Cohesive Subgroups (Answer Sheet)

For this lab we will use three datasets:

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

PV504 is a 504-actor network of consultants working for an R&D consulting firm. The data are symmetric and valued and represent the number of days that pair of individuals worked on a project together.

EXERCISES:

1) Hierarchical Clustering using UCINET with ZACKAR

a) This section uses the ZACHE dataset (you may have to unpack ZACKAR using Data | Unpack 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.

	ucinetlog2 - Notepad	- 🗆 ×
File Edit Format View Help		
UNPACK		^
Created dataset:	ZACHE (C:\Users\lowol_000\Documents\UCINET data\ZACHE	
Created dataset:	ZACHC (C:\Users\lowol_000\Documents\UCINET data\ZACHC	
Running time: 00:00:01 seconds.		
Output generated: 29 May 15 14:18:06	nalvtic Technologies	
Cliner 0.575 copyright (c) 1552-2015 p	marytic recimorogres	
		~

b) Now, run SINGLE_LINK method Hierarchical Clustering

(Tools|Cluster|Hierarchical) on the ZACHC matrix (specifying the appropriate kind of data). We are using ZACHC. What does the output tell you? Why did you get this result? Don't forget to change the type of data (Similarities or Dissimilarities) into "Similarities" because the values in the matrix mean the number of interactions (similarities or intimacy) between actors. Default is "Dissimilarities". Single link method hierarchical clustering defines distance between two clusters as largest similarity between members. For example, similarity level 2 means that every actor within a cluster is no more than 2 units distant from at least one other actor in that cluster. Thus, at similarity level 3, it identifies a subgroup that includes 28 actors who are directly or indirectly connected by ties with values greater than 3, leaving other 6 actors as isolates. Since all actors in ZACHE are connected by ties with values greater than 2, whether directly or indirectly, to each other, they all belong to the same cluster at level 2.



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

Weighted average method considers the average distance between pairs and calculates distance between clusters as the average similarity value weighted by cluster size. For example, similarity level 0.1176 means that the average distance among actors within the cluster is 0.1176. Weighted average method is more useful than single link method here because weighted average method brings out the nuance of the subgroup structure beyond just connected/disconnected (single link method). It considers subgroup's intuitive property of internally dense and externally sparse relationship patterns.



	ucinetlog3 - Notepad	_ 0 ×
File Edit Format View Help		
DOHNSON'S HIERARCHICAL CLUSTERING	ιG	^
Method:	WTD_AVERAGE (average between all pairs)	
Type of Data:	Similarities	
Input dataset:	ZACHC (C:\Users\Wookje Sung\Documents\UCINET data\ZACHC)	
HIERARCHICAL CLUSTERING		
1 1111 Level 516778234812	1 2 2 1 1 2 2 2 1 2 3 2 2 3 3 2 1 2 3 3 2 3 4 0 2 0 9 9 1 3 5 7 0 5 8 9 1 2 6 6 4 3 4	
7.0000		
5 0000		
4.5000 XXX XXX	000 · · · · · · · · · · · · · · · · · ·	
4.0000 XXX XXX	XXX XXX XXX . XXXXX	
3.6667 XXX XXX	0000X X0X X0X X0X . X000X	
3.0000 XXX XXXXX XXX XXX		
2.3333 XXX XXXXX XXXXXXX		
1.6250 XXX XXXXX XXXXXXXX		
1.0833 XXX XXXXX XXXXXXXX	XXXXX XXX XXX XXX XXXXXXXXX	
0.8333 XXXXXXXX XXXXXXXX	00000	
0.7500 XXXXXXXX XXXXXXX		
0.5714 XXXXXXXX XXXXXXXXX		
0.5500 XXXXXXXX XXXXXXXXXX	0000000	
0.5000 XXXXXXXXX XXXXXXXXXXXXX	00000000	
0.4167 XXXXXXXXX XXXXXXXXXX	00000000	
0.3846 XXXXXXXX XXXXXXXXX		
0.2857 XXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXX		
0.2667 XXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXX	00000000	
0.2000 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	00000000 0000000000000000000000000	
0.1875 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX		
0.11/6 ***********************************		
Measures of sluster edequary		
measures or cruster adequacy		
<		>

d) **[OPTIONAL ADVANCED]** We used ZACHC because Hierarchical Clustering really works best when there is variation in the data, and particularly a range of distinct values on which to run the algorithm. Repeat the Hierarchical Clustering Analysis using ZACHE instead of ZACHC. How do the results differ? Which would you find more useful.

d-1) Single link method

ZACHE is dichotomous dataset containing only 1's and 0's. Thus, at similarity level 1, if any one actor in a cluster has a tie (similarity) to another actor, they belong to the same cluster. Since all actors in ZACHE are connected, whether directly or indirectly, to each other, they all belong to the same cluster at level 1. It doesn't identify subgroups unless there is a complete division in the network (i.e. it identifies a component as a subgroup).

G	Cluster Diagram		- 🗆	×						
	1									
▲ <u>OK</u> Save Print		Zoom Scale: X 1	Y [1						
Die Seite Second Many Hele	ucinetlog3 - Notepad		_ (
JOHNSON'S HIERARCHICAL CLUSTERING				^						
Method: Type of Data: Input dataset:	SINGLE_LINK (minimum distance) Similarities ZACHE (C:\Users\lowol_000\Documents\UCINET o	data\ZACHE)								
HIERARCHICAL CLUSTERING										
1 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX										
Measures of cluster adequacy										
Size of each cluster, expressed as a proportion of the total population clustered										
1										
Actor-by-Partition indicator matrix sav	red as dataset Part									
Running time: 00:00:01 Output generated: 29 May 15 14:27:39										

d-2) Weighted average method

The logic is the same as c), but the result is somewhat different because the data are now dichotomous (1's and 0's) instead of valued. For example, similarity level 0.0727 means that the average distance among actors within the cluster is 0.0727. Again, weighted average method is more useful than single link method in identifying cohesive subgroups.



	ucinetlog1 - Notepad –	□ ×							
File Edit Format View Help									
DOHNSON'S HIERARCHICAL CLUSTERING		^							
Mathad	WTD AVERACE (average between all pairs)								
Type of Data:	Similarities								
Input dataset:	ZACHE (C:\Users\lowol 000\Documents\UCINET data\ZACHE)								
HIERARCHICAL CLUSTERING									
1111 112	222222222222222222222222222222222222222								
Level 576172328431480	2645829701931635904								
1.0000 XXX XXX XXXXXXXXX	. XXX XXX XXX XXX XXX XXX . XXX								
0.8000 XXX XXX XXXXXXXXXX									
0.3333 XXXXXXXX XXXXXXXXXXXXXXX	. XXXXXXX XXX XXX . XXXXXXXXXXXXXX								
0.2857 XXXXXXXX . XXXXXXXXXXXXXXX .	. XXXXXXX XXX XXX XXX XXXXXXXXXXXXXX								
0.2500 XXXXXXXXX . XXXXXXXXXXXXXXXXXXXXX	. XXXXXXXXXX XXX XXXXXXXXXXXXXXXXXXXXXX								
0.2222 XXXXXXXX . XXXXXXXXXXXXXXXXXXXXXX	X XXXXXXXXX XXX XXX XXXXXXXXXXXXXXXXXX								
	XX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX								
0.0727 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	X XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX								
0.0347 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	000000000000000000000000000000000000000								
Measures of cluster adequacy	Measures of cluster adequacy								
		*							

2) Girvan-Newman 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.



Now, open the attribute file, ZACHATTR, using the folder with the A

3	Open Data	a File	- • ×
Name of file to open: C:\Users\lowol_000\Documents\	UCINET data\zachattr.##h		🗸 OK
File format: Ucinet (*.##h,*.##d) VNA (*.vna) DL (*.dl) Pajek Network (*.net) Pajek Partition (*.clu) Pajek Vector (*.vec)	Type of Data: 1-Mode Network(s) Node Attribute(s) Network with Attributes 2-Mode Network	Options Ignore reflexive ties ✓ Ignore missing values ✓ Ignore zeros ✓ Ties have values > -99 but 1E36 <	X Cancel

c) Run the Girvan-Newman analysis (Analysis | Subgroups | Girvan-Newman) specifying a minimum of 2 and a maximum of 40 clusters desired. It should automatically color your nodes so that 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? The change from ngPart_2 to ngPart_3 is trivial, but from ngPart_3 to ngPart_4 may be useful as it identifies fairly good chunk of people as a cluster.



b)



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 Girvan-Newman algorithm predict the affiliation of the club members?

Circles and squares stand for different clubs the actors actually joined, whereas blues and reds are predictions by the Girvan-Newman algorithm. There are only two mismatches in

the middle (blue circles) and the divisions are visually similar. So the Girvan-Newman algorithm predicted fairly well the affiliation of the club members.



3) 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 Girvan-Newman algorithm in terms of predicting the affiliations? How could you display the Girvan-Newman results, the Factions result, AND the club attribute all at the same time?

Circles and squares stand for different clubs the actors actually joined, whereas blues and reds are predictions by the Factions algorithm. There are only two mismatches in the middle (a blue square and a red circle) and the divisions are visually similar. So the Factions algorithm predicted fairly well the affiliation of the club members. You can display all three results at the same time if you use "Nodes" tab in the control area on the right, drag down to choose "ngPart_2" for the attribute and check size. Now, shapes stand for club, colors for the Factions result, and node size for the Girvan-Newman result.



4) Putting it all together

When you ran Hierarchical Clustering, it created an output file called "Part" which has the hierarchical partitioning. This is an actor by partition matrix which is basically like an attribute file, so the actors are down the rows, and each column is one of the hierarchical clustering solutions. The first partition (labeled '1' in column 1), has the most distributed cluster solution (many clusters, few actors put together, many clusters only have one actor in them). Through the last one (the number of columns varies based on the data input), which will put all the actors into one big cluster. Typically, the second to last cluster will have two clusters, the one before that will have three, etc. (Though some data conditions may affect that.)

Because of this, you can load Part as an attribute file with the ZACKAR dataset in NetDraw. Now click on the checkbox by the word 'color' so that the nodes will be colored by whatever attribute you select in the dropdown list at the top of the control region on the Node Tab. Once the checkbox is selected, switch between several cluster solutions (with numbers 1, 2, ... n) to see how the hierarchical clustering grouped nodes. The second to last one should be the solution with two clusters. Select that option and UNCHECK the color checkbox.



Now also shape the nodes by the ngPart_2 solution. Change the rim size (Properties | Nodes | Rims | Size | Attribute based) and select the Factions Solution and set the

minimum and maximum values to 1 and 8, respectively. And, finally, SIZE the nodes based on the attribute CLUB. Do this using the Properties | Nodes | Size | Attribute based to bring up a dialog box. Select "CLUB" as the attribute and set maximum size to 18 and minimum size to 8.

The size tells us where each member actually went when the club split, and the shape, color, and rim weight tell us what the three different algorithms determined. To what extend did the three algorithms agree with each other (same shape, color, and rim weight)? How well, in general, did they do (same shape, color, rims weight, and node size)? Remember, these algorithms only know what the patterns of relationships are. Do they predict actual behavior (which club they joined) well?

The three algorithms largely agree with each other, though each offers slightly different solution. All of them predict actual behavior fairly well except some of the actors in the interface between the two clubs.

5) Cliques using UCINET and NetDraw with **KAPFTS2**

a) If you have not done so before, unpack the KAPTAIL using Data | Unpack. Be sure to eliminate any prefix so your filenames match what I list below.

	ucinetlog1 - Notepad	- '	x
File Edit Format View Help			
UNPACK			\sim
Created dataset:	KADETS1 (C:\Usens\Weekie Sung\Decuments\UCINET data\KADETS1		
Created dataset:	KAPETS2 (C:\Users\Wookje Sung\Documents\UCINET data\KAPETS2		
Created dataset:	KAPFTI1 (C:\Users\Wookje Sung\Documents\UCINET data\KAPFTI1		
Created dataset:	KAPFTI2 (C:\Users\Wookje Sung\Documents\UCINET data\KAPFTI2		
Running time: 00:00:01 seconds. Output generated: 29 May 15 23:58:28 UCINET 6.573 Copyright (c) 1992-2015 A	Malytic Technologies		~
<			> .::

b) In UCINET run Network | Subgroups | Cliques on **KAPFTS2** with a minimum size of 3. How many cliques do you get? How many actors are in this network? How useful is this?

There are 118 cliques found in this network. Since there are only 39 actors in the network, the number of cliques may not provide very useful information.

Visualize KAPFTS2 in NetDraw. Does this help us identify clique

structures?

There are quite many nodes and even more edges (ties) between them. So it is not easy to identify clique structures visually now.

d) 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 between actors in cliques in this network? (Redraw the network when you think you have filtered the data enough.) What does this visualization mean?

The ties between actors mean that they have overlap in cliques. The visualization of the network shows that actors with more than 10 shared cliques and includes 15 out of 39 actors, indicating that there is significant overlap between cliques in this network.

e) Now set the filter back down to 0 and open **CliqueSets** in Netdraw and redraw the picture (lightning bolt). 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? Are there actors who seem embedded in a lot of different cliques? What does that mean about those actors?

This picture shows that there are some actors in the core where the actors are embedded in dense relationships and some others are in the periphery. Those actors (red circles) who have outgoing ties to many different cliques (blue squares) seem embedded in a lot of different cliques, which means they are the "core actors" (e.g. Mukubwa).

[OPTIONAL] Try running Analysis | Centrality on the data inside NetDraw specifying that the data are directed. Then, size the nodes based on OutDegree using the attribute it created and loaded into the nodes tab.

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

a) Open **PV504** in NetDraw. Because it is very large, NetDraw does not optimize the layout automatically when opening it. To make the diagram more readable, turn off labels (using the script L button on the icon bar) and arrowheads (the data are symmetric), and then redraw the network. This may take some time, but let it finish. You should begin to see some structure in the network as it draws it.

b) These are valued data about the number of days individuals worked together on projects. Let's increase the filtering to be greater than 3 by clicking on the "+" button toward the bottom of the Rels tab in the control region three times. Now redraw the network by clicking on the lightning bolt. Much more structure should be visible.

c) 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 below the values to step through the k-cores from 0 to 10. What does this tell you about the network?

This visualization tells me that there are multiple distinct large regions within which cohesive subgroups may be found and identifies fault lines across which cohesive subgroups do not span, which is useful in visualizing potential subgroup regions in a large dataset like the PV504 data.

d) Since all nodes of a higher "coreness" are automatically members of the lower cores, we'd like to step down from the highest coreness, to the lowest, but cumulatively. To do this, press the "a" button below the values in the control region to select all the check boxes, then check the "i" button to "inverse" the selection (i.e., uncheck everything that is checked and check everything that is unchecked). This should leave no boxes checked and a blank screen. Now check the box next to the highest value (it should be 10) and look at the graph. Now ALSO check the box next to the second highest value. Repeat until you have checked all boxes. What could you see from stepping "down" the k-cores that was not obvious stepping "up" them?

It is clearer that as the k value increases, the core-ness of the subgroup also increases and that k-cores at a higher number are more cohesive among themselves than at lower numbers,

Hypothesis Testing (Answer sheet)

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.

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.

KRACK-HIGH-TEC & HIGH-TEC-ATTRIBUTES

KRACK-HIGH-TEC is another stacked dataset, containing three dichotomous relations (REPORTS_TO, ADVICE, FRIENDSHIP). HIGH-TEC-ATTRIBUTES contains several attributes about the nodes in KRACK-HIGH-TEC, including Age, Level (CEO, Manager, Staff), Tenure, and Department.

WIRING

This is a stacked dataset that includes many different files. This is a dichotomous adjacency matrix of 14 employees of the bank wiring room of Western Electric used in the famous Hawthorne Studies. Ties are symmetric and represent participation in games during work breaks. RDGAM records people playing games together, RDCON records conflict between people, RDPOS is positive interactions, RDCON is negative interactions.

EXERCISES

Testing dyadic hypothesis

Run Data | Unpack on ZACKAR (if you have not yet), which will create ZACHE and ZACHC. ZACHE has dichotomous data about the ties and ZACHC has valued data (the strength of ties).

Run Tools | Similarities and use the cross-product measure to compute similarities among the rows of ZACHE. (The cross product is a very powerful and common matrix operation that, in this case, will count how many friends each pair of actors have in common.) Call the output CommonFriends.

Similarity Matrix

										1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	з	з	з	з	3
	- 1	2	з	4	5	6	7	8	9	0	1	2	з	4	5	6	7	8	9	0	1	2	з	4	5	6	7	8	9	0	1	2	з	4
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1	16	7	5	5	2	2	2	з	1	1	2	0	1	з	0	0	2	1	0	1	0	1	0	0	1	1	0	1	2	0	2	0	з	4
2	7	9	4	4	1	1	1	з	з	1	1	1	2	з	0	0	0	1	0	1	0	1	0	0	0	0	0	1	1	0	0	1	2	3
- 3	- 5	40	10	4	1	1	1	з	2	0	1	1	2	з	1	1	0	2	1	2	1	2	1	2	1	0	0	0	0	1	з	з	1	6
- 4	- 5	4	4	6	1	1	1	з	2	1	1	1	1	з	0	0	0	2	0	2	0	2	0	0	0	0	0	1	1	0	1	1	1	1
- 5	2	1	1	1	з	з	1	1	1	0	1	1	1	1	0	0	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
6	2	1	1	1	з	4	2	1	1	0	1	1	1	1	0	0	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
7	2	1	1	1	1	2	4	1	1	0	з	1	1	1	0	0	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
8	- 3	з	з	з	1	1	1	4	2	1	1	1	2	4	0	0	0	2	0	2	0	2	0	0	0	0	0	1	1	0	1	1	1	0
9	- 1	з	2	2	1	1	1	2	5	2	1	1	1	з	2	2	0	1	2	2	2	1	2	2	0	0	1	2	2	2	2	з	з	2
10	- 1	1	0	1	0	0	0	1	2	2	0	0	0	2	1	1	0	0	1	1	1	0	1	1	0	0	1	2	2	1	1	1	2	0
11	2	1	1	1	1	1	з	1	1	0	з	1	1	1	0	0	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
12	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
13	- 1	2	2	1	1	1	1	2	1	0	1	1	2	2	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
14	- 3	з	з	з	1	1	1	4	з	2	1	1	2	5	1	1	0	2	1	з	1	2	1	1	0	0	1	2	2	1	2	2	2	0
15	0	0	1	0	0	0	0	0	2	1	0	0	0	1	2	2	0	0	2	1	2	0	2	2	0	0	1	1	1	2	2	2	1	1
16	0	0	1	0	0	0	0	0	2	1	0	0	0	1	2	2	0	0	2	1	2	0	2	2	0	0	1	1	1	2	2	2	1	1
17	2	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	1	2	2	1	1	1	2	1	0	1	1	1	2	0	0	0	2	0	2	0	2	0	0	0	0	0	0	0	0	1	1	0	0
19	0	0	1	0	0	0	0	0	2	1	0	0	0	1	2	2	0	0	2	1	2	0	2	2	0	0	1	1	1	2	2	2	1	1
20	1	1	2	2	1	1	1	2	2	1	1	1	1	з	1	1	0	2	1	з	1	2	1	1	0	0	1	1	1	1	2	2	1	0
21	0	0	1	0	0	0	0	0	2	1	0	0	0	1	2	2	0	0	2	1	2	0	2	2	0	0	1	1	1	2	2	2	1	1
22	1	1	2	2	1	1	1	2	1	0	1	1	1	2	0	0	0	2	0	2	0	2	0	0	0	0	0	0	0	0	1	1	0	0
23	0	0	1	0	0	0	0	0	2	1	0	0	0	1	2	2	0	0	2	1	2	0	2	2	0	0	1	1	1	2	2	2	1	1
24	0	0	2	0	0	0	0	0	2	1	0	0	0	1	2	2	0	0	2	1	2	0	2	5	2	0	2	1	1	2	2	з	2	3
25	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	з	1	0	0	1	0	0	1	1	2
26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	з	0	2	1	1	0	1	2	2
27	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0	1	2	0	0	2	1	1	1	1	1	2	1
28	1	1	0	1	0	0	0	1	2	2	0	0	0	2	1	1	0	0	1	1	1	0	1	1	0	2	1	4	2	2	1	2	з	1
29	2	1	0	1	0	0	0	1	2	2	0	0	0	2	1	1	0	0	1	1	1	0	1	1	1	1	1	2	з	1	1	1	з	1
30	0	0	1	0	0	0	0	0	2	1	0	0	0	1	2	2	0	0	2	1	2	0	2	2	0	1	1	2	1	4	2	2	2	3
31	2	0	3	1	0	0	0	1	2	1	0	0	0	2	2	2	0	1	2	2	2	1	2	2	0	0	1	1	1	2	4	2	2	2
32	0	1	3	1	1	1	1	1	3	1	1	1	1	2	2	2	0	1	2	2	2	1	2	3	1	1	1	2	1	2	2	6	1	2
33	3	2	1	1	0	0	0	1	3	2	0	0	0	2	1	1	0	0	1	1	1	0	1	2	1	2	2	з	3	2	2	11	21	0
34	- 4	3	6	1	0	0	0	0	2	0	0	0	0	0	1	1	0	0	1	0	1	0	1	3	2	2	1	1	1	3	2	21	.01	.7

Note that the number along the diagonal is each actor's outdegree centrality.

Go to Tools | Testing Hypotheses | Dyadic (QAP) | QAP Correlation and browse to include both ZACHC and CommonFriends to be correlated and click okay. What do the results mean?

QAP results for CommonFriends * ZACHC (5000 permutations) Pearson Correlation QAP Correlation QAP Correlations ZACHC Commo 1 ZACHC 1.0000 QAP P-Values ZACHC Commo 2 commonFriends ZACHC Commo 2 commonFriends ZACHC Commo 2 commonFriends 2 ACHC Commo 2 commonFriends 3 A 4 5 Maximum Prop >= 0 Prop <= 0 Prop

QAP statistics saved as datafile QAP Correlation Results

These results tell us that the ZACHC and the commonfriends matrices have a correlation of 0.37. The second matrix tells us that the significance of this correlation is at P < 0.001. In other words there appears to be a significant positive relationship between having friends in common with somebody and having a strong direct tie to that person.

Congratulations, you have just statistically demonstrated the first part of Granovetter's famous "strength of weak ties" theory, which states that I have stronger ties (ZACHC) with those people with whom I share more friends in common (CommonFriends).

Testing multivariate dyadic hypotheses

Unpack the WIRING dataset if you have not done so yet.

Go to Tools | Testing Hypotheses | Dyadic (QAP) | MR-QAP Regression | Double Dekker Semi-Partialling MRQAP. Put RDCON (conflict between members about whether the windows should be open or shut) in as the dependent variable. Put in RDPOS (positive relationships), RDNEG (negative relationships), and RDGAM (playing games together) in as independent variables. Before running it, what do you think would most significantly predict conflict? After running it, are your results what you expected? How would you explain the results? MODEL FIT

	R-Square	Adj R-Sqr	P-Value	obs	Perms
Mode1	0.061	0.045	0.020	182.000	2000.000

REGRESSION COEFFICIENTS

	Un-Stdized	Stdized Coef	P-value	As Large	As Small	Std Err
RDPOS	-0.05286	-0.04551	0.37731	0.62319	0.37731	0.15335
RDNEG	0.00595	0.00595	0.49525	0.49525	0.50525	0.13930
RDGAM	0.23636	0.26840	0.03048	0.03048	0.97001	0.11907
Intercept	0.14237	0.00000	0.00000	0.00000	0.00000	0.00000

Both RDPOS and RDNEG fail to reach a .05 level of significance. This suggests that neither sharing a negative relationship (RDNEG) nor a positive relationship (RDPOS) with another person is a significant predictor of having conflict (RDCON) with him/her. RDGAM, on the other hand, is a significant predictor ($\beta = .268$, P = .030) of RDCON. Thus, playing games with another person is positively associated with having conflict with him/her. There are many potential reasons for this relationship. Perhaps the individuals in this study were very competitive and fought with others when they lost a game to them....

Record the standardized coefficient and significance for any significant predictor, and run the same procedure two more times (still using the default value of 2000 for the number of permutations) and record the same results. Now, run the same procedure three more times setting the number of random permutations set to 100000. Record the same results. How did the parameter affect the results? Why?

You should have found that the standardized coefficient did not change at all during each of your iterations. The P value for this variable (RDGAM) probably did change though. You most likely found that the value changed by a factor of a few thousandths when you used 2,000 permutations. You may have also seen slight changes with 100,000 permutations, but they were probably smaller changes. This is due to the fact that your results become more accurate and more convergent as you increase the number of permutations used.

Testing monadic hypotheses.

You should have already unpacked the KRACK-HIGH-TEC dataset, but if not, do so now. You will get three datasets (REPORTS_TO, ADVICE, FRIENDSHIP). We are going to use the ADVICE dataset. Run Network | Centrality | Degree on this dataset, using the directed version, telling it NOT to treat the data as symmetric. By default, it will name the output FreemanDegree. We are particularly interested in who is sought after for advice, which is captured by InDegree centrality. So, we are going to pull out just that column from the results, but using Data | Filter/Extract | Submatrix. Specify FreemanDegree as your input dataset and that we want to "Keep" "ALL" rows. Then click on the L to the right of the box for "Which Columns" and select the column labeled "InDegree" and call your output ADVISING. This is a measure of how many people said they sought advice from each person.

Display (D) the HIGH-TEC-ATTRIBUTES dataset to determine which columns the AGE and TENURE attributes are in.

Now, it is common wisdom that people look to the "senior" people for advice, but is unclear in an organizational context whether senior is "older" or "longer tenured". You will test if either of these is supported by the data. Run Tools | Testing Hypotheses | Node-Level | Regression specifying ADVISING for your dependent dataset with the appropriate column and HIGH-TEC-ATTRIBUTES and the appropriate columns for your independent dataset (i.e., the columns for Age and Tenure separated by a space), and set the number of permutations to 10000. Which meaning of "senior" do the data support? Number of observations: 21

CORRELATION MATRIX

		1	2	3
		AGE	TENURE	InDegr
1	AGE	1.000	0.489	-0.041
2	TENURE	0.489	1.000	0.542
3	InDegree	-0.041	0.542	1.000

Determinant = 0.76051355

NOTE: All probabilities based on randomization tests.

MODEL FIT

Ad	justed	F Value	One-Tailed
R-square R	-square		Probability
0.417	0.325	6.435	0.031

REGRESSION COEFFICIENTS

Independent	Un-stdized	St'dized	Proportion	Proportion	Proportion
	Coefficient	Coefficient	As Large	As Small	As Extreme
Intercept	11.458937	0.000000	1.000	0.000	1.000
AGE	-0.171299	-0.402398	0.933	0.067	0.146
TENURE	0.373887	0.738914	0.004	0.996	0.006

Looking down the "Proportion as Extreme" column, we see that there is a significant positive relationship between tenure and advising. The relationship between age and advising is not significant (interestingly, the coefficient is negative). Thus, we can conclude that having longer tenure in this organization is related to being sought after for advice.

Why did we use the Regression option of Node-Level instead of T-Test or Anova? When would we use those?

We used regression because our independent variables are continuous. We would use a T-Test for looking at the difference between the means of two categories/groups. ANOVA would be used when you have more than two categories/groups.

Testing Mixed-Dyadic Monadic hypotheses

Since it is only fitting that we end where we started, we shall use the campnet data for these final exercises.

You will run Tools | Testing Hypotheses | Mixed Dyadic/Nodal | Categorical attributes | Anova Density twice. For both, specify CAMPNET as the network matrix, and the gender column of the CAMPATTR matrix as the Actor Attribute. For the first run, choose "Constant Homophily" for your model, and for the second, choose "Variable Homophily". Interpret both sets of results. What do they mean? Is there homophily? Which gender tends to be more homophilous?

Constant Homophily:

Density Table

		1 1	2 2
1	1	0.357	0.050
2	2	0.063	0.278

Density table saved as dataset densitytable

Expected values saved as dataset anovadensity_expectedvalues

Number of permutations performed: 5000

MODEL FIT

R-square	Adj R-Sqr	Probability	# of Obs
0.109	0.109	0.0004	306

REGRESSION COEFFICIENTS

Independent	Un-stdized Coefficient	Stdized Coefficient	Significance	Proportion As Large	Proportion As Small
Intercept	0.056250	0.000000	0.9998	0.9998	0.0004
In-group	0.251969	0.330131	0.0004	0.0004	0.9998

With a beta of .33 and a significance level of .0008, we can conclude from these results that there is gender homophily occurring in this network. In other words, men tend to associate with other men more than they do with women, and women tend to associate with other women more than with men.

Variable Homophily:

Density Table

		1 1	2
1 2	1 2	0.357	0.050

Density table saved as dataset densitytable

Expected values saved as dataset anovadensity_expectedvalues

Number of permutations performed: 5000

MODEL FIT

R-square	Adj R-Sqr	Probability	# of Obs
0.114	0.111	0.0000	306

REGRESSION COEFFICIENTS

Independent	Un-stdized Coefficient	Stdized Coefficient	Significance	Proportion As Large	Proportion As Small
Intercept	0.056250	0.000000	0.9998	0.9998	0.0004
Group 1	0.300893	0.305196	0.0000	0.0000	0.9998
Group 2	0.221528	0.264776	0.0004	0.0004	0.9994

Variable homophily allows us to see the within-group levels of homophily. We again find evidence of homophily in both groups. Group 1 (women) has a slightly higher beta coefficient than Group 2 (men). Thus, we can say that there is a slightly greater tendency for homophily among women in this network.

Using QAP for Mixed Monadic/Dyadic Hypotheses testing.

Using Data | Attribute to matrix, create a matrix of exact matches among the actors in Campnet based on gender.

													-								
Pros	kimity/Ne	two)T)	C I	nat		i.X	9	enx	era	at e	3đ.	f	COL	n (285	np:	st	C.T.	Column	11
											1	1	1	1	1	1	1	÷	1		
		1	2	з	4	5	6	7	R	9	õ	ĩ	2	â	4	5	6	÷	Â		
		Ĥ	ñ	č	p	ě	ă	÷.	Ă	Ň	ň	÷.	ñ	ă.	ñ.	ă	š	ĥ	n		
			1	č	2	2	1	2	2		1	2	1	ž	-	2	č	1	2		
1	HOLLY	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0		
2	BRAZEY	1	ī	ī	1	ī	ī	1	ī	ō	ō	ō	ō	ō	ō	ō	ō	ō	ō		
3	CAROL	1	1	1	1	ī	1	1	1	ö	0	0	ō	0	ō	0	0	ō	0		
4	PAM	1	1	1	1	1	1	1	1	0	0	0	0	0	Ö	0	0	0	0		
5	PAT	1	1	1	1	1	1	1	1	ō	0	0	0	0	ō	0	0	0	0		
6	JENNIE	1	1	1	1	1	1	1	1	ō	0	0	0	0	ō	0	0	0	0		
7	PAULINE	1	1	1	1	1	1	1	1	ō	0	0	ō	0	ō	0	0	ō	0		
8	ANN	1	1	1	1	ī	1	1	ī	ō	ö	ō	ō	Ö	ō	ō	ö	ō	Ö		
9	MICHAEL	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1		
10	BILL	0	0	0	ō	0	0	ō	0	1	1	1	1	1	1	1	1	1	1		
11	LEE	0	ō	0	ö	0	0	ō	0	1	1	1	1	1	1	1	1	1	1		
12	DON	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1		
13	JOHN	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1		
14	HARRY	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1		
15	GERY	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1		
16	STEVE	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1		
17	BERT	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1		
18	RUSS	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1		

Note that we get the blocks of 1s and 0s in this matrix because the actors are ordered by gender. This will usually not happen unless your matrix is ordered based on your attribute of interest as in the case above.

View this new matrix (named CAMPATTR-MAT by default) in Netdraw. What does the diagram show?

As you see, we get two cliques—one of men and one of women. This is because the similarity matrix recorded a 1 if two actors are of the same gender, and a 0 if they were of a different gender.

Use Tools | Testing Hypotheses | QAP Regression to regress the Campnet network on this new matrix of gender similarity, CAMPATTR-MAT. What do the results show?

# of perm Diagonal Random se Dependent Partition Predicted Model fit Model coe Independe	10000 NO 100 CAMPNET CAMPNET-mrpred (C:\Users\Paulo Serodio\Documents\UCI CAMPNET-mrfit (C:\Users\Paulo Serodio\Documents\UCI CAMPNET-mrcoef (C:\Users\Paulo Serodio\Documents\UCI C:\Users\Paulo Serodio\Documents\UCINET data\CAMPAT								
MODEL FIT									
	R-Squ	are Adj	R-Sqr	P-Valu	ie	obs	Pern	IS	
Mod	el 0.	109	0.106	0.00	0	306.000	10000.00	00	
REGRESSIO	N COEFFICIEN	тѕ							
		Un-Stdized	Stdized	Coef	P-val	ue As	Large	As Small	Std Err
CAM	PATTR-mat	0.25197	0.	33013	0.0004	40 0.	00040	0.99970	0.05809

Here we are using a matrix which shows connections between nodes of the same gender only to estimate the connections we observe in campnet. Thus, we are trying to see how well perfect gender homophily explains the variation in terms of connections that we get in campnet. We get basically the same results here as we got when we used the ANOVA Density approach. We again find evidence of gender homophily as the coefficient has a positive sign and is highly significant.

0.00000

0.00000

0.00000

0.00000

0.00000

0.05625

Intercept

Do you prefer this approach of the ANOVA Density Tables? When might you use each of these separate techniques? What research question might involve using Moran's I (or Geary's C) instead of the ANOVA Density Tables? In that case, how would you use QAP to test for Autocorrelation?

The approach above is good when you want to look at the effects of multiple independent variables (i.e., an MRQAP). The "Variable Homophily" ANOVA Density Table is good when you want to compare the differences between groups.

You would use Moran's I or Geary's C when your independent variable is continuous instead of categorical. For example, we would want to use Moran's I if we wanted to examine age homophily, because (presumably) our age variable would be continuous. We just used QAP to test for autocorrelation when we examined the effects of gender homophily!