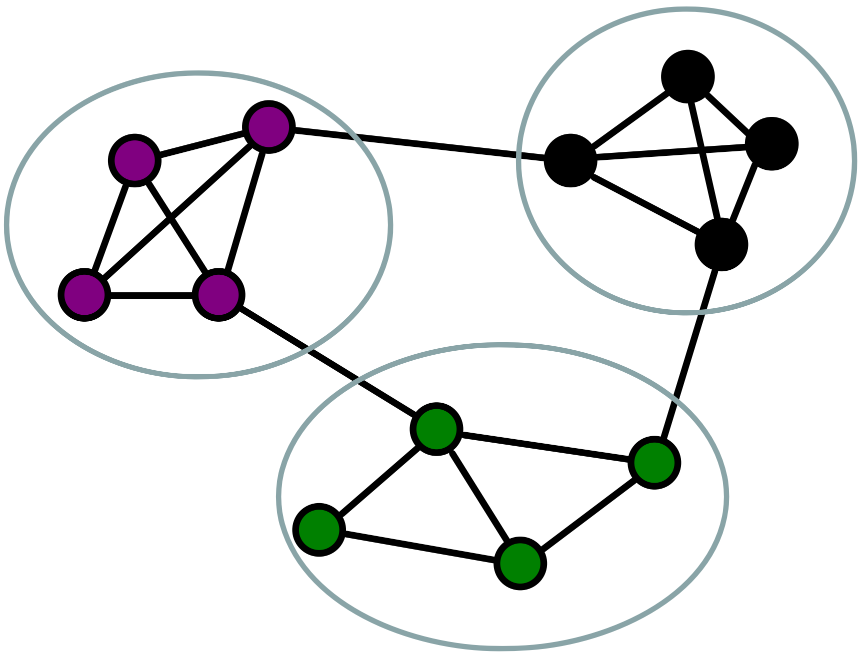
Cohesive Subgroups

Steve Borgatti & Rich DeJordy

Revised March 26, 2008

What are cohesive subgroups?

- Loosely speaking, network “clumps”
  - Regions of high density and low distance



Cohesion family of concepts

	HO LLY	BIL L	DO N	HA RR Y	MI CH AEL	PA M	JEN NIE	AN N	PA ULI NE	PAT	CA RO L	LEE	JO HN	BR AZ EY	GE RY	STE VE	BE RT	RUSS	Sum
HOLLY			1	1	1	1				1									5
BILL			1	1	1														3
DON	1	1		1	1														4
HARRY	1	1	1		1														4
MICHAEL	1	1	1	1											1				5
PAM	1						1	1	1		1								5
JENNIE						1		1		1									3
ANN						1	1		1										3
PAULINE						1		1		1	1		1						5
PAT	1						1		1		1								4
CAROL						1			1	1									3
LEE														1		1	1		3
JOHN									1						1			1	3
BRAZEY												1				1	1		3
GERY					1								1			1		1	4
STEVE												1		1	1		1	1	5
BERT												1		1		1		1	4
RUSS													1		1	1	1		4
Sum	5	3	4	4	5	5	3	3	5	4	3	3	3	3	4	5	4	4	70

0.229 (average)

Concept:
- Relational
- Centrality
- Subgroups
- Network

Why we care

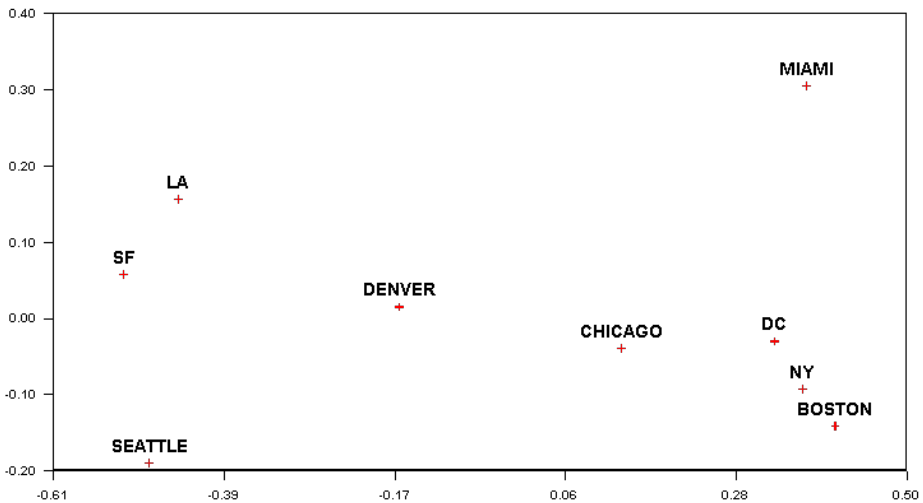
- They exist!
  - Real networks often clumpy. Clumps are subgroups
- They affect (social) processes we care about.
  - Groups are powerful influencers of members
    - Create homogeneity
  - Output is group membership variable (eg segments)
- Need for data reduction
  - Can analyze each group separately
  - Can aggregate subgroups into “supernodes”
    - Build simplified model of complex network

Two approaches

- Mathematical ethnography
  - Take squishy sociological ideas like “primary group” and translate into formal|mathematical language
  - Theoretically informed models
- Methodological
  - Define algorithms that will cluster dyadic cohesion matrix to find clumps
    - Data mining, clustering, neural nets
  - Atheoretical clusters

Johnson’s Hierarchical Clustering

- Output is a set of nested **partitions**, starting with identity partition and ending with the complete partition
  - Partition represented as 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.

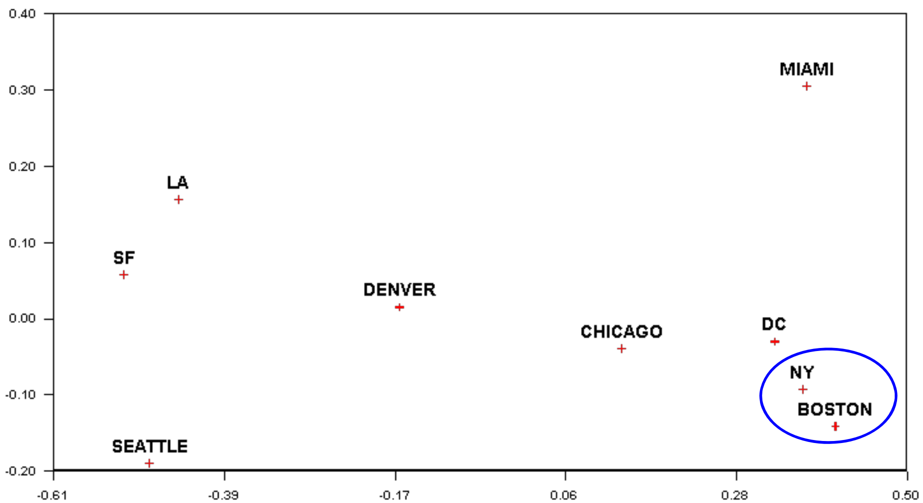



```
        M S     B     C D
        I E S L O N D H E
        A A F A S Y C I N

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

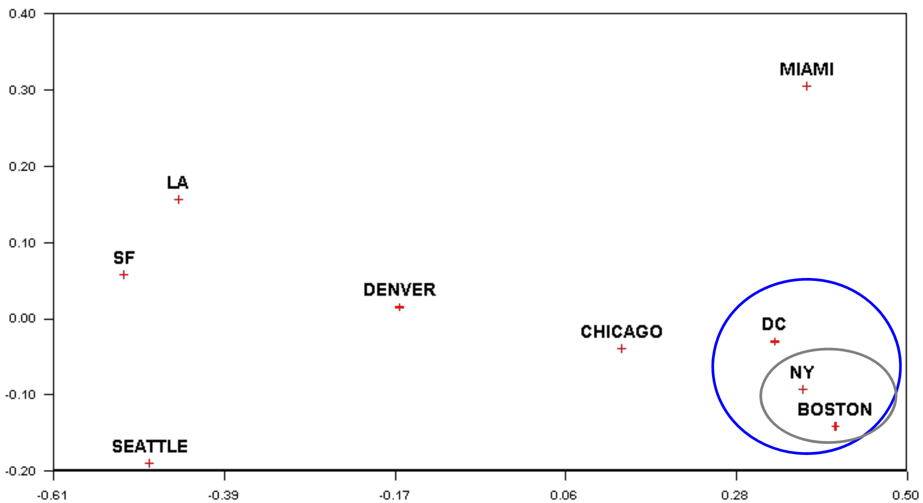
Closest distance is NY-BOS = 206, so merge these.




```
        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

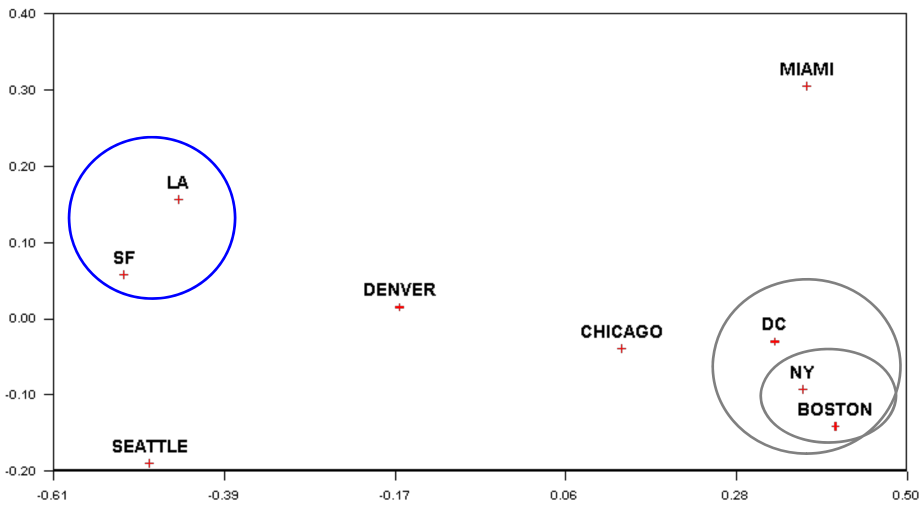
Closest pair is DC to BOSNY combo @ 233. So merge these.




```
         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
```

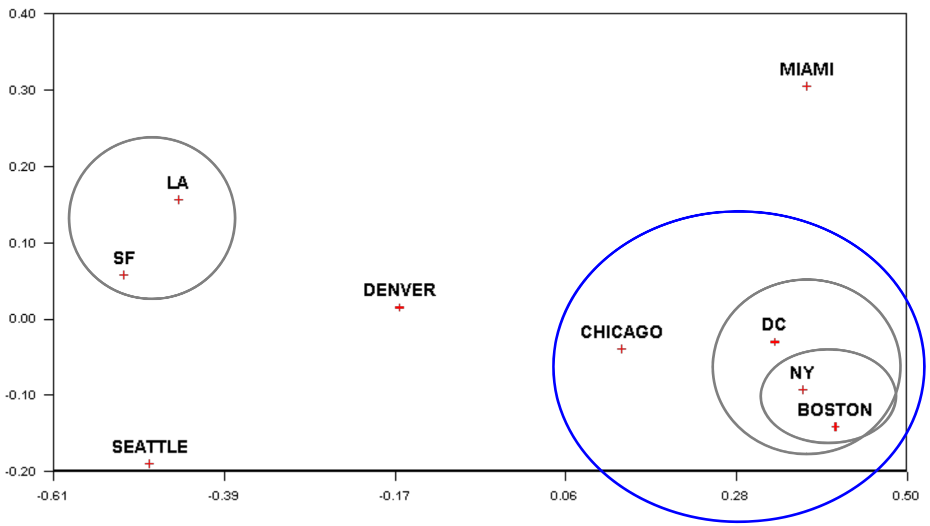
	BOS/ NY/ DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY DC	0	1075	671	2684	2799	2631	1616
MIA	1075	0	1329	3273	3053	2687	2037
CHI	671	1329	0	2013	2142	2054	996
SEA	2684	3273	2013	0	808	1131	1307
SF	2799	3053	2142	808	0	379	1235
LA	2631	2687	2054	1131	379	0	1059
DEN	1616	2037	996	1307	1235	1059	0




```
          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
```

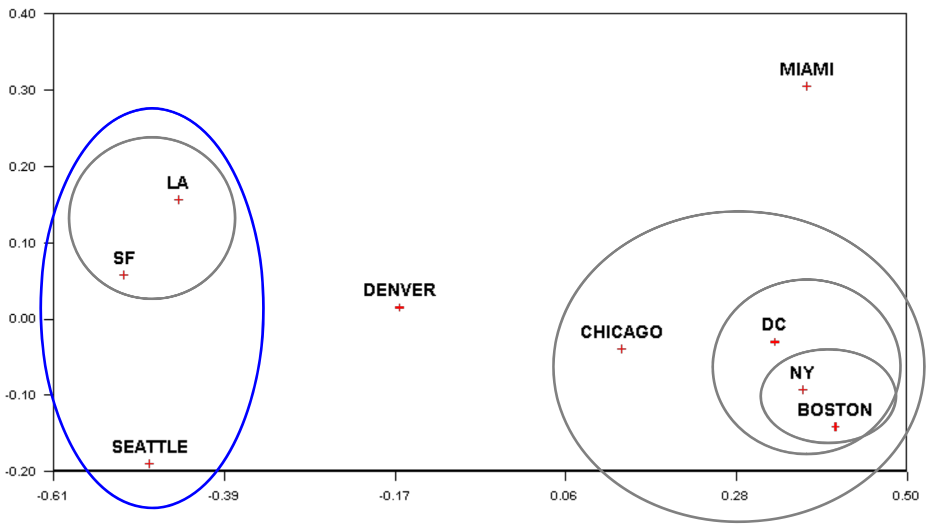
	BOS/ NY/DC	MIA	CHI	SEA	SF/LA	DEN
BOS/NY/DC	0	1075	671	2684	2631	1616
MIA	1075	0	1329	3273	2687	2037
CHI	671	1329	0	2013	2054	996
SEA	2684	3273	2013	0	808	1307
SF/LA	2631	2687	2054	808	0	1059
DEN	1616	2037	996	1307	1059	0




```
              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
```

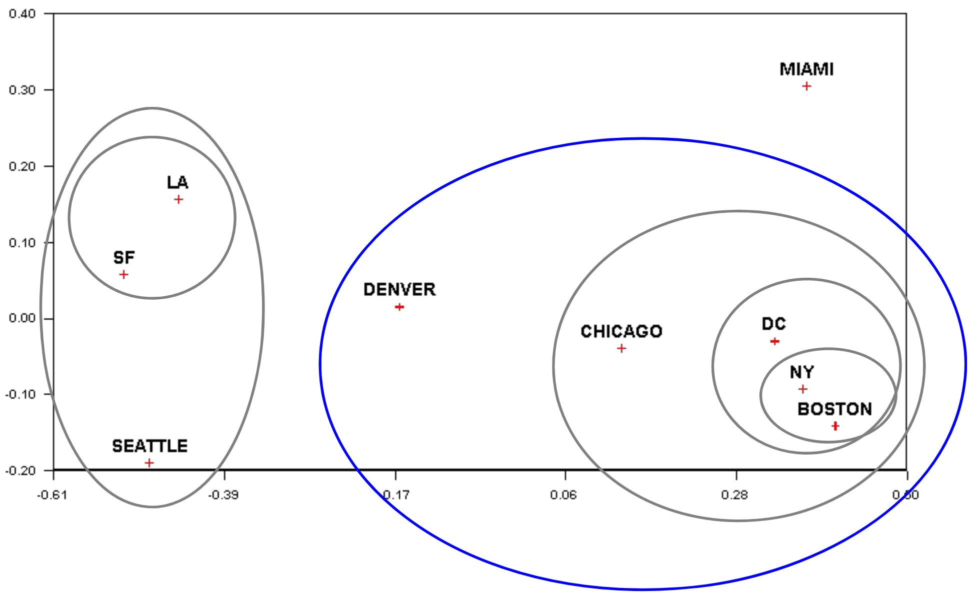
	BOS/NY/DC/CHI	MIA	SEA	SF/LA	DEN
BOS/NY/DC/CHI	0	1075	2013	2054	996
MIA	1075	0	3273	2687	2037
SEA	2013	3273	0	808	1307
SF/LA	2054	2687	808	0	1059
DEN	996	2037	1307	1059	0




```
       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
```

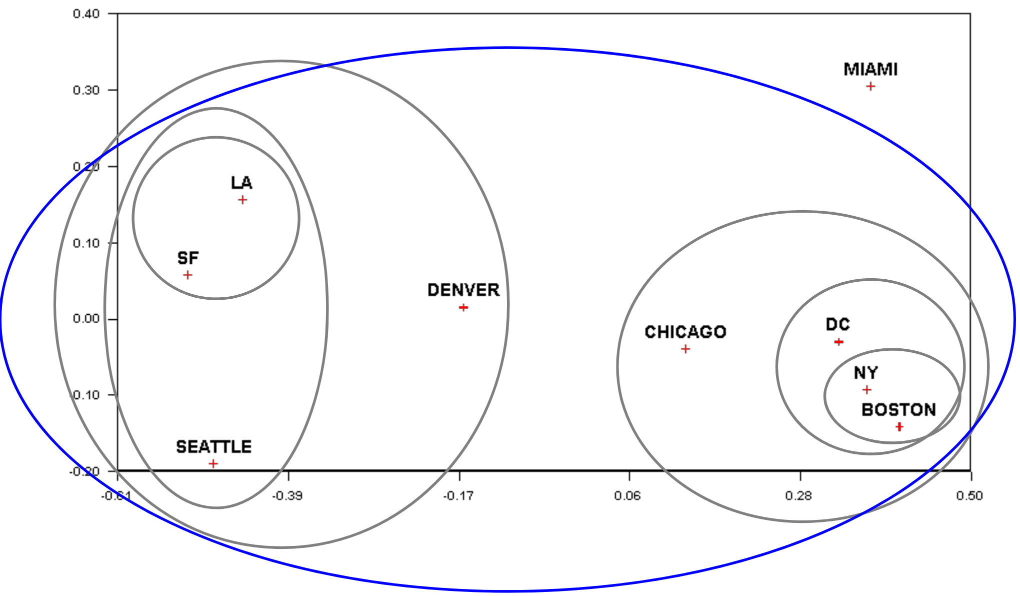
	BOS/NY/DC/CHI	MIA	SF/LA/SEA	DEN
BOS/NY/DC/CHI	0	1075	2013	996
MIA	1075	0	2687	2037
SF/LA/SEA	2054	2687	0	1059
DEN	996	2037	1059	0




```
       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	MIA	SF/LA/SEA
BOS/NY/DC/CHI/DEN	0	1075	1059
MIA	1075	0	2687
SF/LA/SEA	1059	2687	0




```
         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

Applying HiClus to Network Data

- Compute geodesic distances first, then cluster the distance matrix

- Or cluster the structural equivalence matrix

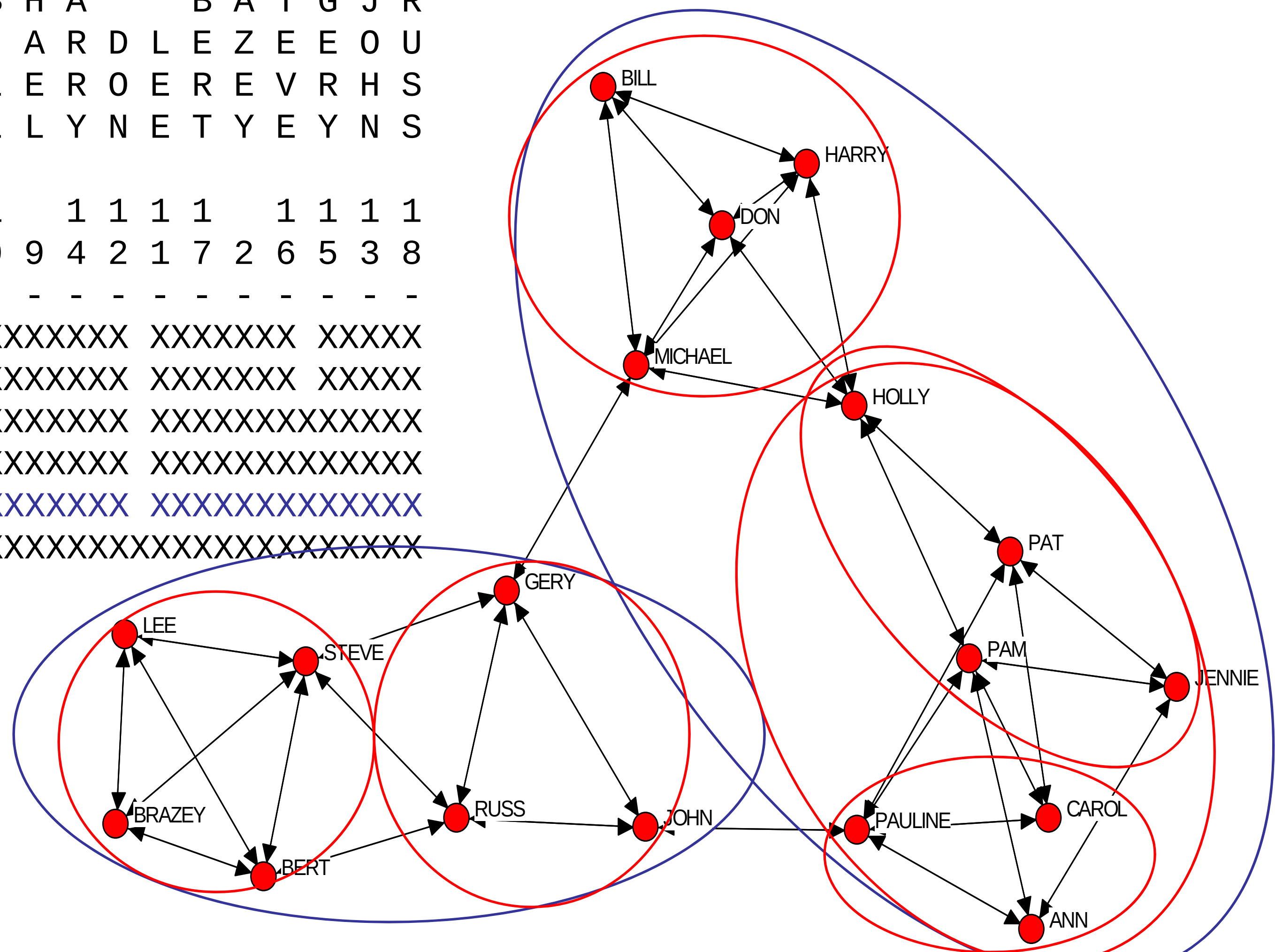
Geodesic Distances

		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

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
```

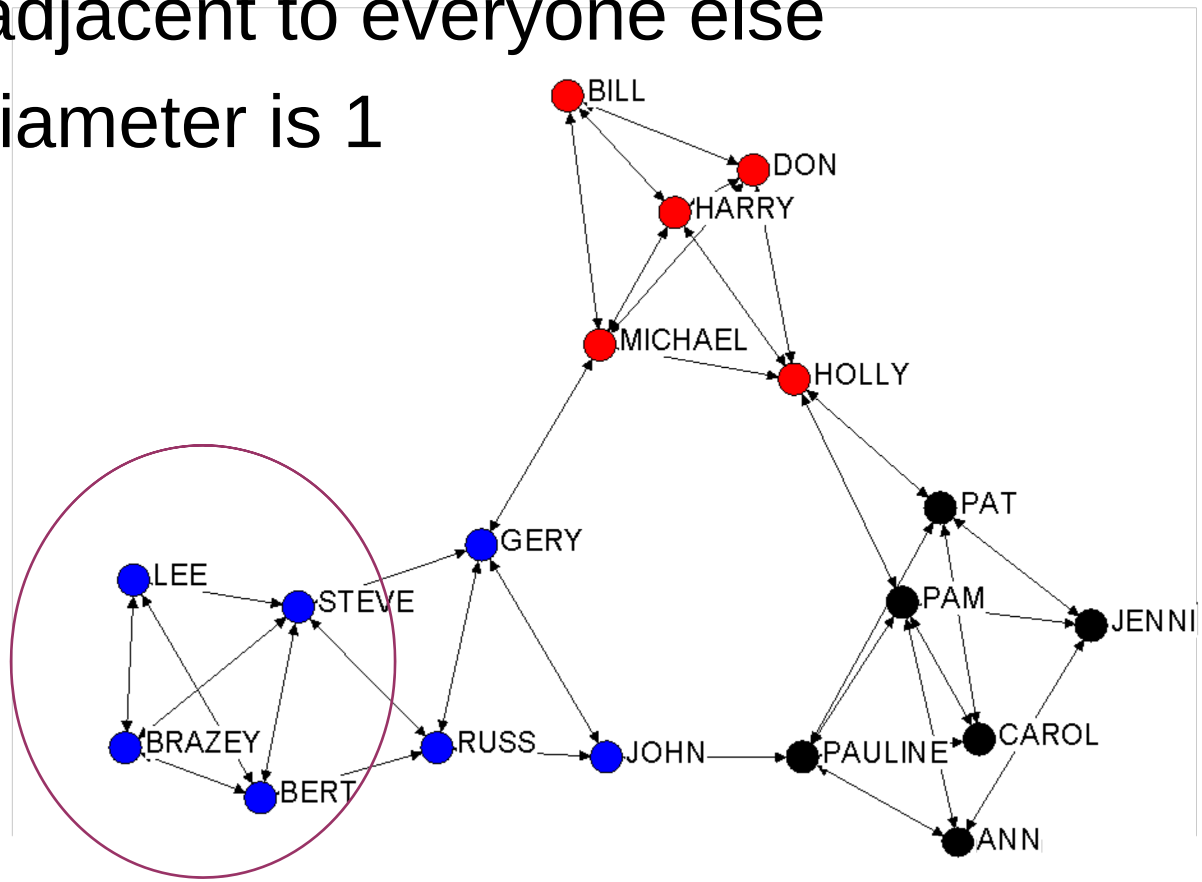


Newman-Girvan

- Calculate edge-betweenness for graph
- Remove edge with highest edge betweenness
- If number of components increases, record partition
- Recalculate edge betweenness & repeat until all nodes are isolates or maximum number of clusters reached/exceeded

Clique

- A maximal **complete** subgraph
  - – Everyone is adjacent to everyone else
  - – Distance & Diameter is 1
  - – Density is 1
- Limitations
  - – Undirected
  - – 3+ nodes



Factions

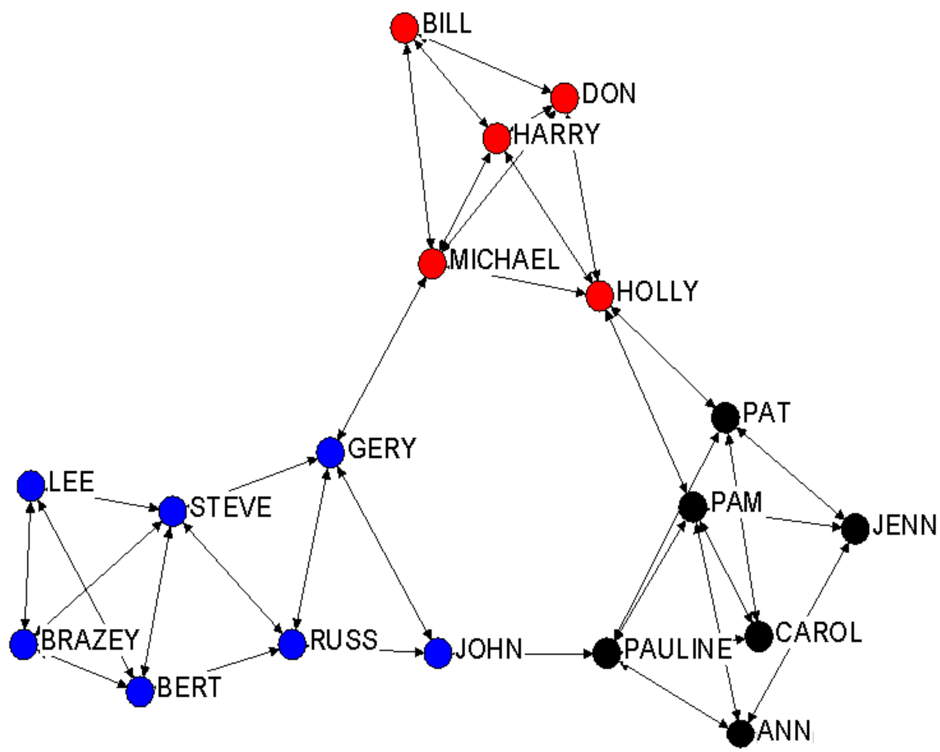
- Computationally arrange nodes into mutually exclusive groups such that some predefined criteria is optimized
  - For example, make groups that maximize density of internal ties and minimize density of external ties

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

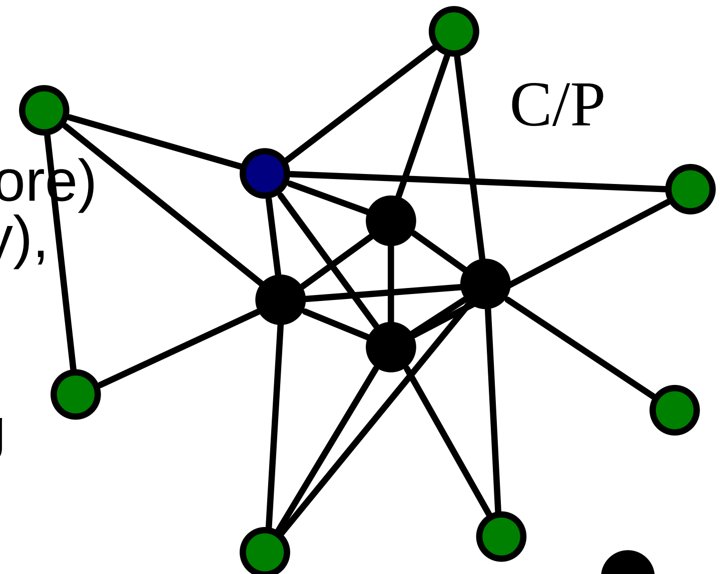
		1 H	10 B	12 D	14 H	9 M	4 P	6 J	8 A	7 P	5 P	3 C	11 L	13 J	2 B	15 G	16 S	17 B	18 R
1	HOLLY	1		1			1				1								
10	BILL		1	1	1	1													
12	DON	1		1	1	1													
14	HARRY	1		1	1	1													
9	MICHAEL	1		1	1	1													
4	PAM						1	1	1	1									
6	JENNIE						1	1	1		1								
8	ANN						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

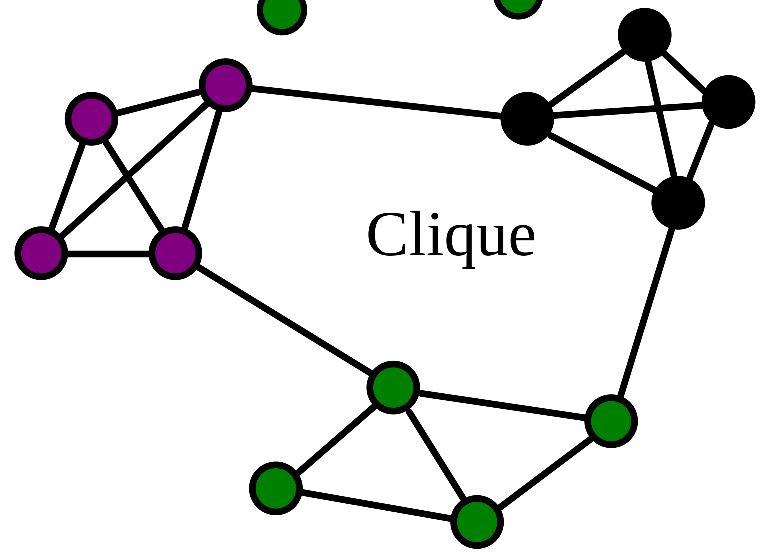


GROUP level of analysis

Core/Periphery Structures

- Core/Periphery
  - Network consists of single group (a core) together with hangers-on (a periphery),
    - Core connects to all
    - Periphery connects only to the core
  - Short distances, good for transmitting information, practices
  - Identification with group as whole
  - E.g., structure of physics
- Clique structure
  - Multiple subgroups or factions
  - Identity with subgroup
  - Diversity of norms, belief
  - E.g., structure of social science

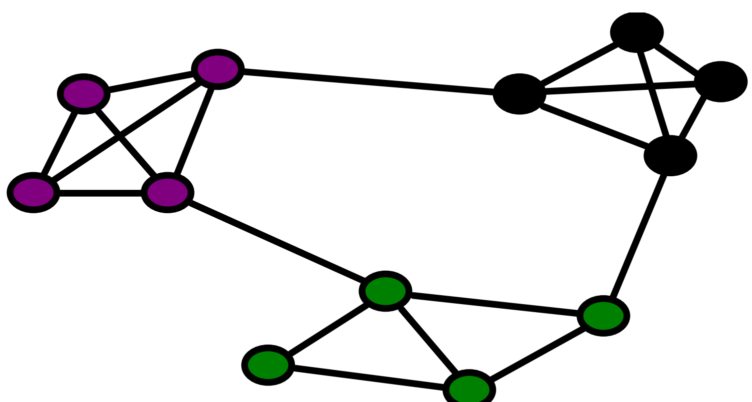




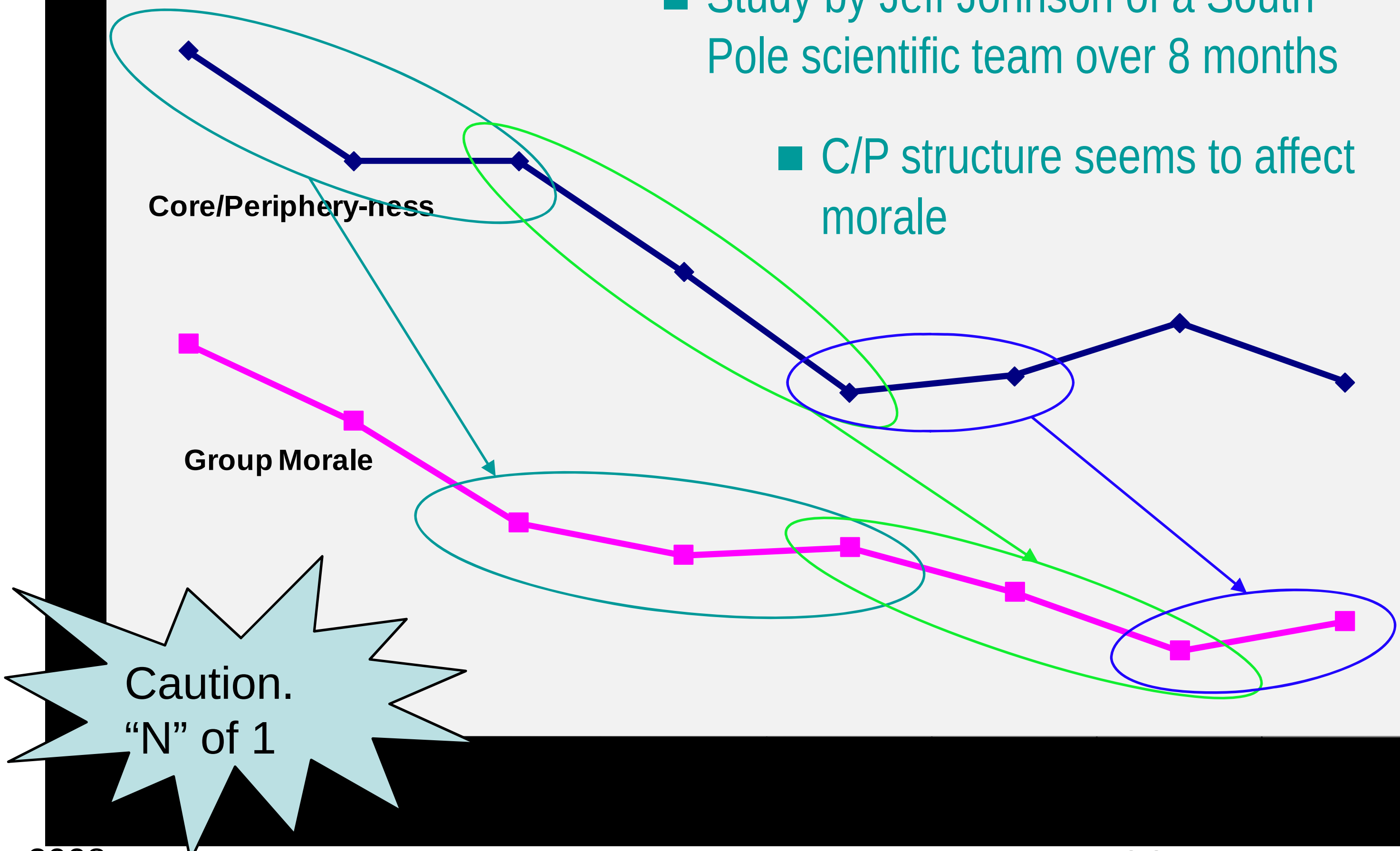
On Innovation and Network Structure

"I would never have conceived my theory, let alone have made a great effort to verify it, if I had been more familiar with major developments in physics that were taking place. Moreover, my initial ignorance of the powerful, false objections that were raised against my ideas protected those ideas from being nipped in the bud."

– Michael Polanyi (1963), on a major contribution to physics



Case Study: Johnson's study of morale at the South Pole



- Study by Jeff Johnson of a South Pole scientific team over 8 months
  - C/P structure seems to affect morale

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)