Centrality II

What is centrality?

- “prominence” or structural importance
- Influence, power, status, control, independence, information

Minimum criteria

- Sabidussi
  - Adding a tie to node cannot reduce centrality
  - Adding a tie anywhere in network cannot reduce centrality of a given node
  - Etc
- Freeman
  - Must achieve maximum value for the center of a star

Involvement in path structure

- Borgatti and Everett

Assumptions of std measures

- Degree
  - Only paths of length 1 considered
- Closeness & betweenness
  - Only shortest paths counted
- Flow betweenness
  - Edge-independent paths of all lengths
- Eigenvector, katz, hubbell, bonacich etc.
  - Unrestricted walks

Dimensions of similarity / difference

- Traversal type: geodesics, paths, trails, walks, independent paths etc
- Summarization type: sums, averages, minimums, etc.
- Traversal property: frequency or length?
  – The no. of traversals of various kinds that a node is involved in
  – The length of traversals that involve a node
- Node position: radial or medial?
    - Walks emanating from / terminating with a node
    - Walks passing through a node

Classification of Measures

- Note: summarization type suppressed

Paths
Trails
Walks

Units	**Radial** (emanating to/from node)	**Medial** (passing thru node)
Frequency	(a) degree, k-path centrality, reach, eigenvector, Hubbell, GPI	(c) betweenness, flow betweenness, proximal betweenness
	Katz, Bonacich power, Alpha Centrality	
Length	(b) closeness, information, current flow closeness	(d) < no well-known measures >

Defining centrality – cont.

- Borgatti and Everett argued that centralities measure the involvement of nodes in the paths of the network
  - Radial measures count paths originating from (or terminating) at a node
  - Medial measures count paths passing through a node
  - Within these classes, measures differ based on what kinds of paths are examined
    - Shortest paths; Independent paths; Paths of length 1, etc

Expected values of flow outcomes

How do the assumptions of the measures match different kinds of real flow processes?

What are some things that flow through networks?

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

Borgatti, S.P. 2005. Centrality and network flow. *Social Networks*. 27(1): 55-71.

Letters

- 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
  - (hopefully) travels along shortest paths (geodesics)

Used Goods

- Canonical example:
  - passing along 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 along graph-theoretic trails

Money Exchange Process

- Examples:
  - specific dollar bill moving through the economy
  - Erdös 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

Viral Infection Process

- Example:
  - virus which activates effective immunological response (including preventing carrying) or which kills host
- Multiple copies may exist simultaneously
- Cannot revisit a node
    - A--B--C--E--D--F...
    - Travels along graph-theoretic paths

Homeless Relative

- Examples
  - Obnoxious homeless relative who visits for six months until kicked 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

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.

Flow typology

goods

information

<table>
<tr><th></th><th>parallel duplication</th><th>serial duplication</th><th>transfer</th></tr>
<tr><td>geodesics</td><td rowspan="2">internet name-server</td><td>mitotic reproduction</td><td>package delivery</td></tr>
<tr><td>paths</td><td>viral infection</td><td>homeless relative</td></tr>
<tr><td>trails</td><td>e-mail broadcast</td><td>gossip</td><td>used goods</td></tr>
<tr><td>walks</td><td>attitude influencing</td><td>emotional support</td><td>money exchange</td></tr>
</table>

Markov

Which processes are off-the-shelf centrality measures appropriate for?

Degree: No. of edges incident upon a node
Closeness: Sum of geodesic distances to all other nodes
Betweenness: Share of geodesics that pass through given node
Eigenvector: No. of walks emanating from node, wtd inversely by length

	parallel duplication	serial duplication	transfer
geodesics	Closeness	Closeness	Closeness Betweenness
paths			
trails			
walks	Eigenvector		Random Walk Betweenness; Degree

"Mind the gap"

Two questions

- What if we use a centrality measure that is compatible with one kind of flow in a situation involving a different flow? E.g.,
  - Suppose you use betweenness, but what you are studying doesn't flow via shortest paths only?
  - What if what you are studying flows along multiple paths at the same time? Betweenness assumes a single path …
- How do the standard measures relate to our theoretical variables
  - The expected amount of time until arrival of flow at a node
  - How likely (how often) the flow reaches a given node

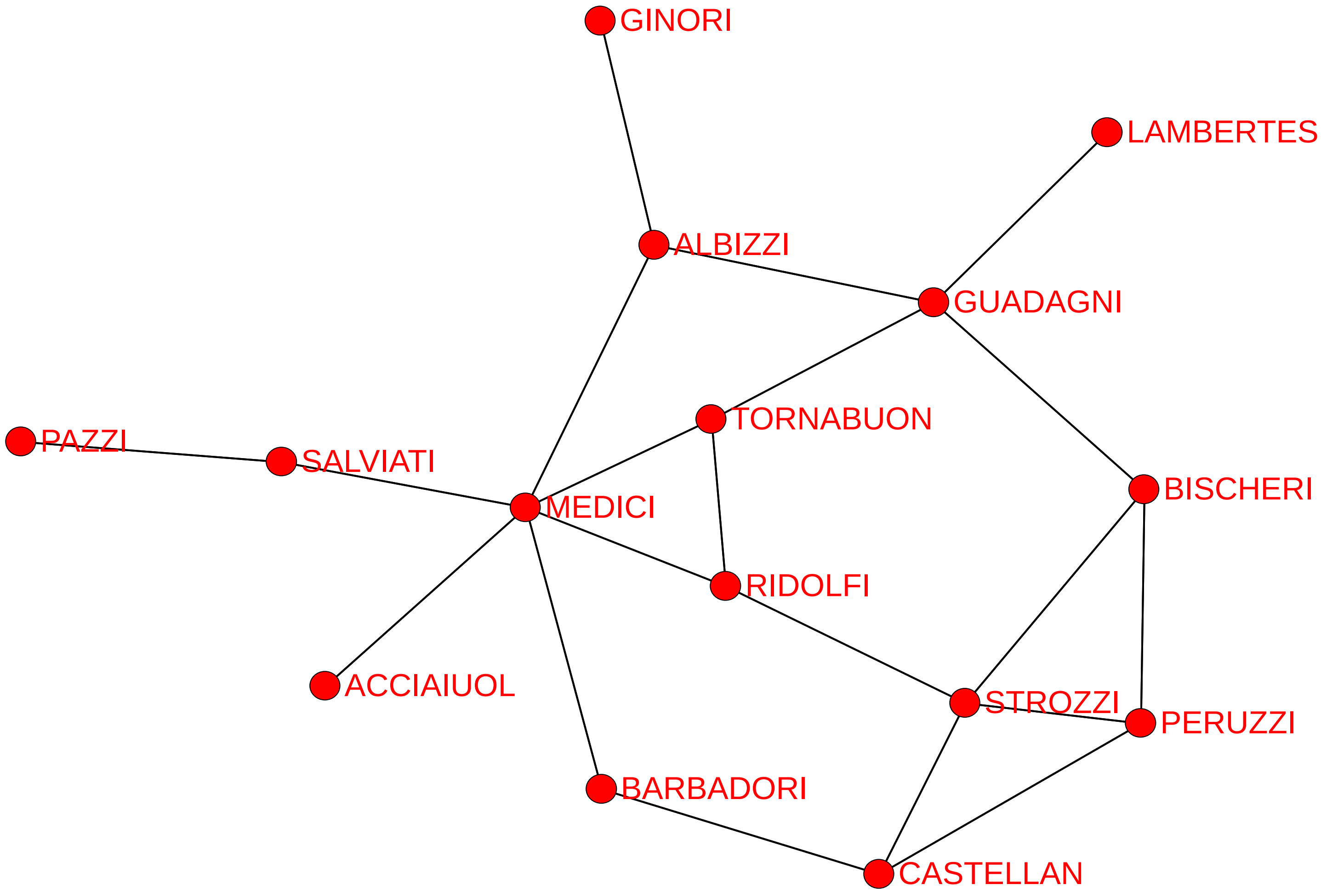
Motivation

- Centrality often used to predict performance
  - More central nodes have better access to information, resources – whatever flows through network
  - "better" means
    - More likely to receive it
    - Receive it sooner
- Can we use standard measures of centrality for this?

Simulation Experiment

- Given a network along which something flows
- Repeat 10,000 times:
  - Let traffic flow according to the rules of a given flow process
  - For each node, measure
    - Time. Time of first arrival at every node
    - Frequency. No of times arriving at each node
- Compare with standard centrality measures
- Repeat for different kinds of flow

Illustrative Dataset



Padgett & Ansell (1991). Marriage ties among Florentine families during the Renaissance

Simulation Results

Frequency of Visits

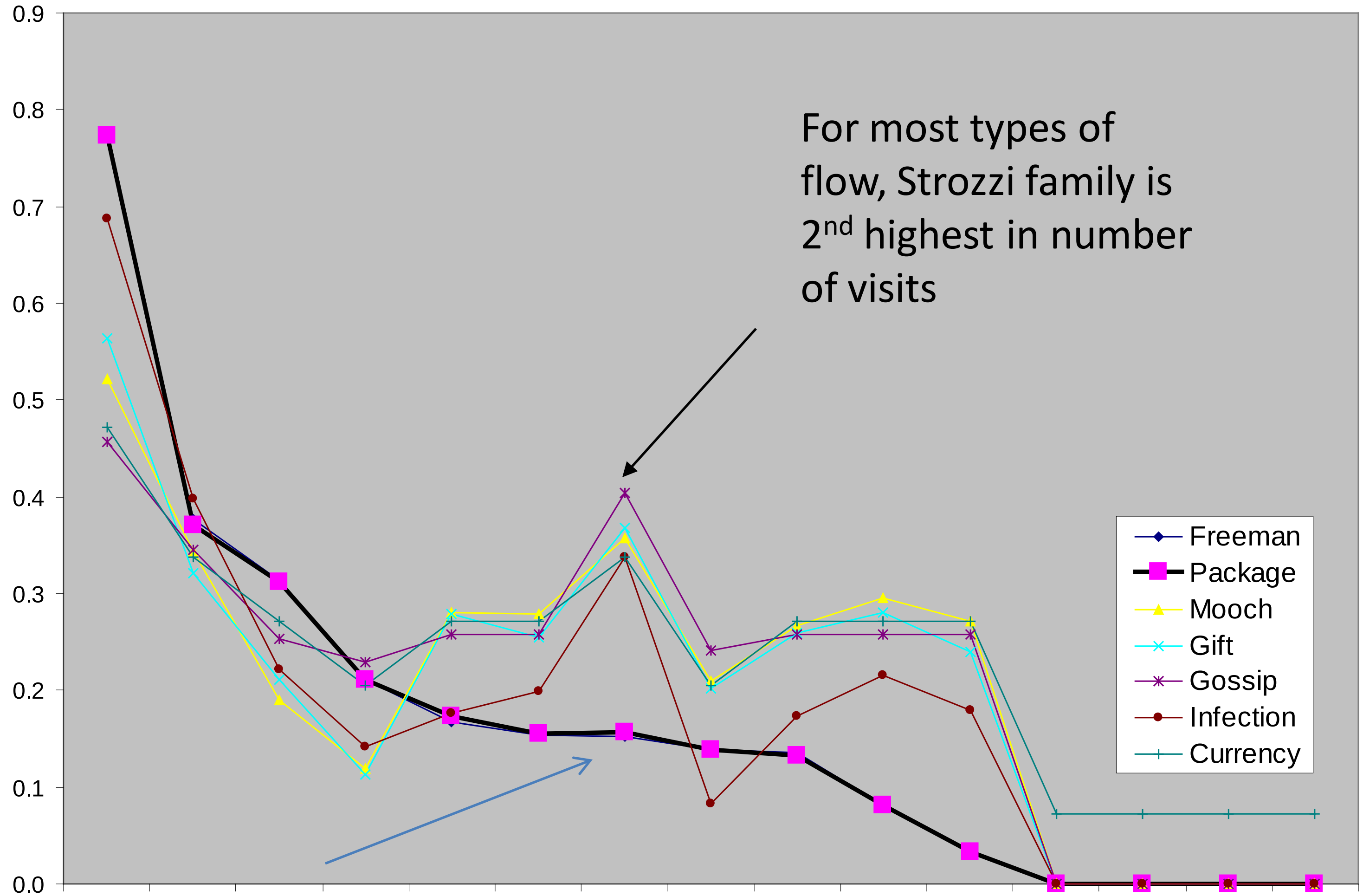
Proportional to degree

Exact match

Node	Freeman Betweenness	Package	Homeless	Used Goods	Gossip	Virus	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 each node en route from source to target

Betweenness / Freq of Visits



Freeman betweenness underestimates importance of Strozzi family

Frequency of Arrivals

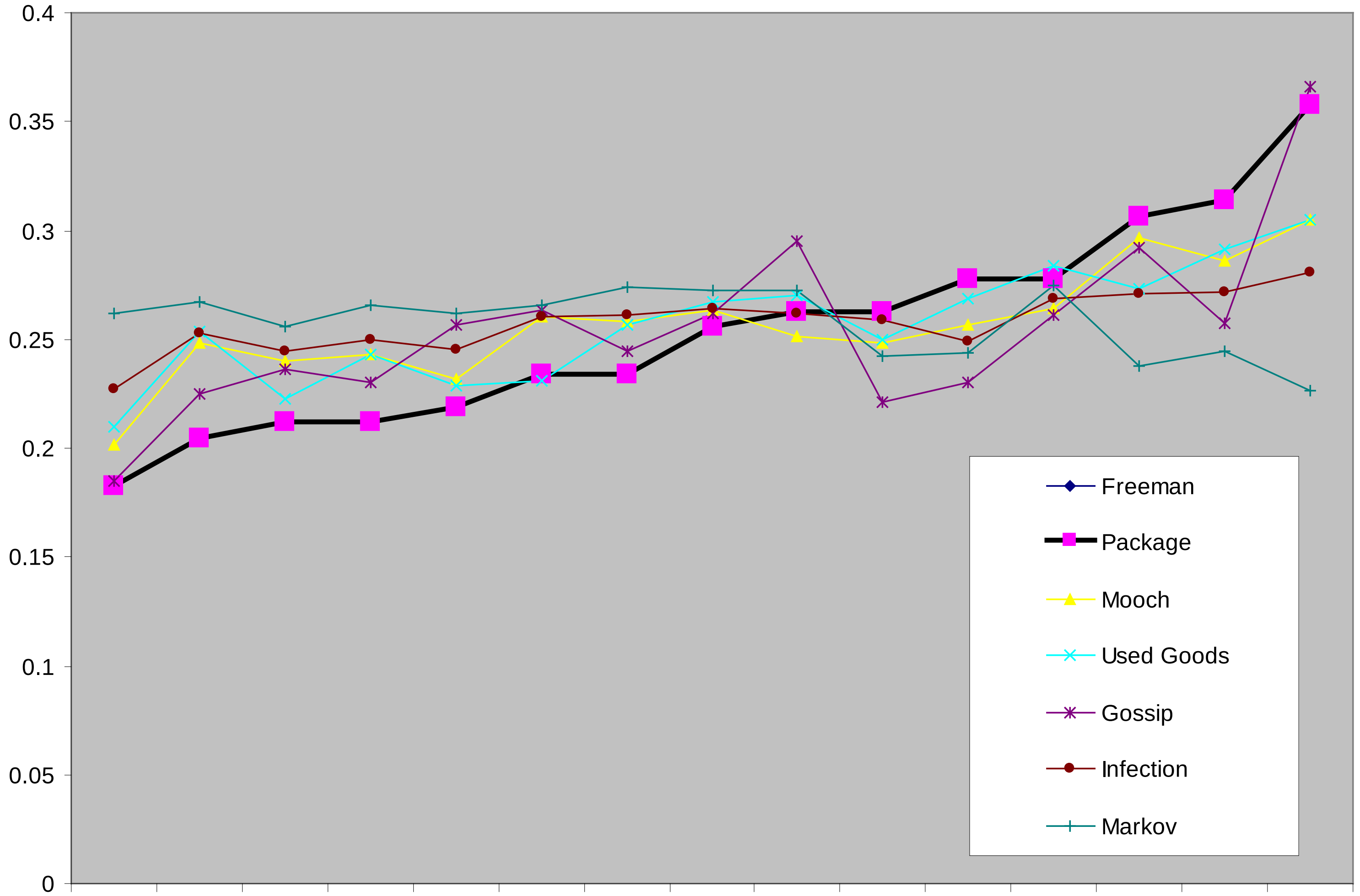
- Freeman betweenness definition gives exact expected values for frequency of visits in *package delivery* process (transfer+geodesics)
  - And **only** the package delivery process
- Other kinds of flow have different outcomes
  - Strozzi family strongly undervalued by Freeman measure
  - Misidentification of topmost central actors
- Also as predicted, *money exchange* process (transfer+walks) yields scores exactly proportional to degree centrality
  - For that process, degree and betweenness are indistinguishable concepts

Closeness / Time to Arrival

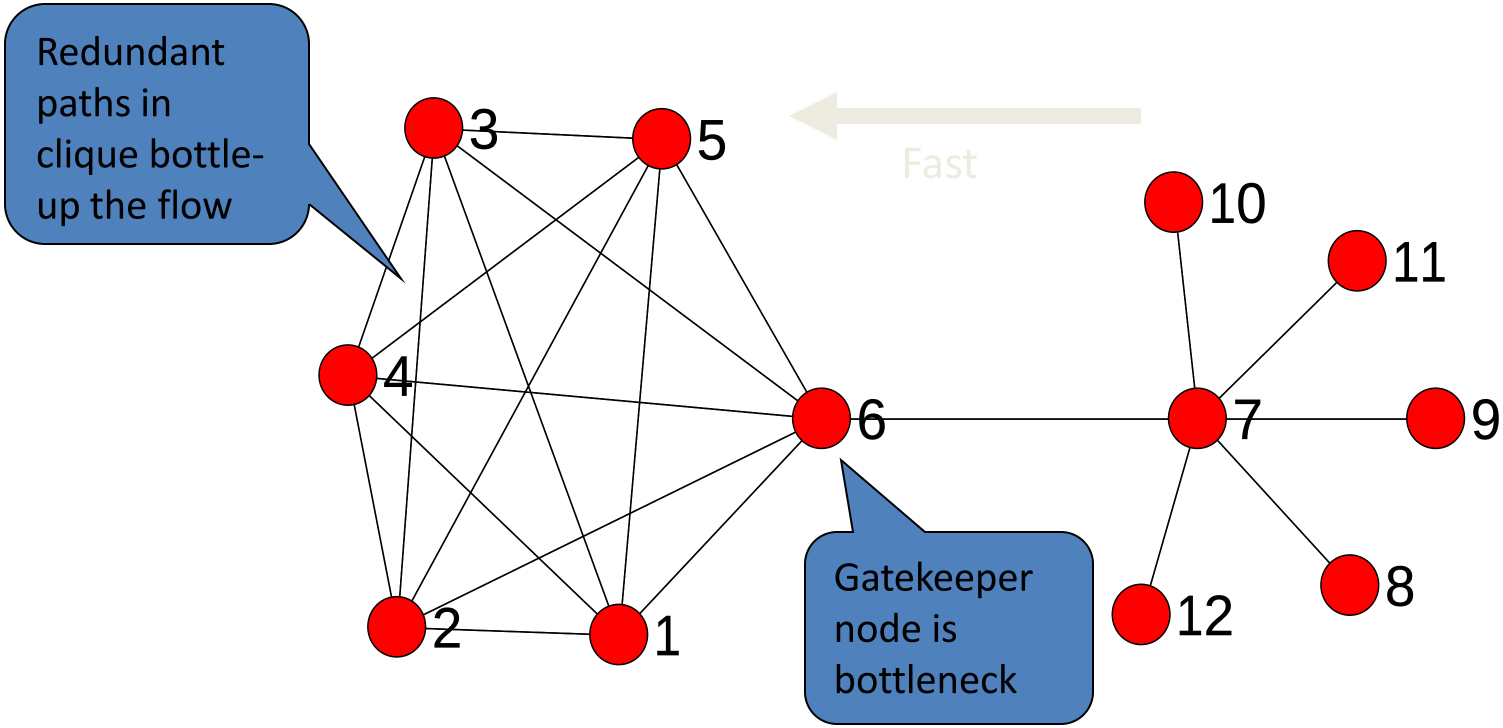
Node	Freeman	Package	Homeless	Used Goods	Gossip	Virus	Money
MEDICI	25	25.0	46.7	50.1	78.9	63.7	575.2
RIDOLFI	28	28.0	57.5	60.6	95.7	70.8	587.7
ALBIZZI	29	29.0	55.7	53.3	100.7	68.6	562.3
TORNABUON	29	29.0	56.4	58.1	98.2	70.0	584.8
GUADAGNI	30	30.0	53.7	54.8	109.3	68.8	575.3
BARBADORI	32	32.0	60.5	55.3	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 for first time

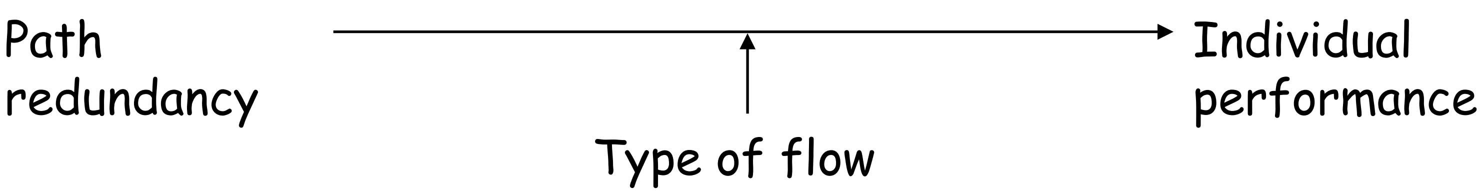
First Arrival Times



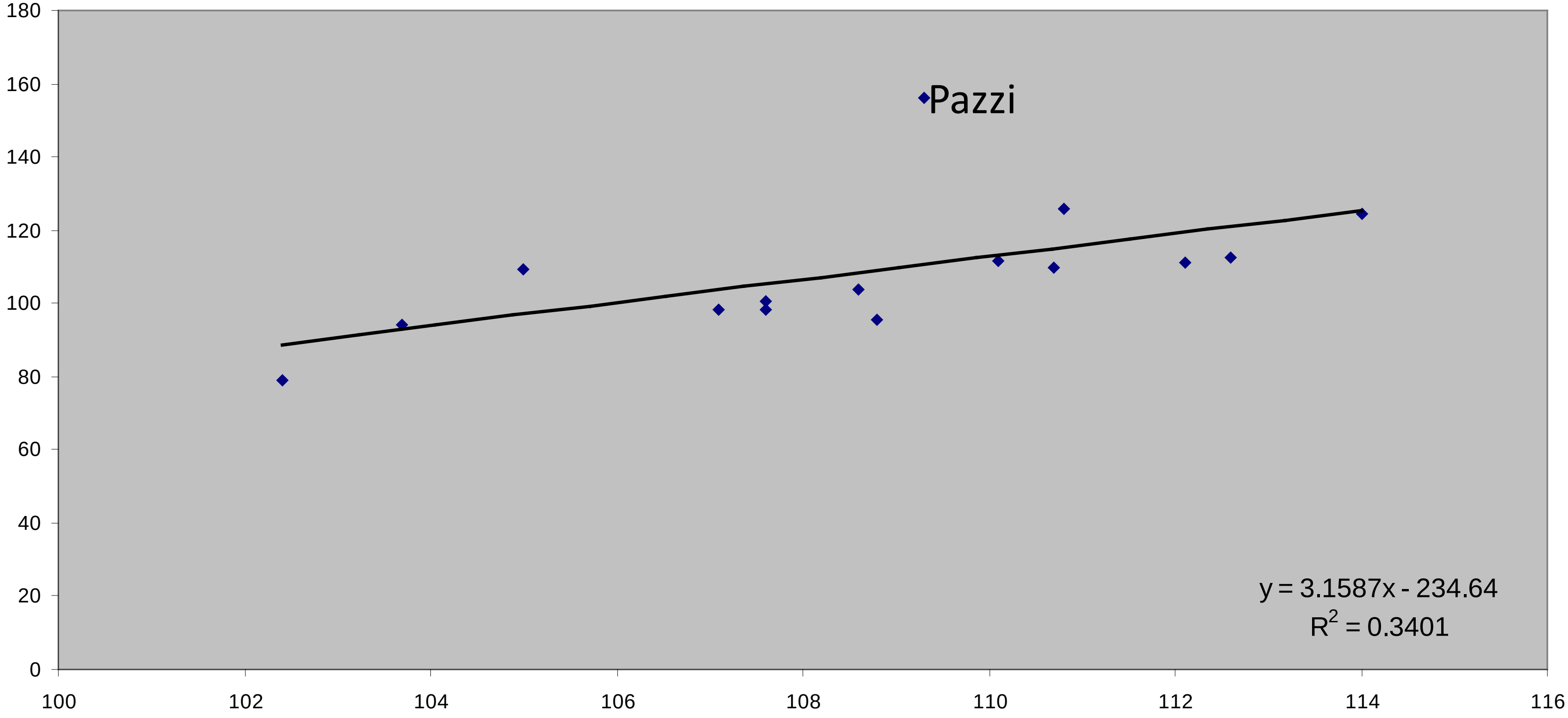
Closeness Asymmetry



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



Comparing in-flow and out-flow



Arrival Times

- Like betweenness, Freeman closeness measure gives correct values in package delivery process, but not other processes
- Centrality measures on undirected graphs necessarily give same prediction for time until arrival as time to reach others, but in reality these are not the same
  - Proximity to hub is better for spreading than receiving

Which processes are off-the-shelf centrality measures appropriate for?

Degree: No. of edges incident upon a node
Closeness: Sum of geodesic distances to all other nodes
Betweenness: Share of geodesics that pass through given node
Eigenvector: No. of walks emanating from node, wtd inversely by length

	parallel duplication	serial duplication	transfer
geodesics	Closeness	Closeness	Closeness Betweenness
paths			
trails			
walks	Eigenvector		Random Walk Betweenness; Degree

“Mind the gap”

Centralities as Statistical Models

- Given explicit model of flow process, centrality measures can be seen as expected values for node outcomes, e.g.,
  - first arrival times
  - freq of arrivals
- Off-the-shelf measures of centrality only appropriate for certain flow processes
- Analytic formulas for all flow processes not currently available
  - But can use simulation to estimate values

Answer:

- If what flows does so
  - through shortest paths only , and
  - can only follow one path at a time
- Then
  - The expected time until arrival at node k is proportional to the closeness centrality of node k
  - The expected number of times that node k is visited is proportional to betweenness centrality

POWER VERSUS CENTRALITY

DIRECTED DATA

Degree Centrality

- Concept
  - Number of ties a node has
- Directed case
  - Indegree: colums sums of adjacency matrix
  - Outdegree: row sums
- Scatter plot:

Indegree ↑

Authority	High involvement
Low involvement	Apprentice

Outdegree →

	Mary	Bill	John	Larry	Out
Mary	0	1	1	1	3
Bill	1	0	1	0	2
John	0	0	0	1	1
Larry	0	0	0	0	0
In degree	1	1	2	2	6

Closeness Centrality

- Concept
  - Distance from/to all other nodes
- Directed
  - Row and column sums of the distance matrix
- Problems
  - Directed graphs usually not connected. Many distances undefined
- Alternative
  - Sum reciprocals the distance matrix instead. Substitute zeros whenever a distance is undefined
  - Or count number of nodes reached

Betweenness

- Concept
  - How often a node lies along a geodesic path between two others

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

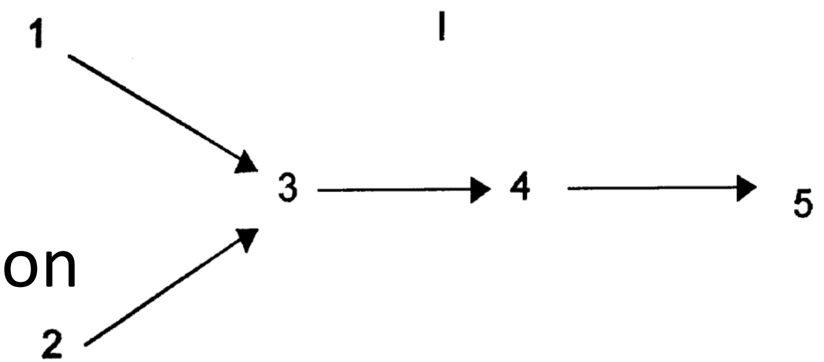
- Directed graphs
  - No adjustment needed

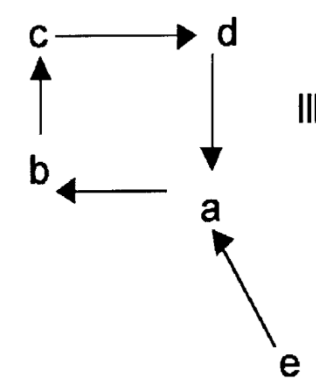
Eigenvector

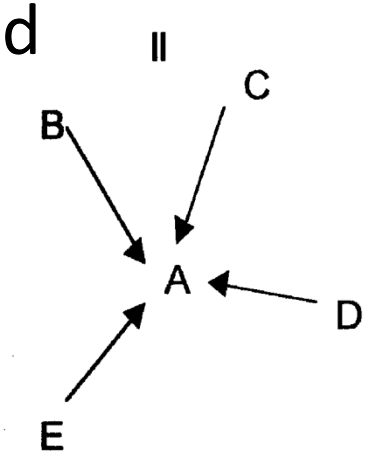
- Concept
  - A person is central to the extent they are connected to many people who are well connected (to people who are well … etc)
- Directed graphs
  - (columns) A person has high status to the extent that they are nominated by many people who are themselves frequently nominated
    - Left eigenvector $\mathbf{x}'A = \lambda\mathbf{x}$ or $A'\mathbf{x} = \lambda\mathbf{x}$
  - (rows) A person has influence to the extent they influence many who themselves influence many
    - Right eigenvector $A\mathbf{x} = \lambda\mathbf{x}$

Eigenvector for Directed graphs

- Often not calculable
- Can give useless answers
  - Nets I and II give all zeros on left eigenvec for all nodes
    - Nodes with 0 indegree have no status to pass along …
  - In net III, nodes *a, b, c* and d have same score, even though *a* has greater indegree







Figures from Bonacich and LLoyd

Alpha Centrality

- Same as eigenvector when applied to symmetric matrices, but better results when applied to non-symmetric matrices
- Basically same as measures by Katz and Hubbell
  - Right alpha centrality: $\boldsymbol{x} = \alpha A\boldsymbol{x} + \boldsymbol{e} = (I - \alpha A)^{-1}\boldsymbol{e}$
    - Assume **e** is vector of 1s
  - left alpha centrality: $\boldsymbol{x} = \alpha A^T\boldsymbol{x} + \boldsymbol{e} = (I - \alpha A^T)^{-1}\boldsymbol{e}$
- In left (right) alpha centrality …
  - If α is positive then a person gets a high score for receiving ties from (sending ties to) people with high scores
  - If α is negative, then a person gets a high score for receiving ties from (sending ties to) people with low scores

Katz Influence

- If i does not have a tie to j, i can still influence j by influencing someone who influences someone … who influences j.
  - more chains from I to j, the more certain the influence,
  - but also the longer the chains the weaker the influence
- Given adjacency matrix R, the number of chains of length k is given by R^k , so we need a sum like this: $\mathbf{R}^1 + \mathbf{R}^2 + \mathbf{R}^3 + ...$ except we want to weight the longer chains less
- A parameter α^k (smaller than 1) can be introduced which goes to zero as k approaches infinity
  - $\mathbf{Q} = \alpha^1\mathbf{R}^1 + \alpha^2\mathbf{R}^2 + \alpha^3\mathbf{R}^3 + ... \alpha^\infty\mathbf{R}^\infty$
  - The row sums of Q give the total influence of a node on the network
- It turns out that when $\alpha < 1/ \lambda_1$ where λ_1 is the largest eigenvalue of R, this series converges to $\mathbf{Q} = (\mathbf{I}\text{-}\alpha\mathbf{R})^{-1} – 1$, which leads to a row sum that is just 1 less than alpha centrality

Singular Value Decomposition (SVD)

- Every matrix A can be decomposed as follows:

$$A_{n\times m} = U_{n\times m} D_{m\times m} V^T_{m\times m}$$

D is a diagonal matrix of singular values

- We can approximate A with lower dimensionality k << m

$$A_{n\times m} = U_{n\times k} D_{k\times k} V^T_{m\times k}$$

- A 1-dimensional solution:

$$A = u\lambda^{1/2}v'$$

- The u-scores and column scores can be written in terms of each other

$$u_i = \lambda^{-1/2}\sum_j a_{ij}v_j$$

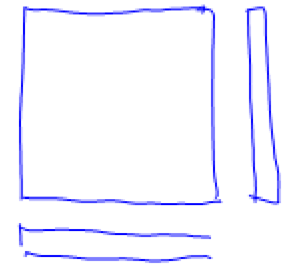
$$v_j = \lambda^{-1/2}\sum_i a_{ij}u_i$$

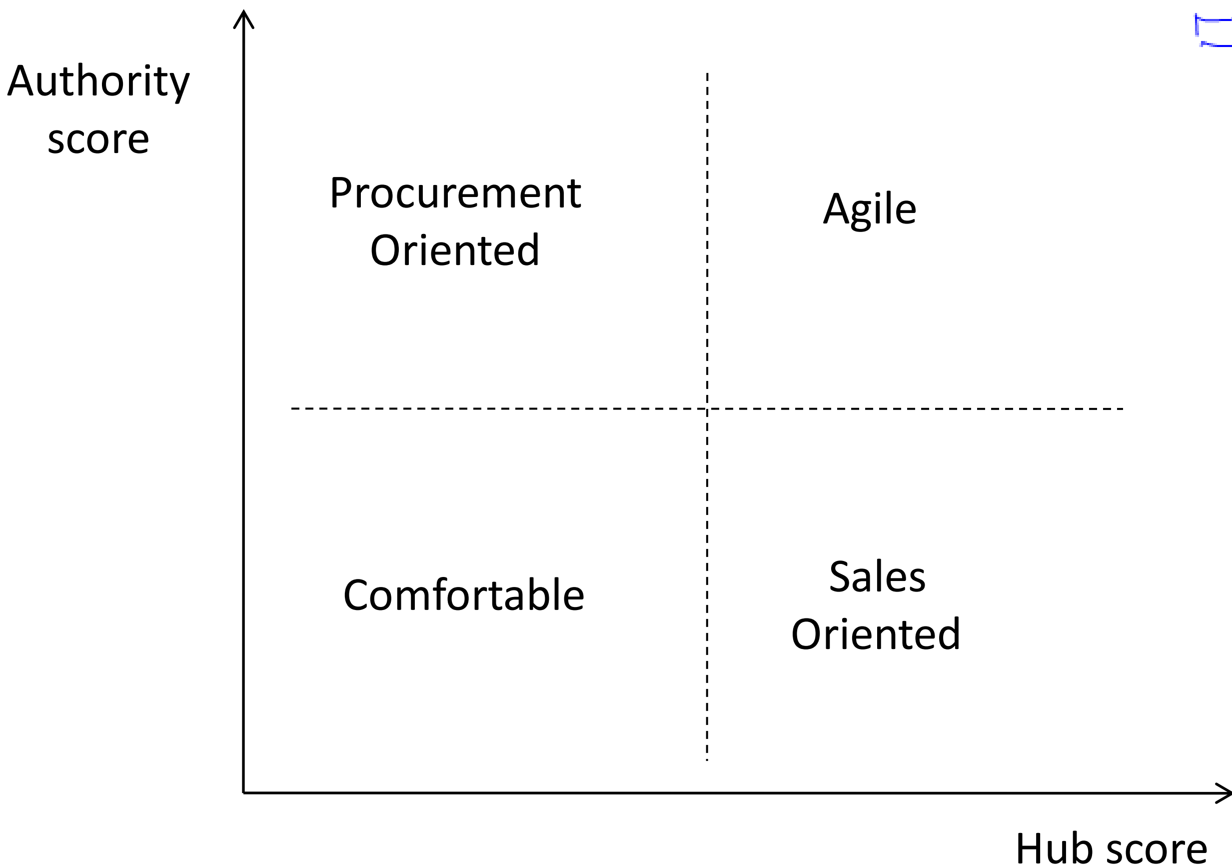
Hubs and Authorities

- Run an SVD on an adjacency matrix A, and retain only the first dimension
$$A = u\lambda^{1/2}v'$$
- The u and v scores measure the extent to which a node is playing the role of a hub or authority respectively
  - The u-score (hub) measures the extent to which the node sends ties to nodes that have high v-scores (are authorities)
  - The v-score (authority) measures the extent to which the node receives ties from nodes with high u-scores (are hubs)

Supply chain example

- Seller by buyer matrix





KEY PLAYERS

Key Player Project

Who are the key players in a network?

- It depends on …
  - whether you are looking for individuals or ensembles
  - the purpose
- On the value of problem-centered research

Funded by the
Office of Naval Research
Thanks Rebecca Goolsby!

Borgatti, S.P. 2006. Identifying sets of key players in a network. *Computational, Mathematical and Organizational Theory*. 12(1): 21-34

Borgatti, S.P. 2003. The Key Player Problem. Pp. 241-252 in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, R. Breiger, K. Carley, & P. Pattison, (Eds.), National Academy of Sciences Press.

Why do we want to know who the key players are?

We want to **remove them** – to maximally **disrupt** the network	**DISRUPT**
We want to **help** them – in order to make network as a whole **function better**	**ENHANCE**
We want to identify key opinion leaders – to influence the network	**INFLUENCE**
We want to know who is in the know – so we can question or surveil them	**LEARN**
We want to remove them – to redirect flows in the network toward more convenient players -- pruning	**REDIRECT**

Key Player Needs by Field

	DISRUPT	PROTECT	INFLUENCE	LEARN	REDIRECT
SECURITY	Who to **arrest or discredit** to disrupt ops	Who to **protect** among allied group	Who to turn or plant info with	Who is best positioned to know most	Who to remove to redirect flows
PUBLIC HEALTH	Who to **immunize or quarantine**		Who to select as PHAs for interventions	Who to study explain spread	
MANAGE MENT	Who to hire away from competitor	Who to give more of a stake in org to avoid turnover	Who to get on board before launching reorg		Who to add/replace to remove drag on good emps
MARKETING	Identify key critics to silence	Which happy users to empower	Identify key mavens to sell on your stuff	Identify key informants for focus	

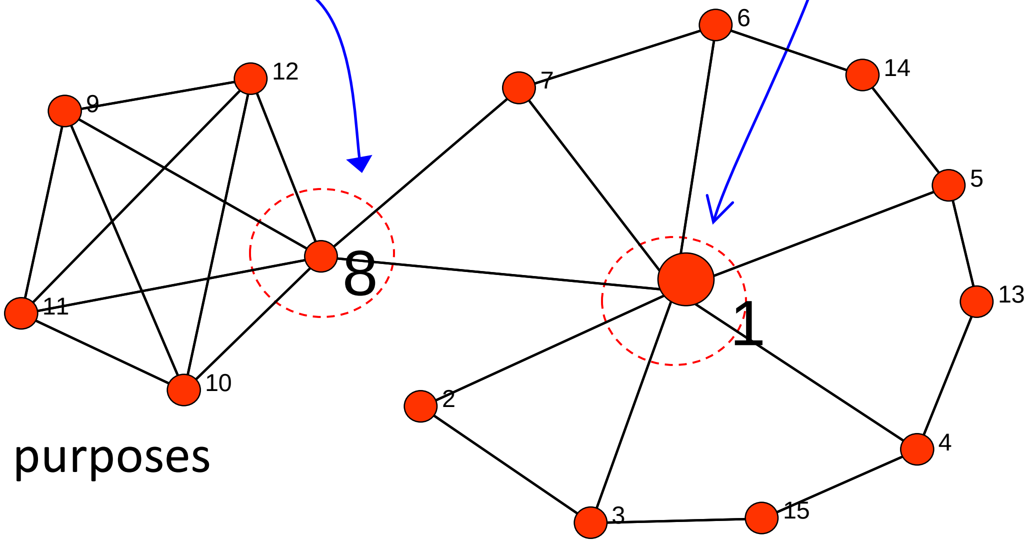
KeyPlayer Research Objectives

- Develop metrics to quantify potential disruption, influence, surveillance etc.
  - Off-the-shelf SNA measures not optimized for these tasks
- Develop combinatorial optimization algorithms and fast heuristics for maximizing metrics given solution parameters
- Predict what happens to the network post-intervention

DISRUPTION

The Design Issue

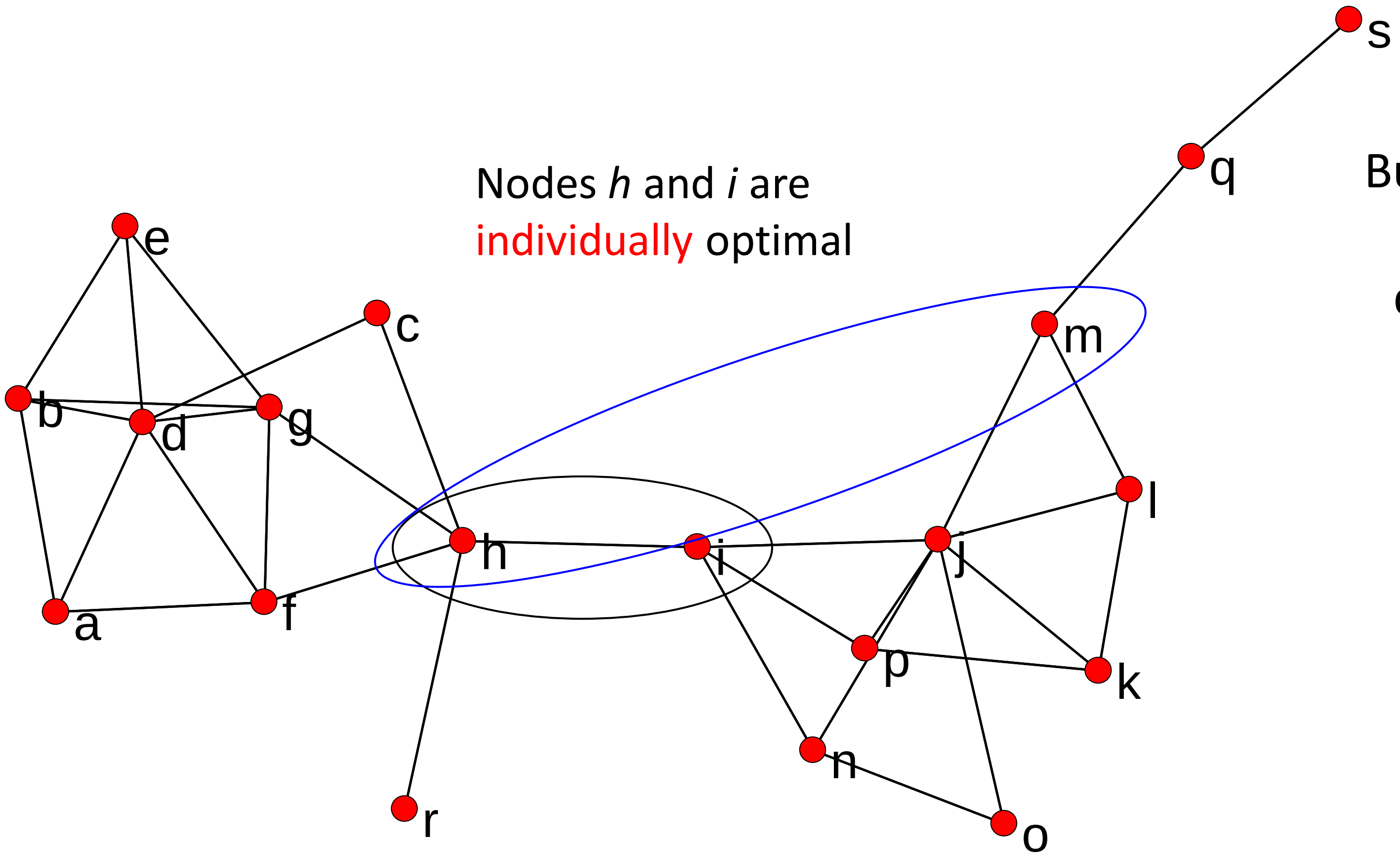
- By standard off-the-shelf measures of node centrality, node 1 is the most important player, but deleting it …
  - does not disconnect the network
- In contrast, deleting node 8 breaks network into two components
  - Yet node 8 is not highest in centrality
- No off-the-shelf centrality measure is optimal for the purpose of disrupting networks
  - Nor any of the other specific purposes



DISRUPTION

The Ensemble Issue

Structural redundancy creates need for choosing complementary nodes



Nodes *h* and *i* are individually optimal

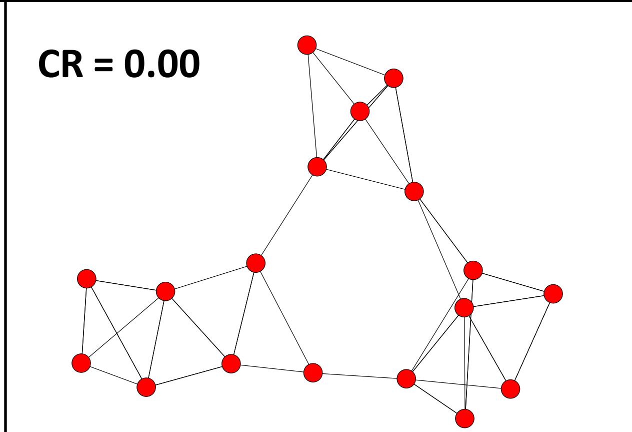
But deleting both is no better than deleting *h* alone -- *h* and *i* are redundant

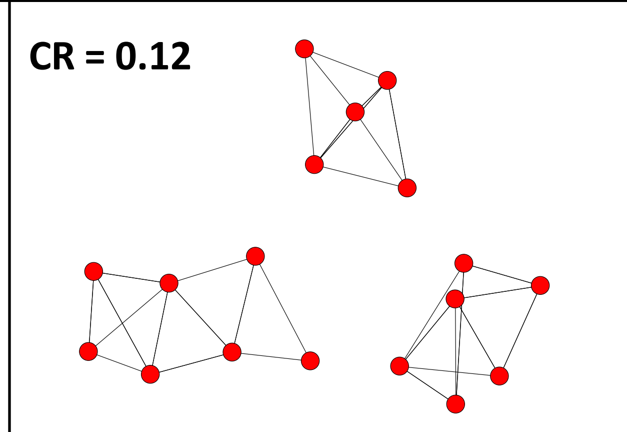
In contrast, {*h*,*m*} splits graph into 4 fragments (is optimal)

- Choosing optimal **set** of *k* players is not same as choosing the *k* best players

No. of components

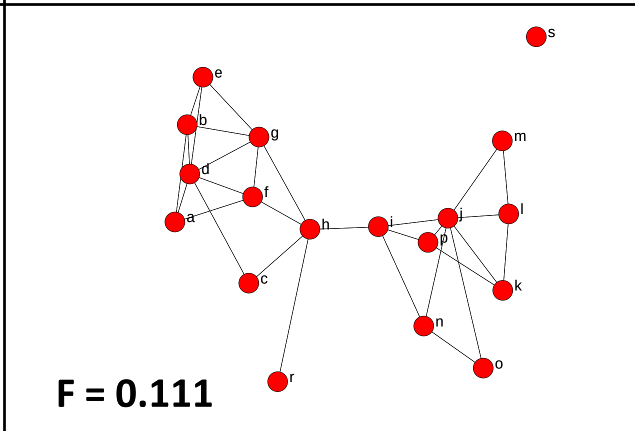
$$CR = \frac{c-1}{n-1}$$

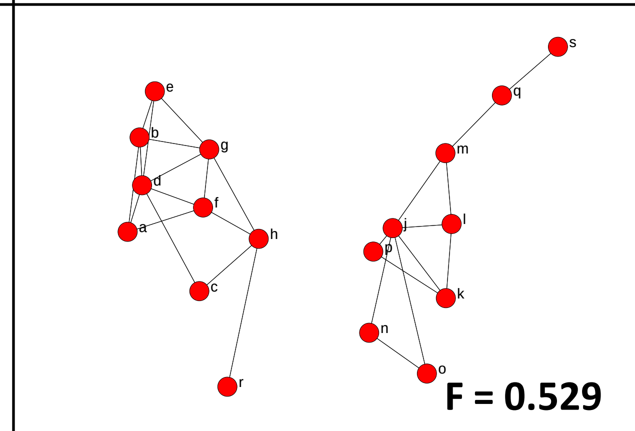




No. of disconnected pairs

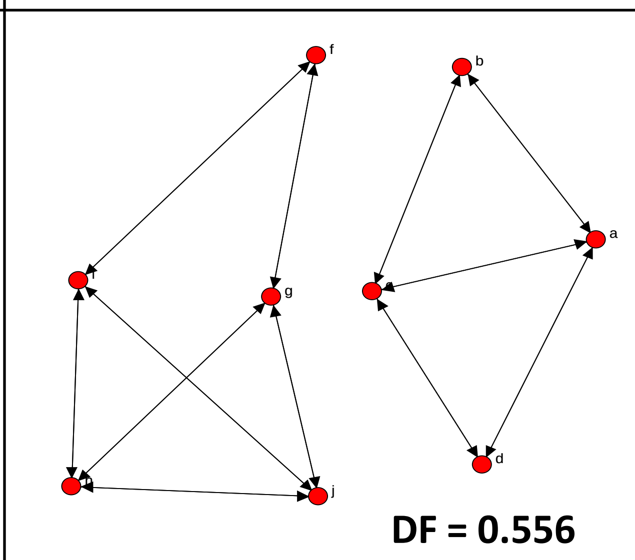
$$F = 1 - \frac{2\sum_{j<i} r_{ij}}{n(n-1)}$$

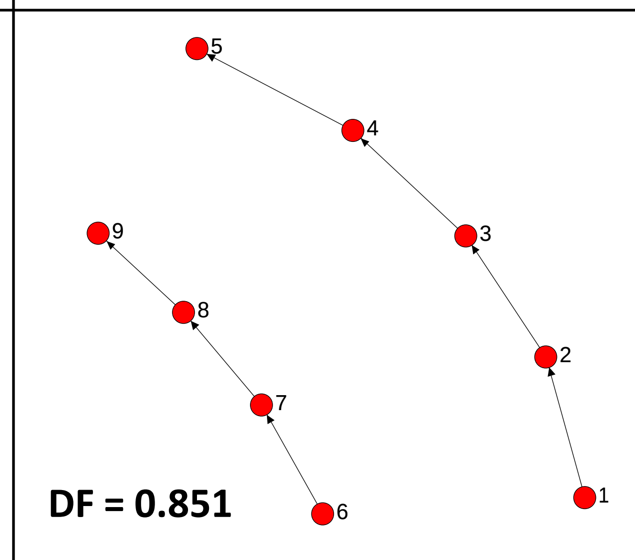




Distance-weighted fragmentation

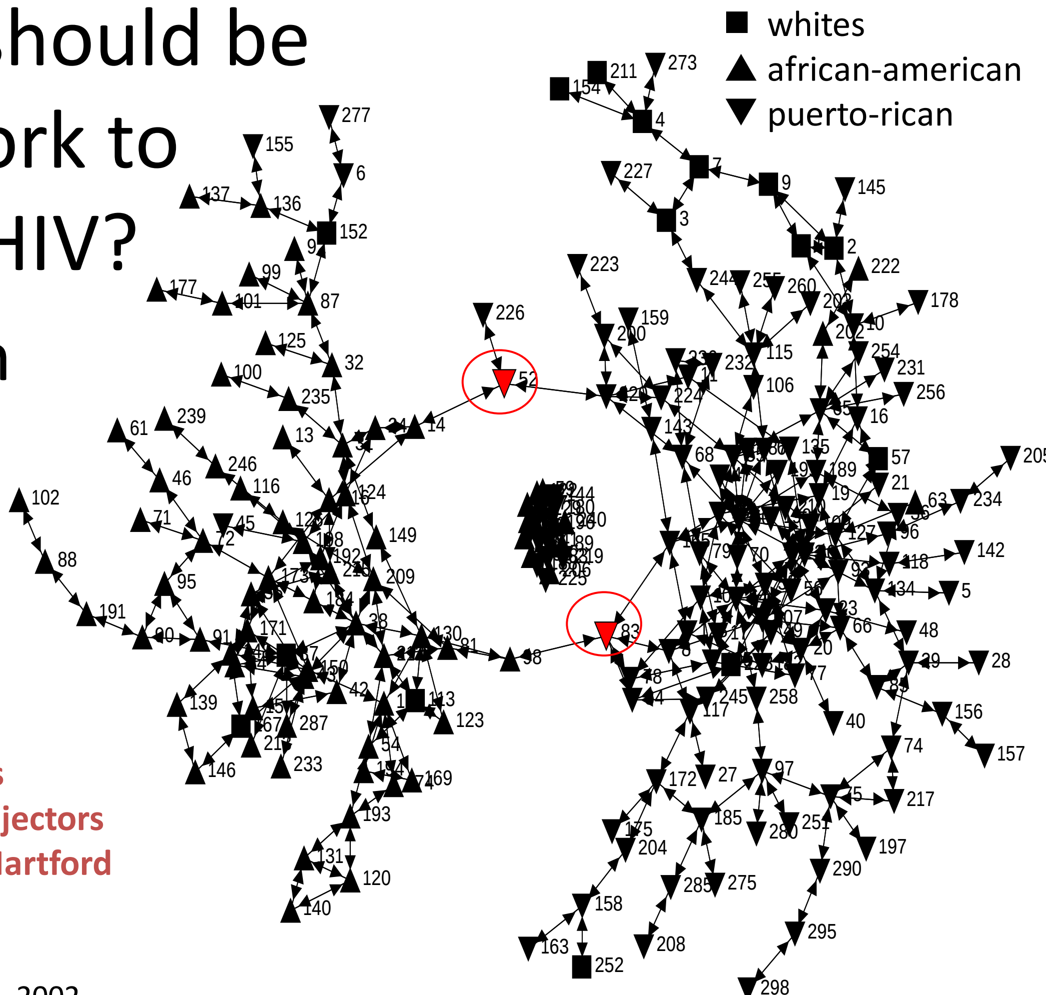
$$dwF = 1 - \frac{2\sum_{i>j} \frac{1}{d_{ij}}}{n(n-1)}$$





Disruption Example – health context

- Which <u>two</u> people should be isolated from network to slow the spread of HIV?
  - KeyPlayer algorithm identifies the two red nodes

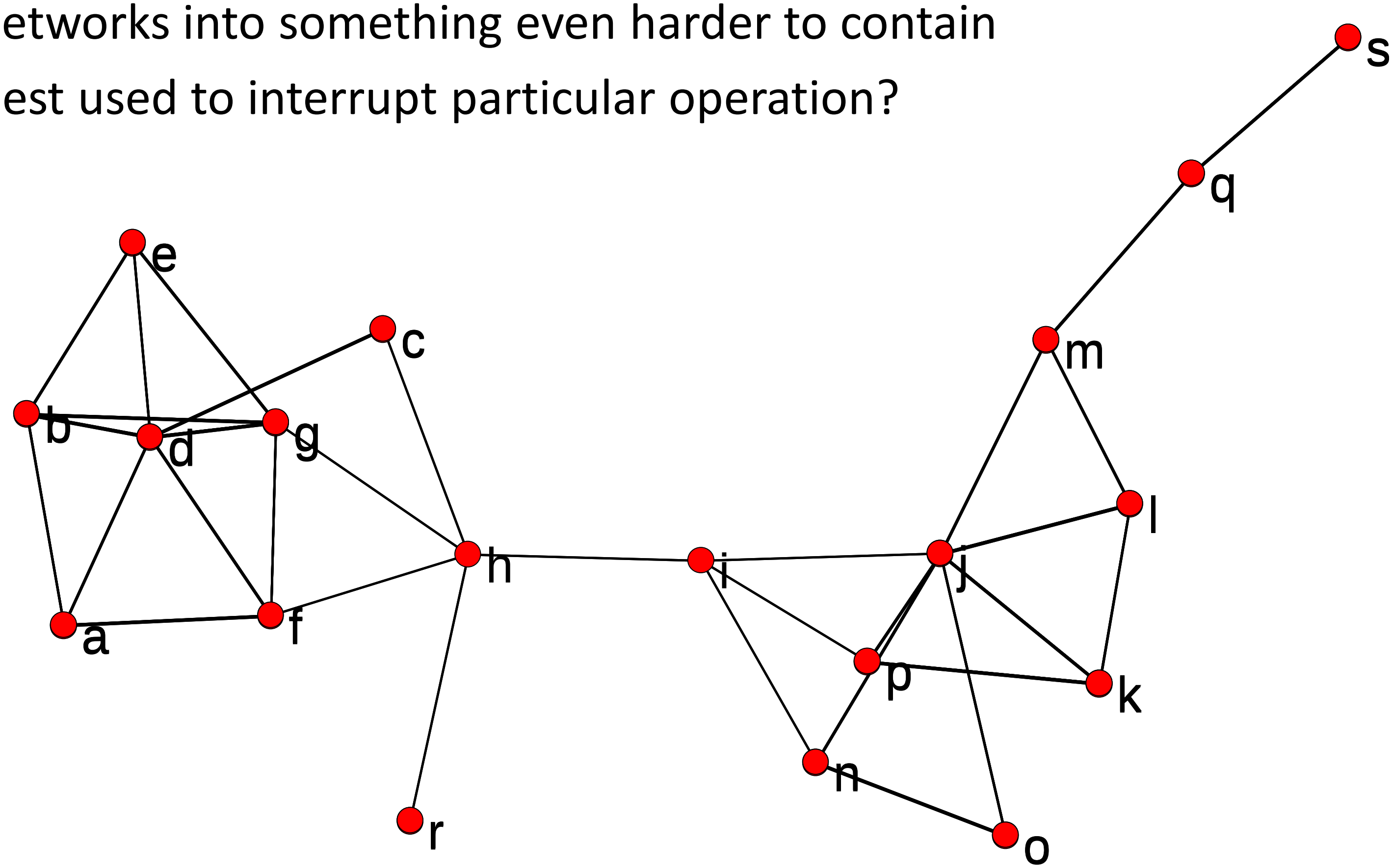


Friendship ties among drug injectors on streets of Hartford

Weeks, M.R., Clair, S., Borgatti, S.P., Radda, K., and Schensul, J.J. 2002.
Social networks of drug users in high risk sites: Finding the connections. *AIDS and Behavior* 6(2): 193-206

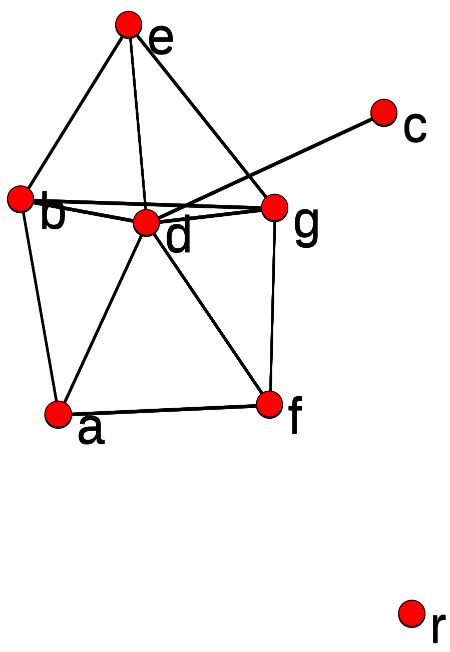
Caveats

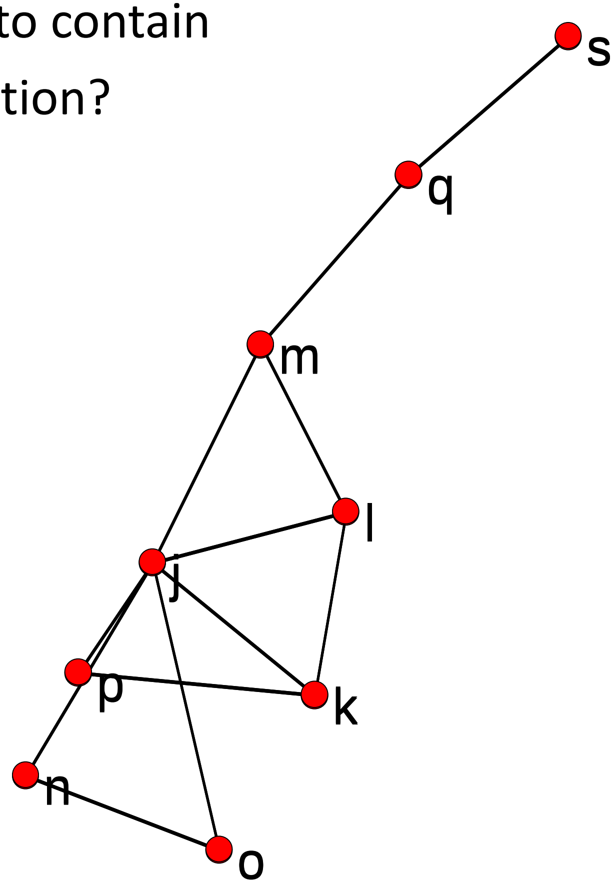
- Strategy of disrupting networks by removing key nodes may be dangerous long-term
  - Ties grow back. Fragmentation strategy may effectively shape enemy networks into something even harder to contain
  - Best used to interrupt particular operation?



Caveats

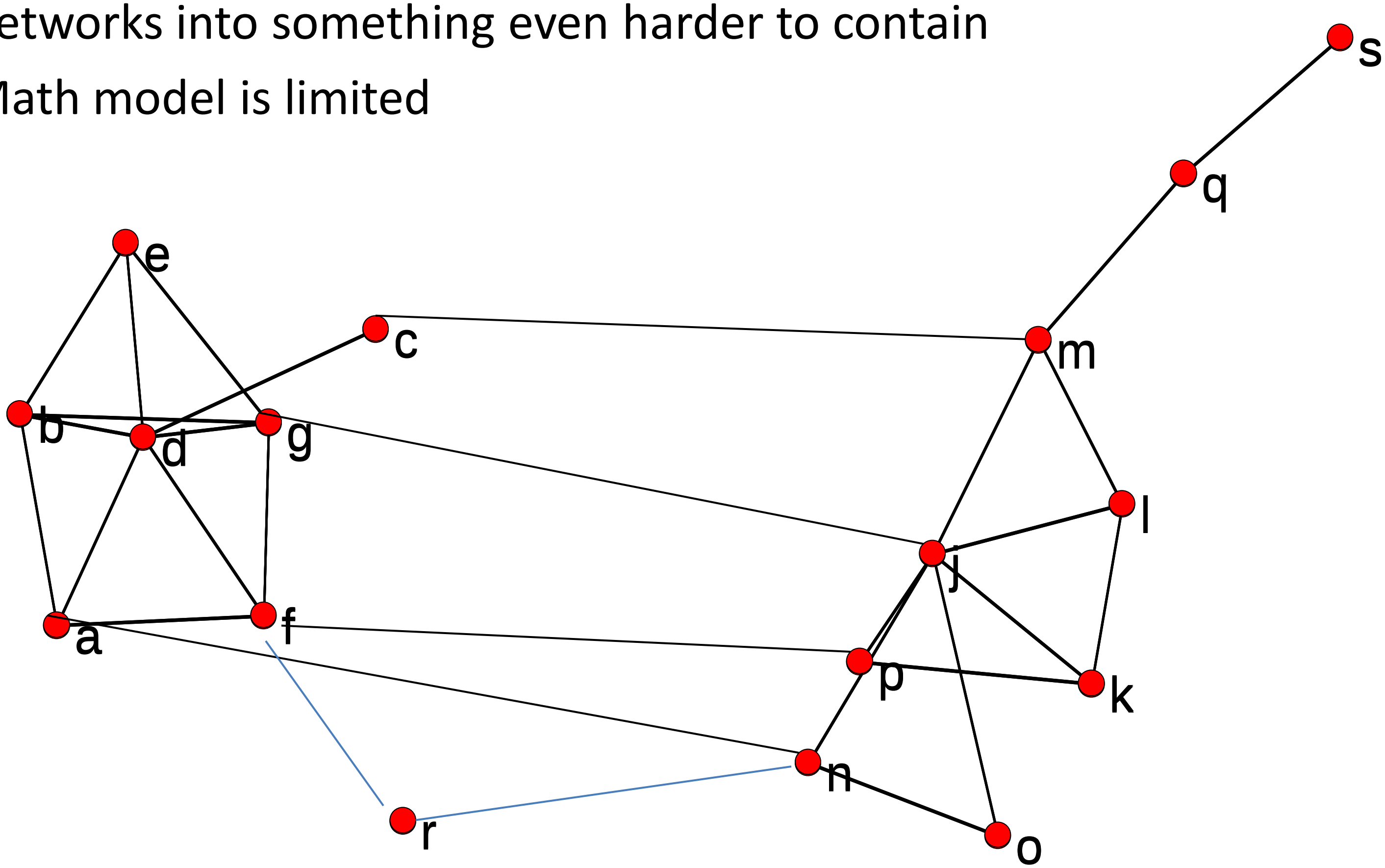
- Strategy of disrupting networks by removing key nodes may be dangerous long-term
  - Ties grow back. Fragmentation strategy may effectively shape enemy networks into something even harder to contain
  - Best used to interrupt particular operation?





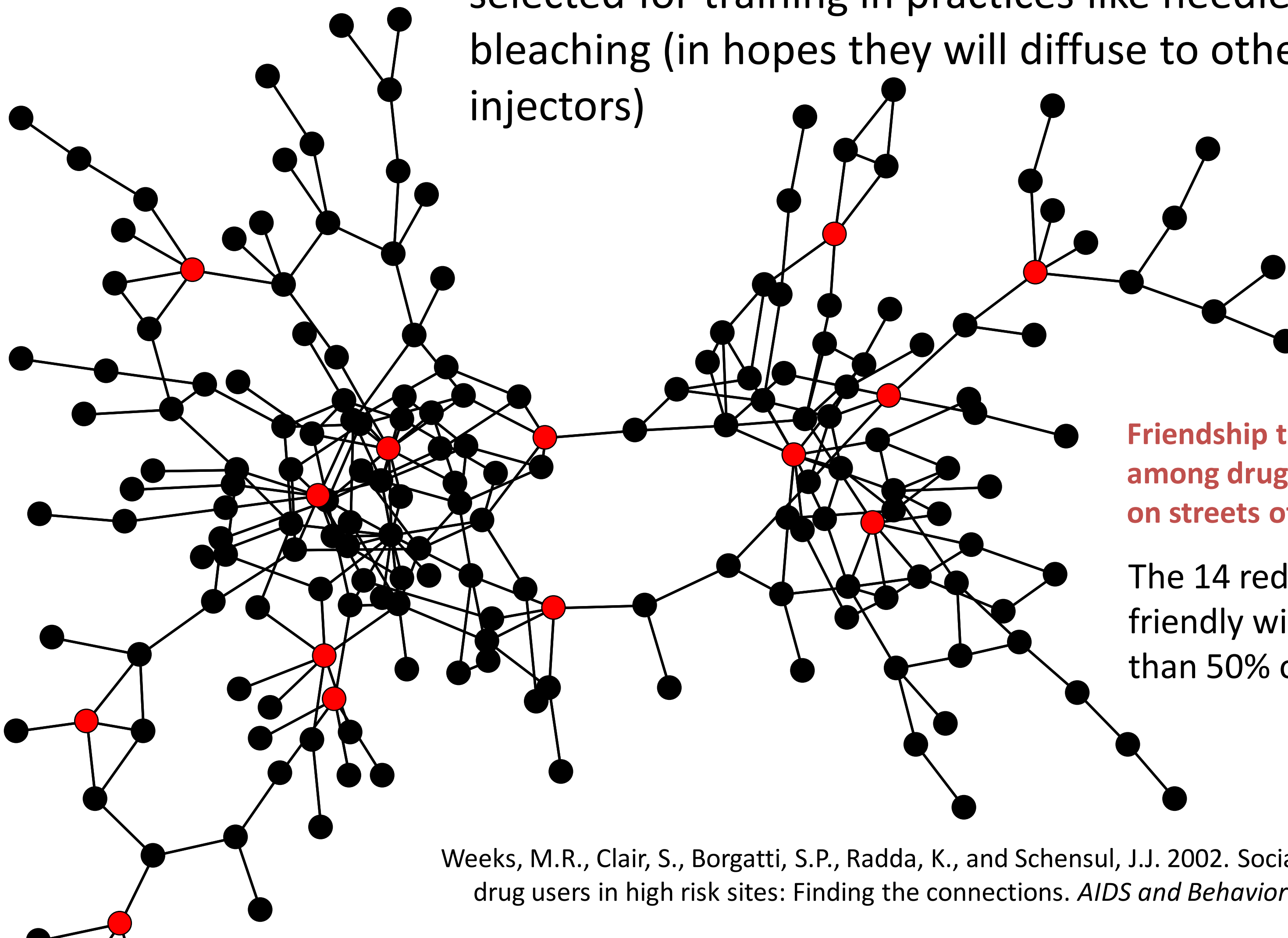
Caveats

- Strategy of disrupting networks by removing key nodes may be dangerous long-term
  - Ties grow back. Fragmentation strategy may effectively shape enemy networks into something even harder to contain
  - Math model is limited



Influence Example – health context

Which small set of drug injectors should be selected for training in practices like needle bleaching (in hopes they will diffuse to other injectors)



Friendship ties among drug injectors on streets of Hartford

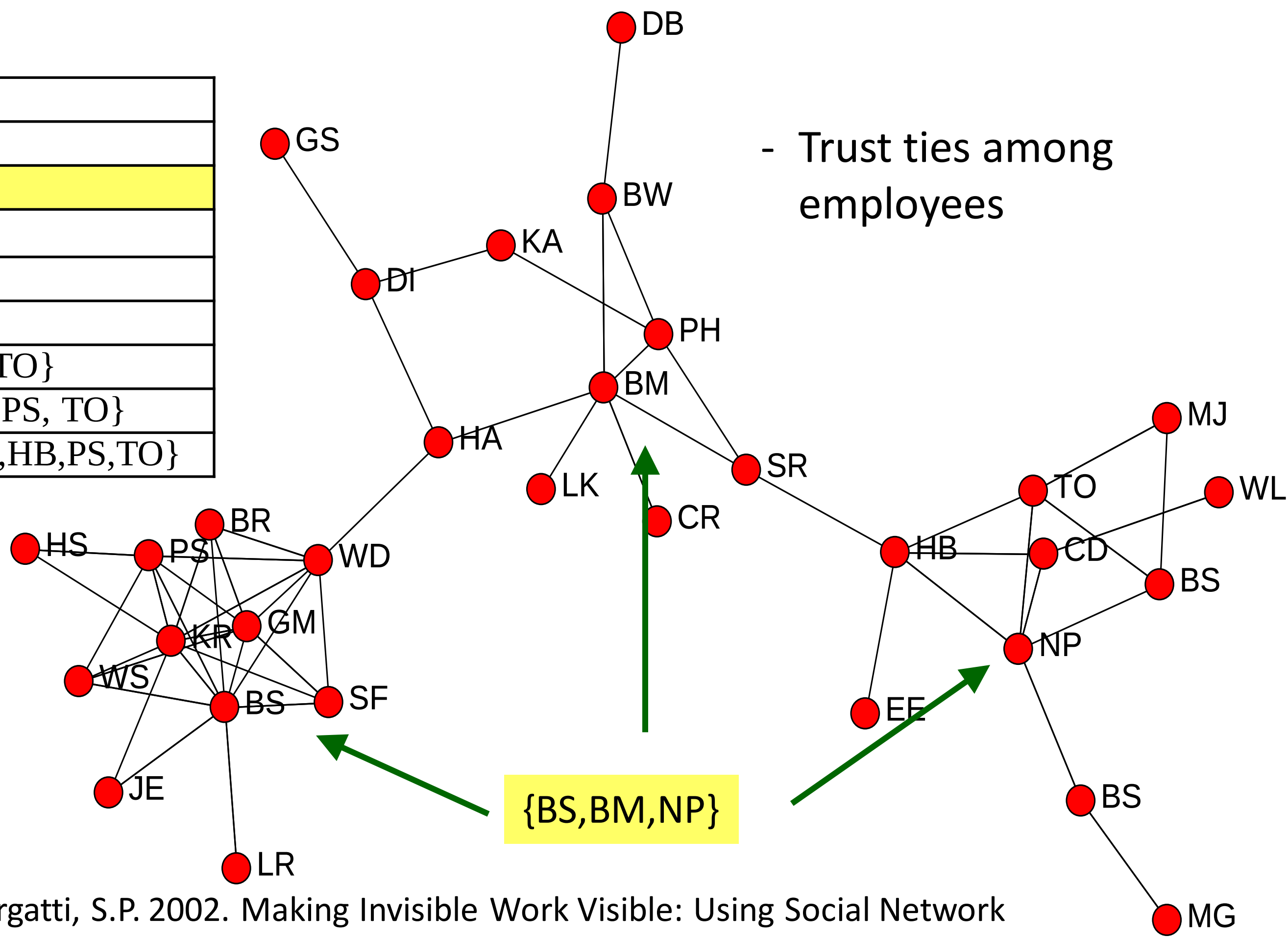
The 14 red nodes are friendly with more than 50% of network

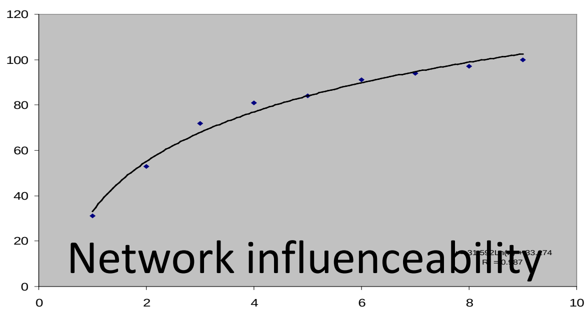
Weeks, M.R., Clair, S., Borgatti, S.P., Radda, K., and Schensul, J.J. 2002. Social networks of drug users in high risk sites: Finding the connections. *AIDS and Behavior* 6(2): 193-206

Influence Example – mgmt context

- Major change initiative is planned. Which small set of employees should we select for intensive indoctrination? in hopes they will diffuse positive attitude/knowledge to others

K	%	KP-Set
1	31	{KR}
2	53	{BM,BS}
3	72	{BM,BS,NP}
4	81	{BM,BS,DI,NP}
5	84	{BM,BS,DI,KR,NP}
6	91	{BM,BS,DI,HB,KR,TO}
7	94	{BM,BS,BS2,DI,HB,PS,TO}
8	97	{BM,BS,BS2,CD,DI,HB,PS, TO}
9	100	{BM,BS,BW,BS2,CD,DI,HB,PS,TO}





Data from: Cross, R., Parker, A., & Borgatti, S.P. 2002. Making Invisible Work Visible: Using Social Network Analysis to Support Strategic Collaboration. *California Management Review*. 44(2): 25-46

Prospects and Levers

- Objective
  - Use network influence models to maximize persuasive efforts
  - Illustrate how network perspective can be used to work with/through networks rather than against them
- Assumptions:
  - All nodes can be measured with respect to friendliness or unfriendliness to our cause (can be yes/no as well)
  - We know who influences whom
    - E.g., among physicians we have who receives referrals from whom

Borgatti, S.P. and Plant, E. 2008. Prospects and Levers. To be submitted to *Social Networks*

Prospects

- Prospects are "unfriendly" nodes that are surrounded by (influenced by) "friendlies"
  - By activating the nearby friendlies, we can try to "turn" the prospect
- Simplest formulation: $p_i = u_i \sum_j \underbrace{a_{ji} f_j}_{}$ Friendliness of neighborhood
  - u_i refers to unfriendliness of prospect *i*, a_{ji} indicates extent that *j* influences *i*, f_j gives the friendliness of node *j*. A node *i* gets a high score if currently unfriendly but surrounded by many friendlies
- Metrics of prospectness provide a way of prioritizing who to go after first
  - Identifying the low hanging fruit

Levers

- Levers are friendly nodes that have influence ties to unfriendly nodes.
  - If activated, can be directed to try to "turn" the unfriendlies who are influenced by them
  - Metrics identify who to activate (e.g., by incentivizing) in order maximize contagion effect per resource dollars
- Simplest formulation: $l_i = f_i \sum_j a_{ij} u_j$
- Incorporating indirect influence: $l_i = f_i \sum_j \alpha^{d_{ij}} a_{ij} u_j$

u_i refers to unfriendliness of prospect *i*, a_{ji} indicates extent that *j* influences *i*, f_j gives the friendliness of node *j*. d_{ij} is the length of the shortest path from *i* to *j*. α is a constant controlling attenuation of influence across long paths.