Quadratic Assignment Procedures (QAP)

Testing hypotheses among networks

Manuel S. González Canché & Cecilia Rios Aguilar
msgc@email.arizona.edu
Cecilia.Rios-Aguilar@cgu.edu

University of Arizona
Claremont Graduate University

INCHER, University of Kassel

June 26, 2012

Change of plans

Decided to change topics. We are now focusing on topics that show an integration of SNA with other fields

- QAP correlation and Regression
- Affiliation Networks (Real World Data)
- Key Actor Analysis (Regression using centrality measures)
- Text Mining techniques using Social Media Data

Purpose

Why Hypothesis testing?

- Techniques were designed to test theories about social interactions.
- Provide a sense of how much confidence we can have that the pattern we see in the data we've collected is actually typical of some larger population.
- Yet, inferential statistics in SNA are perhaps **the most sub-utilized set of techniques**

Purpose **At the end of the session the participants are expected to understand and replicate the procedures followed to conduct QAP correlation and regression to test hypothesis about networks using UCINET.**

Outline

Today we will see the following:

Outline

Today we will see the following:

1. QAP correlation
   - Motivation
   - UCINET implementation
2. QAP Regression
   - Data and preparation
   - UCINET implementation

Testing hypothesis

QAP correlation

- Correlation between two networks with the same actors (we need one-mode networks).
- **Motivation**: Tests the following statements
  - Professors who are friends also publish together.
  - Professors who publish together also are friends.
  - Countries send diplomats to countries from which they import food
- **More generally**: If there is a **tie** between two particular actors in **one relation**, how likely would it be to observe a **tie** between them in **another relation**?
- If two actors have a strong tie of one type, are they also likely to have a strong tie of another type?

Non-technical approach to QAP

QAP correlation...

- Tests the association between networks in two steps:
  1. Five measures of correlation coefficient are calculated between corresponding cells of the two matrices (called dyadic hypothesis).
  2. Hundreds of random permutations of rows and columns are computed while recalculating the correlation measures.
- From the second step the proportion of times that a random correlation measure is larger than or equal to the observed correlation measure is calculated.
- A lower proportion in step two (< 0.05), compared to step one, suggests a strong relationship between the matrices that is unlikely to have occurred by chance.

Intuition behind QAP

Publish together? Observed friendship

$$\begin{pmatrix} 2 & 1 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 6 \end{pmatrix} \text{ and } \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} =$$

Observed correlation measures
High is good!

Publish together? Permuted friendship

$$\begin{pmatrix} 2 & 1 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 6 \end{pmatrix} \text{ and } \begin{pmatrix} ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{pmatrix} =$$

Random correlation measures
High is BAD!

Intuition behind QAP

Publish together? Observed friendship

$$\begin{pmatrix} 2 & 1 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 6 \end{pmatrix} \text{ and } \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} =$$

Observed correlation measures
High is good!

Publish together? Permuted friendship

$$\begin{pmatrix} 2 & 1 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 6 \end{pmatrix} \text{ and } \begin{pmatrix} ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{pmatrix} =$$

Random correlation measures
High is BAD!

Applications

- One of the most influential papers in SNA is Granovetter's (1973) "The strength of weak ties." Today we will test the first part this theory:

  It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another (1973, p. 1,360)

  Hypothesis: The more common friends actors A and B have, the stronger the friendship tie between A and B is.
- As usual, to conduct a test in statistics we need to prepare the data.
- The data to be used is called ZACKAR and is located in UCINET.

Zachary (1977) gives to square matrices:

Observed a Karate club members for three years

- In **ZACHE** a tie is given if two individuals are friends (Binary network)
- In **ZACHC** A valued network measuring the total number of contexts in which two individuals are friends, thus measuring the strength of the friendship tie.
- These two matrices are **packed** in UCINET, meaning that the first step would be to unpack them.
- To unpack them go to **data, unpack** and select ZACKAR.

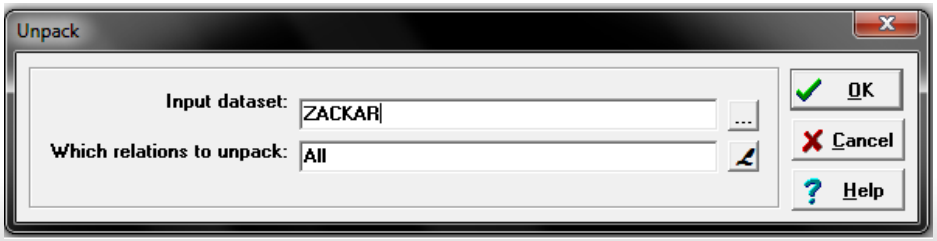


Figure 1: Unpacking procedure: outcome ZACHE and ZACHC

Matrix Algebra in SNA to obtain friends in common

Matrix multiplication creates counts from binary data

- $A * A^T$ in a **square undirected matrix** renders the number of people a given pair of actors are mutually connected to. The overlap in their personal networks

$ZACHE$ ZACHE^T Friends in common

$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} * \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

- This is implemented in **tools,similarities** select ZACHE

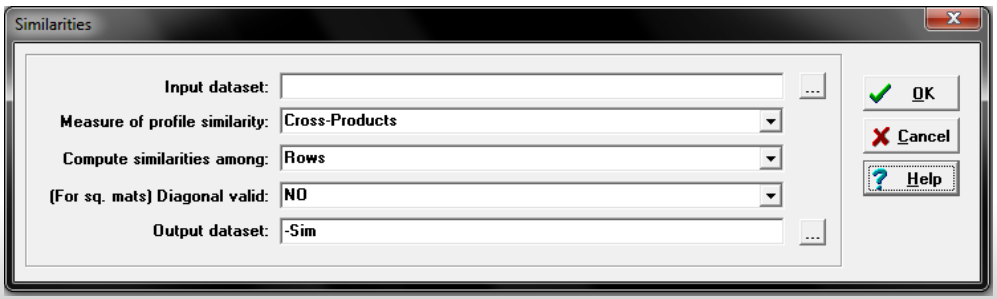


Figure 2: Similarities using matrix multiplication: outcome Friends in common

Data is ready now!

Using Friends in common and ZACHC

- We will statistically test if strength of tie (ZACH) is associated with the number of friends actors have in common

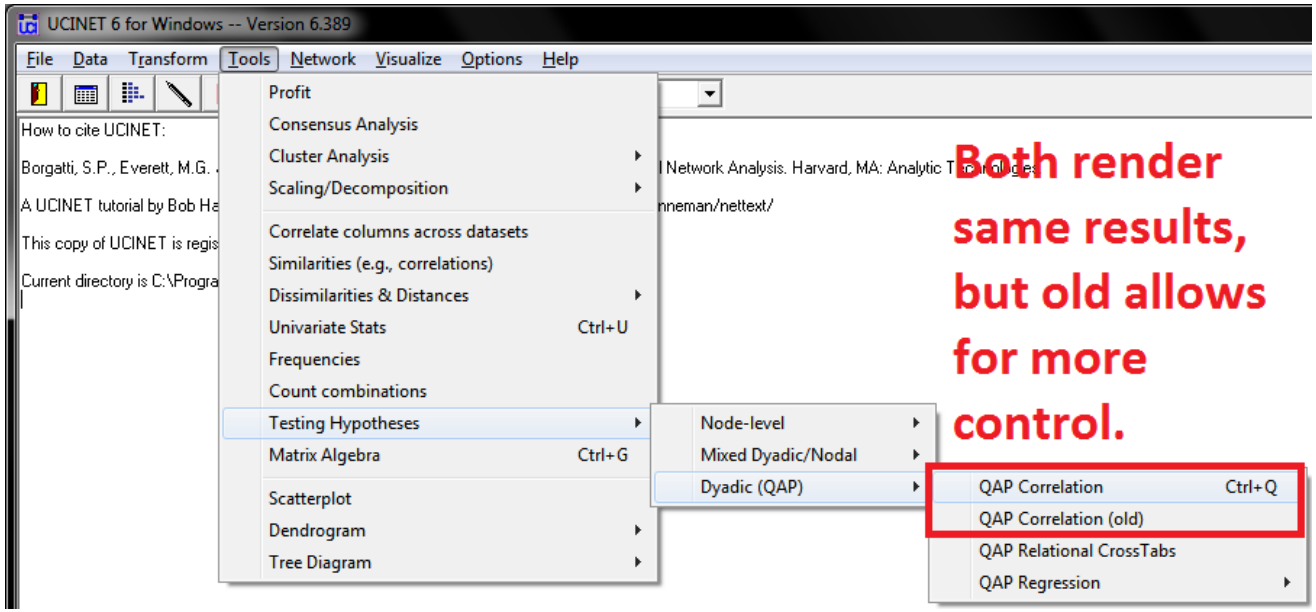
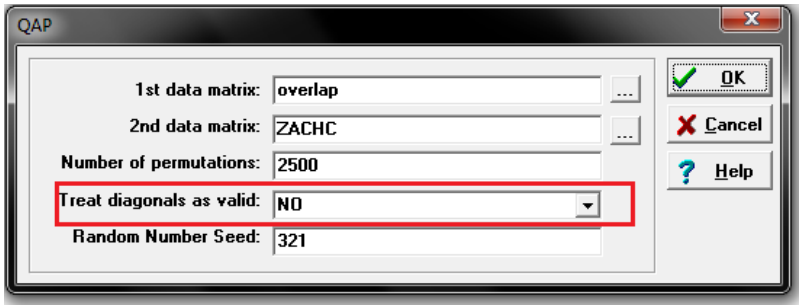


Figure 3: Old and new QAP procedures

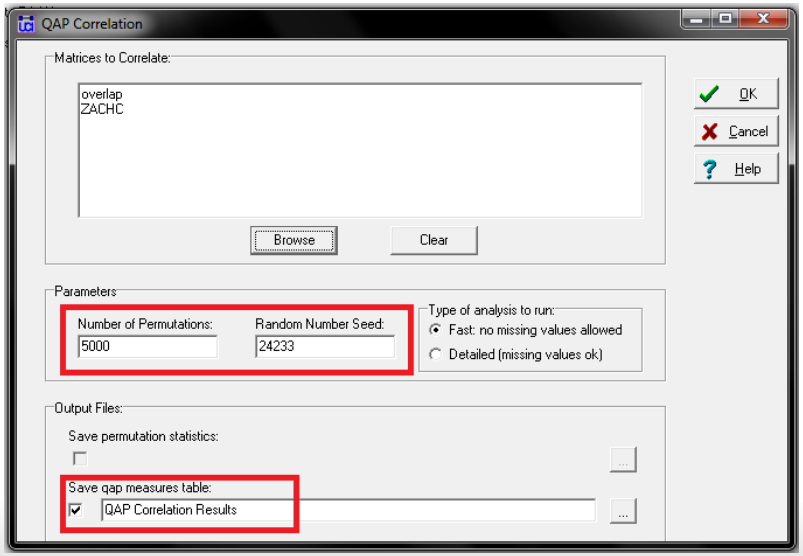
Old and New QAP procedures. . . Conclusions?



(a) Old Procedure, more flexibility

```
Bivariate Statistics

                                    1       2       3       4       5       6       7
                                Value   Signif     Avg      SD P(Large) P(Small)   NPerm
 ------------------------------ ------- ------- ------- ------- ------- ------- -------
 1    Pearson Correlation:       0.366   0.000  -0.000   0.054   0.000   1.000 2500.000
 2      Simple Matching:         0.408   0.000   0.365   0.014   0.000   1.000 2500.000
 3    Jaccard Coefficient:       0.195   0.000   0.127   0.016   0.000   1.000 2500.000
 4  Goodman-Kruskal Gamma:       0.667   0.000  -0.000   0.156   0.000   1.000 2500.000
 5       Hamming Distance:     664.000   0.000 712.025  19.303   1.000   0.000 2500.000

---------------------------------- Every measure was significant
Running time:  00:00:01
Output generated:  21 Jun 12 10:52:38
UCINET 6.389 Copyright (c) 1992-2012 Analytic Technologies
```

(b) More detailed measures: Pearson: numeric, Gamma: Ordinal, Simple matching and Jaccard: binary, Hamming: measure of dissimilarity



(c) New Procedute: Pearson

```
QAP results for ZACHC * overlap (5000 permutations)

                        1       2       3       4       5       6       7       8
                  Obs Value Significa Average  Std Dev Minimum Maximum Prop >= O Prop <= O
 ---------------- --------- --------- ------- ------- ------- ------- --------- ---------
 Pearson Correlation    0.3659    0.0002 -0.0007  0.0543 -0.1609  0.2071    0.0002    1.0000

QAP Correlations

                1     2
              overl ZACHC
              ----- -----
  1 overlap  1.000 0.366
  2   ZACHC  0.366 1.000

QAP P-Values

                1     2
              overl ZACHC
              ----- -----
  1 overlap  0.000 0.000
  2   ZACHC  0.000 0.000

QAP statistics saved as datafile QAP Correlation Results
```

(d) Easier to see, but no attention to distributional form or variables?

QAP correlation can be extended to fit regression models

Motivation: We may want to predict one relation knowing the other

- Which one is true?
  - Professors who publish together are friends
  - Professors are friends because they publish together

- Correlation is a symmetric association between the relations
- Regression is an asymmetric association between the relations
- **Advantage**: we can add more than one independent variable
- QAP regression uses more information to compute R-squares and Standard errors.

Regression Procedure

Several options, but most reliable is Multiple Regression Double Dekker

- The **2-step** procedure is similar as in QAP correlation.
  1. A standard multiple regression across corresponding cells of the dependent and independent matrices is computed.
  2. Randomly permutes rows and columns (together) of the matrices and recomputes the regression, storing resultant values of r-square and all coefficients.

Observed Outcome A Observed Predictor B Observed Predictor C

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Regression Procedure

Several options, but most reliable is Multiple Regression Double Dekker

- The **2-step** procedure is similar as in QAP correlation.
  1. A standard multiple regression across corresponding cells of the dependent and independent matrices is computed.
  2. Randomly permutes rows and columns (together) of the matrices and recomputes the regression, storing resultant values of r-square and all coefficients.

Observed Outcome A Random Predictor B Random Predictor C

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{pmatrix} + \begin{pmatrix} ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{pmatrix}$$

Data and data preparation

We will use Krack-High-Tec and the High-Tec-Attributes files

- Krackhardt's High-Tech network data (1992)
- These matrices describe three different types of ties found **among 21 managers** in a high-technology firm.
- Krack-High-Tec file contains three matrices: Advice, Friendship, and Reports To
- The High-Tec-Attributes file contains the following attributes of each of the 21 managers: Age, Tenure, Level
- Once again, we have to unpack Data, Unpack
- Select the Krack-High-Tec rendering three matrices: ADVICE, FRIENDSHIP, and REPORTS_TO

Questions to be addressed

We select Advice as the dependent variable

- Managers tend to ask advice only to other managers they are friends with
- Managers tend to ask advice only to other managers they have to report to

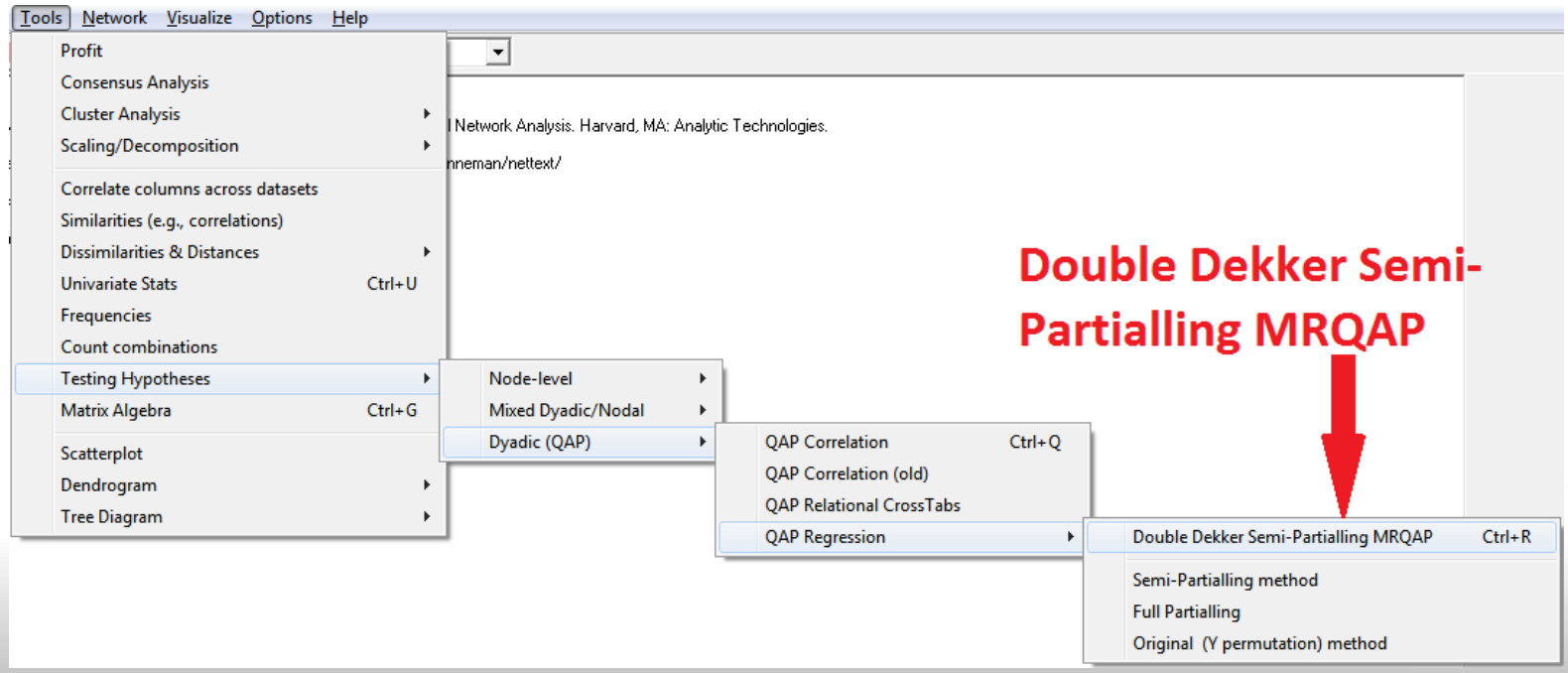
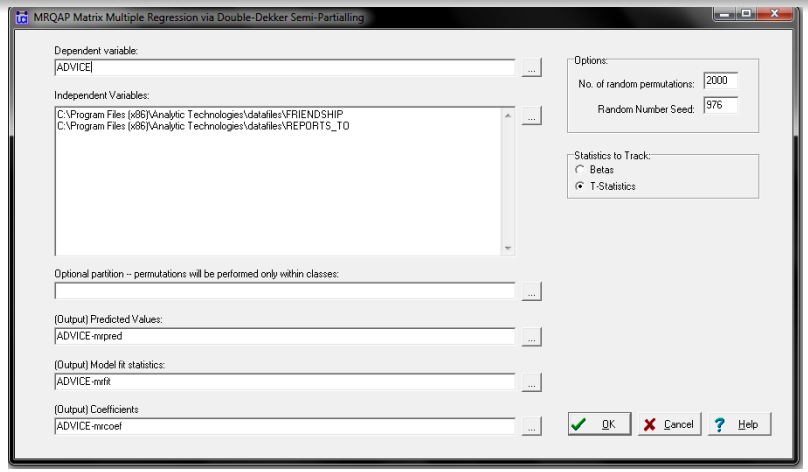


Figure 4: Follow this path or pres $ctrl + R$

QAP regression results



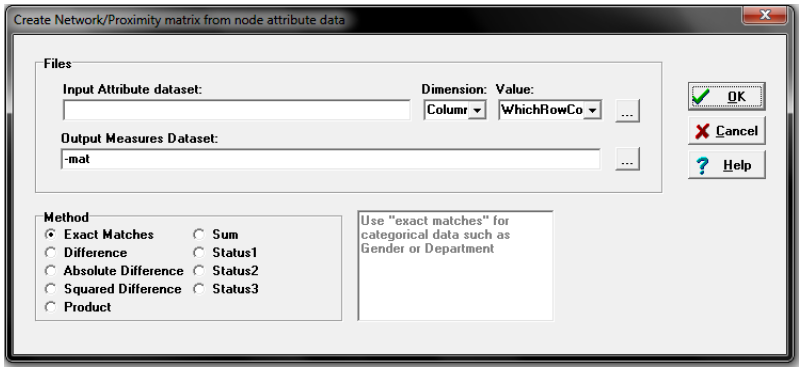
(a) Options to be selected

```
MULTIPLE REGRESSION QAP VIA DOUBLE DEKKER SEMI-PARTIALLING
------------------------------------------------------------------------------

# of permutations:                  2000
Diagonal valid?:                    NO
Random seed:                        976
Dependent variable:                 ADVICE
Partition variable (if any):
Predicted values:                   ADVICE-mrpred (C:\Program Files (x86)\Analytic Technol
Model fit stats:                    ADVICE-mrfit (C:\Program Files (x86)\Analytic Technolo
Model coefficients:                 ADVICE-mrcoef (C:\Program Files (x86)\Analytic Technol
Independent variables:              C:\Program Files (x86)\Analytic Technologies\datafiles
                                    C:\Program Files (x86)\Analytic Technologies\datafiles


MODEL FIT

          R-Square    Adj R-Sqr      P-Value          Obs        Perms
---------- ---------- ---------- ---------- ---------- ----------
Model         0.063      0.059      0.001    420.000    2000.000

REGRESSION COEFFICIENTS

             Un-Stdized Stdized Coef      P-value      As Large      As Small       Std Err
---------- ---------- ---------- ---------- ---------- ---------- ----------
FRIENDSHIP    0.13582    0.11701    0.05447    0.05447    0.94603    0.08425
REPORTS_TO    0.47157    0.20177    0.00050    0.00050    1.00000    0.12217
 Intercept    0.39694    0.00000    0.00000    0.00000    0.00000    0.00000
```

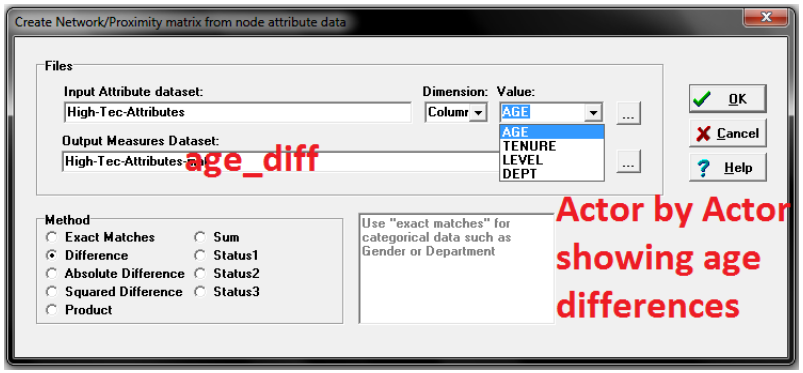
Ask advice to people they report to, not to friends

(b) QAP regression output

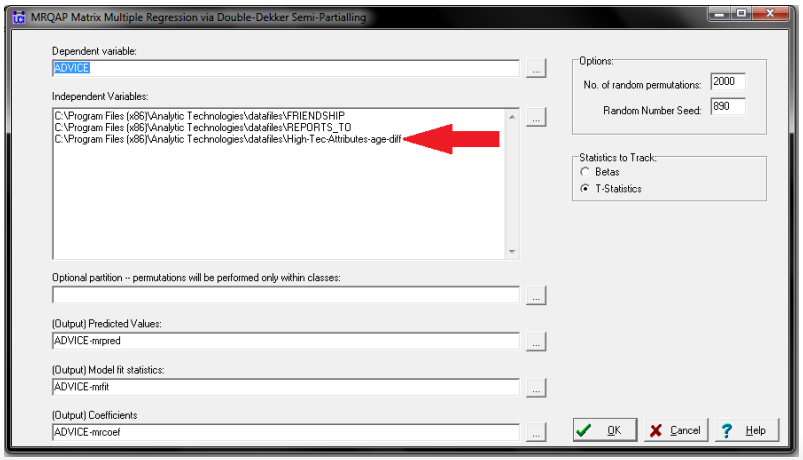
We can also add attributes



(c) **Data, attribute to matrix**



(d) Select age and Differences



(e) New Model Including age differences

REGRESSION COEFFICIENTS

	Un-Stdized	Stdized Coef	P-value
FRIENDSHIP	0.13440	0.11579	0.05447
REPORTS_TO	0.46809	0.20028	0.00050
High-Tec-Attributes-age-diff	-0.00098	-0.02658	0.40080
Intercept	0.39745	0.00000	0.00000

ıg time: 00:00:01 seconds.
: generated: 21 Jun 12 16:46:37

(f) Output, what do we conclude?

Testing hypothesis...

Could probably be the most sub-utilized method of SNA

- The potential benefits of understanding relationships at this level of analysis are not being exploited
- New projects should be designed to test for these hypotheses, although the samples are small, the data gathering plays an important role
- New models are being developed, such as Exponential-family Random Graph Models (ERGM) implemented in the package R.

Thank you!

Contact information

msgc@email.arizona.edu
Cecilia.Rios-Aguilar@cgu.edu