Cohesion & Cohesive Subgroups

Wednesday Afternoon

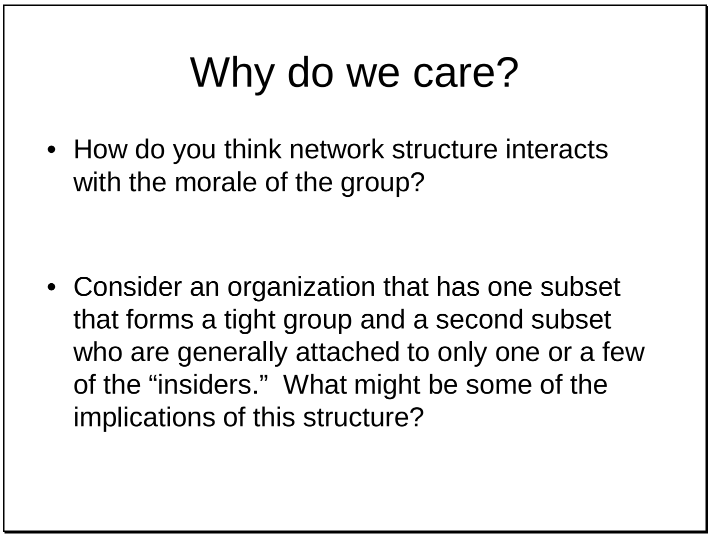
Cohesion, generally, deals with issues connectedness. However, within that concept of connectedness, there are multiple approaches to cohesion, and multiple ways in which social network analysts posit connectedness. Under this heading, we study "cohesive subgroups" (finding regions of the network in which actors are more connected to each other than they are to those outside that region), "whole network cohesion" (measures that describe the cohesion of the entire network), and "dyadic cohesion" (measures that quantify the connectedness of pairs of actors within the network).

Objectives:

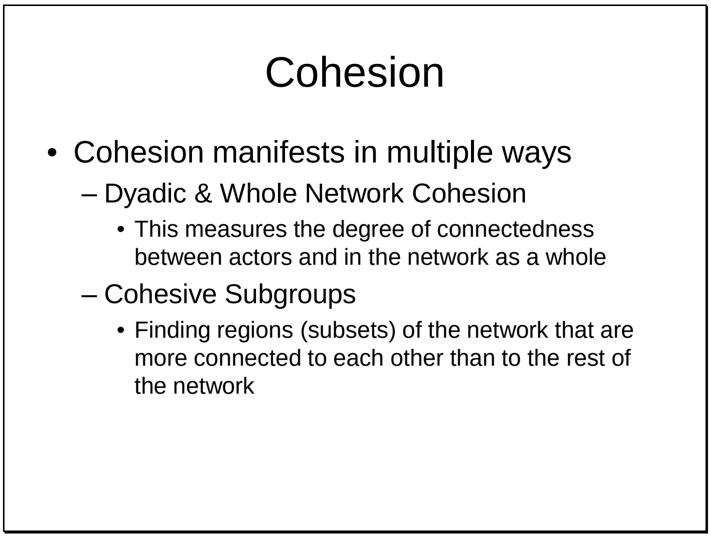
After this section, you should be able to:

- Describe the concept of cohesion from a social network perspective
- Differentiate between distance and density based measures of cohesion
- Name at least two measures of dyadic cohesion
- Name at least three measures of whole network cohesion
- Identify the different approaches to identifying cohesive subgroups
- Recognize core-periphery and centralized network structures
- Use UCINET to:
  - Create a geodesic distance matrix
  - Calculate density and fragmentation
  - Run hierarchical clustering on a geodesic matrix
  - Assess core-periphery structures in a network
  - Identify cliques, n-cliques, and k-plexes in a network
- Use NetDraw to:
  - Visualize cohesion measures from a network
  - Calculate and visualize k-cores
  - Run Newman-Girvan to identify cohesive subgroups

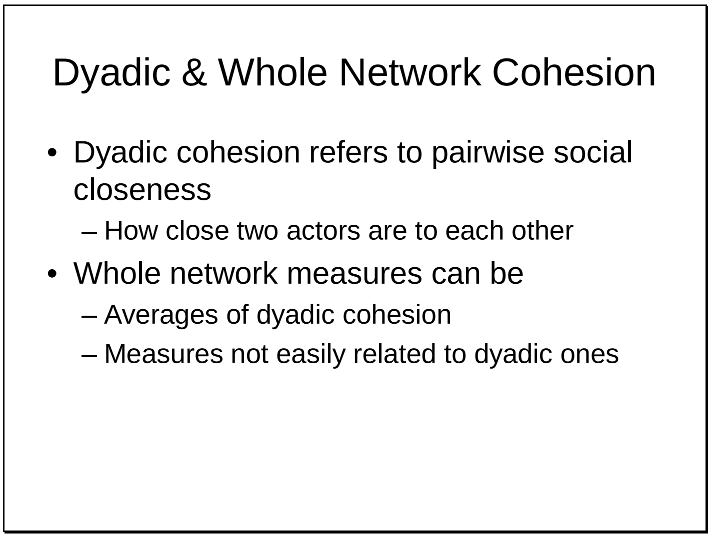
Slide 2


Why do we care?
• How do you think network structure interacts with the morale of the group?
• Consider an organization that has one subset that forms a tight group and a second subset who are generally attached to only one or a few of the “insiders.” What might be some of the implications of this structure?

Slide 3


Cohesion
• Cohesion manifests in multiple ways
– Dyadic & Whole Network Cohesion
• This measures the degree of connectedness between actors and in the network as a whole
– Cohesive Subgroups
• Finding regions (subsets) of the network that are more connected to each other than to the rest of the network

Slide 4


Dyadic & Whole Network Cohesion
• Dyadic cohesion refers to pairwise social closeness
– How close two actors are to each other
• Whole network measures can be
– Averages of dyadic cohesion
– Measures not easily related to dyadic ones

Slide 5

Dyadic Cohesion

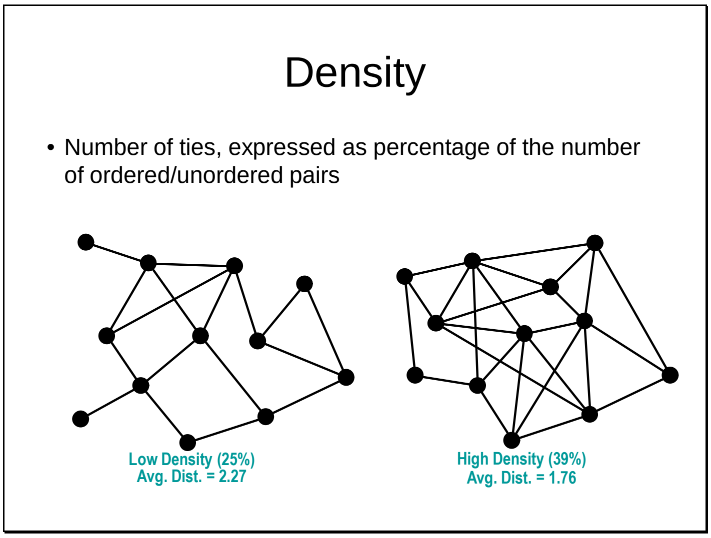
- Adjacency
  - Strength of tie
- Reachability
  - A path exists or does not (usually as $1/d_{ij}$)
- Distance
  - Length of shortest path between two nodes
- Multiplexity
  - Number of ties of different relations linking two nodes
- Number of paths linking two nodes
  - Vertex independent number is Point Connectivity
  - Edge independent number is Maximum Flow

Slide 6

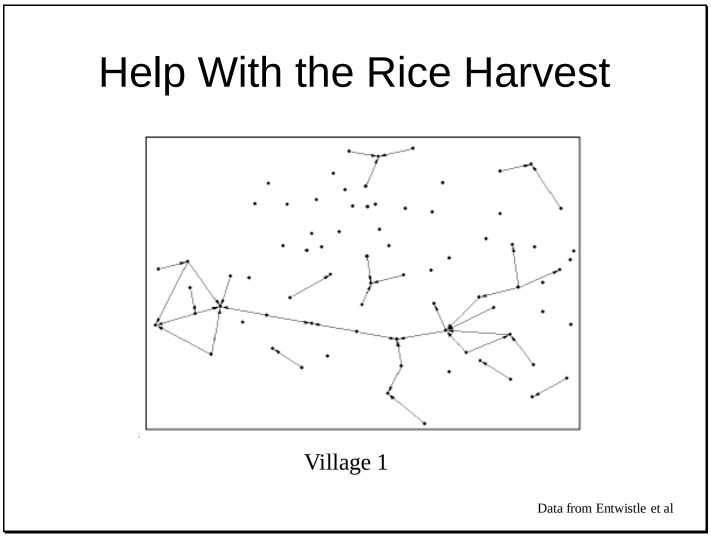
Measures of Whole Network Cohesion

- Density & Average degree
- Average Distance and Diameter
- Number of components
- Fragmentation
- Distance-weighted Fragmentation
- Centralization

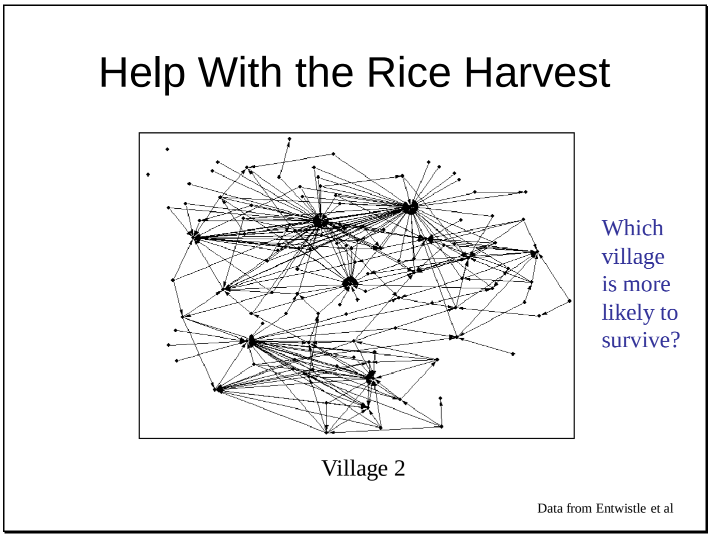
Slide 7


Density
• Number of ties, expressed as percentage of the number of ordered/unordered pairs
Low Density (25%)
Avg. Dist. = 2.27
High Density (39%)
Avg. Dist. = 1.76

Slide 8


Help With the Rice Harvest
Village 1
Data from Entwistle et al

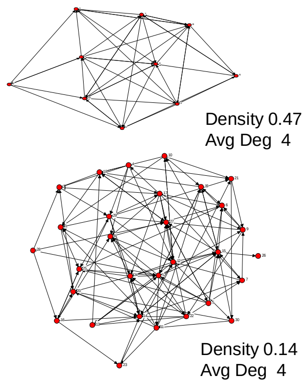
Slide 9


Help With the Rice Harvest
Which
village
is more
likely to
survive?
Village 2
Data from Entwistle et al

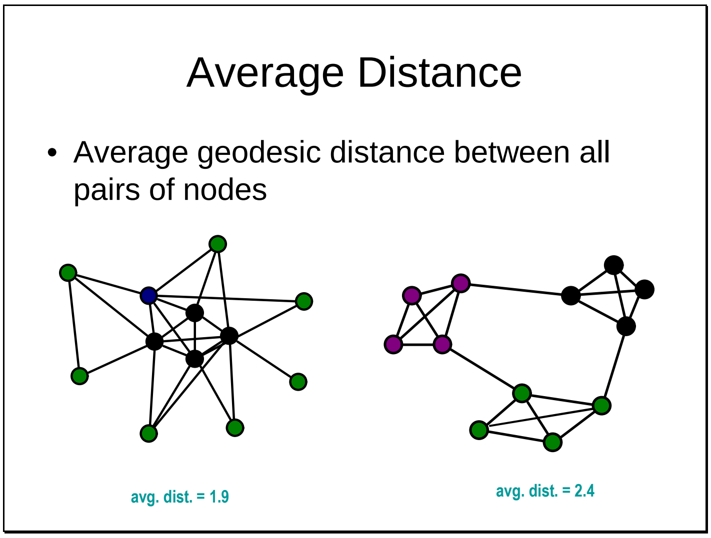
Slide 10

Average Degree

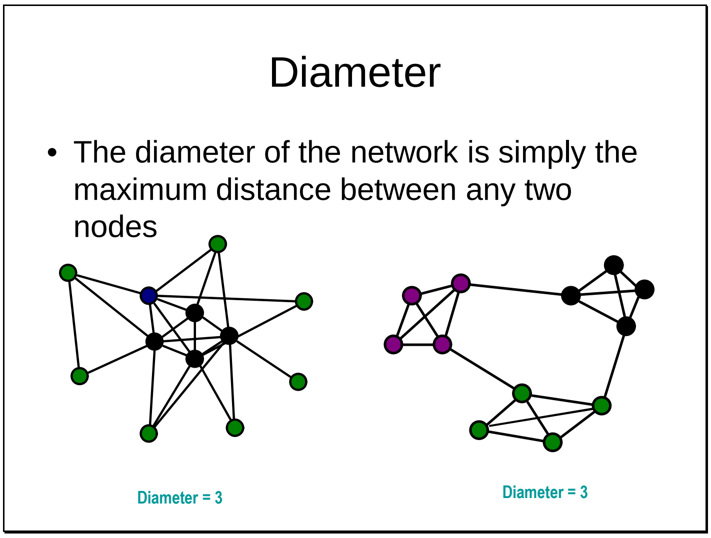
- Average number of links per person
- Is same as density*(n-1), where n is size of network
  - Density is just normalized avg degree
  - Often more intuitive than density
- Which network has the higher density?



Slide 11


Average Distance
• Average geodesic distance between all pairs of nodes
avg. dist. = 1.9
avg. dist. = 2.4

Slide 12


Diameter
• The diameter of the network is simply the maximum distance between any two nodes
Diameter = 3
Diameter = 3

Slide 13

Fragmentation Measures

- Component ratio
- F measure of fragmentation
- Distance-weighted fragmentation ^{D}F

Slide 14

F Measure of Fragmentation

- Proportion of pairs of nodes that are unreachable from each other

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

r_{ij} = 1 if node i can reach node j by a path of any length
r_{ij} = 0 otherwise

- If all nodes reachable from all others (i.e., one component), then F = 0
- If graph is all isolates, then F = 1

Slide 15

Computation Formula for F Measure

- No ties across components, and all reachable within components, hence can express in terms of size of components

$$F = 1 - \frac{\sum_k s_k (s_k - 1)}{n(n-1)}$$

S_k = size of k^{th} component

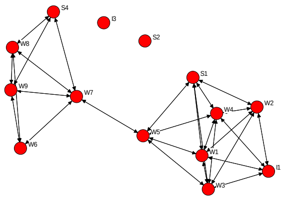
Slide 16

Computational Example

Games Data

Comp	Size	Sk(Sk-1)
1	1	0
2	1	0
3	12	132
	14	132
		0.2747

= 1 – (132/(14*13)) = F



Slide 17

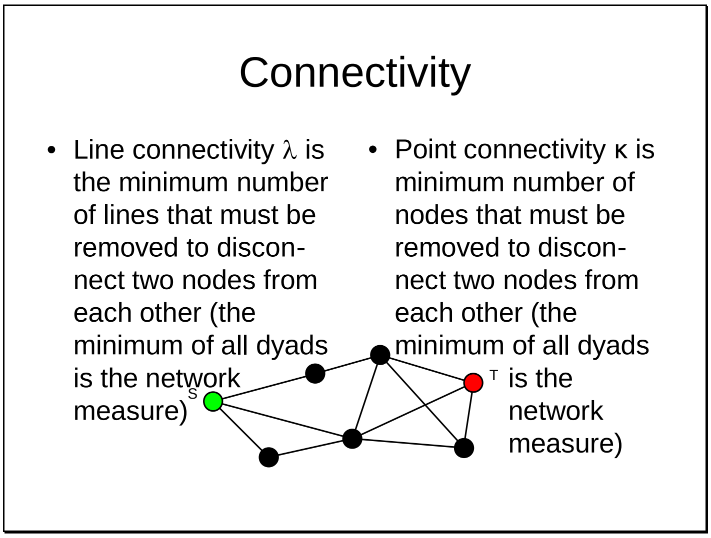
Distance-Weighted Fragmentation

- Use average of the reciprocal of distance
  - letting 1/∞ = 0

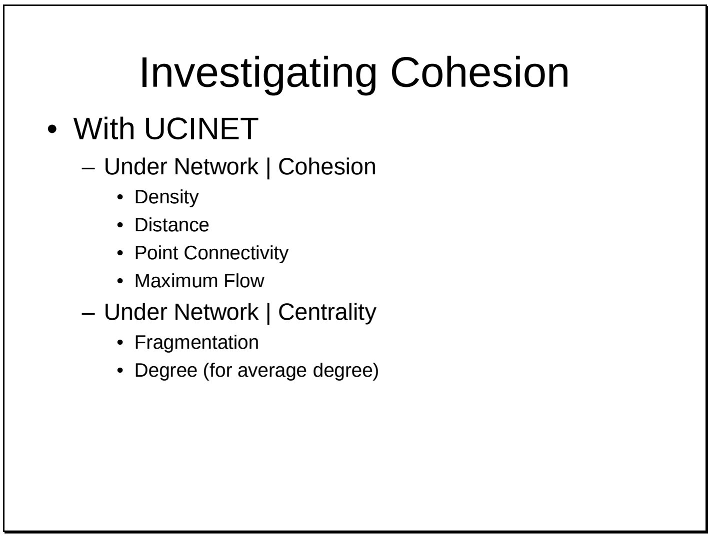
$$^{D}F = 1 - \frac{\sum_{i \neq j} \frac{1}{d_{ij}}}{n(n-1)}$$

- Bounds
  - lower bound of 0 when every pair is adjacent to every other (entire network is a clique, a “Complete” graph)
  - upper bound of 1 when graph is all isolates

Slide 18


Connectivity
• Line connectivity λ is the minimum number of lines that must be removed to discon-nect two nodes from each other (the minimum of all dyads is the network measure)
• Point connectivity κ is minimum number of nodes that must be removed to discon-nect two nodes from each other (the minimum of all dyads is the network measure)
S
T

Slide 19


Investigating Cohesion
• With UCINET
– Under Network | Cohesion
• Density
• Distance
• Point Connectivity
• Maximum Flow
– Under Network | Centrality
• Fragmentation
• Degree (for average degree)

Slide 20

Part II: Subgroups – A Typology

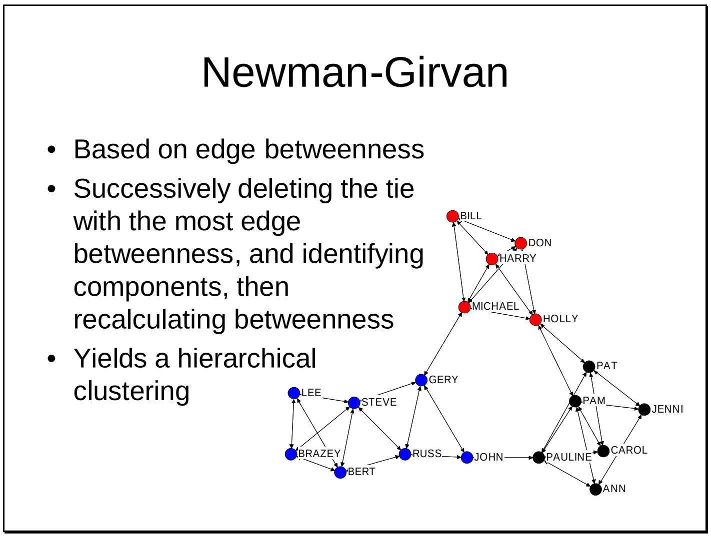
	Found by algorithm based on	Found by finding sets with properties of
Network / Graph theory	*Graph-theoretic data driven algorithms* Newman-Girvan	*Formal definitions of sociological groups* *{mathematical ethnography}* Clique, n-clique, n-clan, n-club, k-plex, ls-set, lambda-set, k-core, component
Proximities / Clustering	*Multivariate clustering analysis methods* Johnson's Hierarchical clustering; k-means; MDS	*Formal definitions of abstract clusters* Combinatorial optimization

Slide 21

Groups defined based algorithm using graph theoretic properties

	Group by defined Algorithm	Group by defined Characteristics
Network/ Graph Theory	Newman-Girvan	**Distance**: Component, Clique, n-clique, n-clan, n-club **Density**: Clique, k-core, k-plex, ls-set, lamba set
Proximity /Distance	Hierarchical Clustering MDS K-Means	Factions (Core/Periphery) (Combinatorial Optimization)

Slide 22


Newman-Girvan
• Based on edge betweenness
• Successively deleting the tie with the most edge betweenness, and identifying components, then recalculating betweenness
• Yields a hierarchical clustering
BILL
DON
HARRY
MICHAEL
HOLLY
PAT
GERY
LEE
STEVE
PAM
JENNI
BRAZEY
RUSS
JOHN
PAULINE
CAROL
BERT
ANN

Slide 23

Groups w/specified characteristics, based on Graph Theoretic Measures

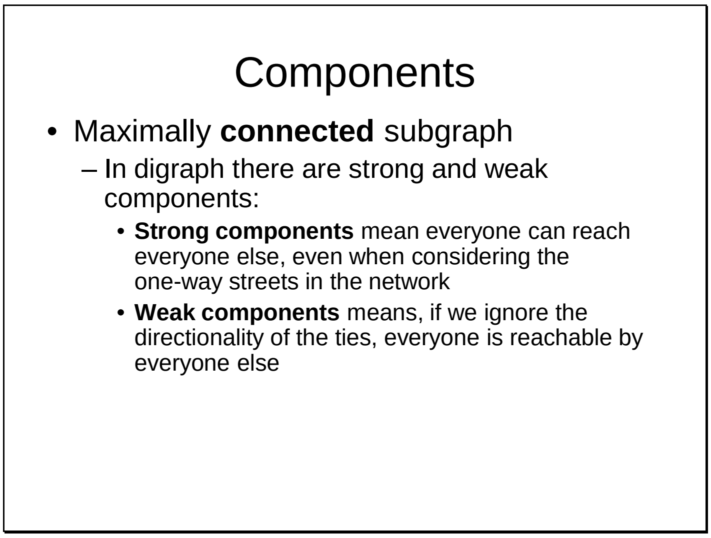
	Group by defined Algorithm	Group by defined Characteristics
Network/ Graph Theory	Newman-Girvan	**Distance**: Component, Clique, n-clique, n-clan, n-club **Density**: Clique, k-core, k-plex, *ls-set, lamba set*
Proximity /Distance	Hierarchical Clustering MDS K-Means	Factions (Core/Periphery) (Combinatorial Optimization)

Slide 24

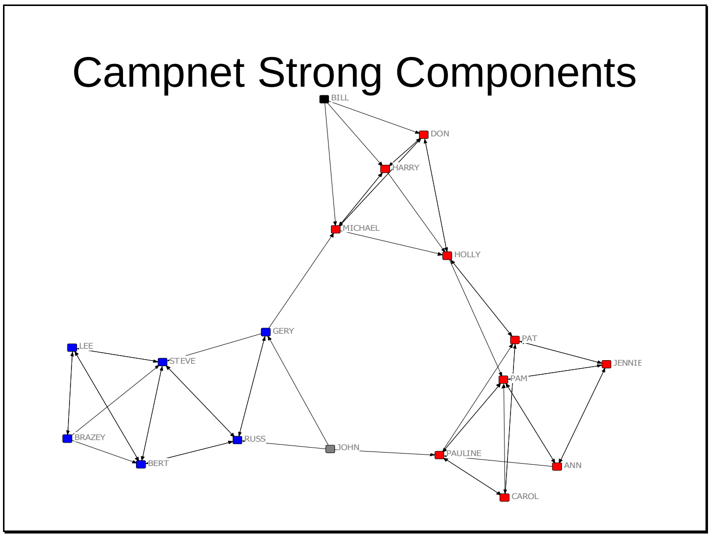
Defined by graph-theoretic characteristics of resultant sets

- Most Common
  - Components
  - Clique (n-clique)
  - k-Plex
  - k-Core
- More advanced variations
  - n-clans
  - n-clubs
  - *Lamba sets*
  - *LS Sets*

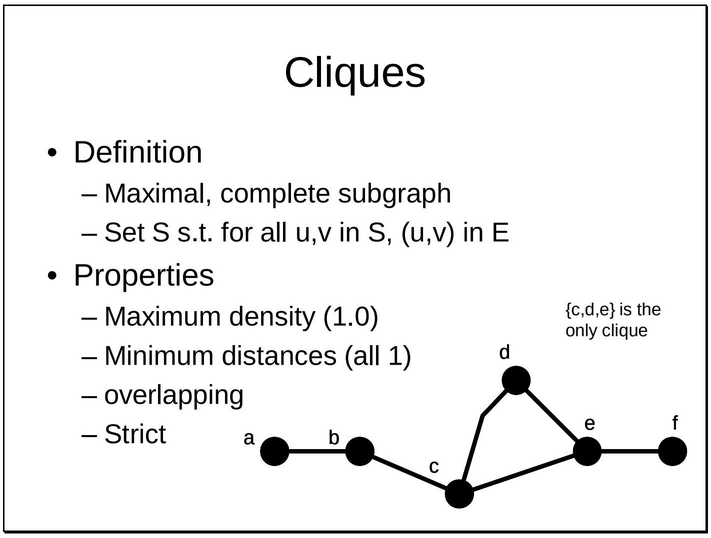
Slide 25


Components
• Maximally **connected** subgraph
– In digraph there are strong and weak components:
• **Strong components** mean everyone can reach everyone else, even when considering the one-way streets in the network
• **Weak components** means, if we ignore the directionality of the ties, everyone is reachable by everyone else

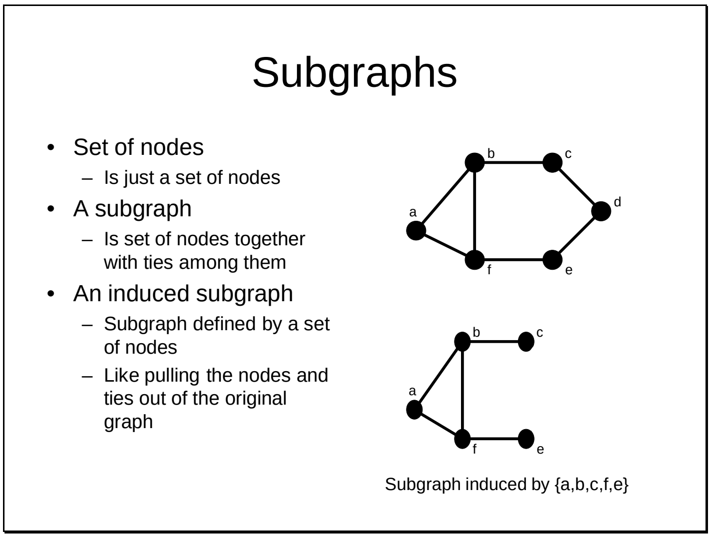
Slide 26


Campnet Strong Components
BILL
DON
HARRY
MICHAEL
HOLLY
GERY
LEE
STEVE
BRAZEY
BERT
RUSS
JOHN
PAULINE
PAT
PAM
JENNIE
ANN
CAROL

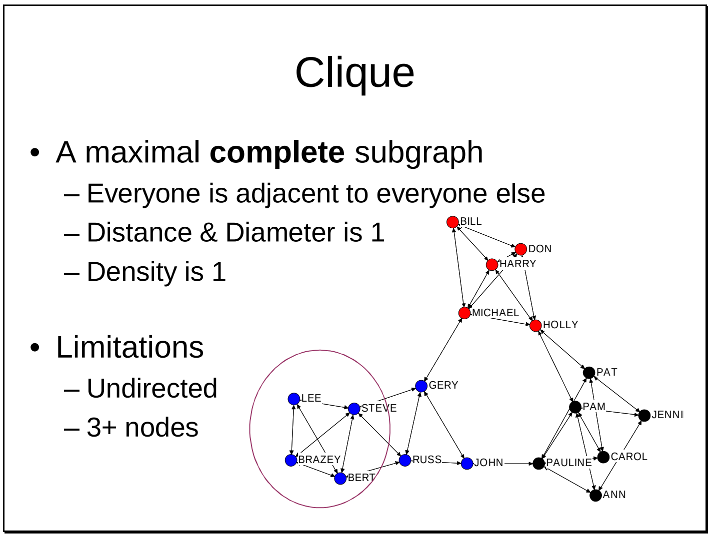
Slide 27


Cliques
• Definition
– Maximal, complete subgraph
– Set S s.t. for all u,v in S, (u,v) in E
• Properties
– Maximum density (1.0)
– Minimum distances (all 1)
– overlapping
– Strict
{c,d,e} is the only clique
a
b
c
d
e
f

Slide 28


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
a
b
c
d
e
f
a
b
c
e
f
Subgraph induced by {a,b,c,f,e}

Slide 29

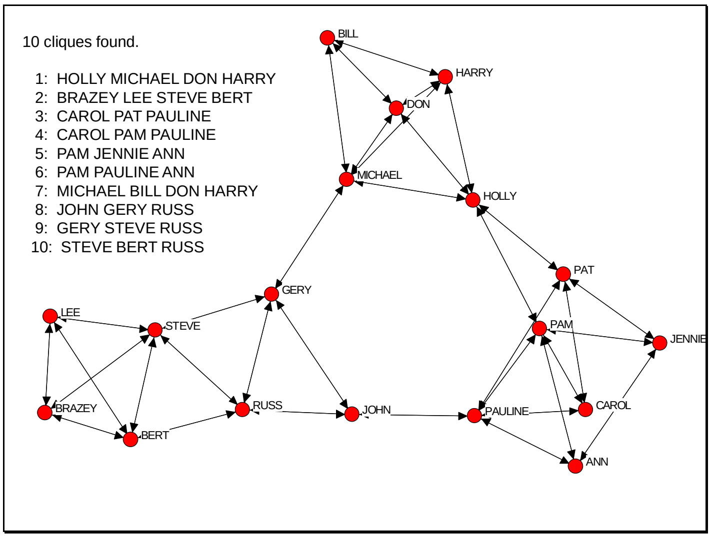

Clique
• A maximal **complete** subgraph
– Everyone is adjacent to everyone else
– Distance & Diameter is 1
– Density is 1
• Limitations
– Undirected
– 3+ nodes
BILL
DON
HARRY
MICHAEL
HOLLY
PAT
GERY
LEE
STEVE
PAM
JENNI
BRAZEY
RUSS
JOHN
PAULINE
CAROL
BERT
ANN

Slide 30

Problems with Cliques

- Can be too many or too few
- If too many:
  - Can put minimum on size
  - Can look at overlap
- If too few, relax requirements in terms of
  - Distance:
    - n-cliques, n-clans, n-clubs
    - This may also bring DOWN the number of cliques
  - Density
    - k-cores, k-plexes, ls-sets, lamba sets

Slide 31

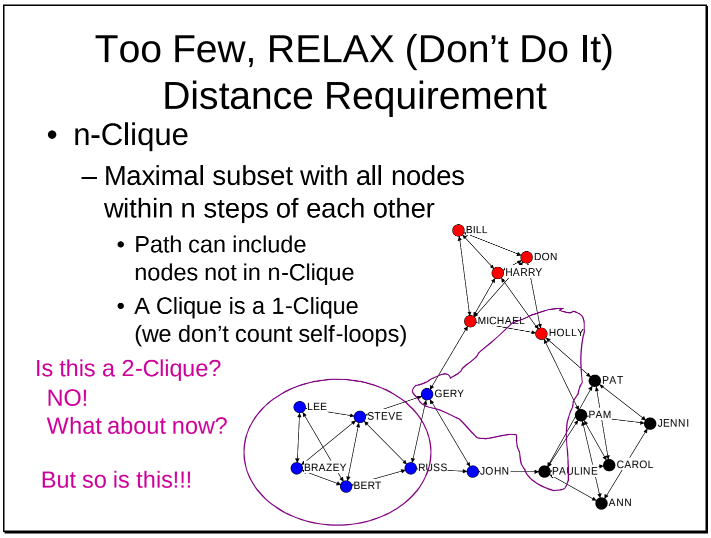

10 cliques found.
1: HOLLY MICHAEL DON HARRY
2: BRAZEY LEE STEVE BERT
3: CAROL PAT PAULINE
4: CAROL PAM PAULINE
5: PAM JENNIE ANN
6: PAM PAULINE ANN
7: MICHAEL BILL DON HARRY
8: JOHN GERY RUSS
9: GERY STEVE RUSS
10: STEVE BERT RUSS
BILL
HARRY
DON
MICHAEL
HOLLY
PAT
PAM
JENNIE
GERY
LEE
STEVE
BRAZEY
BERT
RUSS
JOHN
PAULINE
CAROL
ANN

Slide 32

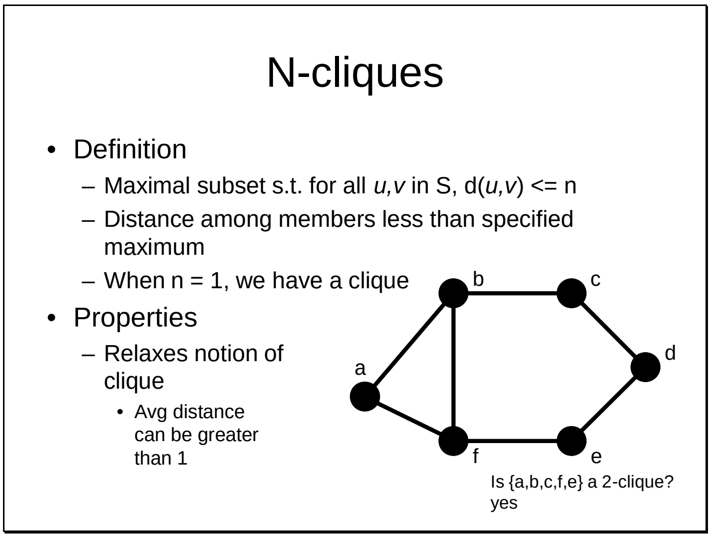
Types of Relaxations

- Distance (length of paths)
  - N-clique, n-clan, n-club
- Density (number of ties)
  - K-plex, k-core, component, *ls-set, lambda set,*

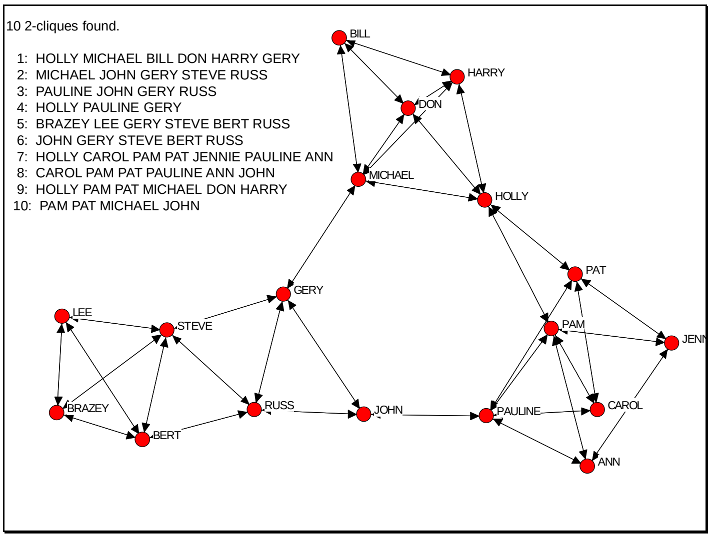
Slide 33


Too Few, RELAX (Don't Do It) Distance Requirement
• n-Clique
– Maximal subset with all nodes within n steps of each other
• Path can include nodes not in n-Clique
• A Clique is a 1-Clique (we don't count self-loops)
Is this a 2-Clique?
NO!
What about now?
But so is this!!!
BILL
DON
HARRY
MICHAEL
HOLLY
PAT
GERY
LEE
STEVE
PAM
JENNI
BRAZEY
RUSS
JOHN
PAULINE
CAROL
BERT
ANN

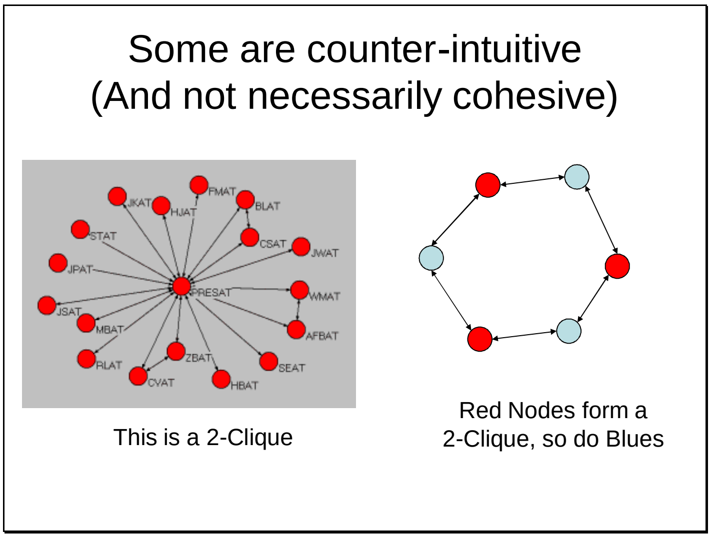
Slide 34


N-cliques
• Definition
– Maximal subset s.t. for all *u,v* in S, d(*u,v*) <= n
– Distance among members less than specified maximum
– When n = 1, we have a clique
• Properties
– Relaxes notion of clique
• Avg distance can be greater than 1
a
b
c
d
e
f
Is {a,b,c,f,e} a 2-clique?
yes

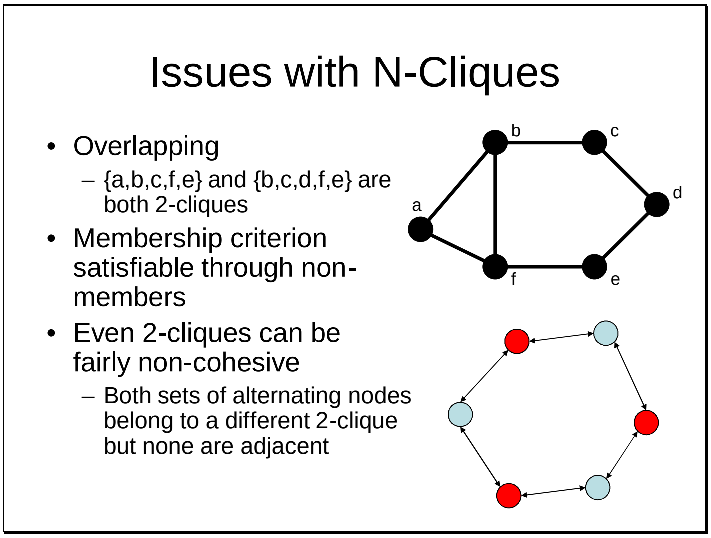
Slide 35


10 2-cliques found.
1: HOLLY MICHAEL BILL DON HARRY GERY
2: MICHAEL JOHN GERY STEVE RUSS
3: PAULINE JOHN GERY RUSS
4: HOLLY PAULINE GERY
5: BRAZEY LEE GERY STEVE BERT RUSS
6: JOHN GERY STEVE BERT RUSS
7: HOLLY CAROL PAM PAT JENNIE PAULINE ANN
8: CAROL PAM PAT PAULINE ANN JOHN
9: HOLLY PAM PAT MICHAEL DON HARRY
10: PAM PAT MICHAEL JOHN
BILL
HARRY
DON
MICHAEL
HOLLY
PAT
GERY
LEE
STEVE
PAM
BRAZEY
BERT
RUSS
JOHN
PAULINE
CAROL
ANN

Slide 36


Some are counter-intuitive
(And not necessarily cohesive)
JKAT
HJAT
FMAT
BLAT
STAT
CSAT
JWAT
JPAT
PRESAT
WMAT
JSAT
MBAT
AFBAT
RLAT
ZBAT
SEAT
CVAT
HBAT
This is a 2-Clique
Red Nodes form a
2-Clique, so do Blues

Slide 37

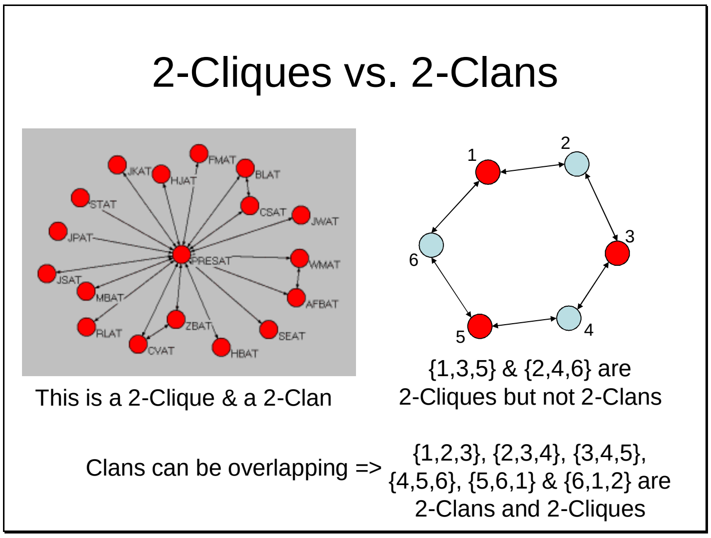

Issues with N-Cliques
• Overlapping
– {a,b,c,f,e} and {b,c,d,f,e} are both 2-cliques
• Membership criterion satisfiable through non-members
• Even 2-cliques can be fairly non-cohesive
– Both sets of alternating nodes belong to a different 2-clique but none are adjacent
a
b
c
d
e
f

Slide 38

So, we can force more cohesion

- n-Clan is an n-Clique whose diameter in the subgraph induced from the nodes in the n-Clique is <= n
  - Don’t allow paths to go outside subset

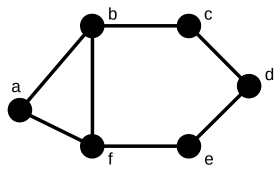
Slide 39


2-Cliques vs. 2-Clans
JKAT
HJAT
FMAT
BLAT
STAT
CSAT
JWAT
JPAT
PRESAT
WMAT
JSAT
MBAT
AFBAT
RLAT
ZBAT
SEAT
CVAT
HBAT
This is a 2-Clique & a 2-Clan
1
2
3
4
5
6
{1,3,5} & {2,4,6} are
2-Cliques but not 2-Clans
Clans can be overlapping =>
{1,2,3}, {2,3,4}, {3,4,5},
{4,5,6}, {5,6,1} & {6,1,2} are
2-Clans and 2-Cliques

Slide 40

But, n-Clans have issues, too

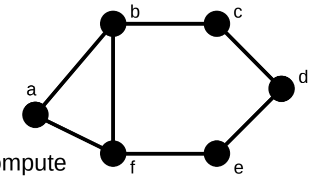
- The n-Clique requirement is restrictive, so there are few found in the data
- Is {a,b,c,f} a 2-Clan?
- How many 2-Clans are there in this graph?



Slide 41

Loosening the restriction

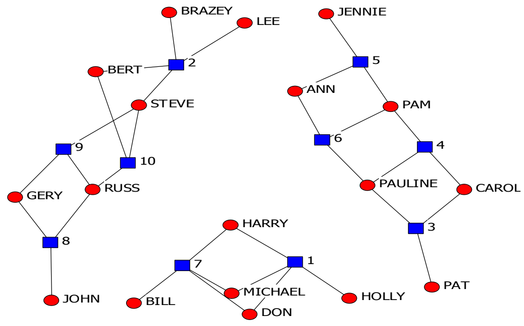
- n-Clubs are, effectively, n-Clans that do not have the n-Clique requirement, or…
  - A maximal subset S such that the graph induced by the nodes S has a diameter <= n
  - Now {a,b,c,f} is a 2-Club, so is {a,b,e,f}
- Properties:
  - Painful (impossible) to compute
  - More plentiful than n-Clans
  - Overlapping



Slide 42

Many of these are (too) plentiful

- One way to process the information is to look at CliqueSets as a two-mode network



Slide 43

Loosen the density restriction

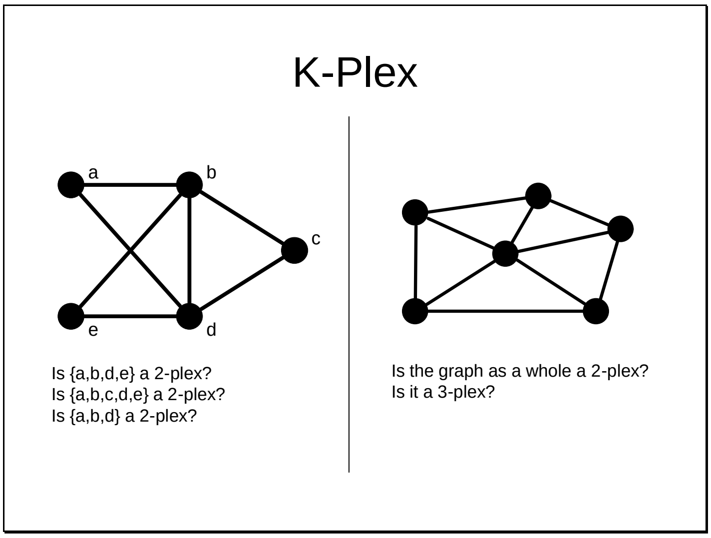
- n-Cliques, n-Clans, and n-Clubs all start from the definition of Cliques and relax the distance requirement (all distances = 1) in varying ways
- But, Cliques also have maximum density (d = 1), and we can relax that definition instead…
- But for this, we must define the alpha operator, α, such that $\alpha(u,G)$ is the number of lines from node u to nodes in graph G

Slide 44

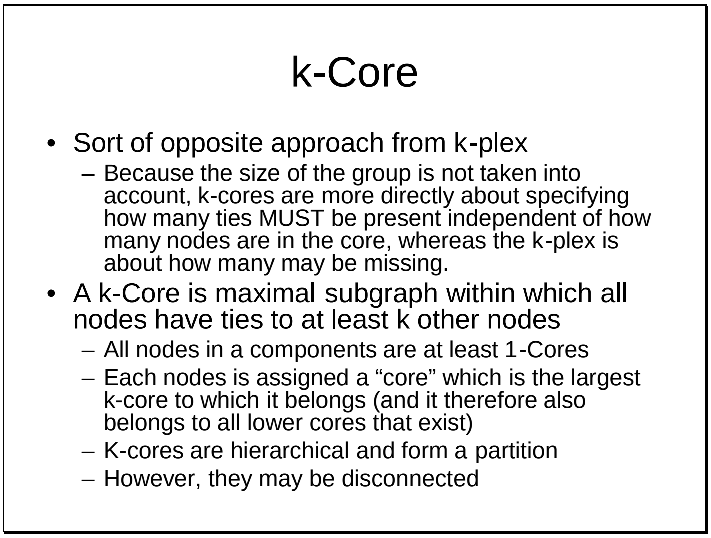
Relaxing the Density Requirement

- k-Plex
  - A clique where members don't have to be connected to everyone else, just all but *k* members, or…
  - a [maximal] subset S s.t. for all *u* in S, $\alpha(u,S)$ >= |S|-k, where |S| is size of set S
    - All subsets of k-plexes are k-plexes (if non-maximal)
    - Get distance for free based on S, k.
      - If k < (|S|+2)/2 then diameter <= 2
    - Numerous & Overlapping
    - May be more intuitive than distance-based measures
    - A Clique is a 1-plex (We assume it not tied to itself)

Slide 45


K-Plex
a
b
c
e
d
Is {a,b,d,e} a 2-plex?
Is {a,b,c,d,e} a 2-plex?
Is {a,b,d} a 2-plex?
Is the graph as a whole a 2-plex?
Is it a 3-plex?

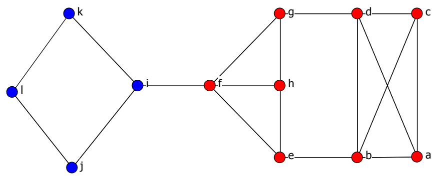
Slide 46


k-Core
• Sort of opposite approach from k-plex
– Because the size of the group is not taken into account, k-cores are more directly about specifying how many ties MUST be present independent of how many nodes are in the core, whereas the k-plex is about how many may be missing.
• A k-Core is maximal subgraph within which all nodes have ties to at least k other nodes
– All nodes in a components are at least 1-Cores
– Each nodes is assigned a "core" which is the largest k-core to which it belongs (and it therefore also belongs to all lower cores that exist)
– K-cores are hierarchical and form a partition
– However, they may be disconnected

Slide 47

Another definition

- A k-core is a maximal subgraph such that for all u in S, α(u,S) >= k



  – All nodes are 2-core (and 1-core)
    Red nodes are 3-core.
- Great for analyzing large networks

Slide 48

Groups w/specified characteristics, based on Proximities

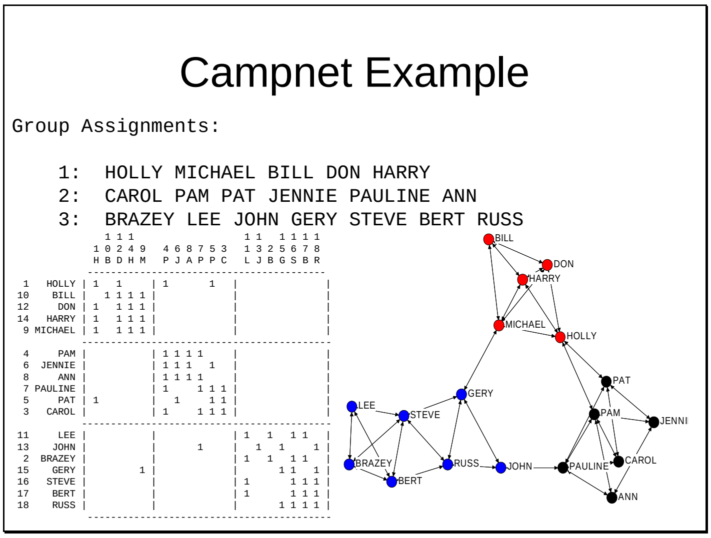
	Group by defined Algorithm	Group by defined Characteristics
Network/ Graph Theory	Newman-Girvan	**Distance**: Component, Clique, n-clique, n-clan, n-club **Density**: Clique, k-core, k-plex, ls-set, lamba set
Proximity /Distance	Hierarchical Clustering MDS K-Means	Factions (Core/Periphery) (Combinatorial Optimization)

Slide 49

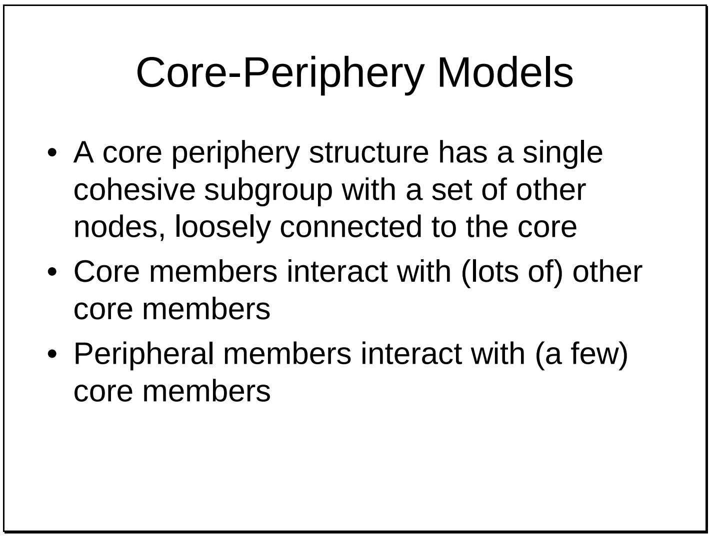
Factions

- Computationally arrange nodes into mutually exclusive groups such that some predefined criteria is optimized
  - For example, make groups that maximize density of internal ties and minimize density of external ties

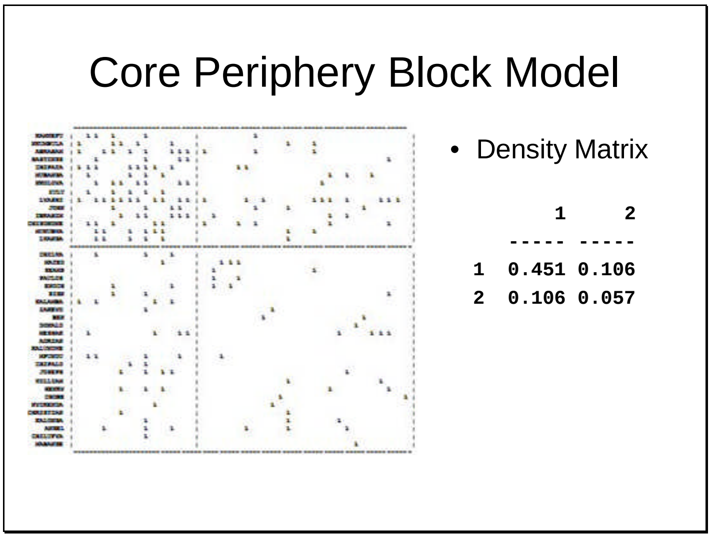
Slide 50


Campnet Example
Group Assignments:
1: HOLLY MICHAEL BILL DON HARRY
2: CAROL PAM PAT JENNIE PAULINE ANN
3: BRAZEY LEE JOHN GERY STEVE BERT RUSS
  1 1 1   1 1   1 1 1 1
1 0 2 4 9   4 6 8 7 5 3   1 3 2 5 6 7 8
H B D H M   P J A P P C   L J B G S B R
1   HOLLY | 1   1 | 1   1 | |
10   BILL | 1 1 1 1 | | |
12   DON | 1   1 1 1 | | |
14 HARRY | 1   1 1 1 | | |
9 MICHAEL | 1   1 1 1 | | |
4   PAM | | 1 1 1 1 | |
6 JENNIE | | 1 1 1   1 | |
8   ANN | | 1 1 1 1 | |
7 PAULINE | | 1   1 1 1 | |
5   PAT | 1 | 1   1 1 | |
3   CAROL | | 1   1 1 1 | |
11   LEE | | | 1   1   1 1 |
13   JOHN | | 1 | 1   1   1 |
2 BRAZEY | | | 1   1   1 1 |
15   GERY | 1 | | 1 1   1 |
16 STEVE | | | 1   1 1 1 |
17   BERT | | | 1   1 1 1 |
18   RUSS | | | 1 1 1 1 |
BILL
DON
HARRY
MICHAEL
HOLLY
PAT
GERY
LEE
STEVE
PAM
JENNI
BRAZEY
RUSS
JOHN
PAULINE
CAROL
BERT
ANN

Slide 51


Core-Periphery Models
• A core periphery structure has a single cohesive subgroup with a set of other nodes, loosely connected to the core
• Core members interact with (lots of) other core members
• Peripheral members interact with (a few) core members

Slide 52

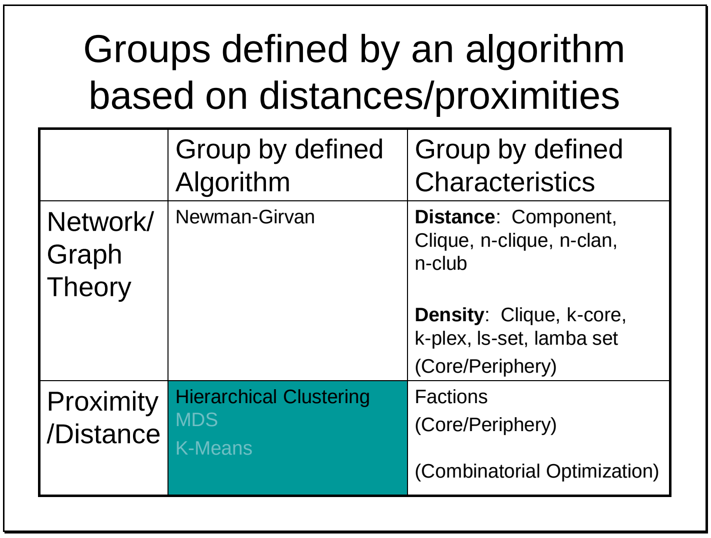

Core Periphery Block Model
• Density Matrix
1 2
----- -----
1 0.451 0.106
2 0.106 0.057

Slide 53

Finding Core/Periphery Structures

- Two ways to deal with it…
  - One is a special case of factions, which maximizes density of core-to-core relations and minimizes all others (categorical)
  - Another is a continuous model that calculates a “coreness” which is how much this node looks like a core node (continuous).

Slide 54



Groups defined by an algorithm based on distances/proximities

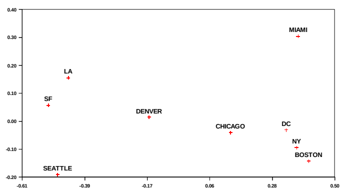
	Group by defined Algorithm	Group by defined Characteristics
Network/ Graph Theory	Newman-Girvan	**Distance**: Component, Clique, n-clique, n-clan, n-club **Density**: Clique, k-core, k-plex, ls-set, lamba set (Core/Periphery)
Proximity /Distance	Hierarchical Clustering MDS K-Means	Factions (Core/Periphery) (Combinatorial Optimization)

Slide 55

Johnson's Hierarchical Clustering

- Output is a set of nested **partitions**, starting with identity partition and ending with the complete partition
  - A "PARTITION" is a vector that associates each node with one and only one "group" (mutually exclusive)
- Different flavors based on how distance from a cluster to outside point/node is defined
  - Single linkage; connectedness; minimum
  - Complete linkage; diameter; maximum
  - Average, median, etc.

Slide 56



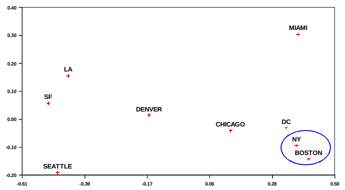

```
                M S       B     C D
                I E S L O N D H E
                A A F A S Y C I N

Level           4 6 7 8 1 2 3 5 9
-----           - - - - - - - - -
  206           . . . . XXX . . .
  233           . . . . XXXXX . .
  379           . . XXX XXXXX . .
  671           . . XXX XXXXXXX .
  808           . XXXXX XXXXXXX .
  996           . XXXXX XXXXXXXXX
 1059           . XXXXXXXXXXXXXXX
 1075           XXXXXXXXXXXXXXXXX
```

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

Closest distance is NY-BOS = 206, so merge these.

Slide 57



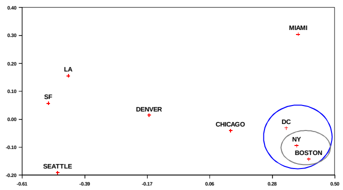

```
          M S       B     C D
          I E S L O N D H E
          A A F A S Y C I N

Level     4 6 7 8 1 2 3 5 9
-----     - - - - - - - - -
  206     . . . . XXX . . .
  233     . . . . XXXXX . .
  379     . . XXX XXXXX . .
  671     . . XXX XXXXXXX .
  808     . XXXXX XXXXXXX .
  996     . XXXXX XXXXXXXXX
 1059     . XXXXXXXXXXXXXXX
 1075     XXXXXXXXXXXXXXXXX
```

	BOS NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS/ NY	0	233	1308	802	2815	2934	2786	1771
DC	233	0	1075	671	2684	2799	2631	1616
MIA	1308	1075	0	1329	3273	3053	2687	2037
CHI	802	671	1329	0	2013	2142	2054	996
SEA	2815	2684	3273	2013	0	808	1131	1307
SF	2934	2799	3053	2142	808	0	379	1235
LA	2786	2631	2687	2054	1131	379	0	1059
DEN	1771	1616	2037	996	1307	1235	1059	0

Closest pair is DC to BOSNY combo @ 233. So merge these.

Slide 58

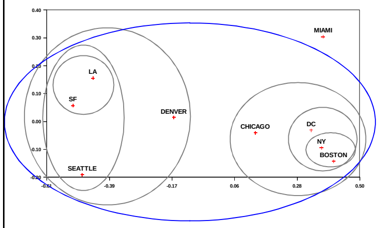



```
         M S       B     C D
         I E S L O N D H E
         A A F A S Y C I N

Level    4 6 7 8 1 2 3 5 9
-----    - - - - - - - - -
  206    . . . . XXX . . .
  233    . . . . XXXXX . .
  379    . . XXX XXXXX . .
  671    . . XXX XXXXXXX .
  808    . XXXXX XXXXXXX .
  996    . XXXXX XXXXXXXXX
 1059    . XXXXXXXXXXXXXXX
 1075    XXXXXXXXXXXXXXXXX
```

	BOS/ NY/ DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY DC	0	1075	671	2684	2799	2631	1616
MIA	1075	0	1329	3273	3053	2687	2037
CHI	671	1329	0	2013	2142	2054	996
SEA	2684	3273	2013	0	808	1131	1307
SF	2799	3053	2142	808	0	379	1235
LA	2631	2687	2054	1131	379	0	1059
DEN	1616	2037	996	1307	1235	1059	0

Slide 63




```
           M S     B       C D
           I E S L O N D H E
           A A F A S Y C I N

Level      4 6 7 8 1 2 3 5 9
-----      - - - - - - - - -
  206      . . . . XXX . . .
  233      . . . . XXXXX . .
  379      . . XXX XXXXX . .
  671      . . XXX XXXXXXX .
  808      . XXXXX XXXXXXX .
  996      . XXXXX XXXXXXXXX
 1059      . XXXXXXXXXXXXXXX
 1075      XXXXXXXXXXXXXXXXX
```

	BOS/NY/DC/CHI/DEN/SF/LA/SEA	MIA
BOS/NY/DC/CHI/DEN/SF/LA/SEA	0	1075
MIA	1075	0

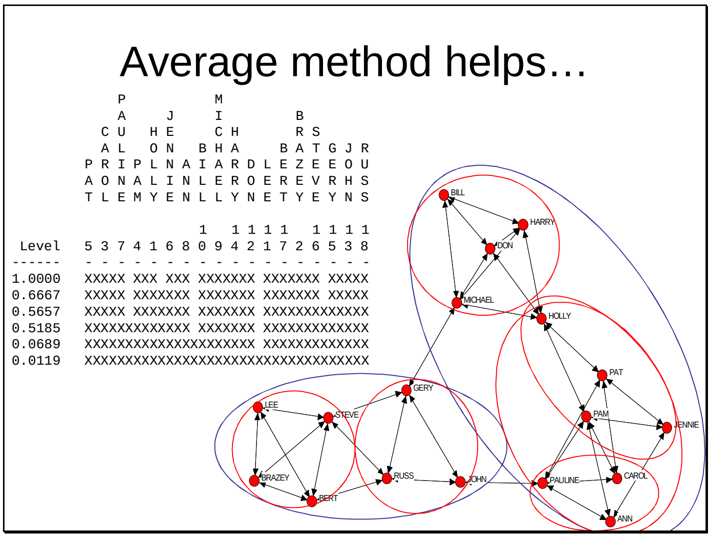
Slide 64

Clustering Network Data

- Clustering requires symmetric data
  - It will symmetrize if you don't
- Does better with a range of values

```
                    P   M
          B       J A   I
        H R C     E U   C         H   S
        O A A     N L   H B     J A G T B R
        L Z R P P N I A A I L D O R E E E U
        L E O A A I N N E L E O H R R V R S
        Y Y L M T E E N L L E N N Y Y E T S
                          1 1 1 1 1 1 1 1 1
Level   1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8
-----   - - - - - - - - - - - - - - - - - -
    0   XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Slide 65


Average method helps…
P M
A J I B
C U H E C H R S
A L O N B H A B A T G J R
P R I P L N A I A R D L E Z E E O U
A O N A L I N L E R O E R E V R H S
T L E M Y E N L L Y N E T Y E Y N S
1 1 1 1 1 1 1 1 1
Level 5 3 7 4 1 6 8 0 9 4 2 1 7 2 6 5 3 8
------ - - - - - - - - - - - - - - - - - -
1.0000 XXXXX XXX XXX XXXXXXX XXXXXXX XXXXX
0.6667 XXXXX XXXXXXX XXXXXXX XXXXXXX XXXXX
0.5657 XXXXX XXXXXXX XXXXXXX XXXXXXXXXXXXX
0.5185 XXXXXXXXXXXXX XXXXXXX XXXXXXXXXXXXX
0.0689 XXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXX
0.0119 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
BILL
HARRY
DON
MICHAEL
HOLLY
PAT
PAM
JENNIE
GERY
LEE
STEVE
BRAZEY
BERT
RUSS
JOHN
PAULINE
CAROL
ANN

Slide 66

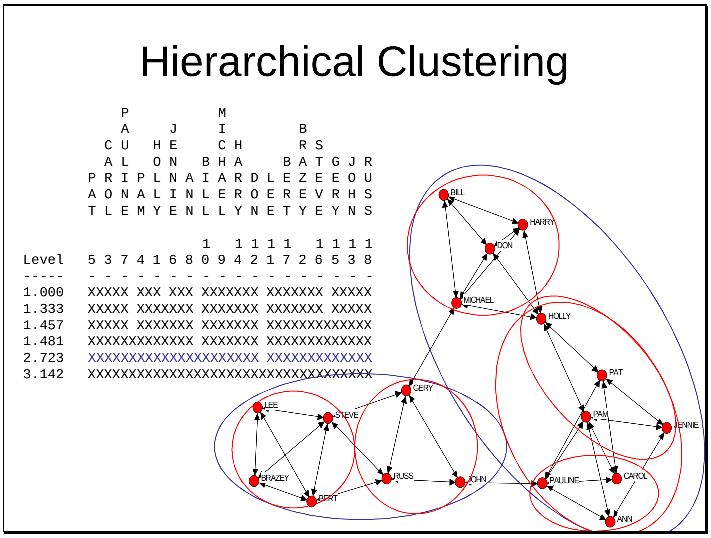
Applying HiClus to Network Data

- BETTER: Compute geodesic distances first, then cluster the distance matrix
- Or cluster the structural equivalence matrix

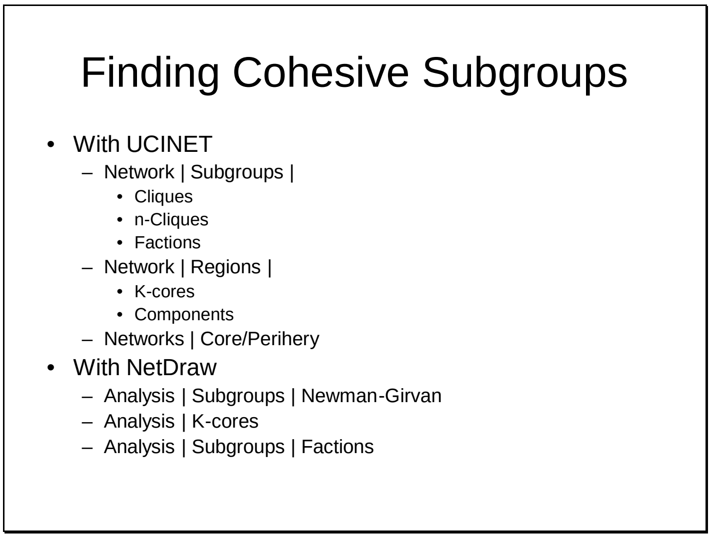
Geodesic Distances

		1 H	2 B	3 C	4 P	5 P	6 J	7 P	8 A	9 M	10 B	11 L	12 D	13 J	14 H	15 G	16 S	17 B	18 R
1	HOLLY	0	4	2	1	1	2	2	2	1	2	4	1	3	1	2	3	4	3
2	BRAZEY	4	0	5	5	5	6	4	5	3	4	1	4	3	4	2	1	1	2
3	CAROL	2	5	0	1	1	2	1	2	3	4	5	3	2	3	3	4	4	3
4	PAM	1	5	1	0	2	1	1	1	2	3	5	2	2	2	3	4	4	3
5	PAT	1	5	1	2	0	1	1	2	2	3	5	2	2	2	3	4	4	3
6	JENNIE	2	6	2	1	1	0	2	1	3	4	6	3	3	3	4	5	5	4
7	PAULINE	2	4	1	1	1	2	0	1	3	4	4	3	1	3	2	3	3	2
8	ANN	2	5	2	1	2	1	1	0	3	4	5	3	2	3	3	4	4	3
9	MICHAEL	1	3	3	2	2	3	3	3	0	1	3	1	2	1	1	2	3	2
10	BILL	2	4	4	3	3	4	4	4	1	0	4	1	3	1	2	3	4	3
11	LEE	4	1	5	5	5	6	4	5	3	4	0	4	3	4	2	1	1	2
12	DON	1	4	3	2	2	3	3	3	1	1	4	0	3	1	2	3	4	3
13	JOHN	3	3	2	2	2	3	1	2	2	3	3	3	0	3	1	2	2	1
14	HARRY	1	4	3	2	2	3	3	3	1	1	4	1	3	0	2	3	4	3
15	GERY	2	2	3	3	3	4	2	3	1	2	2	2	1	2	0	1	2	1
16	STEVE	3	1	4	4	4	5	3	4	2	3	1	3	2	3	1	0	1	1
17	BERT	4	1	4	4	4	5	3	4	3	4	1	4	2	4	2	1	0	1
18	RUSS	3	2	3	3	3	4	2	3	2	3	2	3	1	3	1	1	1	0

Slide 67


Hierarchical Clustering
P M
A J I B
C U H E C H R S
A L O N B H A B A T G J R
P R I P L N A I A R D L E Z E E O U
A O N A L I N L E R O E R E V R H S
T L E M Y E N L L Y N E T Y E Y N S
1 1 1 1 1 1 1 1 1
Level 5 3 7 4 1 6 8 0 9 4 2 1 7 2 6 5 3 8
----- - - - - - - - - - - - - - - - - - -
1.000 XXXXX XXX XXX XXXXXXX XXXXXXX XXXXX
1.333 XXXXX XXXXXXX XXXXXXX XXXXXXX XXXXX
1.457 XXXXX XXXXXXX XXXXXXX XXXXXXXXXXXXX
1.481 XXXXXXXXXXXXX XXXXXXX XXXXXXXXXXXXX
2.723 XXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXX
3.142 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
BILL
HARRY
DON
MICHAEL
HOLLY
PAT
PAM
JENNIE
PAULINE
CAROL
ANN
GERY
LEE
STEVE
BRAZEY
BERT
RUSS
JOHN

Slide 68


Finding Cohesive Subgroups
• With UCINET
– Network | Subgroups |
• Cliques
• n-Cliques
• Factions
– Network | Regions |
• K-cores
• Components
– Networks | Core/Perihery
• With NetDraw
– Analysis | Subgroups | Newman-Girvan
– Analysis | K-cores
– Analysis | Subgroups | Factions

Cohesion Lab

For this lab we will use four datasets:

CAMPNET:

This is a dichotomous adjacency matrix of 18 participants in a qualitative methods class. Ties are directed and represent that the ego indicated that the nominated alter was one of the three people with which s/he spent the most time during the seminar.

KAPTAIL:

This is a stacked dataset containing four dichotomous matrices. There are two adjacency matrices each for social ties (indicating the pair had social interaction) and instrumental ties (indicated the pair had work-related interaction). The two pairs of matrices represent two different points in time. The names of the datasets encode the type of tie in the sixth letter, and the time period in the seventh. Thus, the dataset KAPFTS1 is social ties at time 1 and KAPFTI2 is instrumental ties at time 2, etc.

ZACKAR & ZACHATTR:

ZACKAR is another stacked dataset, containing a dichotomous adjacency matrix, ZACHE, which represents the simple presence or absence of ties between members of a Karate Club, and ZACHC, which contains valued data counting the number of interactions between actors. ZACHATTR is a rectangular matrix with three columns of attributes for each of the actors from the ZACKAR datasets.

PV504:

This is a symmetric, valued dataset. It has data for 504 actors that are employees of a consulting organization, with values representing the number of days on which each pair worked together on projects.

EXERCISES:

1) Cohesion using UCINET with **CAMPNET**

a. Calculate the following measures of cohesion using Network | Cohesion
   - Density
   - Distance
   - Maximum Flow

b) Distance produces a matrix. Using this matrix, how would you determine the network diameter?

c) Using your Netdraw visualization, verify a couple entries in the distance, point connectivity, and maximum flow matrices produced.

2) Fragmentation using UCINET and **KAPTAIL**

Using the **KAPFTS1** dataset (you may have to unpack **KAPTAIL** if you have not already done so), calculate its fragmentation under Network | Centrality using both measures in the options portion of the dialog box and compare the results. Why did they generate different results? Which one is more useful for this network? When would you choose to use one or the other?

3) Hierarchical Clustering using UCINET with **ZACKAR**

a. This section uses the ZACHE dataset (you may have to unpack **ZACKAR** to create **ZACHE**) and the **ZACHATTR** attribute dataset. Check to make sure you have both, and let one of the facilitators know if you do not.

b. Now, run SINGLE_LINK method Hierarchical Clustering (Tools|Cluster|Hierarchical) on the **ZACHE** adjacency matrix (specifying the appropriate kind of data). What does the output tell you? Why did you get this result?

c. Now re-run using the AVERAGE method on the same data. Why did you get a different result? Which one is more useful in identifying cohesive subgroups from these data?

d. Now, create the geodesic distance matrix from these data (Network|Cohesion|Distance) and run that matrix through Hierarchical clustering with the **appropriate parameters,** using both the AVERAGE and SINGLE_LINK methods. How did your results (and parameters) vary from using the adjacency matrix?

4) Newman-Girvan using NetDraw with **ZACKAR**

a. Open the **ZACKAR** stacked dataset in NetDraw. It should open to displaying the relation **ZACHE** but if not, make sure it does.

b. Now, open the attribute file, **ZACHATTR**, using the folder with the A next to it. (It should display a spreadsheet with four columns. Just close that window.)

c. Run the Newman-Girvan analysis (Analysis|Subgroups|Newman-Girvan) specifying a minimum of 2 and a maximum of 40 clusters desired. It should automatically color your nodes so that there are nodes are one of two colors. What it has done behind the scenes is color based on the ngPart_2 partition (a partition with 2 colors). Click on the color palette icon and pull down on the drop down list to select ngPart_3 to see how it partitions it next. And then ngPart_4. How useful are these partitions?

d. Using the color palette, go back to the ngPart_2 partition. Now, click on the shape palette icon, and select "Club" from the list. This will shape the nodes according to which club the members went to after the split. How well did the N-G algorithm predict the affiliation of the club members?

5) Factions using Netdraw with **ZACKAR**

Now run Analysis|Subgroups|Factions selecting 2 for the desired number of groups. This time, instead of using the color palette, use the "Nodes" tab in the control area on the right hand side of the screen and scroll down to the last attribute, which should be called "Factions 2" and then click the "Color" checkbox. How does factions compare with the Newman-Girvan algorithm in terms of predicting the affiliations? How could you display both the Newman-Girvan results and the Factions result at the same time?

6) Factions using UCINET with **KAPTAIL**

a. If you have not already done so, in UCINET, unpack the **KAPTAIL** dataset to make **KAPFTI1**, **KAPFTI2**, **KAPFTS1**, **KAPFTS2**. These are "instrumental" and "social" ties at time 1 & 2 from a tailor shop in Zambia. The instrumental ties are asymmetric, and the social ties are undirected. We will be using the social ties at time 1 (**KAPFTS1)** for this analysis.

b. Run Network|Subgroups|Factions on **KAPFTS1** and specify two factions. Review the results. How well do you think the two factions describe the subgroup structure of the network? Run it a couple more times, increasing the number of factions. How do the results change? Why? How would you interpret those changes?

7) Core-Periphery using UCINET with **KAPTAIL**

Run Network|Core/Periphery on **KAPFTS1** and **KAPFTS2**. How do the results differ? Which time might you expect that there was a successful strike in the Tailor Shop and why?

8) Cliques using UCINET and NetDraw with **KAPTAIL**

a. In UCINET run Network|Subgroups|Cliques on **KAPFTS2** with the a minimum size of 3. How many cliques do you get? How useful is this?

b. Let's visualize the data. Open **KAPFTS2** in NetDraw. Does this help us identify clique structures?

c. What about if we open **CliqueOverlap** (which is an actor-by-actor matrix in which each cell holds the number of different cliques that this pair of actors is in together that was created when we ran Cliques in UCINET). Start increasing the filter at the bottom of the "Rels" tab on the control panel on the right side of the screen up from 1 using the "+" button. Does this indicate there is a significant or minimal overlap of cliques in this structure?

d. Now open **CliqueSets** in Netdraw and set the filter value back down to 0 and redraw the picture. This is a two-mode network were lines indicate actors (typically red circles with names) belong to a specific clique (typically blue squares with numbers). What does this picture convey about the structure of the network?

e. Try deleting the pendants (nodes with only one line, in this case, they would be people who are members of only one clique) by pressing the button labeled ~~**Pen**~~ on the icon bar in NetDraw, and the isolates (those with no lines, in this case not members of any clique) using the ~~**Iso**~~ button. How much did that affect the meaning of the visualization for you?

9) n-Cliques using UCINET and NetDraw with **KAPTAIL**

Now, run Network|Subugroups|n-Cliques on **KAPFTS1** specifying n = 2, and minimum size of 4. How many 3-cliques did you get? Open **nClqSets** in NetDraw and compare that picture with the one from the previous step. What is the difference between these two pictures?

10) K-CORES using NetDraw with **PV504**

a. Open **PV504** in NetDraw. Increase the filtering to only show relations of more than 3 (days together on a projects), turn off labels (using the script L button), and redraw the network.

b. Now run Analysis|K-Cores. It will automatically color the nodes according to their k-Core. Select the Nodes tab, and pull down to the *K-core attribute, and use the "s" button to step through the k-cores from 1 to 10. What does this tell you about the network?