

IMA Workshop

On Identifying **Key Players**  
In the Context of **Network Flows**

Steve Borgatti, Boston College  
[borgatts@bc.edu](mailto:borgatts@bc.edu)

Research supported by the Office of Naval Research & NSF-MKIDS program

# Motivating Questions

## MAXIMUM SPREAD

- If you wanted to diffuse something as quickly or thoroughly as possible through a network, at which node(s) would you begin?

## MINIMUM SPREAD

- If you wanted to maximally disrupt the spread of something through a network, which node(s) would you isolate?

# Applications

Maximize  
Spread

Minimize  
Spread

Public Health	{ Selecting peer health advocates for <b>diffusing</b> safe practices (e.g. bleaching) and material	Who to <b>immunize or quarantine</b> in order to slow spread of infectious disease
Criminal Justice	{ Who to "turn", feed false information to, or surveil	Who to <b>arrest or discredit</b> to disrupt criminal networks
Management	{ Select employees for intervention prior to change initiative	Where is an organization most vulnerable to turnover?

# Research Agenda

1. Centrality to the rescue
2. Issues with centrality
  1. Ensemble issues with centrality
  2. Optimality issues with centrality
  3. Flow assumptions of centrality
3. Better ways to find key players
  1. Group centrality & combinatorial optimization
  2. Designer measures
  3. Flow awareness
4. What we've learned about centrality

# Today's Agenda

- Typology of centrality measures
- Typology of flow processes
- Matching centrality measures to flow processes
- Simulations
- On centrality

# Typology of Centrality Measures

# Degree Centrality

- Number of ties that involve a given node
  - Marginals of symmetric adjacency matrix

	I1	I3	W1	W2	W3	W4	W5	W6	W7	W8	W9	S1	S2	S4	Deg
I1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	4
I3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W1	1	0	0	1	1	1	1	0	0	0	0	1	0	0	6
W2	1	0	1	0	1	1	0	0	0	0	0	1	0	0	5
W3	1	0	1	1	0	1	1	0	0	0	0	1	0	0	6
W4	1	0	1	1	1	0	1	0	0	0	0	1	0	0	6
W5	0	0	1	0	1	1	0	0	1	0	0	1	0	0	5
W6	0	0	0	0	0	0	0	0	0	1	1	1	0	0	3
W7	0	0	0	0	0	0	1	1	0	1	1	0	0	1	5
W8	0	0	0	0	0	0	0	1	1	0	1	0	0	1	4
W9	0	0	0	0	0	0	0	1	1	1	0	0	0	1	4
S1	0	0	1	1	1	1	1	0	0	0	0	0	0	0	5
S2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S4	0	0	0	0	0	0	0	0	0	1	1	1	0	0	3

# Interpreting Degree Centrality

- Index of immediate exposure to what is flowing through the network
  - Gossip network: Central actor more likely to hear a given bit of gossip
- Opportunity to influence & be influenced directly
- Ignores indirect influences?
- Predicts variety of outcomes from virus resistance to power & leadership to job satisfaction to amount of knowledge

# Freeman Closeness Centrality

- Sum of geodesic distances to all other nodes
  - Computed as marginals of symmetric geodesic distance matrix

	I1	I3	W1	W2	W3	W4	W5	W6	W7	W8	W9	S1	S2	S4	Clo
I1	0	?	1	1	1	1	2	4	3	4	4	2	?	4	27
I3	?	0	?	?	?	?	?	?	?	?	?	?	?	?	0
W1	1	?	0	1	1	1	1	3	2	3	3	1	?	3	20
W2	1	?	1	0	1	1	2	4	3	4	4	1	?	4	26
W3	1	?	1	1	0	1	1	3	2	3	3	1	?	3	20
W4	1	?	1	1	1	0	1	3	2	3	3	1	?	3	20
W5	2	?	1	2	1	1	0	2	1	2	2	1	?	2	17
W6	4	?	3	4	3	3	2	0	1	1	1	3	?	2	27
W7	3	?	2	3	2	2	1	1	0	1	1	2	?	1	19
W8	4	?	3	4	3	3	2	1	1	0	1	3	?	1	26
W9	4	?	3	4	3	3	2	1	1	1	0	3	?	1	26
S1	2	?	1	1	1	1	1	3	2	3	3	0	?	3	21
S2	?	?	?	?	?	?	?	?	?	?	?	?	0	?	0
S4	4	?	3	4	3	3	2	2	1	1	1	3	?	0	27

# Interpreting Freeman Closeness

- Inverse measure of centrality
- Index of expected time until arrival for given node of whatever is flowing through the network
  - Gossip network: central player hears things first
- Involves lengths of (shortest) paths emanating from node

# Freeman Betweenness Centrality

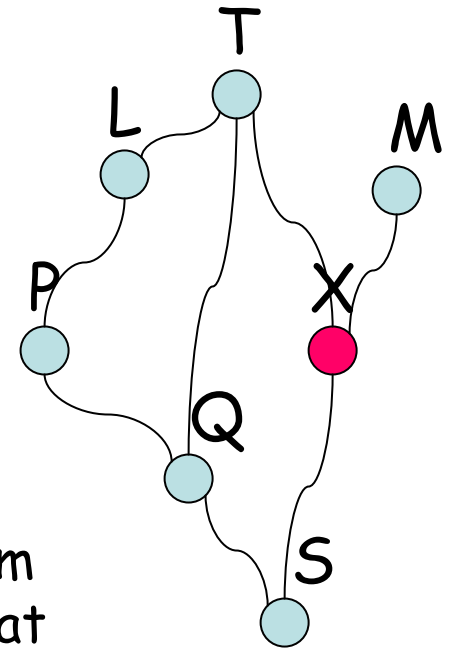
- How often a node lies along the shortest path between two other nodes

- Defined as:

$$b_k = \sum_{i,j} \frac{g_{ikj}}{g_{ij}}$$

- where  $g_{ij}$  is number of geodesic paths from  $i$  to  $j$  and  $g_{ikj}$  is number of those paths that pass through  $k$

- Share of geodesics "controlled" by a node



# Interpreting Freeman Betweenness

- 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
- Involves counts of geodesics passing through node

# Katz, Hubbell, Eigenvector etc

- Katz

$$X = bA + (bA)^2 + (bA)^3 + \dots$$

$$X = (I - bA)^{-1} - I \quad \text{for } 1/b > \lambda$$

$$s = Xu \quad \{u \text{ is vector of ones}\}$$

- Hubbell

$$s = e + Ws$$

$$s = (I - W)^{-1}e \quad \{\text{when invertable}\}$$

- If  $e = u$  then  $\text{katz} = \text{hubbell} - 1$

- Eigenvector

-  $\lambda v = Av$

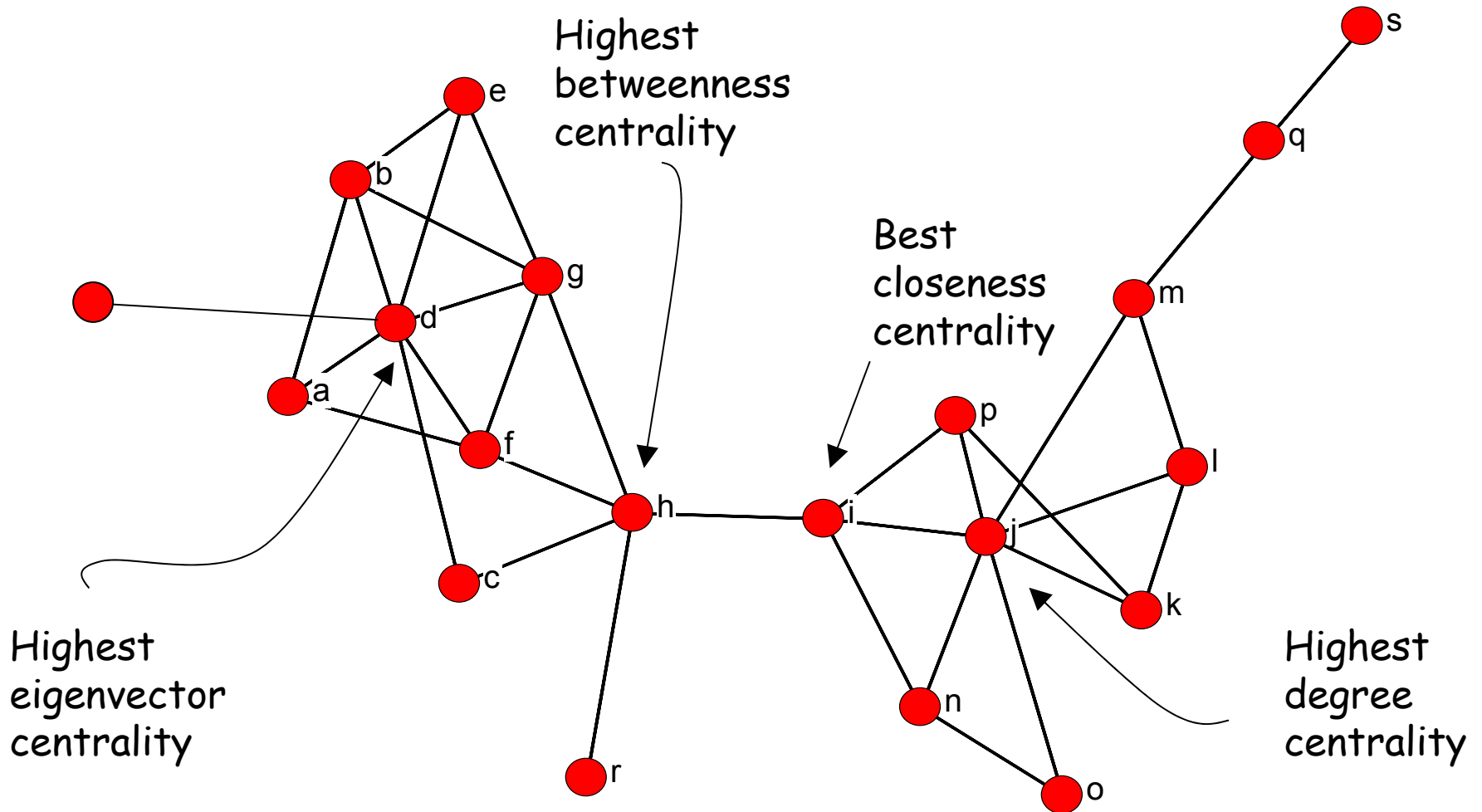
-  $S = (A + \lambda A^2 + \lambda^2 A^3 + \dots + \lambda^2 A^3)u$

All of these involve counts of walks emanating from a node

# Eigenvector Centrality

- Indicator of popularity,
  - "in the thick of things"
- Turbo-charged index of exposure, risk
- Tends to identify centers of large cliques
- Involves counts of walks emanating from node (weighted by length)

# Centrality



# Typology of Centrality Measures

		Position in Walk	
		Radial	Medial
Walk Property	Frequency (no. of walks)	Degree; Eigenvector, Hubbell, Katz, etc.	FreeBet; NewBet; FloBet;
	Distance (length of walks)	FreeClo; Information	

# Typology of Network Flows

Does it matter how things flow through the network? Or is centrality centrality regardless?

# Things that move thru networks

- Used goods
- Money
- Packages
- Personnel
- Gossip / information
- E-mail
- Infections
- Attitudes

# Used Goods Process

- Canonical example:
  - passing along used paperback novel
- Single object in only one place at a time
- Doesn't (usually) travel between same pair twice
- Could be received by the same person twice
  - A--B--C--B--D--E--B--F--C ...
  - Travels in trails

# Mooch Process

- Examples
  - Obnoxious homeless relative who visits for six months until kick out and moves to next relative
  - Personnel flows between firms
- In just one place at a time
- Doesn't repeat a node (bridges burned)
  - Travels along paths

# Money Exchange Process

- Examples:
  - specific dollar bill moving through the economy
  - Erdos itinerary
  - Any markov process
- Single object in only one place at a time
- Can travel between same pair more than once
  - A--B--C--B--C--D--E--B--C--B--C ...
  - Travels along unconstrained walks

# Gossip Process

- Example:
  - Confidential story moving through informal network
- Multiple copies exist simultaneously
- Person tells only one person at a time\*
- Doesn't travel between same pair twice
- Can reach same person multiple times

\* More generally, they tell a very limited number at a time.

# E-Mail Process

- Example:
  - forwarded jokes and virus warnings
  - e-mail viruses themselves
- Multiple copies exist simultaneously
- All (or many) connected nodes told simultaneously
  - Except, perhaps, the immediate source

# Influence Process

- Example:
  - attitude formation
- Multiple "copies" exist simultaneously
- Multiple simultaneous transmission, even between the same pairs of nodes

# Infection Process

- Example:
  - virus which activates effective immunological response
- Multiple copies may exist simultaneously
- Cannot revisit a node
  - A--B--C--E--D--F...

# Package Delivery Process

- Example:
  - package delivered by postal service
- Single object at only one place at one time
- Map of network enables the intelligent object to select only the shortest paths to all destinations

# Properties of Flow Processes

- Sequence type: path, trail, walk
  - path: can't revisit node nor edge (tie)
  - trail: can revisit node but not edges
  - walk: can revisit edges & nodes
- Deterministic vs non-deterministic
  - blind vs guided
  - always chooses best route; aware of map
- Combine into 4-way "traversal type" property:
  - geodesics, paths, trails, walks

# Properties -- cont.

- Duplication vs transfer (copy vs move)
  - transfer/move: only one place at one time
  - duplication/copy: multiple copies exist
- Serial vs parallel duplication
  - serial: only one transmission at a time
  - parallel: broadcast to all surrounding nodes
- Combine into "method" 3-way property:
  - parallel dup., serial dup., transfer

# Simplified Typology

	parallel duplication	serial duplication	transfer
geodesics	<no process>	mitotic reproduction	package delivery
paths	internet name-server	viral infection	mooch
trails	e-mail broadcast	gossip	used goods
walks	attitude influencing	emotional support	money exchange

\*Note: Names not to be taken too seriously.

Markov

# Matching Measures to Flows

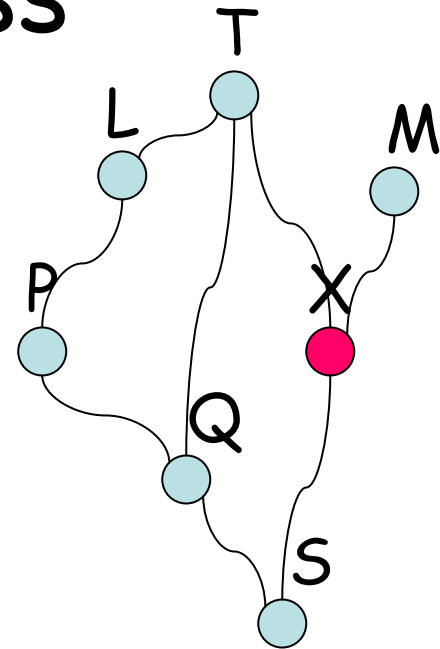
# Freeman Closeness

- Sums lengths of shortest paths
- Time until arrival of traffic
- Consistent processes
  - non-deterministic (e.g., delivery)
  - parallel duplication (e.g., e-mail, name-server)

	parallel duplication	serial duplication	transfer
geodesics	<no process>	FreeClo	FreeClo
paths	FreeClo		
trails	FreeClo		
walks	FreeClo		

# Freeman Betweenness

- Counts proportion of shortest paths from  $i$  to  $j$  that pass through  $k$
- Share of traffic passing through node
- Consistent processes
  - Geodesic transfer process (e.g., delivery)

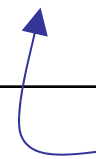


	parallel duplication	serial duplication	transfer
geodesics	<no process>	FreeClo	FreeClo FreeBet
paths	FreeClo		
trails	FreeClo		
walks	FreeClo		NewmanBet

# Eigenvector (& Katz, Hubbell, etc)

- Counts walks
- Consistent processes
  - Walk-based parallel process (e.g., influence)

	parallel duplication	serial duplication	transfer
geodesics	<no process>	FreeClo	FreeClo FreeBet
paths	FreeClo		
trails	FreeClo		
walks	FreeClo Eigenvector		NewBet Degree



“Mind the gap”

# Simulations

# Research Questions

- How far off are the off-the-shelf centrality measures when applied to the “wrong” processes?
  - Like applying freeman's measures in STI or information flow context
- Can we extend centrality measures to other processes?
  - Need to abstract out the essence of the measure
  - Treat measures as models

# What are the measures really about?

- Closeness
  - Time until arrival of something flowing through network
  - Speed; aggregate commute time
- Betweenness
  - How often a given node is along the way between a source and a target
  - Frequency of visitation; hit rates

Freeman's centrality measures give expected values for time until arrival and frequency of visitation under model of geodesic+transfer process (e.g., package delivery).

# Centrality as Model

- Centrality “measures” are best viewed as formulas for the expected values of certain kinds of node outcomes in a particular model of traffic flow
  - Which node outcomes are of interest is what defines the fundamental character or purpose of a measure
  - It is the essence of the measure that needs to be abstracted out of existing formula in order to be made applicable to different flow models

# Simulation Methodology

- Given network of ties along which traffic flows
- Let traffic flow according to the rules of a given flow process
- For each node, measure time-until-arrival and frequency of pass-throughs
  - Yields "realized" closeness and betweenness
- Repeat 10,000 times and report avg meas

# Stopping Rule

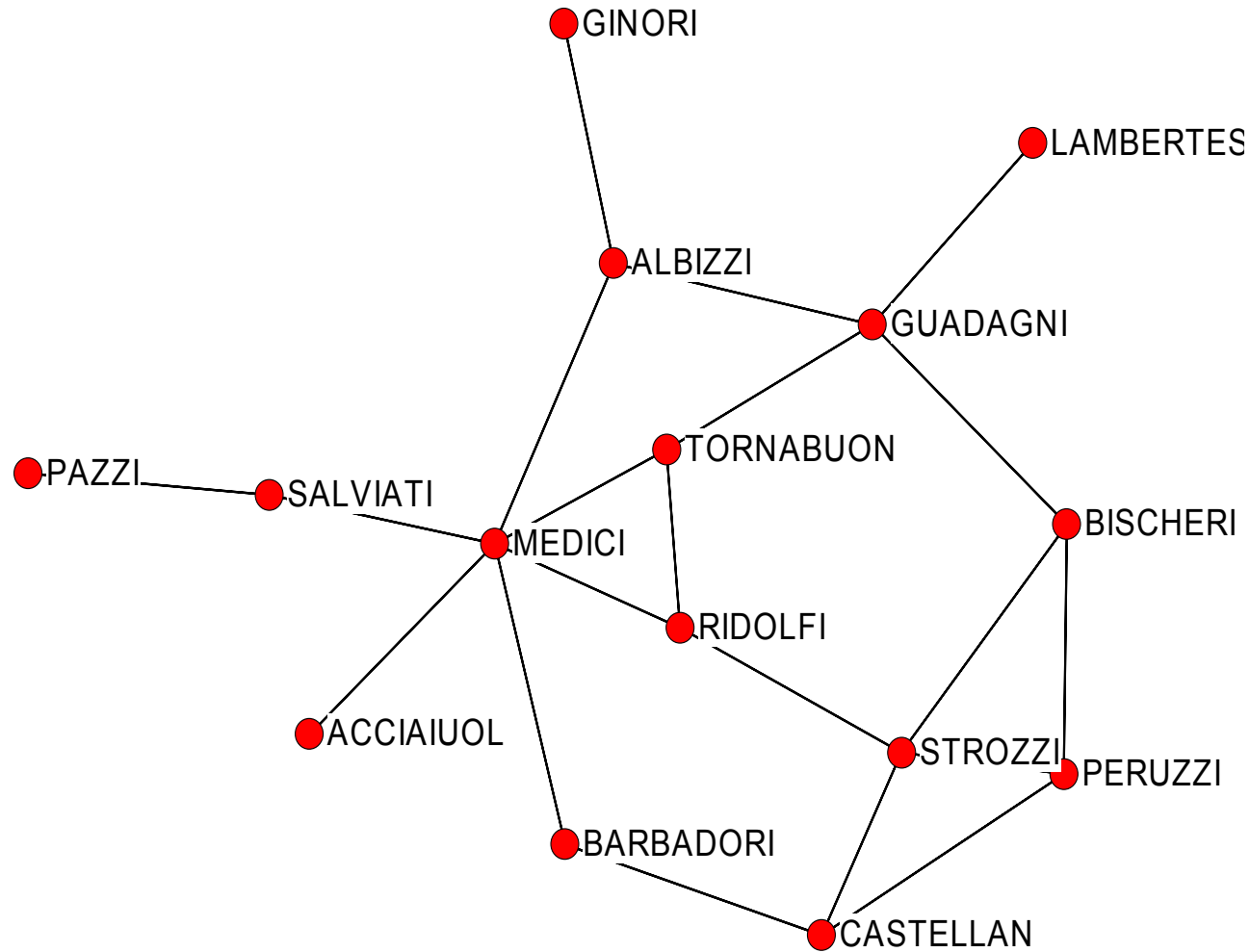
- Only geodesic processes, like package delivery, have target
  - Target provides stopping rule
- Other processes, like gossip, wander randomly until they hit cul-de-sac ... or don't
  - Walk processes

# Simulation Algorithm

- For comparability with Freeman measures, assume traffic has target

```
For k = 1 to 10,000 {independent trials}
.....For i = 1 to n {source nodes}
.....For j = 1 to n {target nodes} if i <> j then
.....Repeat
.....Start traffic at i and don't stop until it reaches j
.....or reaches cul-de-sac
.....Until j is actually reached
.....Compute node statistics for this trial
Average the node statistics across trials
```

# Illustrative Dataset



Padgett & Ansell (1991).

Marriage ties among Florentine families during the Renaissance

# Betweenness / Freq of Visits

Node	Freeman	Package	Mooch	Used Goods	Gossip	Infect	Money
MEDICI	47.5	47.5	113.7	129.8	334.3	887.03	1155.1
GUADAGNI	23.2	22.8	74.9	73.8	252.2	513.35	827.9
ALBIZZI	19.3	19.2	41.5	48.5	185.0	285.37	665.9
SALVIATI	13.0	13.0	26.0	26.0	168.0	182.00	503.3
RIDOLFI	10.3	10.7	61.3	64.2	189.0	227.89	665.4
BISCHERI	9.5	9.5	60.9	58.6	189.0	257.23	664.7
STROZZI	9.3	9.7	78.1	84.8	295.6	435.10	827.5
BARBADORI	8.5	8.5	45.8	46.5	176.0	107.65	503.5
TORNABUON	8.3	8.2	58.2	59.8	189.0	222.97	666.1
CASTELLAN	5.0	5.0	64.5	64.7	188.7	277.20	665.3
PERUZZI	2.0	2.0	59.1	55.1	189.0	232.30	664.7
ACCIAIUOL	0.0	0.0	0.0	0.0	0.0	0.00	176.9
GINORI	0.0	0.0	0.0	0.0	0.0	0.00	176.8
LAMBERTES	0.0	0.0	0.0	0.0	0.0	0.00	176.6
PAZZI	0.0	0.0	0.0	0.0	0.0	0.00	177.2

Number of times token passed through node en route from source to target



# Betweenness Summary

- Freeman betweenness definition gives exact expected values for frequency of pass-throughs in package delivery process
  - And only the package delivery process
- Other processes gives different results
  - Different nodes are the ones with bigger flows
  - Freeman formula would misidentify key players
- Money exchange process yields scores proportional to degree.
  - Degree and betweenness are kin

# Closeness / Time to Arrival

---

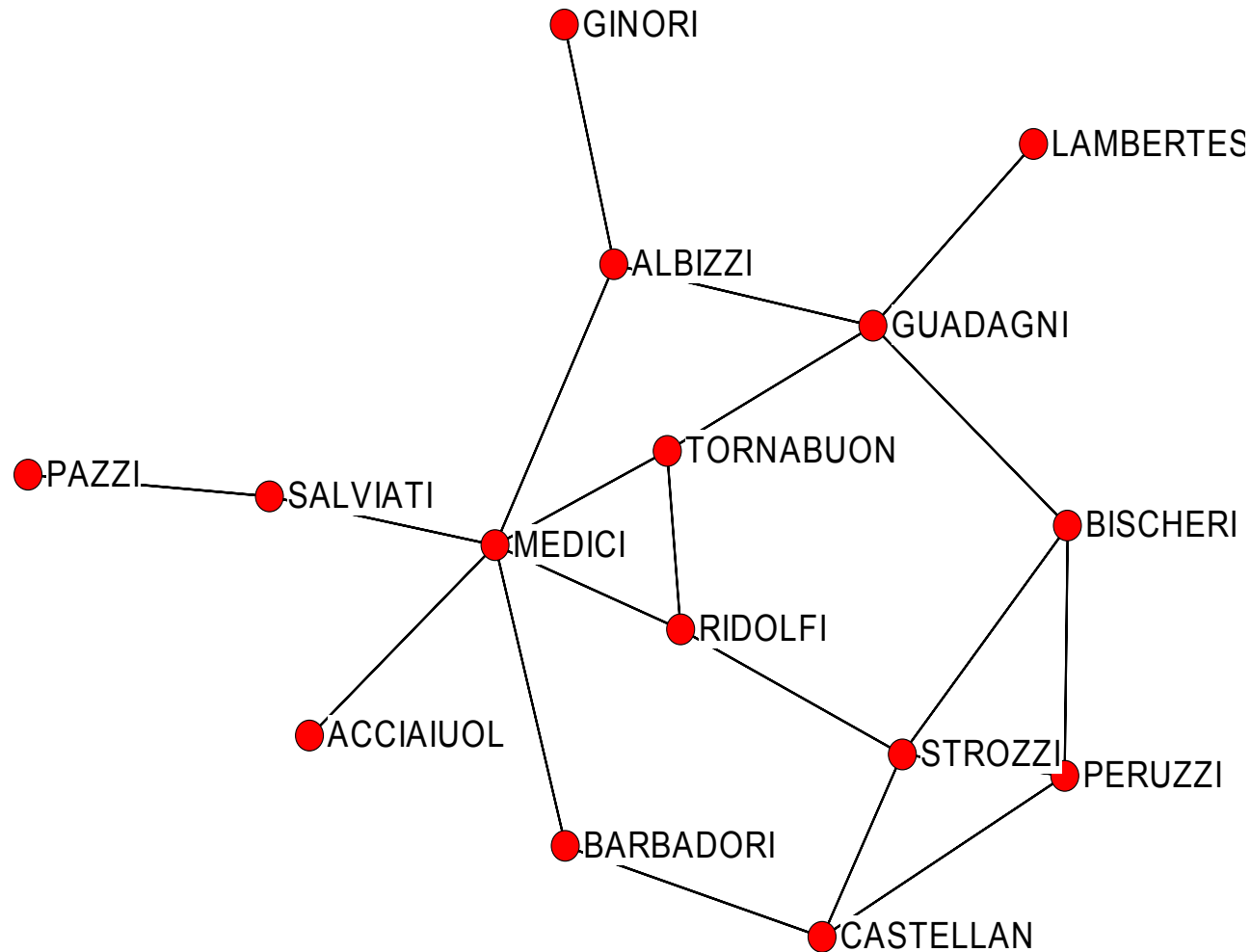
Node	Freeman	Package	Mooch	Used Goods	Gossip	Infect	Money
MEDICI	25	25.0	46.7	50.1	78.9	63.7	575.2
RIDOLFI	28	28.0	57.5	60.6	95.7	70.8	587.7
ALBIZZI	29	29.0	55.7	53.3	100.7	68.6	562.3
TORNABUON	29	29.0	56.4	58.1	98.2	70.0	584.8
GUADAGNI	30	30.0	53.7	54.8	109.3	68.8	575.3
BARBADORI	32	32.0	60.5	55.3	112.3	73.1	584.4
STROZZI	32	32.0	59.9	61.3	104.0	73.3	602.9
BISCHERI	35	35.0	61.1	63.9	111.6	74.1	599.0
CASTELLAN	36	36.0	58.3	64.6	125.8	73.3	599.2
SALVIATI	36	36.0	57.6	59.9	94.3	72.7	533.0
ACCIAIUOL	38	38.0	59.5	64.3	98.2	69.8	536.3
PERUZZI	38	38.0	61.3	67.9	111.3	75.4	603.7
GINORI	42	42.0	68.9	65.3	124.5	75.9	523.2
LAMBERTES	43	43.0	66.4	69.8	109.6	76.1	538.2
PAZZI	49	49.0	70.7	72.9	155.9	78.8	497.8

---

Units of time passed until node received token

Check these #s

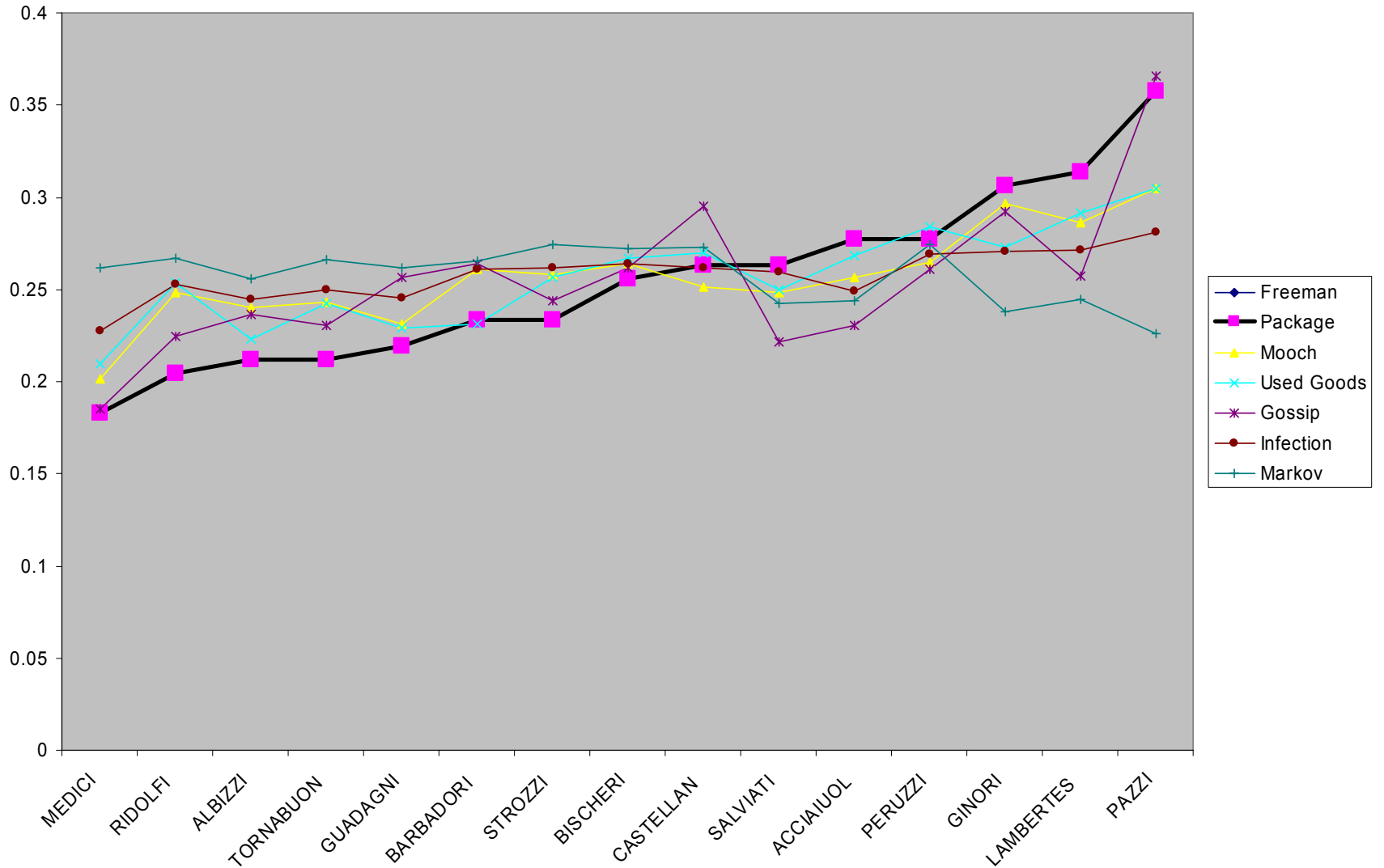
# Illustrative Dataset



Padgett & Ansell (1991).

Marriage ties among Florentine families during the Renaissance

# Closeness Measures



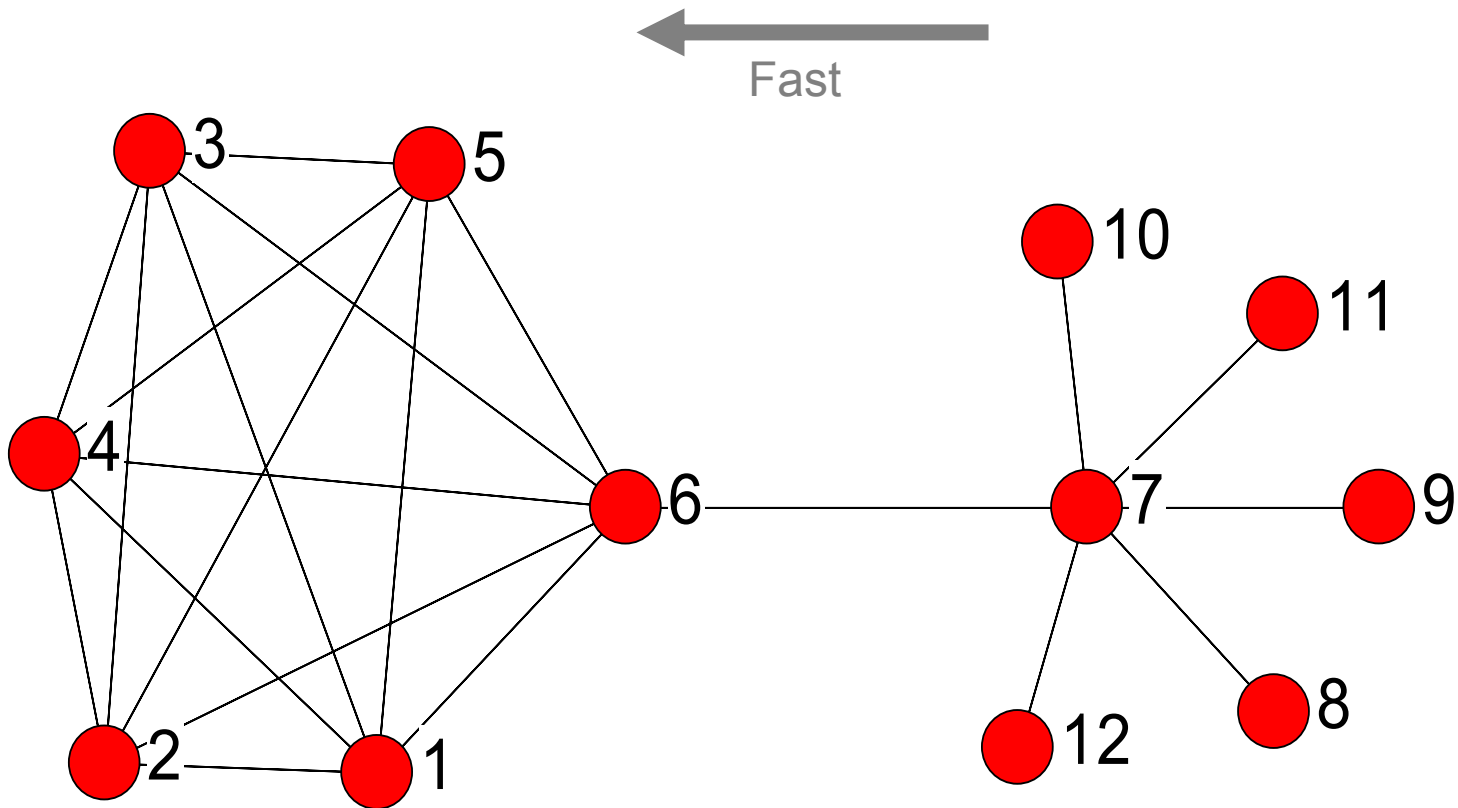
# Out-Closeness / Agg. Reach Time

---

Node	Freeman	Package	Mooch	Used Goods	Gossip	Infect	Money
MEDICI	25	25.0	43.9	42.7	102.4	40.4	172.5
RIDOLFI	28	28.0	54.9	54.5	108.8	58.6	287.0
ALBIZZI	29	29.0	57.3	58.2	107.6	61.8	387.9
TORNABUON	29	29.0	55.0	54.5	107.1	58.7	292.3
GUADAGNI	30	30.0	54.8	53.5	105.0	52.8	285.0
BARBADORI	32	32.0	63.0	60.1	112.6	69.4	446.0
STROZZI	32	32.0	52.0	53.0	108.6	58.8	296.1
BISCHERI	35	35.0	57.9	59.0	110.1	62.5	350.6
CASTELLAN	36	36.0	57.3	57.9	110.8	73.4	390.8
SALVIATI	36	36.0	57.7	67.2	103.7	85.8	777.4
ACCIAIUOL	38	38.0	59.8	71.4	107.6	90.4	870.0
PERUZZI	38	38.0	59.4	60.2	112.1	74.2	408.5
GINORI	42	42.0	78.1	72.0	114.0	104.0	1084.5
LAMBERTES	43	43.0	72.0	77.7	110.7	94.3	981.2
PAZZI	49	49.0	70.9	80.1	109.3	99.2	1473.2

---

# Closeness Asymmetry

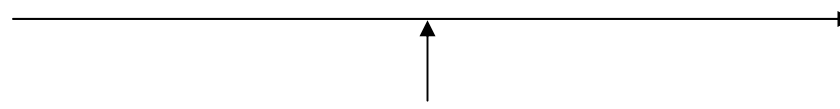


When traffic does not follow shortest paths, nodes on the right will on average reach the nodes on the left more quickly than the other way around

# Closeness Summary

- Like betweenness, Freeman measure identifies key players in appropriate processes, but not all processes
- Time until arrival not same as time to reach others, even if graph is undirected
  - Proximity to hub is better for spreading than receiving
- Cliques create global slowness

Path  
redundancy



Individual  
performance

Type of flow

Conclusion

# Identifying Key Players

(immediate motivation)

- Naïve approach of picking most central actors fails
  - Ensemble issue: avoiding redundancy
  - Design issue: centrality not designed for specific tasks
    - Question not asked enough: what do they do?
  - Flow issue: which node is central depends on type of flow
    - Interaction between rules of flow & position in generating node outcomes
- Can now choose better - combine new measures with combinatorial optimization and simulation
  - Work needed to develop analytic solutions for all flows

# Understanding Centrality

(long term goal)

- Off-the-shelf centrality measures make assumptions about the way things flow
  - When applied to the “wrong” flows, they get the “wrong” answers
    - Future research: which structures are most problematic?
- Few existing measures are appropriate for the kinds of flows we are most interested in
  - Gossip & information
  - Infections
- Flow characteristics interact with node position to determine node participation

# Understanding Centrality

- Advocate seeing centrality from modeling perspective
  - Formulas yield expected values of key node outcomes under specific models of how things flow
- Two such outcomes identified:
  - Time: Who gets things early (closeness)
  - Frequency: Who gets a lot of traffic (betweenness)
- Clarity on what is being measured allows ..
  - Comparing formulas with observed values
  - Constructing new formulas appropriate for other flows
  - Testing hypothesis about unobserved flow parameters
- Can use simulations to construct centrality scores for each kind of flow
  - But analytical solutions (e.g., Newman) would be better

} Parallel  
w/ Markov  
world

# Some Next Steps

- Typology of centrality measures in terms of what is really being measured
  - Are medial versus radial measures truly different?
  - Do the averaging operations of influence measures constitute a different class of measures?
- Developing analytical solutions for various flow processes
- Understanding how network structure interacts with flow characteristics
- Abandoning targets in simulations

Enough.

# Ensemble Issues

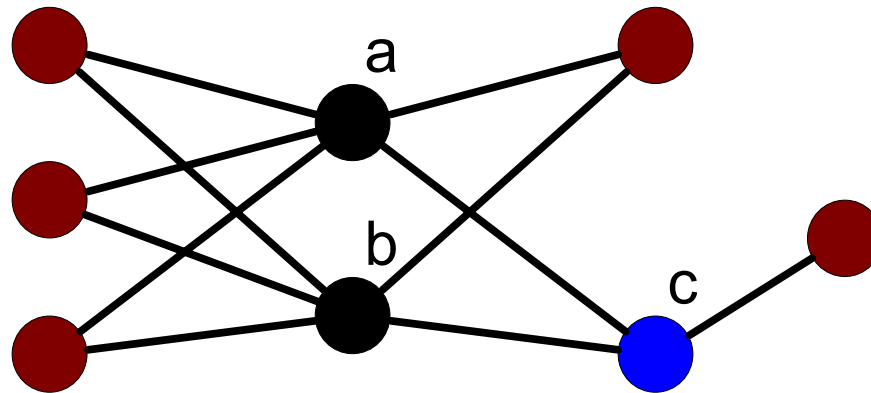
What if you wanted to pick a *set* of key players?

# Naive Strategy

- Measure centrality of each node, using appropriate measure of centrality
- Pick the top  $k$  nodes for your key player set

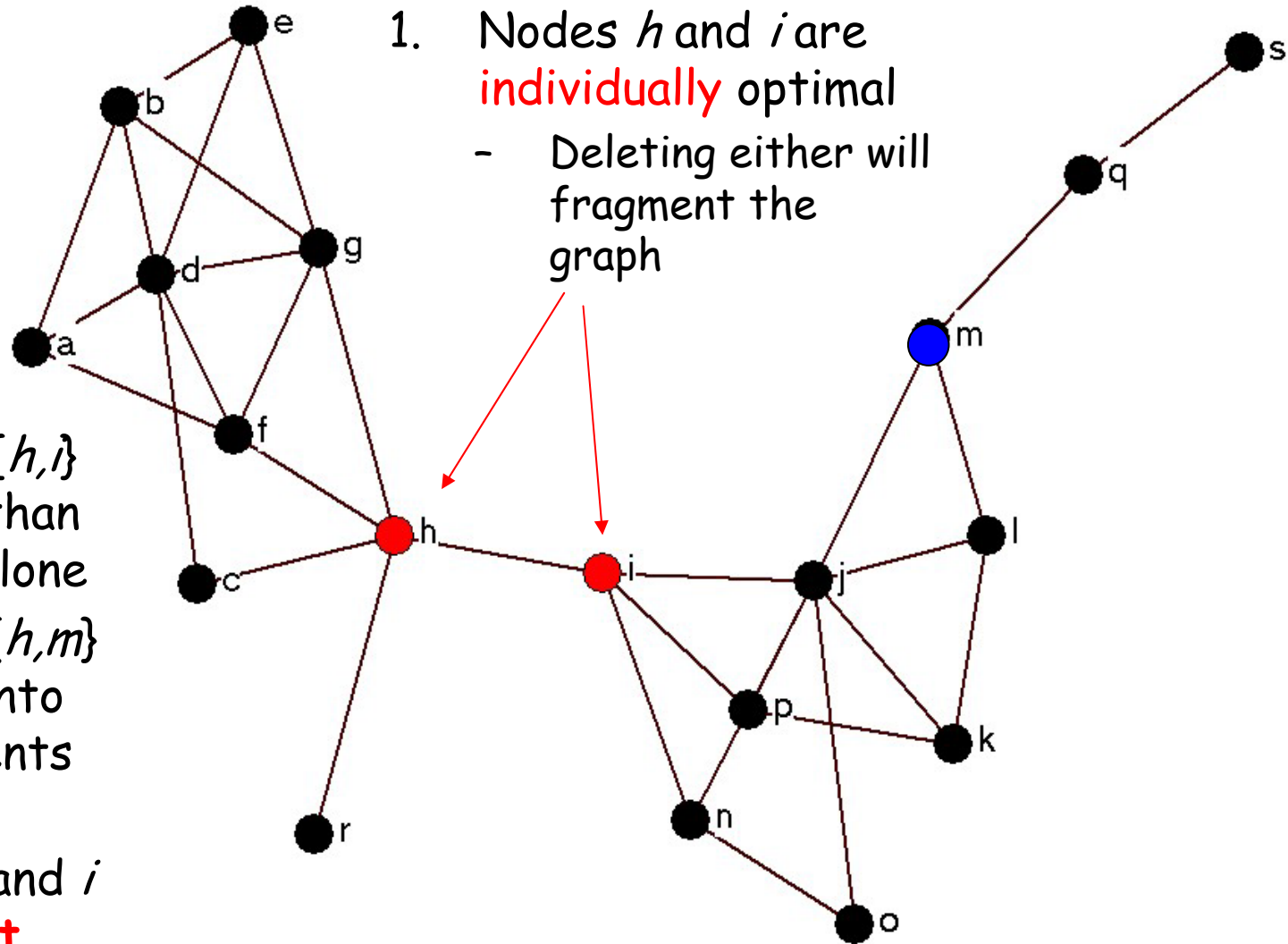
# Maximizing Spread

- Node **a** and Node **b** are individual the best connected
  - Each reaches 5 others in one step



- But the set  $\{a, b\}$  is no better than **a** or **b** alone
  - Problem is that **a** and **b** are redundant in sense of structural equiv
- In contrast, the set  $\{a, c\}$  reaches all nodes in the network
  - Even though **c** is not well connected

# Minimizing Spread



1. Nodes  $h$  and  $i$  are **individually** optimal

- Deleting either will fragment the graph

2. But deleting  $\{h, i\}$  is **no better** than deleting  $\{h\}$  alone

3. In contrast,  $\{h, m\}$  splits graph into three fragments (is optimal)

4. Problem is  $h$  and  $i$  are **redundant**

# Sidebar

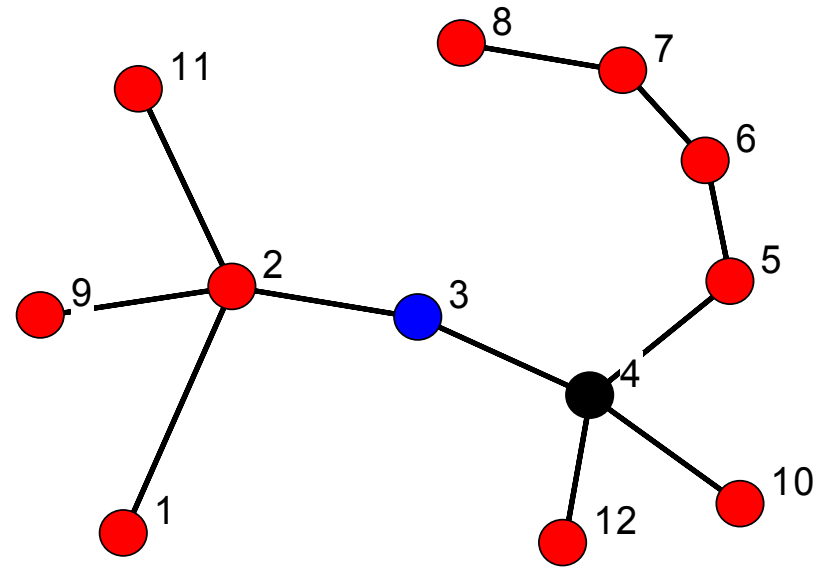
- The two kinds of redundancy worth exploring
- Need for further development on group centrality ala Everett & Borgatti (1999) and Vito (in progress)

# Design Issues

Are off-the-shelf centrality  
measures *optimal*?

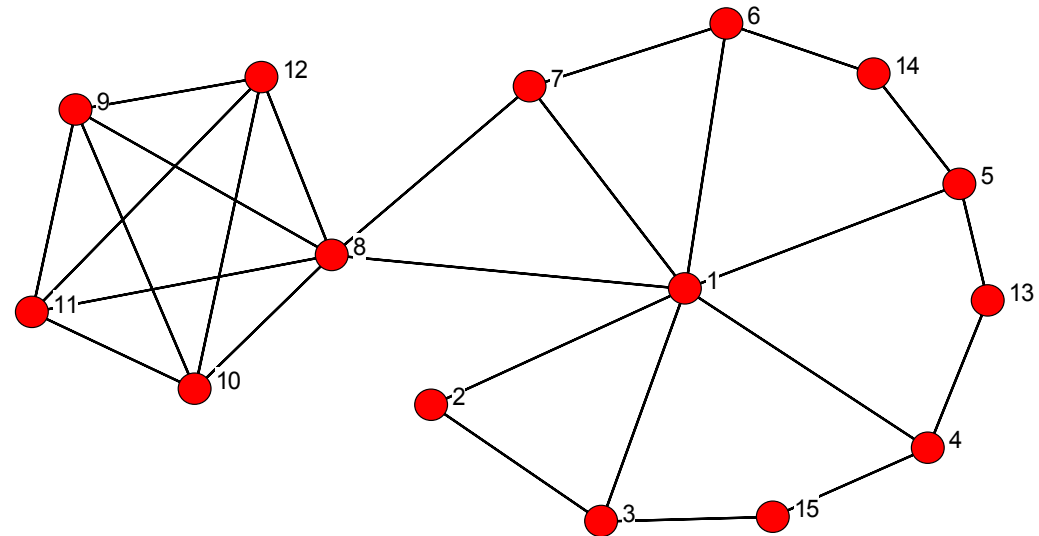
# Maximizing Spread

- Closeness centrality seems best suited
- Node 4 has greatest closeness centrality
- But if we want to reach the most nodes within 2 links distance, we should select node 3
  - Node 4 reaches six others
  - Node 3 reaches eight others



# Minimizing Spread

- Betweenness seems best suited
- Node 1 has highest betweenness, but deleting it ...
  - does not disconnect the network, and
  - increases distances only marginally
- In contrast, deleting node 8 breaks network into 2 components
  - Yet node 8 is not highest in centrality
  - So centrality is not optimal for this purpose



Now I'm really done.

# Degree

- Is proportional to number of visits in random walk through network
- Consistent with walk-based transfer processes such as money exchange process

	parallel duplication	serial duplication	transfer
geodesics	<no process>	FreeClo	FreeClo Betweenness
paths	FreeClo Degree		
trails	FreeClo Degree		
walks	FreeClo Eigenvector Degree		Degree