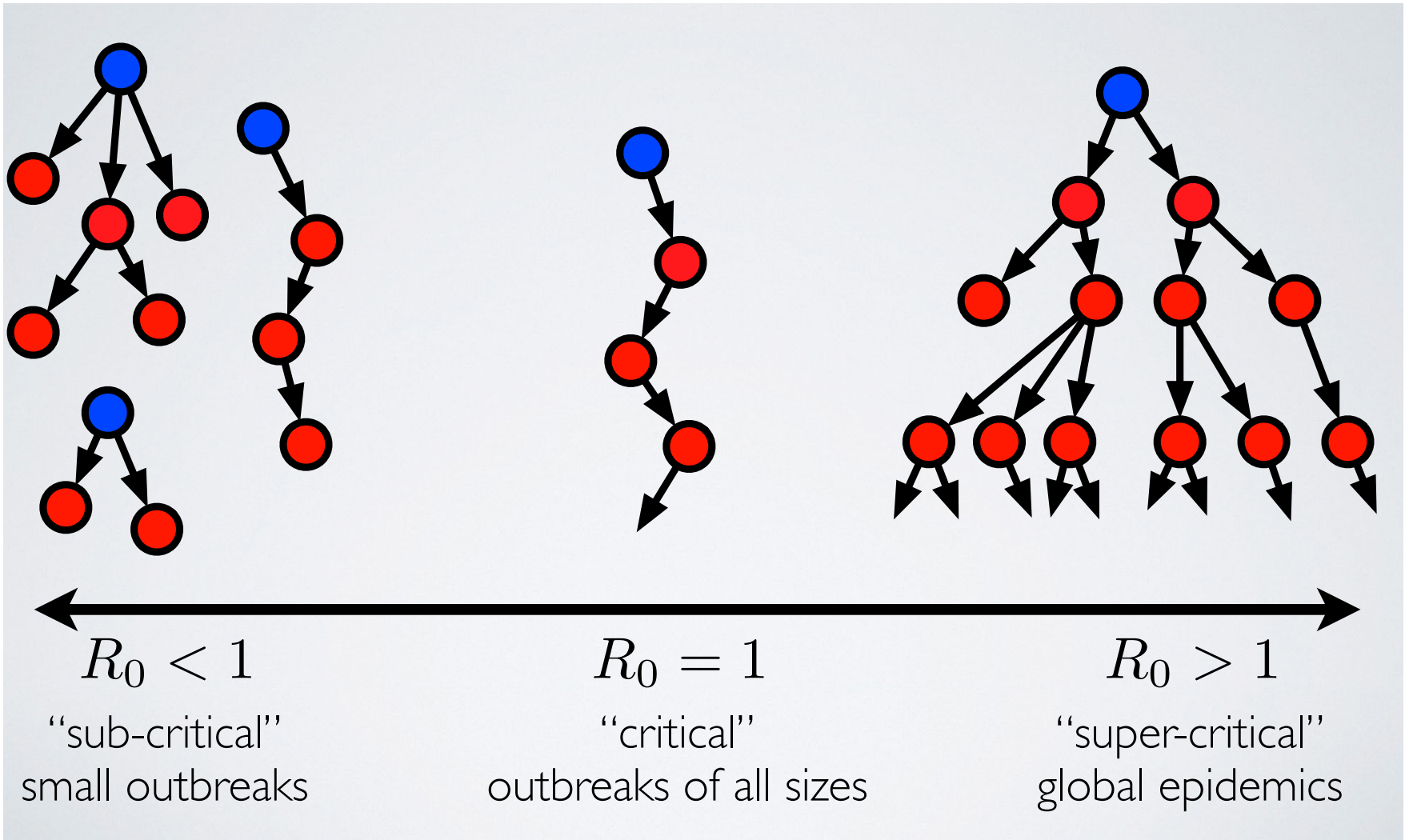
cascade  
epidemic  
branching process  
spreading process

$$R_0 = \text{net reproductive rate} \\ = \text{average degree } \langle k \rangle$$

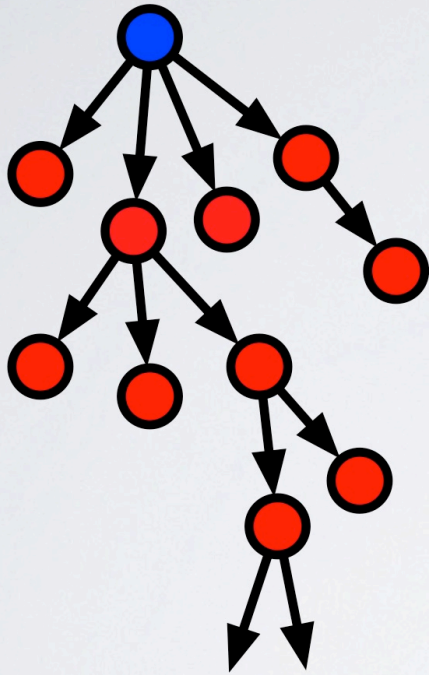
### **caveat:**

ignores network structure,  
dynamics, etc.

## describing networks



## describing networks



### disease

Measles

Chicken pox

Polio

Smallpox

H1N1  
influenza

### R0

5-18

7-12

5-7

1.5-20+

1.0-3.0

### vaccination minimum

90-95%

85-90%

82-87%

70-80%

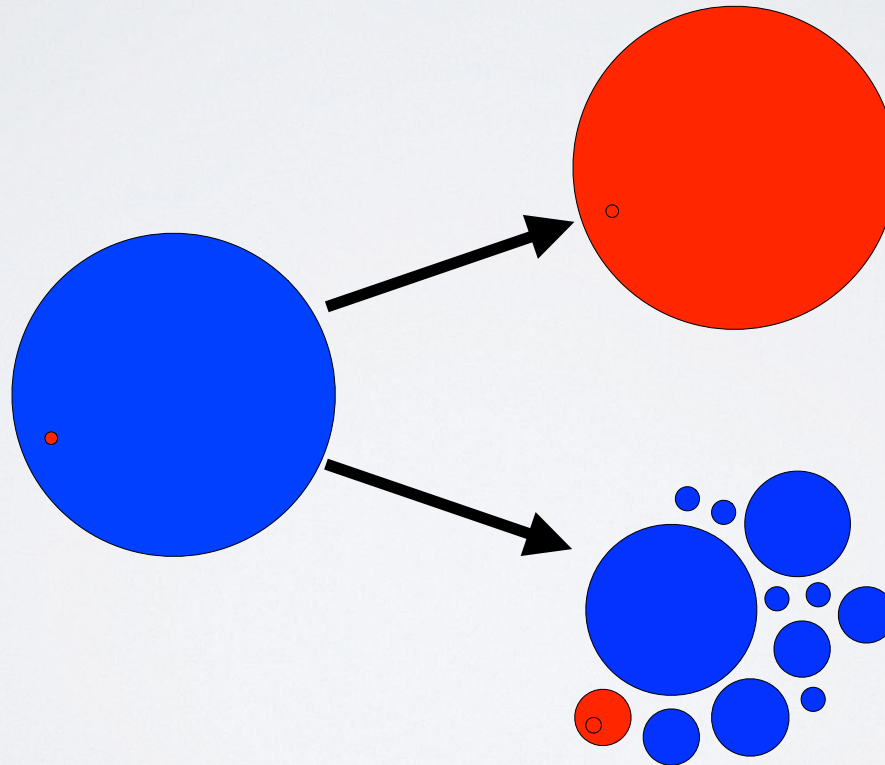
all super-critical



## describing networks

### how could we halt the spread?

- break network into disconnected pieces

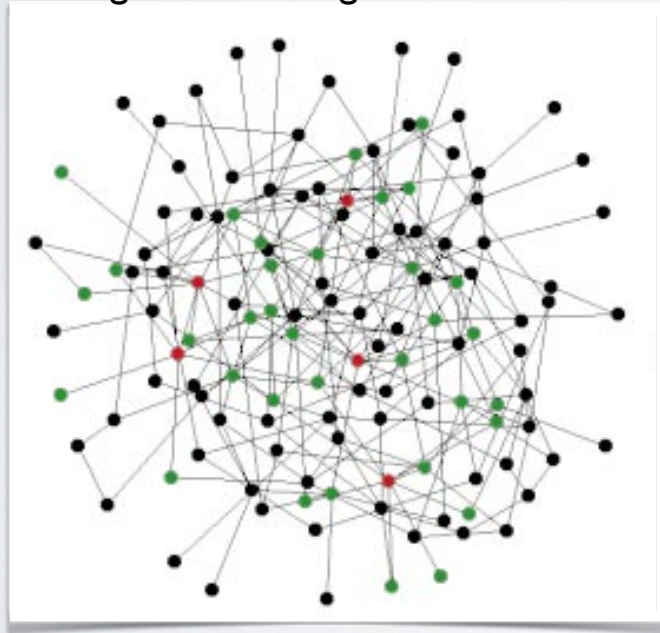


# describing networks

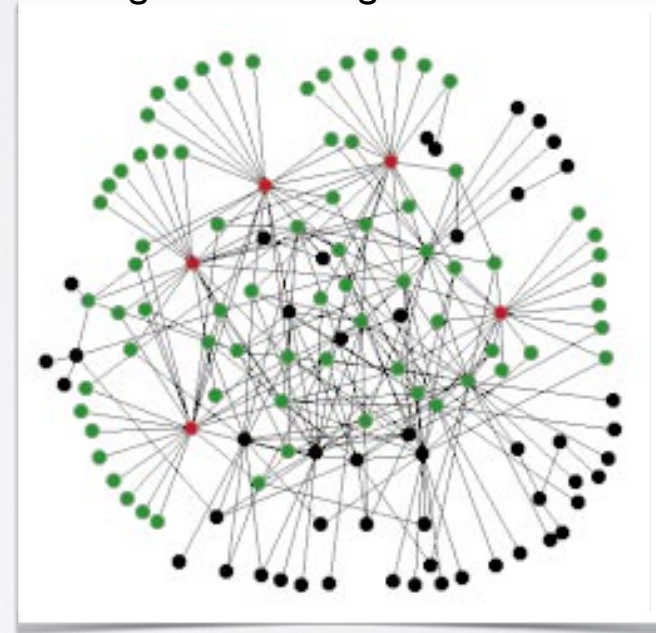
## what promotes spreading?

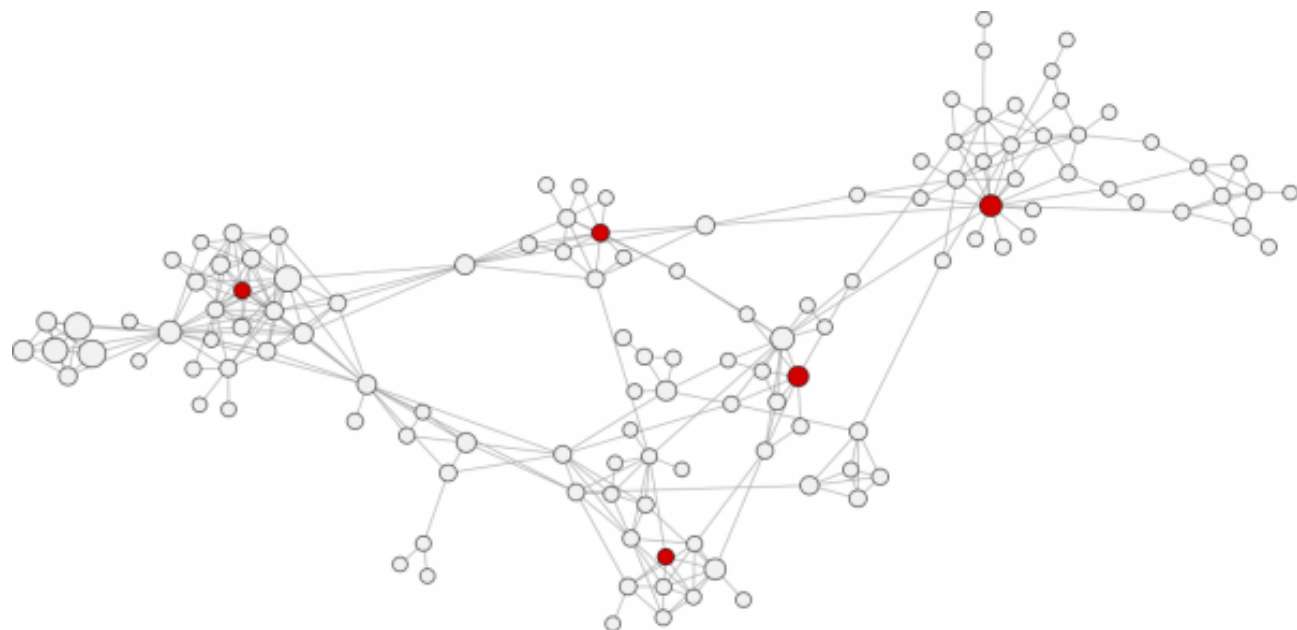
- high-degree vertices\*
- centrally-located vertices

homogeneous in degree

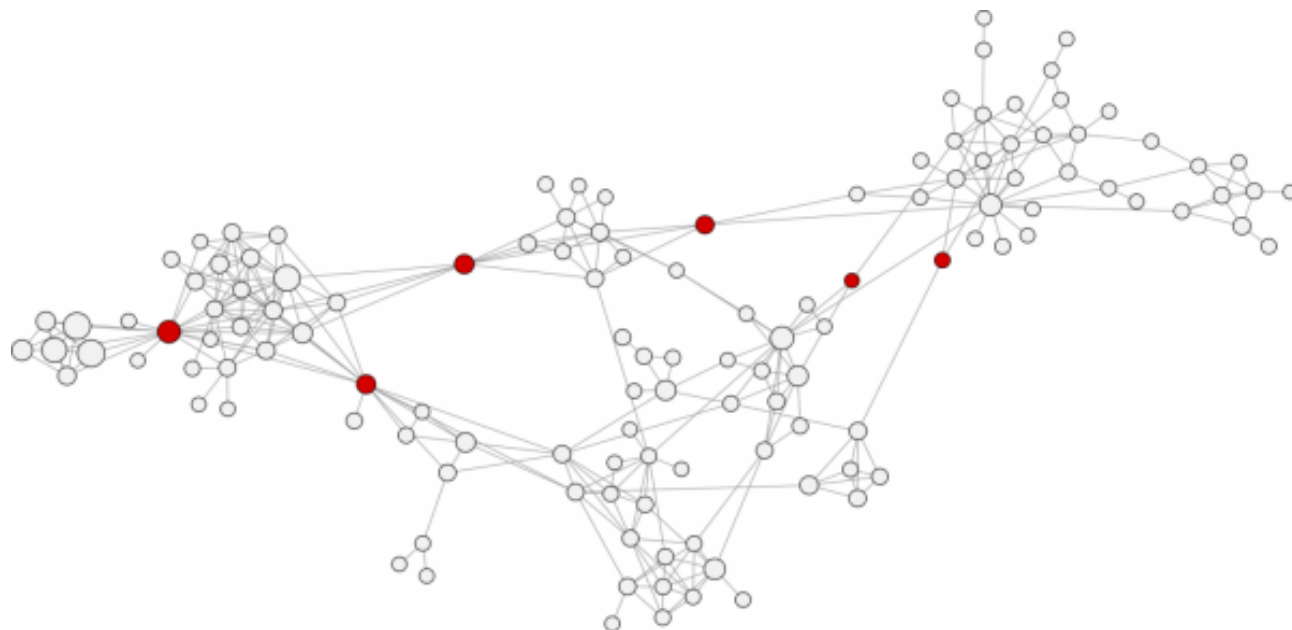


heterogeneous in degree







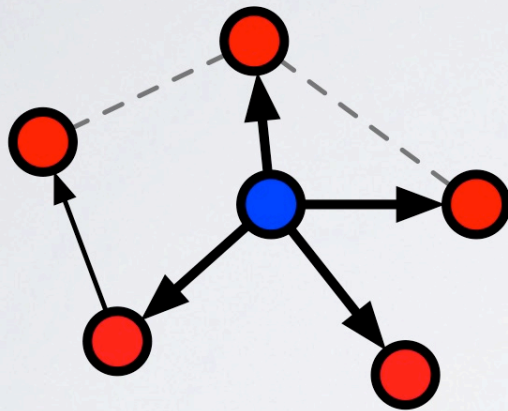




# Centrality (node-level)

- A measure of how network structure and position contributes to a node's importance
- Value associated with every node
- Many different measures which capture different aspects
- Can be characterized by the nature of the flow

# describing networks



**position = centrality:**  
measure of positional  
“importance”

geometric

harmonic centrality

closeness centrality

betweenness centrality

connectivity

degree centrality

eigenvector centrality

PageRank

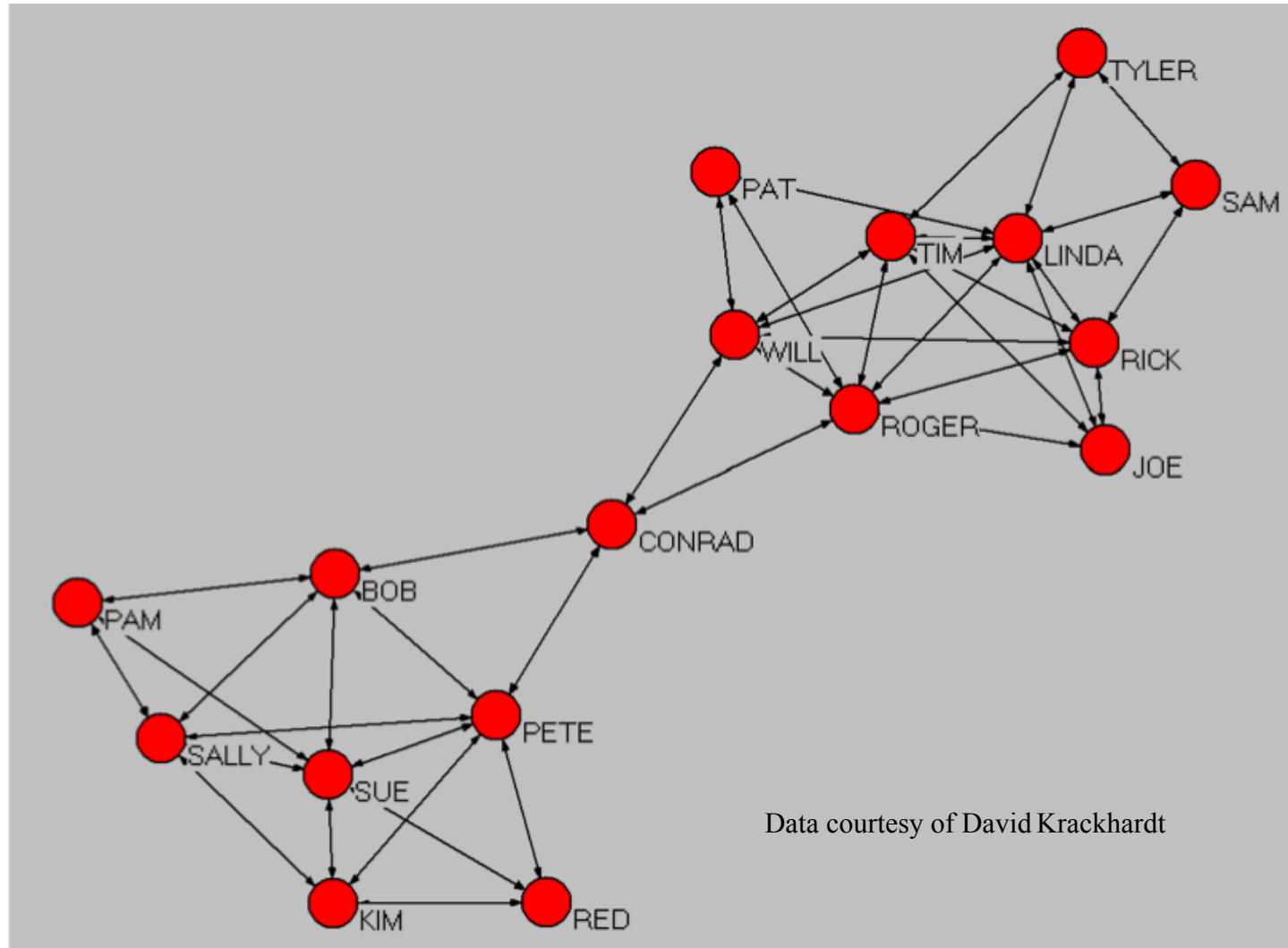
Katz centrality

many many more...

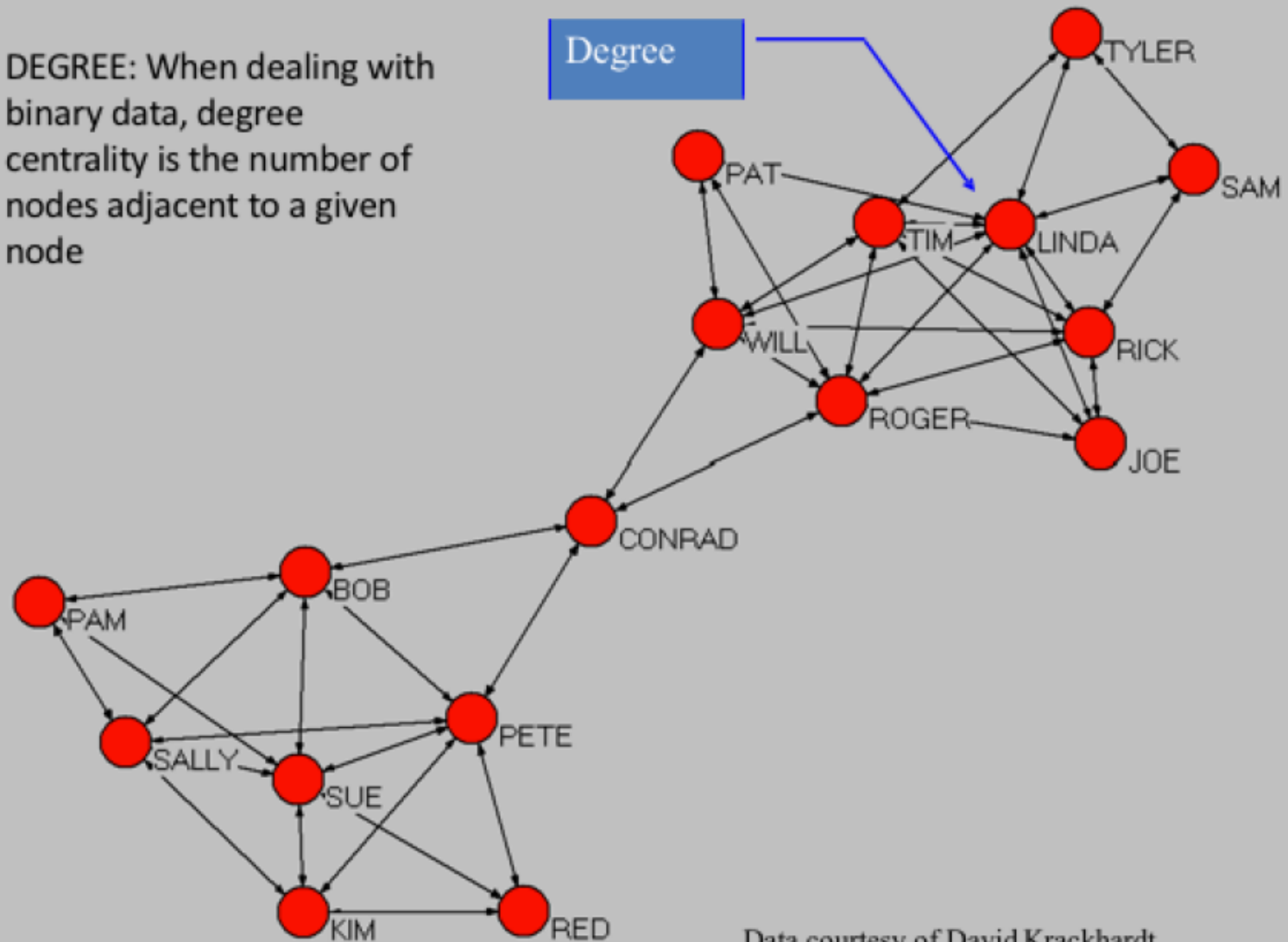
# Centrality measures

- Degree
  - how well connected; direct influence
- Closeness
  - how far from all others
  - how long information takes to arrive
- Betweenness
  - brokerage, gatekeeping, control of info
- Eigenvector
  - being connected to the well connected (a popularity & power measure)

# Who's Important in this network?



DEGREE: When dealing with binary data, degree centrality is the number of nodes adjacent to a given node



Data courtesy of David Krackhardt

# Degree Centrality

- Index of exposure to what is flowing through the network
- Interpreted as opportunity to influence & be influenced directly
- Predicts variety of outcomes from virus resistance to power & leadership to job satisfaction to knowledge



# Degree

	Bob	Con	Joe	Kim	Lin	Pam	Pat	Pete	Red	Rick	Rgr	Sally	Sam	Sue	Tim	Tylr	Will	
Bob	0	1	0	0	0	1	0	1	0	0	0	1	0	1	0	0	0	5
Con	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	4
Joe	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	4
Kim	0	0	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	4
Lin	0	0	1	0	0	0	1	0	0	1	1	0	1	0	1	1	1	8
Pam	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	3
Pat	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	3
Pete	1	1	0	1	0	0	0	0	1	0	0	1	0	1	0	0	0	6
Red	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	3
Rick	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0	1	5
Rgr	0	1	1	0	1	0	1	0	0	1	0	0	0	0	1	0	1	7
Sally	1	0	0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	5
Sam	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	3
Sue	1	0	0	1	0	1	0	1	1	0	0	1	0	0	0	0	0	6
Tim	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1	1	5
Tylr	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	3
Will	0	1	0	0	1	0	1	0	0	1	1	0	0	0	1	0	0	6
	5	4	4	4	8	3	3	6	3	5	7	5	3	6	5	3	6	

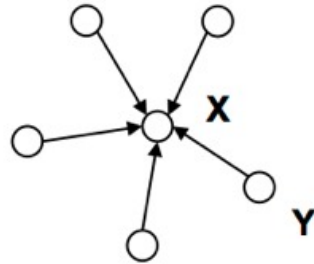
# Degree Centrality with Directed Data

- Indegree- The number of ties directed to the node
- Outdegree- The number of ties that the node directs to others

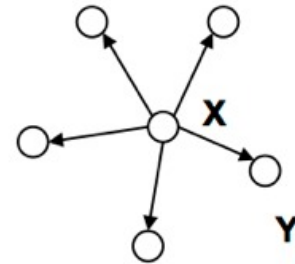
# Degree Centrality with Valued Data

										OUTDEGREE
		MT6	MT71	MT72	MT83	MT93	MT210	MT215	MT272	
MT6		0	100	500	1600	1100	300	2450	1500	7550
MT71		0	0	0	0	0	0	0	0	0
MT72		0	0	0	0	0	0	0	0	0
MT83		0	0	0	0	0	0	0	0	0
MT93		0	0	0	0	0	0	0	0	0
MT210		0	0	0	0	0	0	0	0	0
MT215		0	0	0	0	0	0	0	0	0
MT272		0	0	0	0	0	0	0	0	0
INDEGREE		0	100	500	1600	1100	300	2450	1500	0

NOTE: some software may binarize networks before calculating degree with valued data.



indegree



outdegree

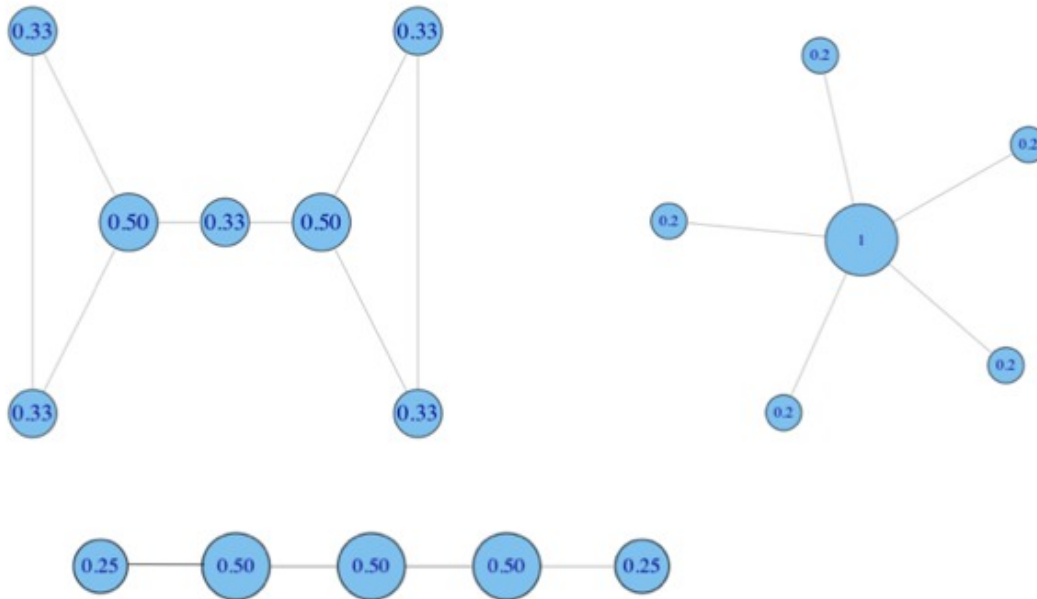
Best measure if importance means:

- how popular you are
- how many people you know

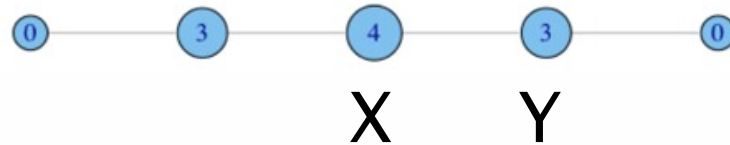
**It is a local measure!**

formula for degree (normalized)

$$C^D(i) = \frac{k_i}{N - 1}$$



degree is not everything...



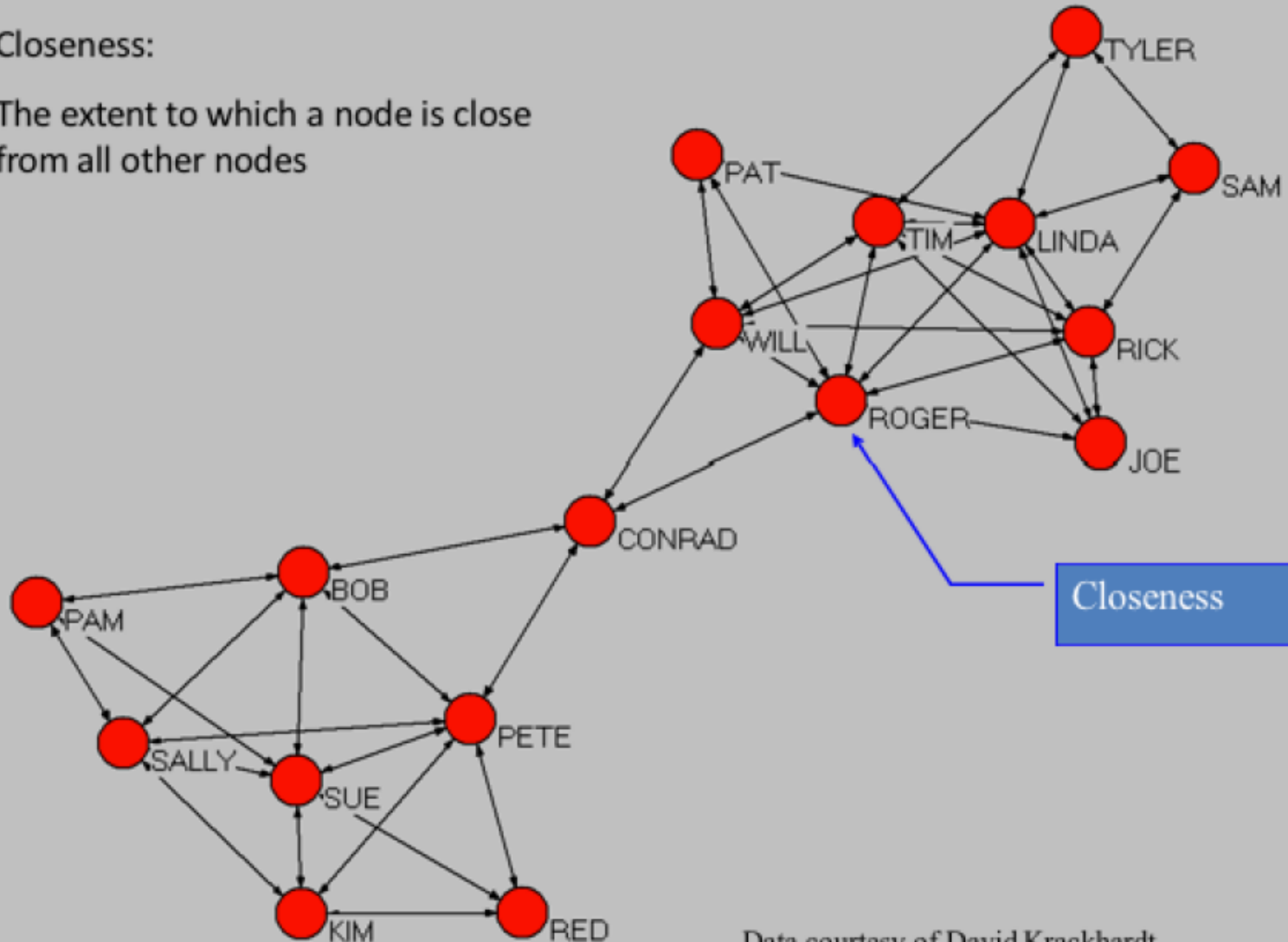
We want to capture:

→ Being close to all nodes

# Closeness

Closeness:

The extent to which a node is close from all other nodes



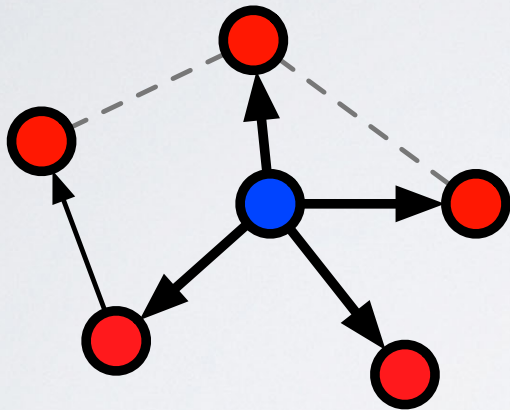
Data courtesy of David Krackhardt

# Closeness Centrality

- Is an inverse measure of centrality
- The extent to which a node is close from all other nodes
- Index of expected time until arrival for a given node of whatever is flowing through the network
  - Gossip network: central player hears things first, on average



# describing networks



**position = centrality:**

harmonic, closeness  
centrality

importance = being in  
“center” of the network

$$\text{harmonic } C_i = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{d_{ij}}$$

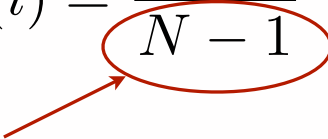
length of shortest path

$$\text{distance: } d_{ij} = \begin{cases} \ell_{ij} & \text{if } j \text{ reachable from } i \\ \infty & \text{otherwise} \end{cases}$$

closeness centrality formula

$$\tilde{C}^C(i) = \left[ \sum_{j=1}^N d(i, j) \right]^{-1}$$

**Normalized**

$$C^C(i) = \frac{\tilde{C}^C(i)}{N-1}$$


All other nodes in the network

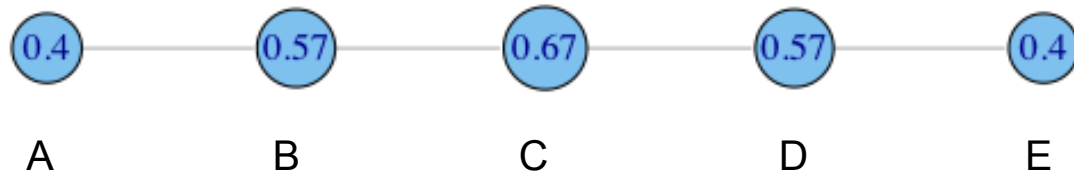
What happens to isolates?

## GEODESIC DISTANCE MATRIX used to Calculate Closeness

	Bob	Con	Joe	Kim	Lin	Pam	Pat	Pete	Red	Rick	Rgr	Sally	Sam	Sue	Tim	Tylr	Will	
Bob	0	1	3	2	3	1	3	1	2	3	2	1	4	1	3	4	2	36
Con	1	0	2	2	2	2	2	1	2	2	1	2	3	2	2	3	1	30
Joe	3	2	0	4	1	4	2	3	4	1	1	4	2	4	1	2	2	40
Kim	2	2	4	0	4	2	4	1	1	4	3	1	5	1	4	5	3	46
Lin	3	2	1	4	0	4	1	3	4	1	1	4	1	4	1	1	1	36
Pam	1	2	4	2	4	0	4	2	2	4	3	1	5	1	4	5	3	47
Pat	3	2	2	4	1	4	0	3	4	2	1	4	2	4	2	2	1	41
Pete	1	1	3	1	3	2	3	0	1	3	2	1	4	1	3	4	2	35
Red	2	2	4	1	4	2	4	1	0	4	3	2	5	1	4	5	3	47
Rick	3	2	1	4	1	4	2	3	4	0	1	4	1	4	2	2	1	39
Rgr	2	1	1	3	1	3	1	2	3	1	0	3	2	3	1	2	1	30
Sally	1	2	4	1	4	1	4	1	2	4	3	0	5	1	4	5	3	45
Sam	4	3	2	5	1	5	2	4	5	1	2	5	0	5	2	1	2	49
Sue	1	2	4	1	4	1	4	1	1	4	3	1	5	0	4	5	3	44
Tim	3	2	1	4	1	4	2	3	4	2	1	4	2	4	0	1	1	39
Tylr	4	3	2	5	1	5	2	4	5	2	2	5	1	5	1	0	2	49
Will	2	1	2	3	1	3	1	2	3	1	1	3	2	3	1	2	0	31

NOTE: if data is directed, you can calculate in-closeness and out-closeness centrality

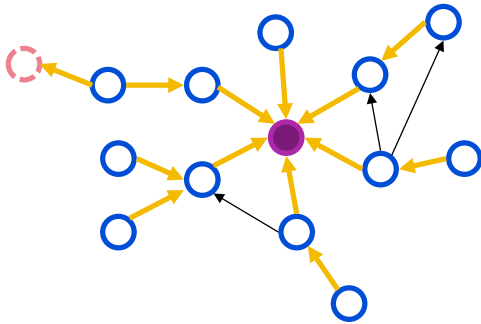
## closeness centrality example



$$C'_c(A) = \left[ \frac{\sum_{j=1}^N d(A, j)}{N-1} \right]^{-1} = \left[ \frac{1+2+3+4}{4} \right]^{-1} = \left[ \frac{10}{4} \right]^{-1} = 0.4$$

# Closeness in directed networks

- choose a direction
  - in-closeness (e.g. prestige in citation networks)
  - out-closeness
- usually consider only vertices from which the node  $i$  in question can be reached

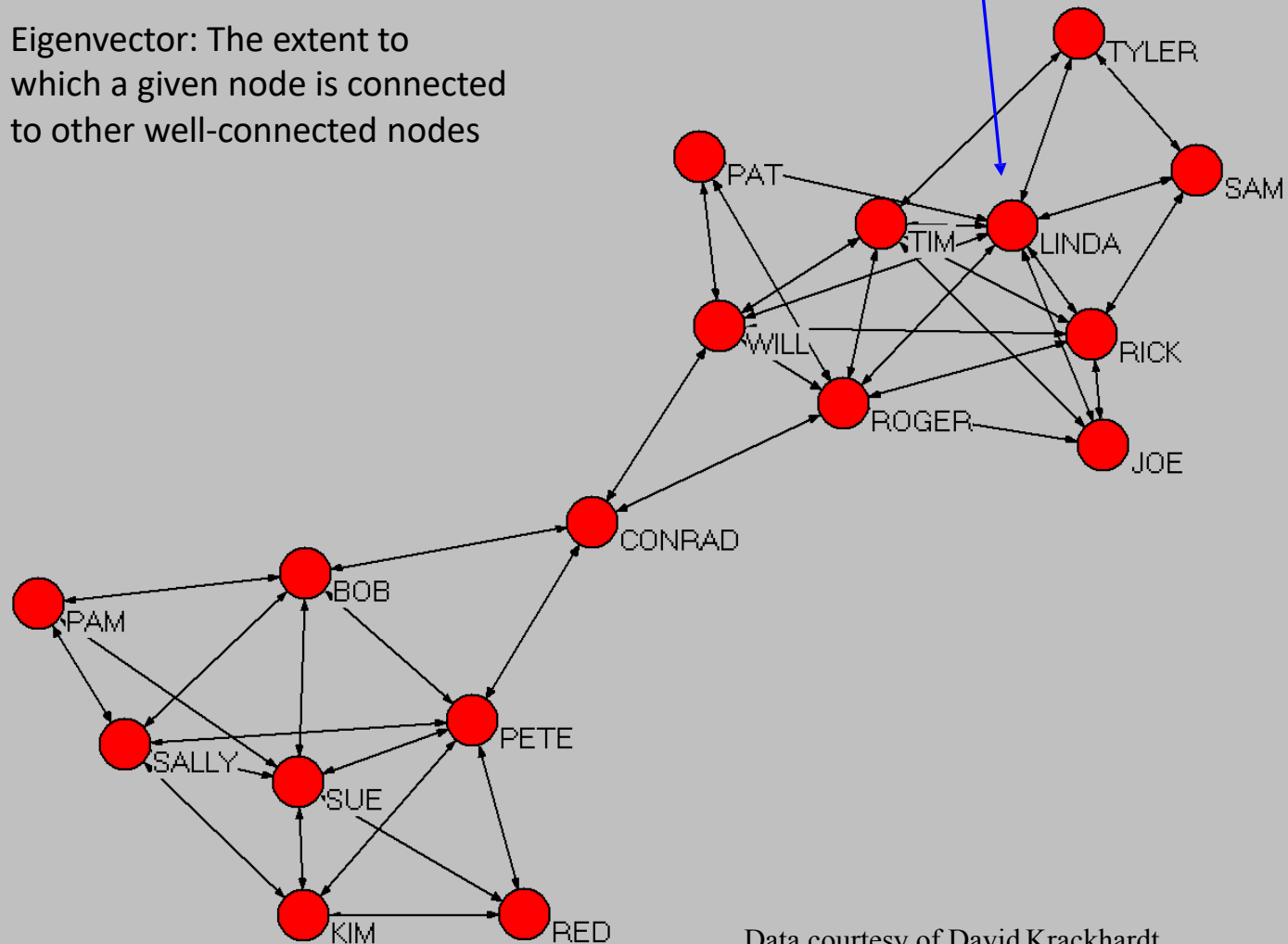


# Closeness in directed networks

- In-Closeness centrality measures the degree to which a node can be easily reached \*from\* other nodes (i.e. using edges coming in towards the node) where easily means shortest distance.
- Out-Closeness centrality measures the degree to which a node can easily reach other nodes (i.e. using edges out from the node), and easily again means shortest distance.
- If there **is no (directed) path** between vertex  $v$  and  $i$  then the total number of vertices is used in the formula instead of the path length.

## Eigenvector

Eigenvector: The extent to which a given node is connected to other well-connected nodes



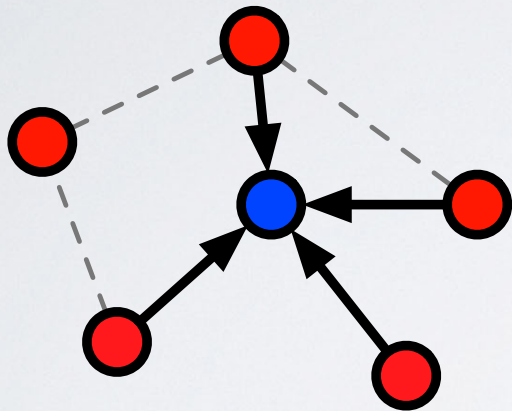
Data courtesy of David Krackhardt

# Eigenvector Centrality

- Node has high score if connected to many nodes that are themselves well connected  
*(you are important to the extent your friends are important)*
- Indicator of popularity,
  - Google Page Rank
- Like degree, is index of exposure, risk
- However, tends to identify centers of large cliques



# eigenvector centrality



## **position = centrality:**

PageRank, Katz, eigenvector centrality

importance = sum of importances\* of nodes that point at you

$$I_i = \sum_{j \rightarrow i} \frac{I_j}{k_j}$$

or, the left eigenvector of

$$\mathbf{Ax} = \lambda \mathbf{x}$$

A node is important if it is connected  
to important nodes

$$X_i = \sum_{j \in \Lambda(i)} X_j \quad X_i = \sum_{j=1}^N A_{ij} X_j \quad \boxed{AX = \lambda X}$$

The solution (when exists) gives the node  
centrality. We take the highest  $\lambda$

Note: **Bonacich eigenvector centrality** includes a parameter  $\beta$  which allows one to adjust how important are neighbours in different path lengths to a node's centrality versus how important is the number of neighbours in path length = 1; high  $\beta$  leads to low attenuation and the global network structure matters; low  $\beta$  yields high attenuation and only the immediate friends matter. When  $\beta = 0$ , equivalent to degree centrality.

## Bonacich eigenvector centrality

$$c_i(\beta) = \sum_j (\alpha + \beta c_j) A_{ji}$$

$$c(\beta) = \alpha(I - \beta A)^{-1} A \mathbf{1}$$

- $\alpha$  is a normalization constant
- $\beta$  determines how important the centrality of your neighbors is
- $\mathbf{A}$  is the adjacency matrix (can be weighted)
- $\mathbf{I}$  is the identity matrix (1s down the diagonal, 0 off-diagonal)
- $\mathbf{1}$  is a matrix of all ones.

## Bonacich Power Centrality: attenuation factor $\beta$

small  $\beta \rightarrow$  high attenuation

only your immediate friends matter, and their importance is factored in only a bit

high  $\beta \rightarrow$  low attenuation

global network structure matters (your friends, your friends' of friends etc.)

$\beta = 0$  yields simple degree centrality

$$c_i(\beta) = \sum_j (\alpha \square) A_{ji}$$

## Bonacich Power Centrality: examples

$\beta = .25$

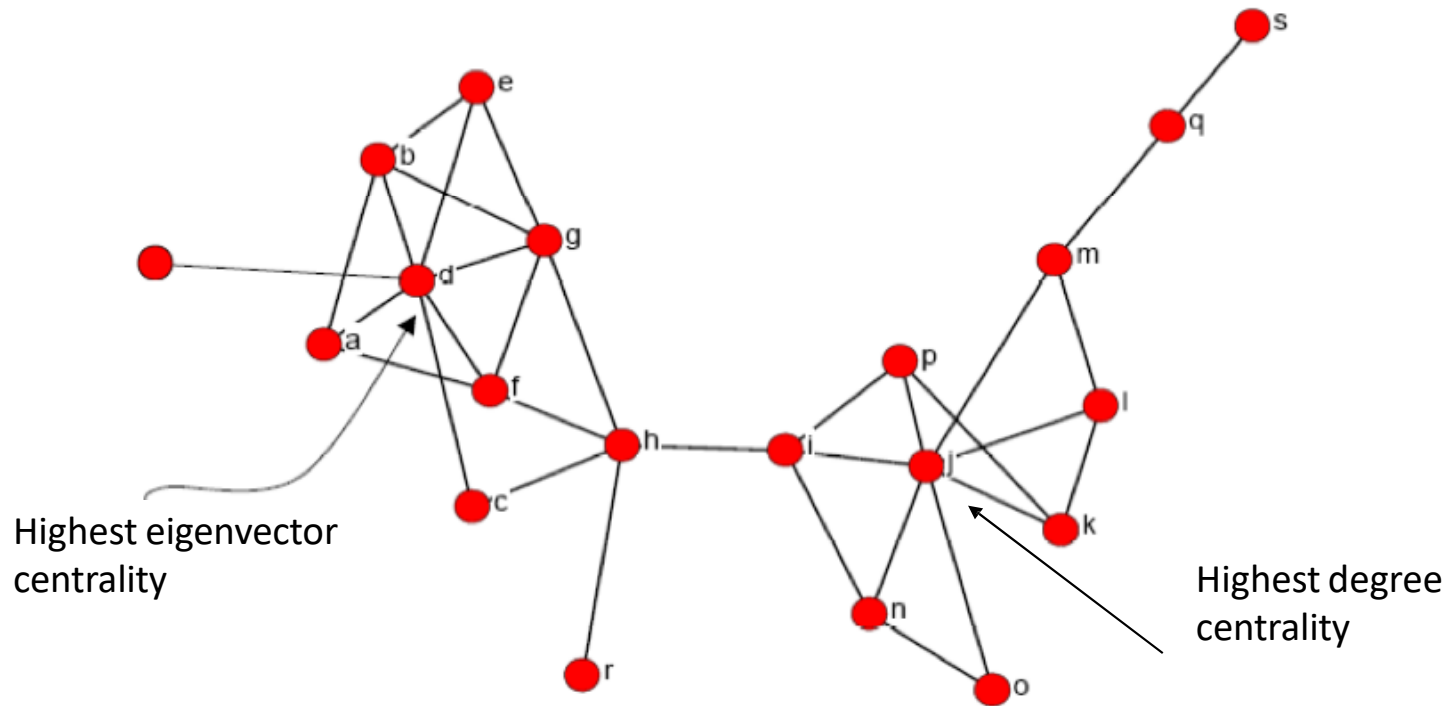


$\beta = -.25$

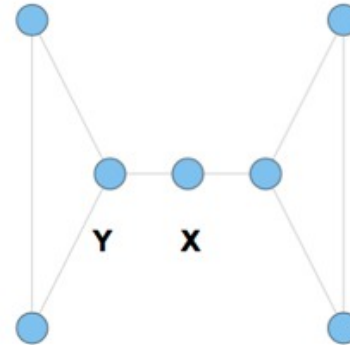
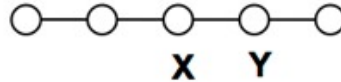
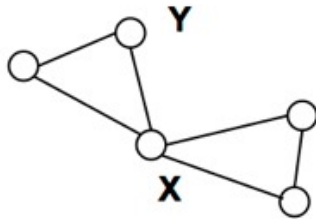


Why does the middle node have lower centrality than its neighbors when  $\beta$  is negative?

NOTE: Node with highest eigenvector centrality is not always node with highest degree centrality



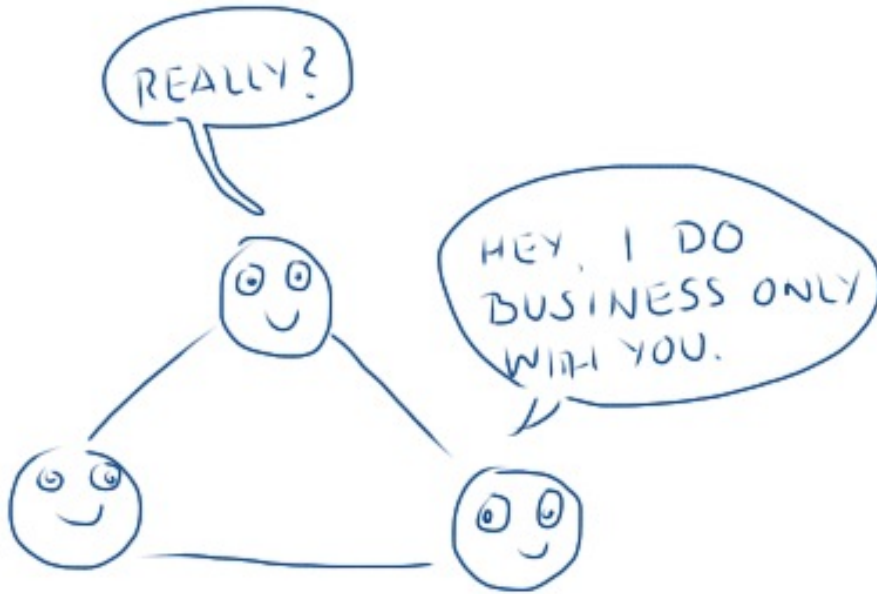
Who is more important in the networks below? X or Y



We want to capture:

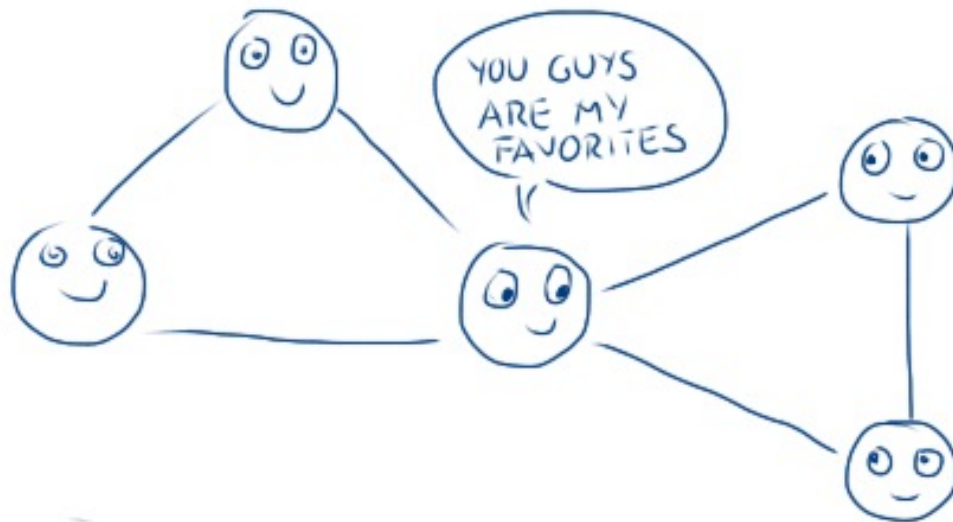
- Ability to broker between groups
- Likelihood that information originating anywhere in the network reaches you

ability to broker between groups



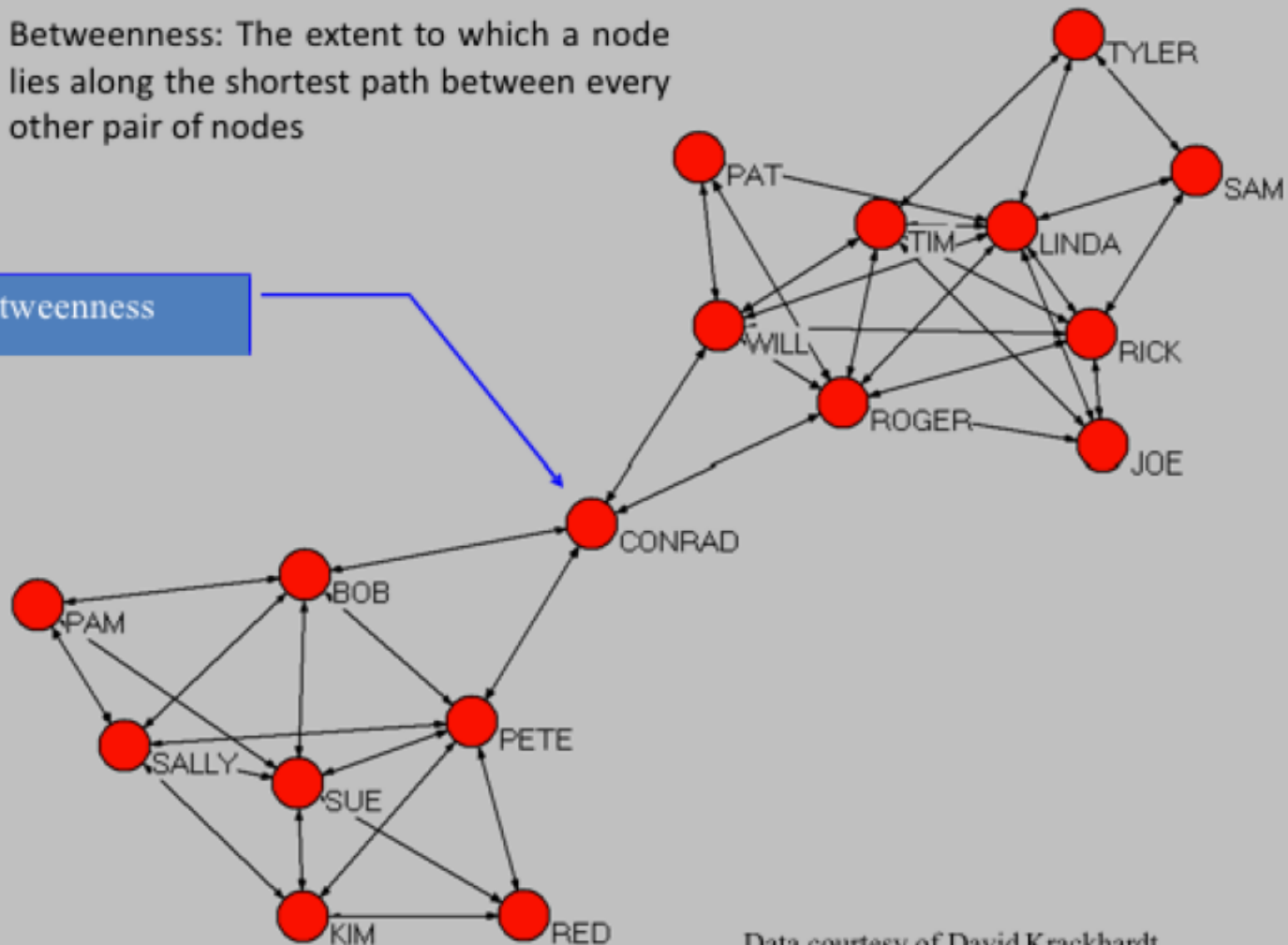


ability to broker between groups



Betweenness: The extent to which a node lies along the shortest path between every other pair of nodes

Betweenness



Data courtesy of David Krackhardt

# Betweenness Centrality

- How often a node lies along the shortest path between two other nodes
- Index of potential for gatekeeping, brokering, controlling the flow, and also of liaising otherwise separate parts of the network
- Interpreted as indicating power and access to diversity of what flows; potential for synthesizing

formula

$$\tilde{C}^B(i) = \sum_{j < k} \frac{d_{jk}(i)}{d_{jk}}$$

$d_{jk}$  # of shortest paths between  $j$  and  $k$   
 $d_{jk}(i)$  # of shortest paths between  $j$  and  $k$  that go through  $i$

Normalized

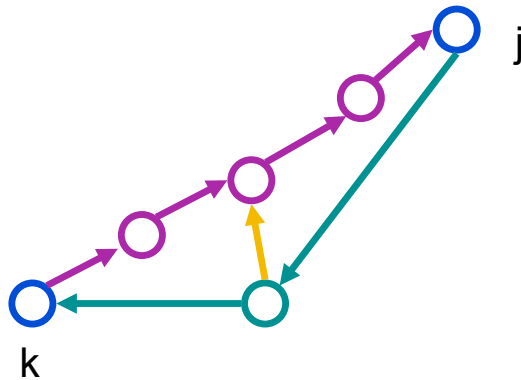
$$C^B(i) = \frac{\tilde{C}^B}{(N-1)(N-2)/2}$$

Number of pairs of vertices excluding  $i$

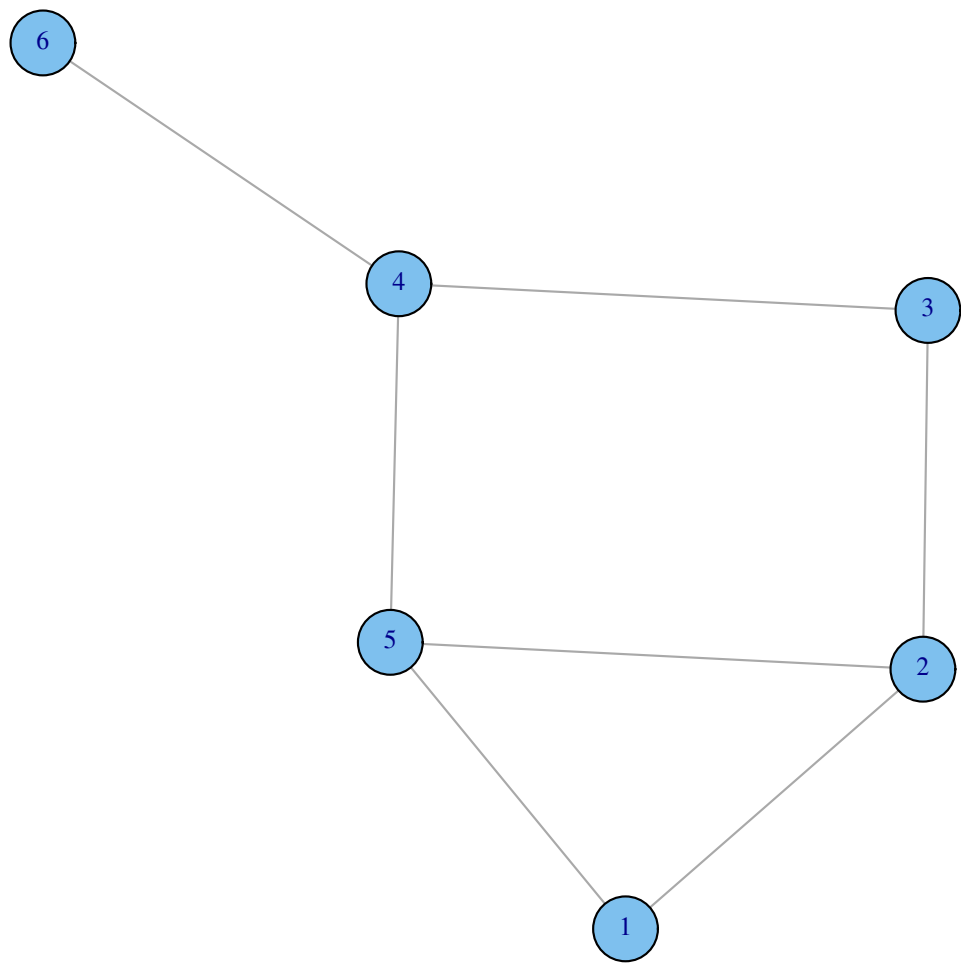
For **directed graphs**: when normalizing, we have  $(N-1)*(N-2)$  instead of  $(N-1)*(N-2)/2$ , because we have twice as many ordered pairs as unordered pairs.

## Betweenness in directed networks

- A node does not necessarily lie on a geodesic (shortest path) from  $j$  to  $k$  if it lies on a geodesic from  $k$  to  $j$



# Betweenness in directed networks

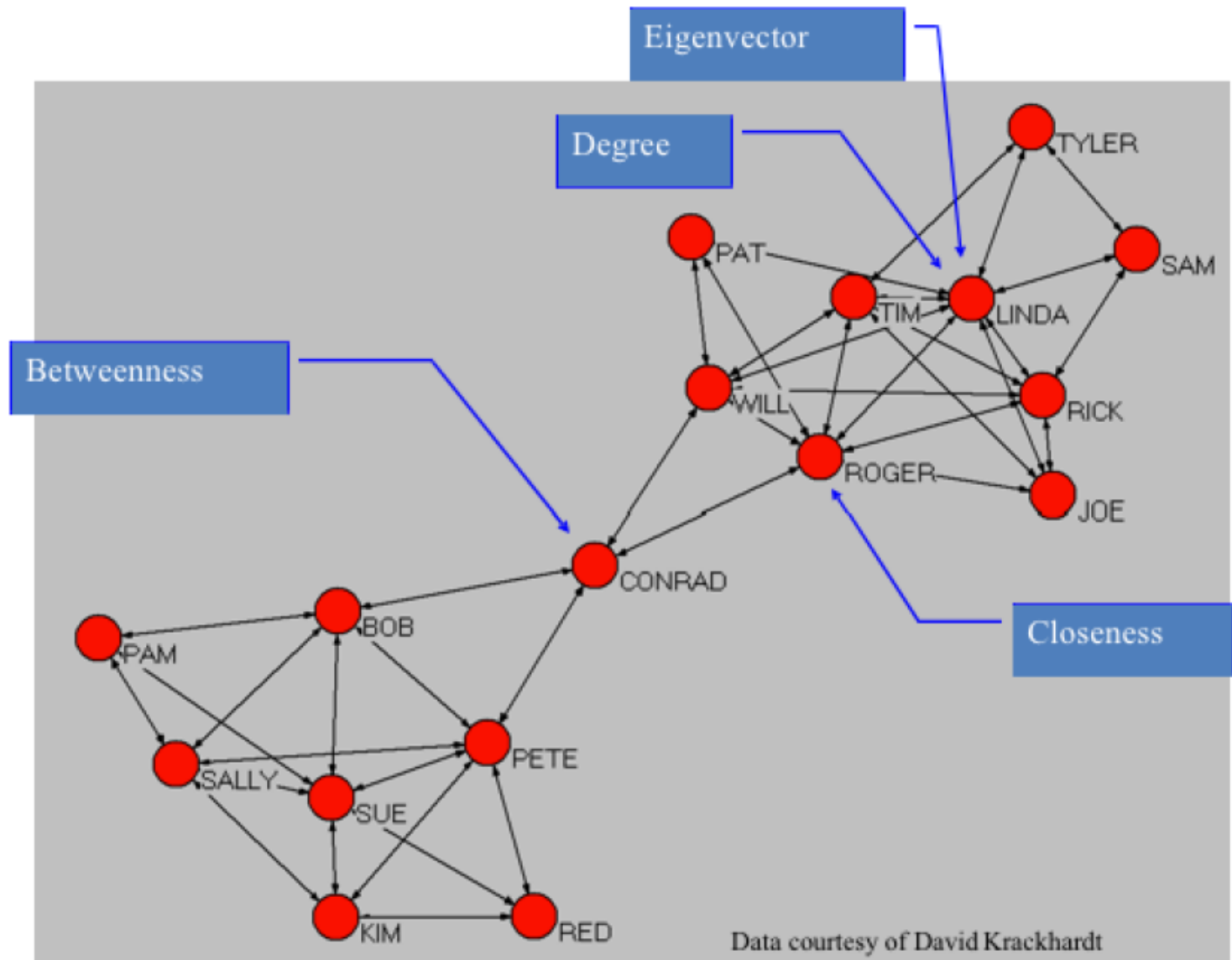


# Betweenness in directed networks

- For example: for node 2, the  $(n - 1)(n - 2)/2 = 5(5 - 1)/2 = 10$  terms in the summation in the order of 13, 14, 15, 16, 34, 35, 36, 45, 46, 56 are

$$\frac{1}{1} + \frac{0}{1} + \frac{0}{1} + \frac{0}{1} + \frac{0}{1} + \frac{1}{2} + \frac{0}{1} + \frac{0}{1} + \frac{0}{1} + \frac{0}{1} = 1.5.$$

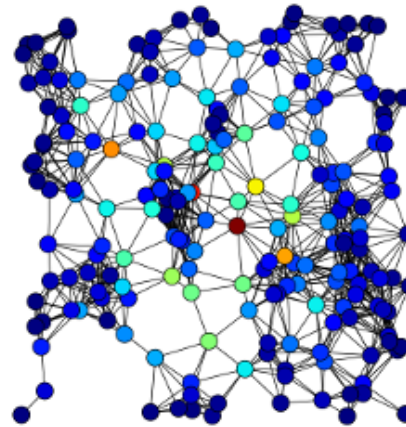
- Here the denominators are the number of shortest paths between pair of edges in the above order and the numerators are the number of shortest paths passing through edge 2 between pair of edges in the above order.



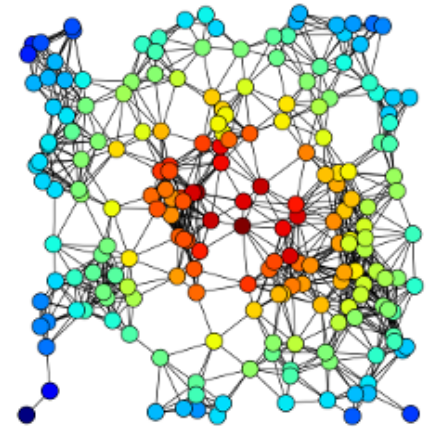


Centrality indices are answers to the question "What characterizes an important node?"

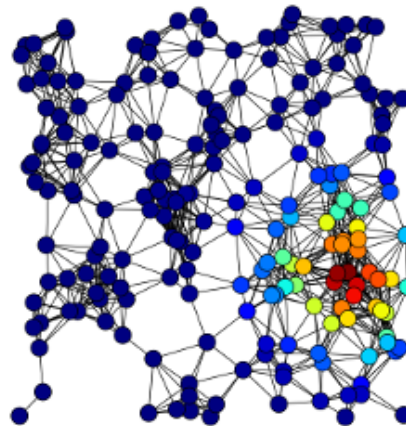
The word "importance" has a wide number of meanings, leading to many different definitions of centrality.



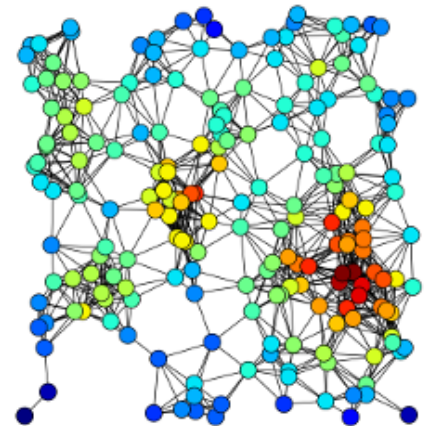
A Betweenness



B Closeness



C Eigenvector



D Degree

# Which nodes are most “central”?

Definition of ‘central’ varies by context/  
purpose.

## Local measure:

→ degree

## Relative to the rest of network:

→ betweenness

→ closeness

→ eigenvector (Bonacich power  
centrality)

# Data Types: Centrality

	<b>Disconnected or Connected</b>	<b>Binary or Valued</b>	<b>Directed or Undirected</b>
<b>Degree</b>	Both	Both	Both
<b>Closeness</b>	Strongly Connected	Binary	Both
<b>Betweenness</b>	Both	Binary	Both
<b>Eigenvector</b>	Connected	Both	<b>Undirected</b>

# check your understanding

- generally different centrality metrics will be positively correlated
- when they are not, there is likely something interesting about the network
- suggest possible topologies and node positions to fit each square

	Low Degree	Low Closeness	Low Betweenness
High Degree			
High Closeness			
High Betweenness			

- generally different centrality metrics will be positively correlated
- when they are not, there is likely something interesting about the network
- suggest possible topologies and node positions to fit each square

	Low Degree	Low Closeness	Low Betweenness
High Degree		Embedded in cluster that is far from the rest of the network	Ego's connections are redundant - communication bypasses him/her
High Closeness	Key player tied to important/active players		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 cell. Would mean that ego monopolizes the ties from a small number of people to many others.	

# Fun Applications of Centrality

- [Oracleofkevinbacon.org](http://Oracleofkevinbacon.org)
  - 6 degrees of Kevin Bacon
  - Can you find anyone with a Bacon score  $> 4$ ?
- [Theyrule.net](http://Theyrule.net)
  - Board overlaps of top corporations
- [Oilmoney.priceofoil.org](http://Oilmoney.priceofoil.org)
  - Tracking petroleum industry campaign contributions

# Global Properties/Graph-level approach:

## Centralization

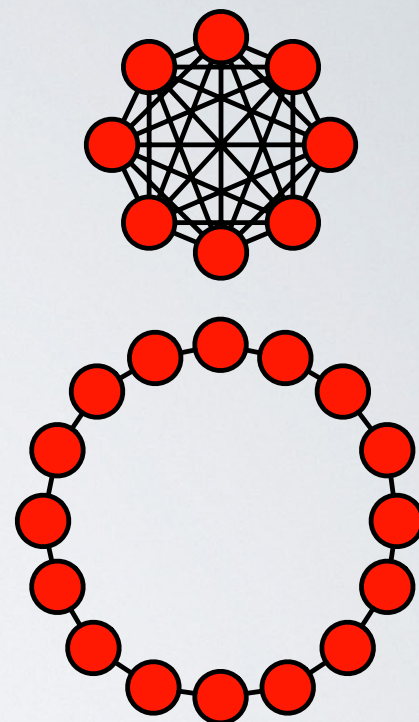
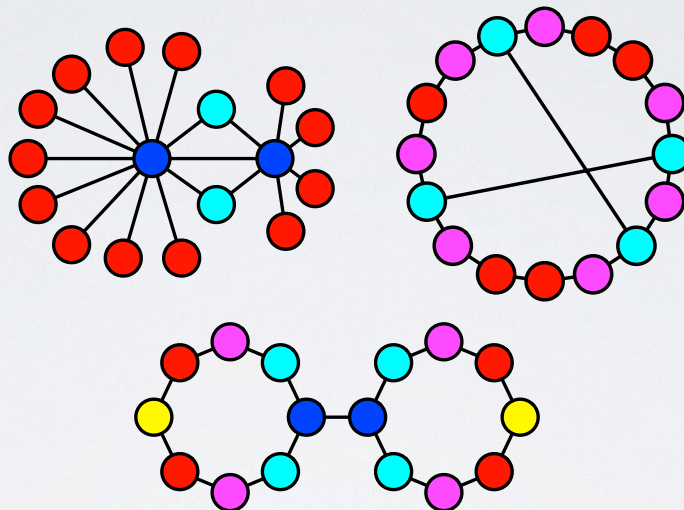
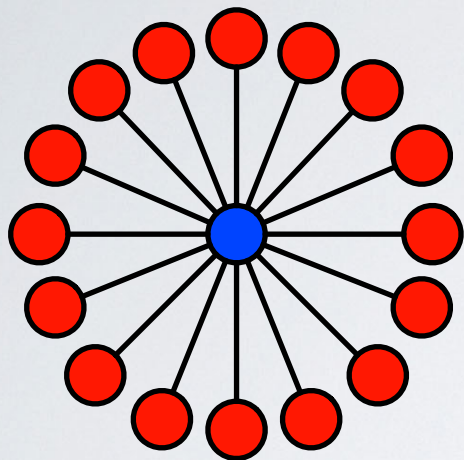
To measure the degree to which the graph as a whole is centralized, we look at *dispersion* of centrality

How much variation is there in the centrality scores among the nodes?

Freeman's general formula for centralization (can use other metrics, e.g. gini coefficient or standard deviation):

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

Diagram illustrating the formula for Centralization ( $C_D$ ). The term  $C_D(n^*)$  is highlighted with a yellow box, and an arrow points to it from a text box labeled "maximum value in the network".



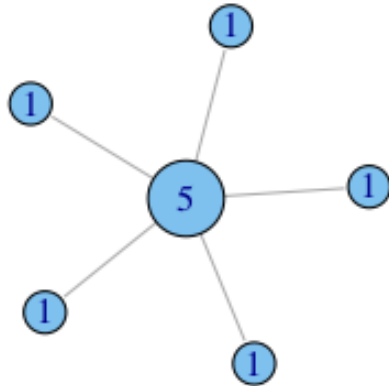
most  
centralized

vast wilderness  
of in-between

most  
decentralized



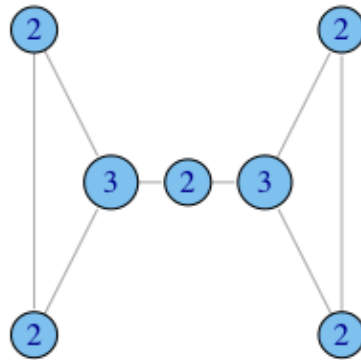
# degree centralization examples



$$C_D = 1.0$$



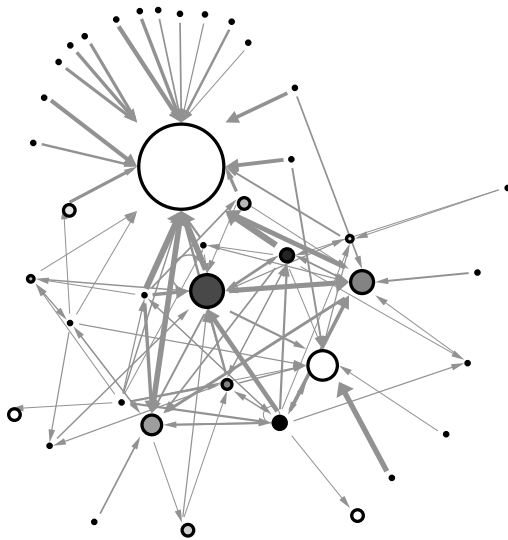
$$C_D = 0.167$$



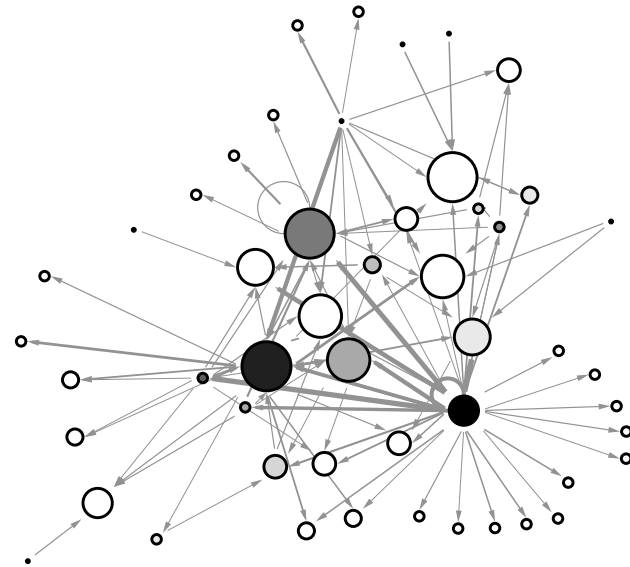
$$C_D = 0.167$$

# real-world networks

## example financial trading networks



high in-centralization:  
one node buying from  
many others



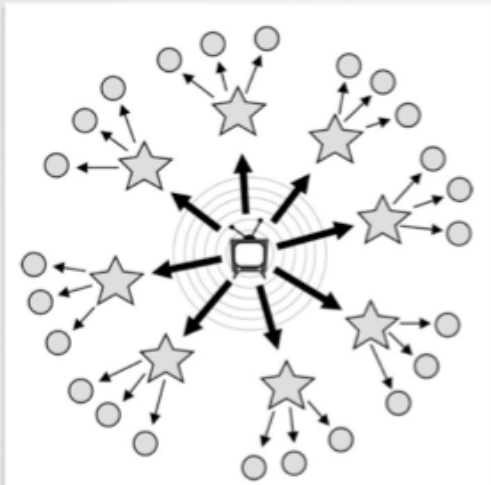
low in-centralization:  
buying is more evenly  
distributed

“model in which opinion flows only from the media to influentials, and then only from influentials to the larger populace is deprecated”

# Influentials, Networks, and Public Opinion Formation

DUNCAN J. WATTS  
PETER SHERIDAN DODDS\*

2007



broadcast influence

- classic information marketing
- message saturation
- **degree** is most important

“large cascades of influence are driven not by influentials, but by a critical mass of easily influenced individuals.”

# Influentials, Networks, and Public Opinion Formation

DUNCAN J. WATTS  
PETER SHERIDAN DODDS\*

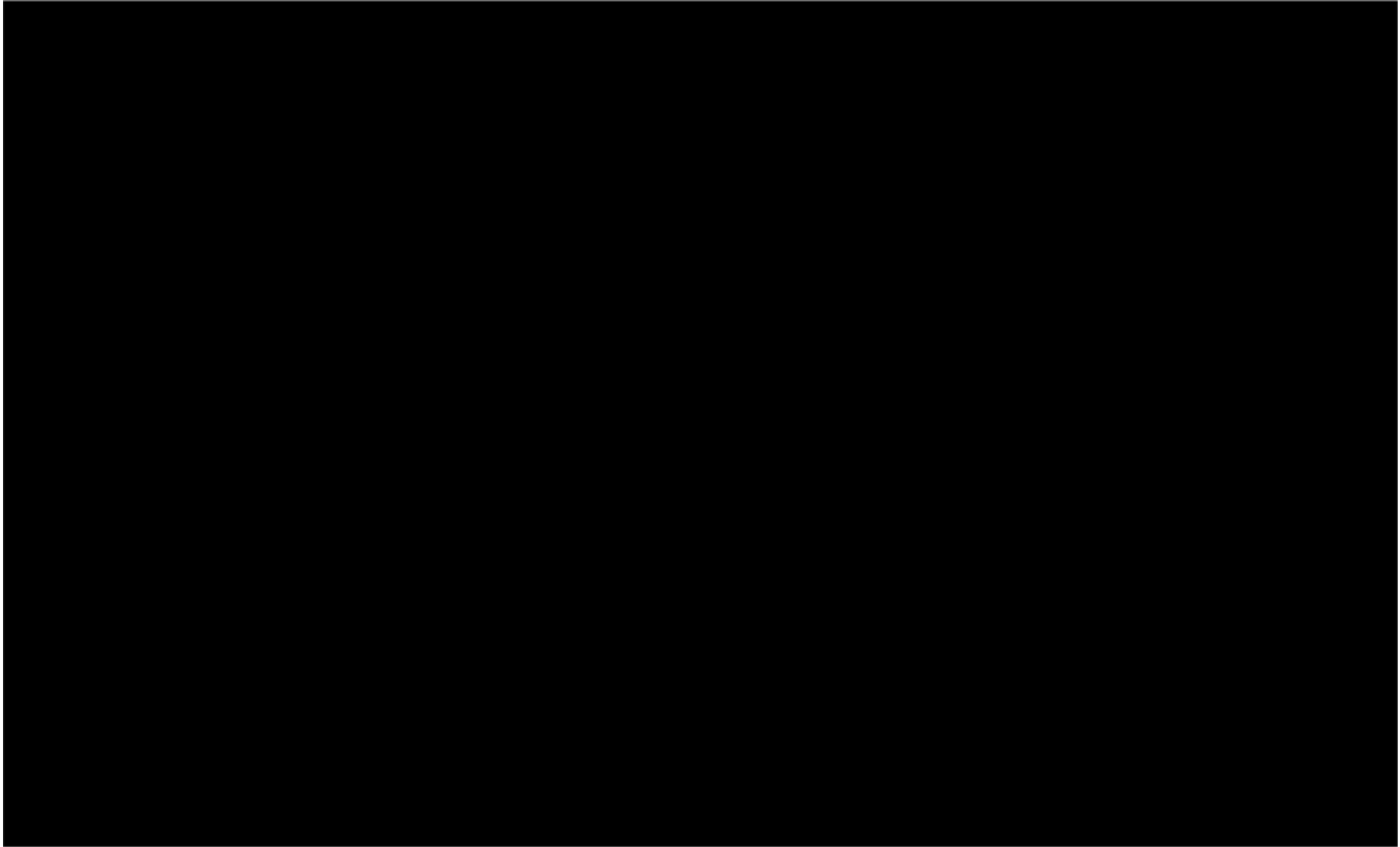
2007



network influence

- “network” (decentralized) marketing
- high-degree = “opinion leader”
- high-degree alone = **irrelevant**
- a cascade requires a legion of *susceptibles* (a system-level property)

“influence is not really about the influencer as much about the susceptibles”



## what have we learnt from it...

Baker & Faulkner (1993): Social Organization of conspiracy

(reconstructs communication networks in three well-known price-fixing conspiracies in the heavy electrical equipment industry to study social organization)

Questions: How are relations organized to facilitate illegal behavior?

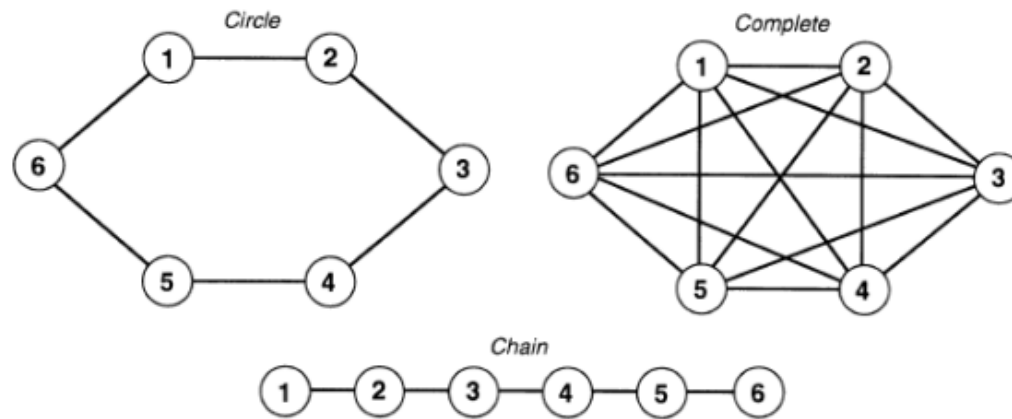
Pattern of communication maximizes concealment, and predicts the criminal verdict.

Inter-organizational cooperation is common, but too much ‘cooperation’ can thwart market competition, leading to (illegal) market failure.

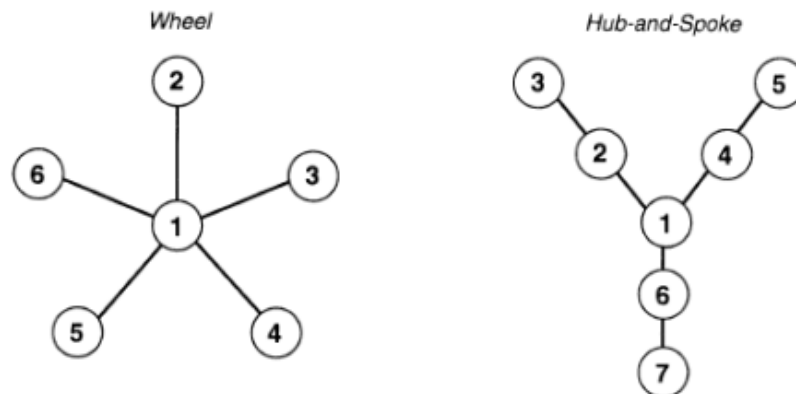
Illegal networks differ from legal networks, in that they must conceal their activity from outside agents. A “Secret society” should be organized to (a) remain concealed and (b) if discovered make it difficult to identify who is involved in the activity

The need for secrecy should lead conspirators to conceal their activities by creating **sparse** and **decentralized** networks.

### Decentralized Networks



### Centralized Networks



- reconstructs communication networks in three well-known price-fixing conspiracies in the heavy electrical equipment industry to study social organization;
- findings:
  - structure of illegal networks is driven by need to maximize concealment, rather than efficiency;
  - structure also contingent on information-processing requirements;
  - person centrality in networks predicts *verdict*, *sentence* and *fine*.

Organization Objective	Information-Processing Requirement	
	High	Low
Concealment	Centralized networks	Decentralized networks
Coordination	Decentralized networks	Centralized networks

**Figure 1. Concealment Versus Coordination: Theoretical Expectations and experimental results**

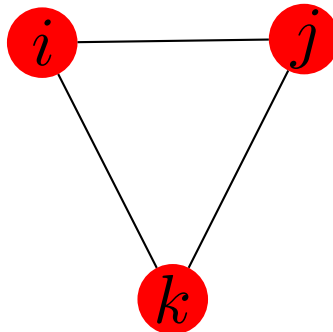


# Clustering

A feature of interest when studying a network is its *transitivity*, i.e., if  $i \sim j$  and  $j \sim k$  then  $i \sim k$ .

If node  $i$  is connected to nodes  $j$  and  $k$ , how often is it the case that  $j$  and  $k$  are also connected?

When  $i$ ,  $j$ , and  $k$  are all connected to each other they form a *triangle*.



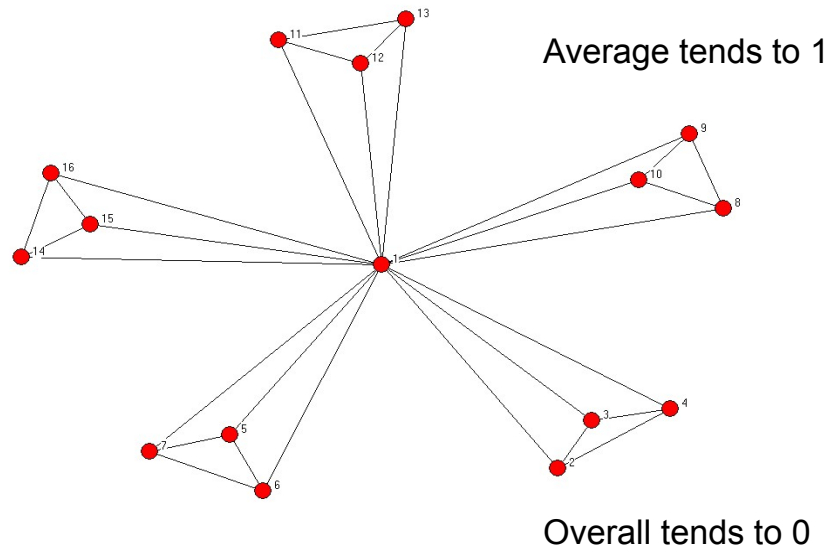
# Clustering

What fraction of my friends are friends of each other?

(1) Calculate clustering for a particular node;

(1) Average individual clustering coefficients across the network (it weights clustering node by node)

(2) Overall clustering: out of all possible triplets in the network, what the frequency with which it is a triangle

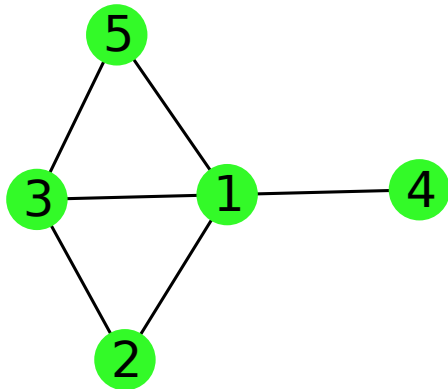


# local clustering coefficient

If  $i$  is a node with  $k_i \geq 2$  then its *local clustering coefficient* is defined as:

$$\begin{aligned} C_i &= \frac{\text{Number of triangles containing } i}{\text{Number of pairs of neighbours of } i}, \\ &= \frac{t_i}{\frac{1}{2}k_i(k_i - 1)}, \end{aligned}$$

where  $t_i = [A^3]_{ii}$ .



Possible triangles including node 1:

$\{(1 - 2 - 3), (1 - 3 - 5), (1 - 2 - 5),$   
 $(1 - 5 - 4), (1 - 2 - 4), (1 - 3 - 4)\}.$

Actual triangles:

$\{(1 - 2 - 3), (1 - 3 - 5)\}.$

$$C_1 = \frac{1}{3}.$$

# global clustering coefficient

There are two alternative definitions of the global clustering coefficient:

Version 1: Average Clustering Coefficient

$$C = \langle C_i \rangle = \frac{1}{N} \sum_{i=1}^N C_i.$$

Version 2: Overall Clustering Coefficient

$$C = \frac{3 \times t}{\text{number of connected triples}}$$

where  $t$  is the total number of triangles. If there are no self-loops then  $t = \frac{1}{3}\text{trace}(A^3)$ .

In adjacency matrix notation,

$$C(v) = \frac{\sum_{u,w \in V} a_{u,v} a_{w,v} a_{u,w}}{\sum_{u,w \in V} a_{u,v} a_{w,v}}.$$

The (*average*) *clustering coefficient* is defined as

$$C = \frac{1}{|V|} \sum_{v \in V} C(v).$$

Note that

$$\sum_{u,w \in V} a_{u,v} a_{w,v} a_{u,w}$$

is the number of triangles involving  $v$  in the graph. Similarly,

$$\sum_{u,w \in V} a_{u,v} a_{w,v}$$

is the number of *2-stars* centred around  $v$  in the graph. The clustering coefficient is thus the ratio between the number of triangles and the number of 2-stars. The clustering coefficient describes how "locally dense" a graph is. Sometimes the clustering coefficient is also called the *transitivity*.

# Social balance theories

Central question in modeling social networks from *structural individualism*:

**how can the global properties of the network be understood from local properties?**

**E.g. theory of clusterability of balanced signed graphs:**

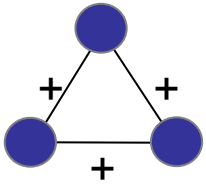
(1) Harary's theorem says that a complete signed graph is balanced iff the nodes can be partitioned into two sets, so that all ties within sets are positive, and all ties between sets are negative;

(2) Heider's work on cognition of social situations (Person-Object-Other), interested in correspondence between P and O, given their beliefs (like/dislike) about Object X [dyads PO, PX, OX];

(3) David & Leinhardt generalized conditions for clusterability of signed graphs and structures of ranked clusters;

These theories pose the problem: how can triadic properties of signed graphs (aggregate properties of all subgraphs of 3 nodes) determine global properties of signed graphs.

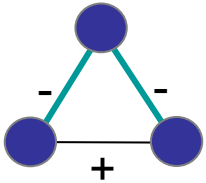
argument that unbalanced triads tended towards balance, which implied all intransitive triads would disappear from the network. **not what we find empirically...**



$$(+)(+)(+) = (+)$$

Balanced

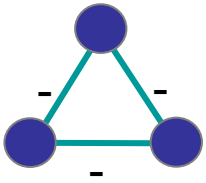
“A friend of a friend is a friend”



$$(-)(+)(-) = (-)$$

Balanced

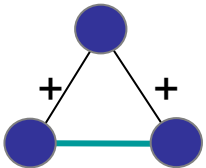
“An enemy of my enemy is a friend”



$$(-)(-)(-) = (-)$$

Unbalanced

“An enemy of my enemy is an enemy”



$$(+)(-)(+) = (-)$$

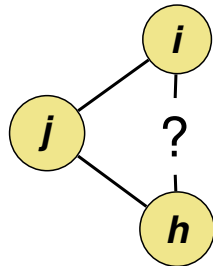
Unbalanced

“A Friend of a Friend is an enemy”

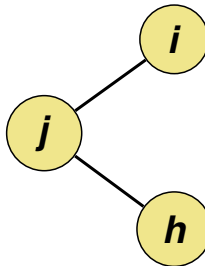
# Transitivity

The tendency for a tie from  $i$  to  $k$  to occur at greater than chance frequencies if there are ties from  $i$  to  $j$  and from  $j$  to  $k$  – the  $i$  to  $j$  tie completes “transitively” the triple consisting of the tie from  $i$  to  $j$  and the tie from  $j$  to  $k$ .

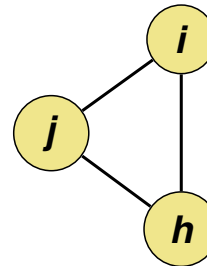
Transitivity depends on *triads*, subgraphs formed by 3 nodes



Potentially  
transitive



Intransitive



Transitive



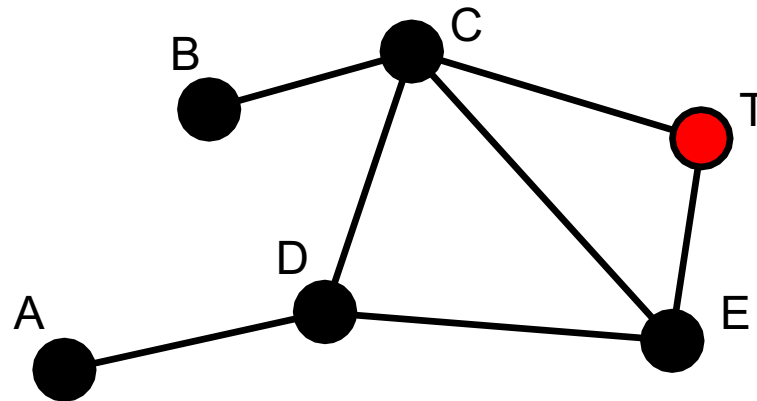
# Simmelian Ties

- These are the ties that make up transitivity
- Simmelian ties are [reciprocated] transitive triples

CD,DE,EC is one set of simmelian ties

CT,TE,EC is another set

All other sets are not simmelian



## measuring transitivity – **clustering index**

A measure for transitivity is the (global) transitivity index, defined as the ratio

$$\text{Transitivity Index} = \frac{\# \text{Transitive triads}}{\# \text{Potentially transitive triads}} .$$

(Note that “ $\#A$ ” means the number of elements in the set  $A$ .)

This also is sometimes called a *clustering* index.

This is between 0 and 1; it is 1 for a transitive graph.

For random graphs, the expected value of the transitivity index is close to the density of the graph (**why?**);

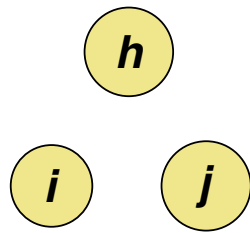
for actual social networks,

values between 0.3 and 0.6 are quite usual.

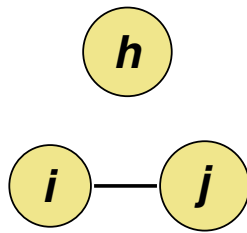
# local structure and triad counts

The studies about transitivity in social networks led Holland and Leinhardt (1975) to propose that the *local structure* in social networks can be expressed by the *triad census* or *triad count*, the numbers of triads of any kinds.

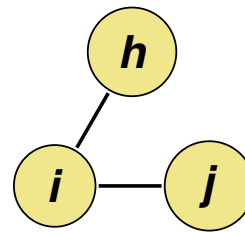
For (nondirected) graphs, there are four triad types:



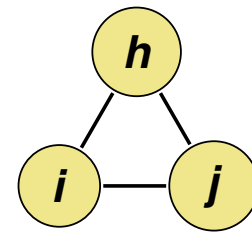
Empty



One edge



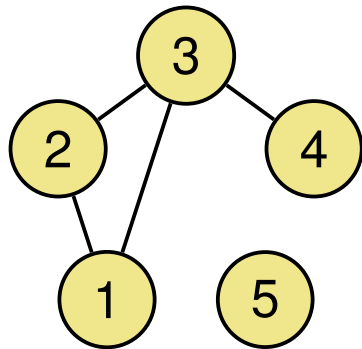
Two-path /  
Two-star



Triangle

# local structure and triad counts

A simple example graph  
with 5 nodes.



<i>i</i>	<i>j</i>	<i>h</i>	triad type
1	2	3	triangle
1	2	4	one edge
1	2	5	one edge
1	3	4	two-star
1	3	5	one edge
1	4	5	empty
2	3	4	two-star
2	3	5	one edge
3	4	5	one edge

In this graph, the triad census is (1, 5, 2, 1)  
(ordered as: empty – one edge – two-star – triangle).

# MAN coding for triad census

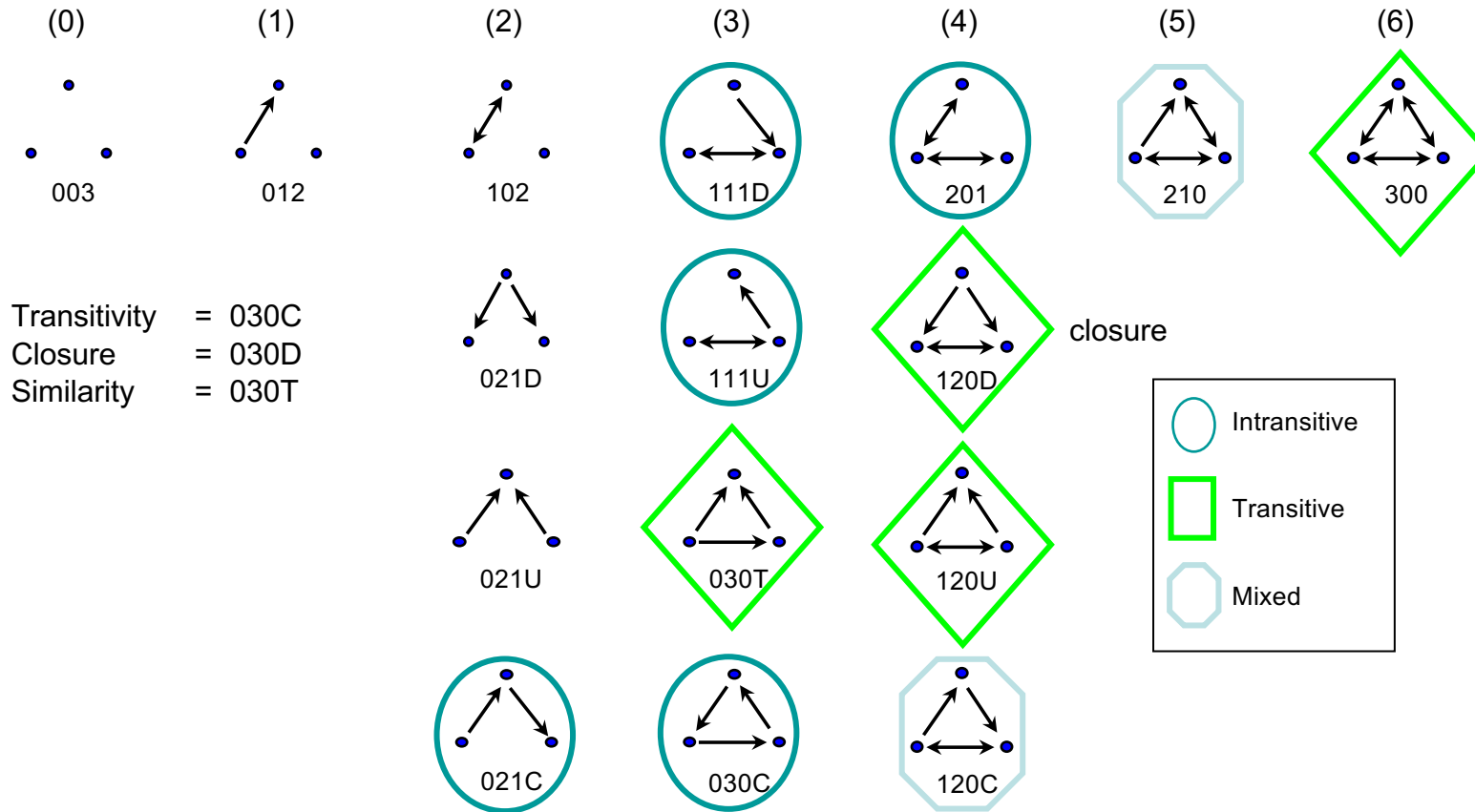
Holland and Leinhardt (1975) proposed the following MAS coding.



the scheme a further identifying letter: Up, Down, Cyclical, Transitive.

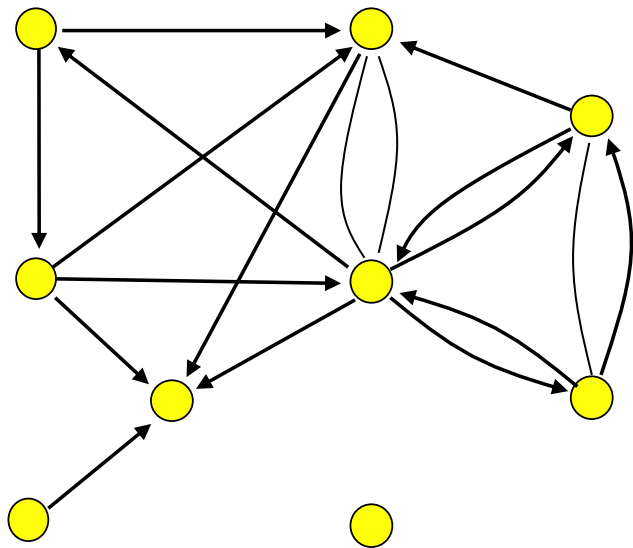
**E.g.** 120 has 1 mutual, 2 asymmetric, 0 null dyads and the Down orientation

# triad census



**Transitivity:** tie *i* to *k* to occur if ties from *i* to *j* and *j* to *k* exist;  
**Closure:** tie *i* to *j* to occur if persons *k* with ties to both *i* and *j* exist;  
**Similarity:** tie *i* to *j* to occur if persons *k* to whom *i* and *j* have ties exist;

# triad census - example



Type	Number of triads
1 - 003	21
2 - 012	26
3 - 102	11
4 - 021D	1
5 - 021U	5
6 - 021C	3
7 - 111D	2
8 - 111U	5
9 - 030T	3
10 - 030C	1
11 - 201	1
12 - 120D	1
13 - 120U	1
14 - 120C	1
15 - 210	1
16 - 300	1
Sum (2 - 16):	63

- triads define behavioral mechanisms: we can leverage the distribution of triads in a network to test whether the hypothesized mechanism is active.
- How?

(1) Count the number of each triad type in a given network

(2) Compare to the expected number, given some (random) distribution of ties in the network;

- Statistical approach proposed by Holland and Leinhardt is now obsolete. Statistical methods have been proposed for probability distributions of graphs depending primarily on triad counts, but complemented with stat counts and nodal variables, along with some higher-order configurations essential for adequate modeling of empirical network data.