Slide 1

Testing Network Hypotheses

Thursday Afternoon

Whether we are using Social Network Analysis as part of a consulting project or in support of academic research, it is important to know if the measures and relationships we see in the data represent a significant phenomenon, or are simply an artifact of the data themselves. But standard methods of calculating significance are inappropriate for network data. In this section we deal with testing hypotheses based on network data.

Objectives:

After completing this module you will be able to:

- Understand and explain why standard statistical methods are inappropriate for network data
- Conceptually explain how the permutation tests for significance work
- Explain the difference between monadic, dyadic, and mixed monadic/dyadic hypotheses
- Explain the concepts of variable and constant homophily
- Run the appropriate hypothesis testing techniques in UCINET

Slide 2

Units of Analysis

- Dyadic (tie-level)
  - The raw data
  - Cases are pairs of actors
  - Variables are attributes of the relationship among pairs (e.g., strength of friendship; whether give advice to; hates)
  - Each variable is an actor-by-actor matrix of values, one for each pair
- Monadic (actor-level)
  - Cases are actors
  - Variables are aggregations that count number of ties a node has, or sum of distances to others (e.g., centrality)
  - Each variable is a vector of values, one for each actor
- Network (group-level)
  - Cases are whole groups of actors along with ties among them
  - Variables aggregations that count such things as number of ties in the network, average distance, extent of centralization, average centrality
  - Each variable has one value per network

Slide 3

Types of Hypotheses

- Dyadic (multiplexity)
  - Friendship ties lead to business ties
  - Social ties between leads to less formal contractual ties (embeddedness)
- Monadic
  - Actors with more ties are more successful (social capital)
- Network
  - Teams with greater density of communication ties perform better (group social capital)
- Mixed Dyadic-Monadic (autocorrelation)
  - People prefer to make friends (dyad level) with people of the same gender (actor level) (homophily)
  - Friends influence each other's opinions

Slide 4

Statistical Issues

- Samples non-random
- Often work with populations
- Observations not independent
- Distributions unknown

Slide 5

Solutions

- Non-independence
  - Model the non-independence explicitly as in HLM
    - Assumes you know all sources of dependence
  - Permutation tests
- Non-random samples/populations
  - Permutation tests

Slide 6

Logic of Permutation Test

- Compute test statistic
  - e.g., correlation or difference in means
  - Correlation between centrality and salary is 0.384 or difference in mean centrality between the boys and the girls is 4.95.
  - Ask what are the chances of getting such a large correlation or such a large difference in means if the variables are actually completely independent?
- Wait! If the variables are independent, why would the correlation or difference in means be anything but zero?
  - Sampling
  - "Combinatorial chance": if you flip coin 10 times, you expect 5 heads and 5 tails, but what you actually get could be quite different

Slide 7

Logic of Permutation Test

- So to evaluate an observed correlation between two variables of 0.384, we want to
  - correlate thousands of variables similar to the ones we are testing that we <u>know</u> are truly independent of each other, and
  - see how often these independent variables are correlated at a level as large as 0.384
    - The proportion of random correlations as large observed value is the p-value of the test
- How to obtain thousands of independent variables whose values are assigned independently of each other?
  - Fill them with random values
    - But need to match distribution of values
  - Permute values of one with respect to the other

Slide 8

Outline of Permutation Test

- Get observed test statistic
- Construct a distribution of test statistics under null hypothesis
  – Thousands of permutations of actual data
- Count proportion of statistics on permuted data that are as large as the observed
  – This is the p-value of the test

Slide 9

Monadic Hypotheses

	Centrality	Grades
bill	10	2.1
maria	20	9.5
mikko	40	7.3
esteban	30	4.1
jean	70	8.1
ulrik	50	8.1
joao	40	6.6
myeong-gu	50	3.3
akiro	60	9.1
chelsea	10	7.2

- This, effectively, is basic social science research
  - However, centrality measures in most network based research are non-independent, so OLS is not appropriate
  - Ego-Net based research, on the other hand, would arguably yield independent measures

Slide 10

Dyadic Hypotheses

- Hubert / Mantel QAP test
  - All variables are actor-by-actor matrices
  - We use one relation (dyadic variable) to predict another
  - Test statistic is $\gamma = \sum_i \sum_j x_{ij} y_{ij}$
  - Significance is $prop(\gamma \geq \gamma^P)$, $\gamma^P = \sum_i \sum_j x_{ij} y_{p(i)p(j)}$
- QAP correlation & MR-QAP multiple regression

Friendship (X)

	Jim	Jill	Jen	Joe
Jim	-	1	0	1
Jill	1	-	1	0
Jen	0	1	-	1
Joe	1	0	1	-

Proximity (Y)

	Jim	Jill	Jen	Joe
Jim	-	3	9	2
Jill	3	-	1	15
Jen	9	1	-	3
Joe	2	15	3	-

Slide 11

Dyadic/Monadic Hypotheses

- One dyadic (relational) variable, one monadic (actor attribute) variable
  - Technically known as autocorrelation
  - But, unlike in OLS, we don't autocorrelation is bad
- Diffusion
  - adjacency leads to similarity in actor attribute
    - Spread of information; diseases
- Selection
  - similarity leads to adjacency
    - Homophily: birds of feather flocking together
    - Heterophily: disassortative mating
- Tom Snijders' SIENA model

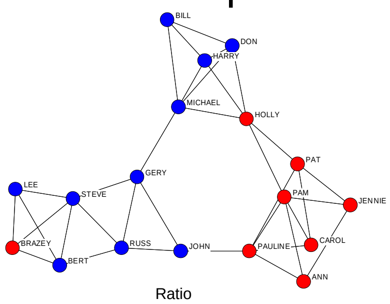
Slide 12

Categorical Autocorrelation

- Nodes partitioned into mutually exclusive categories, e.g., gender or race
- We expect more ties within group than between
  - Boys interact w/ boys, girls w/ girls
  - Cohesive subgroups may be result of autocorrelation
- Count up number of ties between all ordered pairs of groups:
  - boys to boys, boys to girls, girls to boys, girls to girls
- Compare with number expected given independence of interaction and node characteristic
  - i.e., if people choose partners without regard for gender

Slide 13

Campnet Example



Observed

	Female	Male
Female	12	7
Male	7	16

Ratio

	Female	Male
Female	1.87	0.38
Male	0.38	1.55

Expected

	Female	Male
Female	6.4	18.3
Male	18.3	10.3

Slide 14

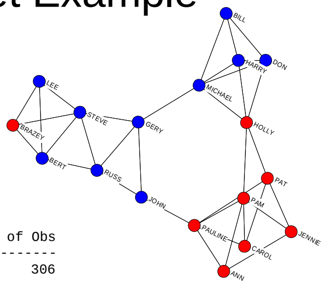
Campnet Example

Density Table

	1 Femal	2 Male
1 Fem	0.429	0.087
2 Mal	0.087	0.356

MODEL FIT

R-square	Adj R-Sqr	Probability	# of Obs
0.127	0.124	0.001	306



REGRESSION COEFFICIENTS

Independent	Un-stdized Coefficient	Stdized Coefficient	Significance	Proportion As Large	Proportion As Small
Intercept	0.087500	0.000000	1.000	1.000	0.001
Group 1	0.341071	0.313982	0.001	0.001	0.999
Group 2	0.268056	0.290782	0.001	0.001	0.999

Slide 15

Continuous Autocorrelation

- Each node has score on continuous variable, such as age or rank
- Positive autocorrelation exists when nodes of similar age tend to be adjacent
  - Friendships tend to be homophilous wrt age
  - Mentoring tends to be heterophilous wrt age
- Can measure similarity via difference or product

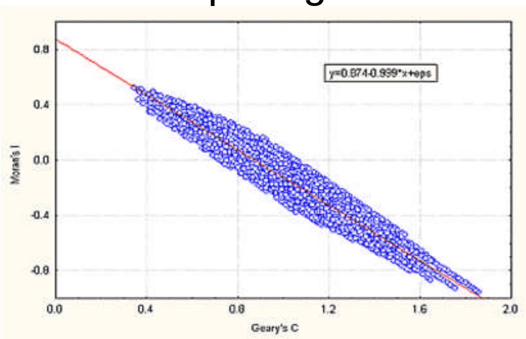
Slide 16

Autocorrelation Measures

- Geary's C
  - Also called Geary's [Contiguity] Ratio
  - Most sensitive to local autocorrelation
- Moran's I
  - Measures autocorrelation not only on variable values or location (adjacency), but rather on both simultaneously
  - More sensitive to global autocorrelatoin
- I is about covariation of pairs, C is about variation in variable values
- Really the differences are probably immaterial

Slide 17

Comparing C & I



This figure suggests a linear relation between Moran's *I* and Geary's *C*, and either statistic will essentially capture the same aspects of spatial autocorrelation.

http://www.lpc.uottawa.ca/publications/moransi/moran.htm

Slide 18

Geary's C

- Let $w_{ij} > 0$ indicate adjacency of nodes i and j, and X_i indicate the score of node i on attribute X (e.g., age)

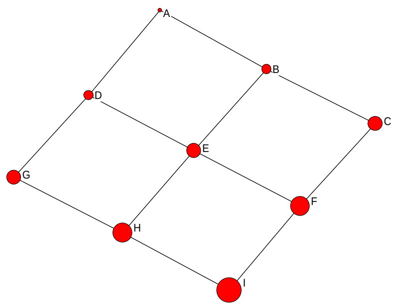
$$C = (n-1)\frac{\sum_i \sum_j w_{ij}(x_i - x_j)^2}{2\sum_{i,j} w_{ij} \sum_i (x_i - \overline{x})^2}$$

- Range of values: 0 <= C <= 2
  - C=1 indicates independence;
  - C > 1 indicates negative autocorrelation;
  - C < 1 indicates positive autocorrelation (homophily)

Slide 19

Positive Autocorrelation

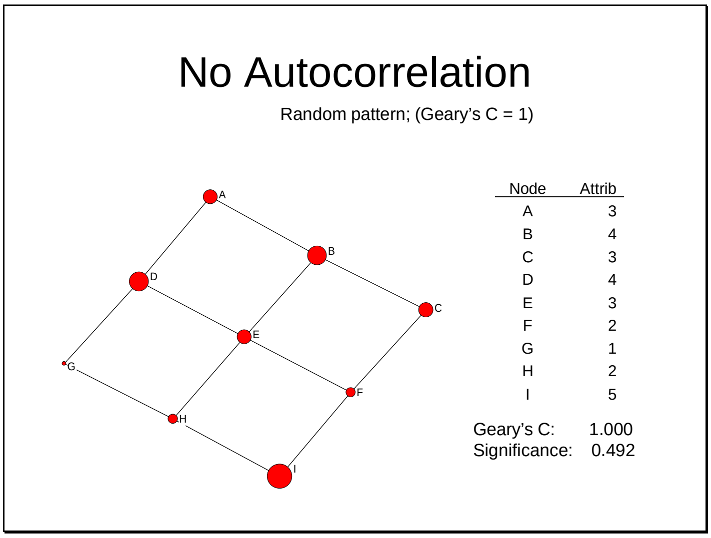
(Similars adjacent; Geary's C < 1)



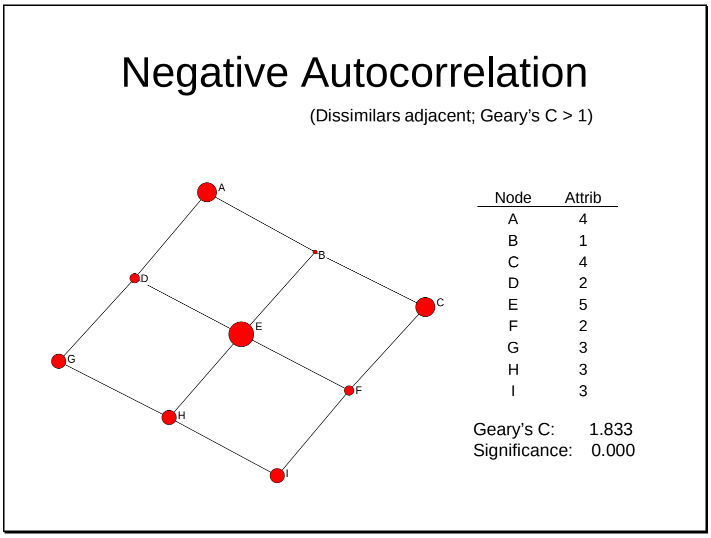
Node	Attrib
A	1
B	2
C	3
D	2
E	3
F	4
G	3
H	4
I	5

Geary's C: 0.333
Significance: 0.000

Slide 20


No Autocorrelation
Random pattern; (Geary's C = 1)
A
B
C
D
E
F
G
H
I
Node Attrib
A 3
B 4
C 3
D 4
E 3
F 2
G 1
H 2
I 5
Geary's C: 1.000
Significance: 0.492

Slide 21



Negative Autocorrelation
(Dissimilars adjacent; Geary's C > 1)
A
B
C
D
E
F
G
H
I
Node Attrib
A 4
B 1
C 4
D 2
E 5
F 2
G 3
H 3
I 3
Geary's C: 1.833
Significance: 0.000

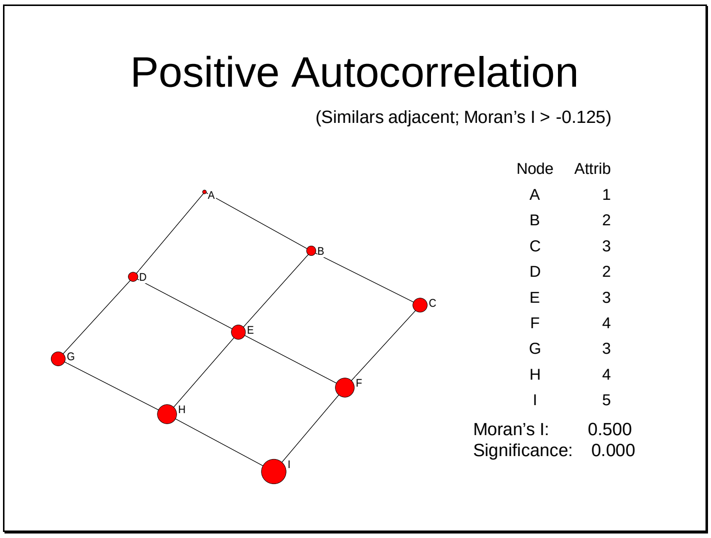
Slide 22

Moran's I

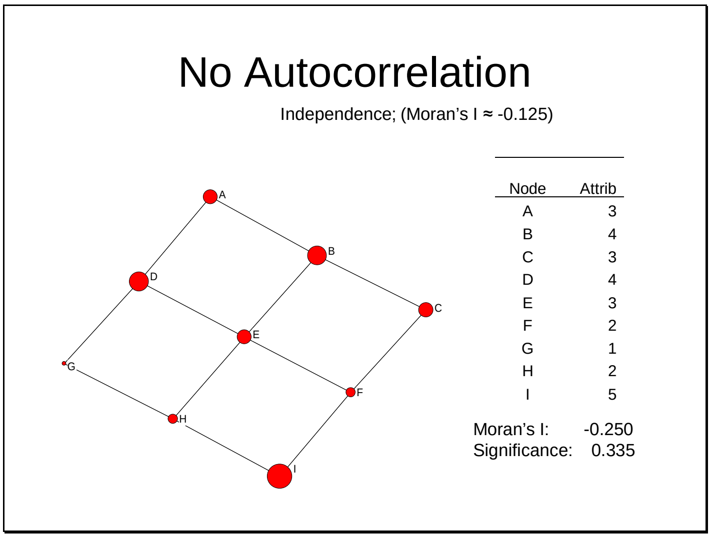
- Ranges between -1 and +1
- Expected value under independence is -1/(n-1)
- I → +1 when positive autocorrelation
- I → -1 when negative autocorrelation

$$I = n\frac{\sum_{i,j} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i,j} w_{ij}\sum_{i}(x_i - \bar{x})^2}$$

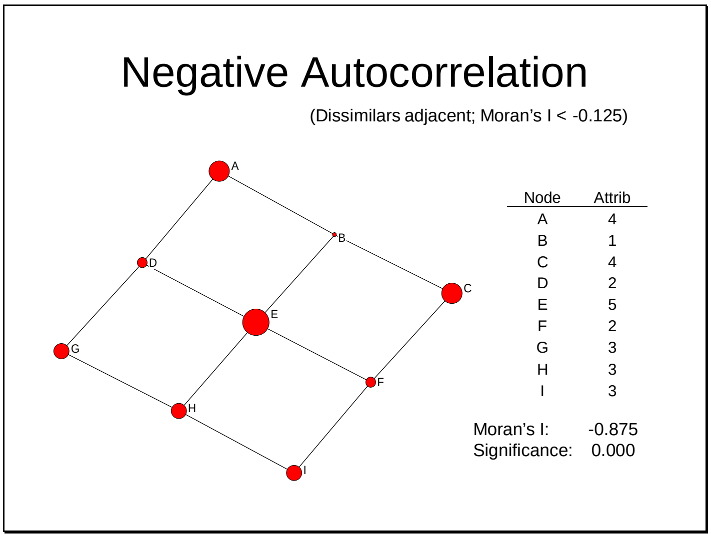
Slide 23


Positive Autocorrelation
(Similars adjacent; Moran's I > -0.125)
A
B
C
D
E
F
G
H
I
Node Attrib
A 1
B 2
C 3
D 2
E 3
F 4
G 3
H 4
I 5
Moran's I: 0.500
Significance: 0.000

Slide 24


No Autocorrelation
Independence; (Moran's I ≈ -0.125)
A
B
C
D
E
F
G
H
I
Node Attrib
A 3
B 4
C 3
D 4
E 3
F 2
G 1
H 2
I 5
Moran's I: -0.250
Significance: 0.335

Slide 25


Negative Autocorrelation
(Dissimilars adjacent; Moran's I < -0.125)
A
B
C
D
E
F
G
H
I
Node Attrib
A 4
B 1
C 4
D 2
E 5
F 2
G 3
H 3
I 3
Moran's I: -0.875
Significance: 0.000

Slide 26

Interpreting Autocorrelation

- With Moran's *I*
  - A value near +1.0 indicates clustering (adjacency tends to accompany similarity along a dimension)
  - A value near -1.0 indicates dispersion (adjacency tends to accompany dissimilarity along a dimension)
  - a value near 0 indicates random distribution
- For Geary's C
  - just substitute 0, 2, and 1 for 1, -1, and 0 above

Slide 27

Another Approach

- Convert the attribute vector into a matrix
  - Use Data | Attribute to Matrix in UCINET
- QAP this new matrix against the adjacency matrix
  - Significances will be the same because it uses same underlying permutation method
  - Values will follow same pattern (but not same values) as Moran's I

Slide 28

Using QAP for Autocorrelation

	Gender
HOLLY	1
BRAZEY	1
CAROL	1
PAM	1
PAT	1
JENNIE	1
PAULINE	1
ANN	1
MICHAEL	2
BILL	2
LEE	2
DON	2
JOHN	2
HARRY	2
GERY	2
STEVE	2
BERT	2
RUSS	2

	HOL	BRA	CAR	PAM	PAT	JEN	PAU	ANN	MIC	BIL	LEE	DON	JOH	HAR	GER	STE	BER	RUS
HOLLY	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
BRAZEY	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
CAROL	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
PAM	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
PAT	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
JENNIE	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
PAULINE	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
ANN	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
MICHAEL	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
BILL	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
LEE	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
DON	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
JOHN	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
HARRY	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
GERY	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
STEVE	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
BERT	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
RUSS	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

This matrix was constructed based on "exact match"
but you can use different transformations

Slide 29

Comparing QAP & Moran's I

Moran's I Output

A value of -0.059 indicates perfect independence.

Autocorrelation:	0.667
Significance:	0.001

QAP Output

Independent	Un-stdized Coefficient	Stdized Coefficient	Significance
Intercept	0.056250	0.000000	0.999
CAMPATTR2-MAT	0.251969	0.330131	0.001

Hypothesis Testing 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.

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.

1) Testing dyadic hypothesis
    a. 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, a count).

    b. Run Tools | Similarities and use the cross-product measure to compute similarities 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 FOF (Friends of Friends).

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

    d. 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 (FOF).

2) Testing multivariate dyadic hypotheses
    a. Unpack the WIRING dataset if you have not done so yet.

    b. Go to Tools | Testing Hypotheses | Dyadic (QAP) | QAP Regression | Full Partialling. 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?

    c. 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 50000. Record the same results. How did the parameter affect the results? Why? When running with 2000 permutations, why did one number change but the other remain constant?

d. Now run Tools | Testing Hypotheses | Dyadic (QAP) | QAP Regression | Double Dekker (MRQAP) still using the same independent and dependent variables, and setting your permutations to 50000. Compare these results with your previous run. Compare the time required to run it.

3) Testing monadic hypotheses.
   a. 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, and calling your output ADVISING. Record which column has InDegree centrality. This is a measure of how many people said they sought advice from each person.

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

   c. 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 than" or "longer tenured than". 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 and the appropriate column, and HIGH-TEC-ATTRIBUTES for your independent dataset and the appropriate columns, and set the number of permutations to 10000. Which meaning of "senior" does the data support?

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

4) Testing Mixed-Dyadic Monadic hypotheses

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

b. 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 CAMPATTR2 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? Who tends to be more homophilous?

5) Using QAP for Mixed Monadic/Dyadic Hypotheses testing.

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

b. View this new matrix (named CAMPATTR2-MAT by default) in Netdraw. What does the diagram show?

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

d. Do you prefer this approach, ANOVA Density Tables, Moran's I, or Geary's C? When might you use each of these separate techniques?