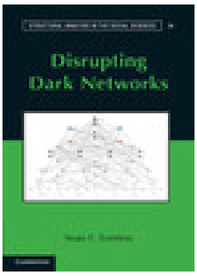


Disrupting Dark Networks

Sean F. Everton

Book DOI: http://dx.doi.org/10.1017/CBO9781139136877

Online ISBN: 9781139136877

Hardback ISBN: 9781107022591

Paperback ISBN: 9781107606685

Chapter

11 - Statistical Models for Dark Networks pp. 343-362

Chapter DOI: http://dx.doi.org/10.1017/CBO9781139136877.016

Cambridge University Press

11

Statistical Models for Dark Networks

11.1 Introduction

It is not uncommon when attempting to tease out the causes of a particular outcome that several potential factors are identified, all of which are correlated with the outcome of interest but may not be actual causes. For example, a strong correlation exists between checking into a hospital and subsequently dying. However, few would argue that checking into a hospital typically increases the chance that someone will die. Instead, the correlation between the two events exists because people who are extremely ill and thus have a higher probability of dying are more likely to check into a hospital than are those who are in good health. This correlation is referred to as spurious because it is due to the presence of a third factor, which, in this example, is the state of one's health.

It is also possible for a variable to appear to have either no effect or a negative effect on a particular outcome, when in fact it has a positive one. To illustrate this, consider the following (purely hypothetical) example (adapted from an example presented in Starbird 2006), which cross tabulates whether members of a particular dark network attended college with whether they are a key player (see Chapter 8). If we relied solely on the data presented in Table 11.1, we would probably conclude that attending college reduces the likelihood that a network member will be a key player: 70 percent of members who have not attended college are key players, whereas only 60 percent of those who have attended college are.

Such a conclusion would be wrong, however. Consider the next set of tables, which breaks the data down into whether members are native Indonesians. The first table (11.2a) presents data for those who are native Indonesians, whereas the second (11.2b) presents data for those who are not. Clearly, the story has changed. Now, attending college appears to have a positive effect on whether someone is a key player. Of the native

Table 11.1. *Crosstab of key players and college education*

Attended college	Key player No	Key player Yes	Key player Total
No	300 (30%)	700 (70%)	1,000 (100%)
Yes	600 (60%)	400 (40%)	1,000 (100%)
Total	900 (45%)	1,100 (55%)	2,000 (100%)

Indonesian members, 100 percent of those who have attended college are key players, whereas only 82.5 percent of those who have not attended college are. The same holds true for non-Indonesian members: 25 percent who have attended college are key players, whereas only 20 percent of those who have not attended college are. In both cases the difference is not huge, but attending college still appears to matter. What appears to matter more, however, is whether a network member is a native Indonesian:

Table 11.2a. *Crosstab of key players and college education (Indonesians)*

Indonesian attended college	Key player No	Key player Yes	Key player Total
No	140 (17.5%)	660 (82.5%)	800 (100%)
Yes	0 (0.0%)	200 (100%)	200 (100%)
Total	140 (14%)	860 (86%)	1,000 (100%)

Table 11.2b. *Crosstab of key players and college education (non-Indonesians)*

Non-Indonesian attended college	Key player No	Key player Yes	Key player Total
No	160 (80%)	40 (20%)	200 (100%)
Yes	600 (75%)	200 (25%)	800 (100%)
Total	760 (76%)	240 (24%)	1,000 (100%)

Table 11.3a. *Crosstab of key players and college education (Afghan vet)*

Afghan vet attended college	Key player No	Key player Yes	Total
No	165 (30.84%)	370 (69.16%)	535 (100%)
Yes	320 (57.66%)	235 (42.34%)	555 (100%)
Total	485 (44.50%)	605 (55.50%)	1,090 (100%)

Table 11.3b. *Crosstab of key players and college education (non–Afghan vet)*

Non–Afghan vet attended college	Key player No	Key player Yes	Total
No	135 (29.03%)	330 (70.97%)	465 (100%)
Yes	280 (62.92%)	165 (37.08%)	445 (100%)
Total	415 (76%)	495 (24%)	910 (100%)

86 percent of Indonesian members are key players but only 24 percent of non-Indonesians are. And, because most Indonesian members have not attended college, the aggregated data presented in Table 11.1 are misleading 11.2b.

Finally, consider Tables 11.3a and 11.3b, which break the sample down into whether network members fought in Afghanistan. The first table (11.3a) presents data for those who did, whereas the second (11.3b) presents data for those who did not. In both tables, the effect of a college education appears to be negative. Regardless of whether they are an Afghan veteran, network members appear to be less likely to be key players if they attended college. Specifically, only 42.34 percent of those who fought in Afghanistan and attended college are key players, whereas 69.16 percent of those who fought in Afghanistan but did not attend college are key players. Similarly, only 37.08 percent of those who did not fight in Afghanistan and attended college are key players, whereas 70.97 percent of those who did not fight and did not attend college are key players. Moreover, it appears that being an Afghan veteran matters: 55.5 percent of Afghan veterans are key players, but only 24 percent of non–Afghan veterans are. It is possible that a high percentage of key

Table 11.4. *Regression of key player on variables of interest*

	Model 1	Model 2	Model 3
Attended College	−1.25***	0.90***	0.90***
Indonesian		3.58***	3.59***
Afghan Veteran			0.16
Intercept	0.85	−1.91	−1.99
Pseudo R^2	6.71	31.63	31.69
AIC	2,571.75	1,888.02	1,888.22
BIC	2,582.95	1,904.83	1,910.62

Note: * = $p < .05$; ** = $p < .01$; *** = $p < .001$

players are both Indonesian and Afghan veterans, and this is confounding or hiding the effect of a college education. To see if this were the case, we could cross tabulate the Afghan veteran and Indonesian variables and then parse the data between those who are key players and those who are not, and then between those who attended college and those who did not. As one can see, however, as the number of independent variables increases, the number of potential cross tabulations that are needed to disentangle genuine from spurious effects increases exponentially and becomes impractical. An easier and more practical approach is to use multivariate regression.

Estimating a regression model sounds more daunting than it actually is. That is not to suggest that the math lying behind it is not computationally intense or that it is impossible to make a mistake in interpreting the results. Rather, it is to highlight the fact that standard statistical packages, such as SPSS, Stata, and SAS as well as UCINET and ORA, make estimating such models relatively painless for analysts.[1]

How can regression help us make sense of these tables?[2] Table 11.4 presents the results of a logistic regression model that regresses whether a member is a key player on the three variables previously explored.[3] Model 1 only includes college attendance as a variable and is consistent with the results presented in Table 11.1; it indicates that by itself there is a negative association (note that the coefficient is negative) between attending college and being a key player. Model 2, however, adds the

[1] Although Pajek allows users to estimate correlations between variables, it does not currently include the ability to estimate regression models.

[2] The following discussion of multivariate regression does not purport nor is it intended to be exhaustive; it merely seeks to introduce the topic in order to set the stage for the discussion of the regression of social network data. For a helpful introduction to regression, see Hamilton (1992).

[3] A logistic regression was used because the dependent variable is binary (i.e., yes or no); if the dependent variable was continuous, then an ordinary least squares (OLS) model would have been used (Long 1997).

Indonesian variable and two aspects of it are of note. One is that the college coefficient is now positive; the other is that not only is the Indonesian coefficient positive (as expected) but also substantially larger than the effect of having a college education. Finally, Model 3 includes the Afghan veteran variable. Interestingly, the effect of being an Afghan veteran, although positive, is not as large as whether someone has attended college or is a native Indonesian. Moreover, it has very little effect on the overall fit of the model as evidenced by the Pseudo R^2 associated with the model,[4] which only climbs from 31.63 to 31.69. The other two models of fit – the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) – support this conclusion as well. For both, the lower the number, the better the fit, but as we can see, both actually increase slightly from Model 2 to Model 3, indicating that the third model, which includes an additional variable, actually does a poorer job in explaining whether a member is key player.

This raises the question whether the Afghan variable has a significant effect on whether someone is a key player. Traditionally, social scientists have used measures of statistical significance to help decide whether a variable has a significant effect. Although social network analysts generally use a different approach for estimating measures of statistical significance, it is sufficient to note that social scientists generally consider a variable's effect to be statistically significant if the probability (p-value) that it could have occurred by random chance falls below a particular threshold, typically .05. If the p-value of a particular variable's coefficient falls below .05 ($p < .05$), then the probability that the result could occurred by random chance is less than 5 percent; put differently, the probability that the result represents a genuine association is greater than 95 percent. Similarly, if $p < .01$, then the probability that the result could have occurred by random chance is less than 1 percent, and if $p < .001$, then the probability is less than 0.1 percent. Looking at Table 11.4, we can see that both the college and Indonesian variables are statistically significant at a p-value of less than .01, indicating that there is a strong probability that the positive association between the variables and whether members are key players is genuine.[5] Note, however, that the Afghan variable is not statistically significant; thus, we cannot conclude

[4] Pseudo R^2 (and the traditional R^2 – aka the coefficient of determination – used with OLS models) estimates the extent to which a model's variables account for the variance in the dependent variable. Thus, Model 2's pseudo R^2 indicates that the two variables included in the model explain about 32 percent of the variation in whether members are key players, which is not bad, but it means that 68 percent of the variation remains unexplained.

[5] In terms of substantive effect, being Indonesian matters more than attending college, of course. Statistical and substantive significance are consistently conflated with one another. Variables can be statistically significant but have no substantive effect, a fact that gets lost in a lot of research (see McCloskey 1995; Ziliak and McCloskey 2008).

with any confidence that it has a positive effect on whether a member is a key player. Unless there are some unknown variables that we have yet to account for, it appears that the correlation between the two variables is spurious. This is true in spite of the results presented in Tables 11.3a and 11.3b, highlighting one of the dangers of relying solely on cross tabulations and one of the reasons for including multivariate regression in one's analysis.

11.2 Statistical Models for Social Network Data

Regression analysis of social network data differs from standard regression models in two important ways. One difference is that standard statistical models are designed to analyze random samples so that researchers can generalize their results to the population at large. As we saw in the previous section, variables that are statistically significant are seen as being unlikely to have occurred by random chance and thus can be generalized to the population from which the sample was drawn. However, because social network analyses do not typically analyze samples of networks – instead they analyze (at least in theory) complete networks – there is no need to generalize to the population at large. A second difference is that standard statistical models assume that observations are independent of one another, but as we discussed at length in Chapter 1, social network analysis assumes that observations (i.e., actors) are tied to one another (i.e., they are not independent of one another) and that these interdependencies influence behavior.

For these reasons, with social network data we should not use standard approaches for estimating statistical significance. Instead, we should turn to a form a nonparametric estimation known as permutation testing (similar to bootstrapping), which entails the random rearrangement of a network's rows and columns thousands of times in order to calculate a sampling distribution of statistics that can then be compared to the statistics generated by the observed (i.e., actual) network. If the observed statistics differ significantly from the randomly generated ones, we can conclude that the observed statistics could not have occurred by random chance and are "statistically significant."[6]

For example, when calculating the level of correlation between two networks, ORA and UCINET first compute the correlation coefficient between corresponding cells. Then, they randomly permute (i.e., rearrange) the rows and columns (together) of one of the networks and recalculate the correlation. This second step is carried out numerous times in order to compute the proportion of times that a random measure is

[6] This is known as the quadratic assignment procedure or QAP (Krackhardt 1987b).

larger than or equal to the observed measure calculated in the first step. A low proportion (e.g., less than 0.05) suggests that a correlation between networks is unlikely to have occurred by chance and thus is considered statistically significant. UCINET and ORA use a similar approach when regressing a dependent network on a series of independent networks. First, they estimate a standard multivariate regression across the corresponding cells of the dependent and independent networks. Then, they randomly permute the rows and columns of the dependent matrix, recalculate the regression, and store the resultant R^2 and coefficient values. This second step is repeated hundreds or thousands of times in order to compute the proportion of times that the randomly generated statistics are larger than or equal to those generated in the first step, and a low proportion is interpreted to mean that the results could not have occurred by random chance. Finally, the same approach is used when calculating the correlation or regression of attribute data (e.g., centrality and education level – see Chapter 7). In the first step, a correlation or standard multivariate regression is estimated across corresponding values of the dependent and independent attributes; in the second, the elements of the dependent attribute data are randomly permuted numerous times, and the distribution of randomly generated results is compared to the actual results to see whether the latter is likely to have occurred by random chance.

Now let's turn to estimating multivariate regression models in UCINET and ORA. Because we explored how to estimate the correlation between various attributes in Chapter 7, there is no reason to cover that ground again. We will, however, explore how correlations between networks are calculated.

11.3 Statistical Models in UCINET and ORA

Multivariate Regression with UCINET

Regression with Attribute Data. Let's begin by estimating a multivariate regression model using attribute data. For this we use UCINET's *Tools>Testing Hypotheses>Node-Level>Regression* command, which calls up a dialog box similar to Figure 11.1 (which we saw in Chapter 7). In this example, the data file that includes the dependent variable is loaded as the dependent dataset (`Alive Operational Network-cent.##h`), and the data file that includes the independent variables is loaded as the independent dataset (`Attributes.##h`). We also need to indicate the columns in which the attribute data can be found. Here we are regressing normalized degree centrality (column 1) on three independent variables: (1) education level (column 1), which is an ordered variable; (2) Indonesian (column 3), which is a binary variable where "1" indicates

[UCINET] Tools>Testing Hypotheses>Node-Level>Regression

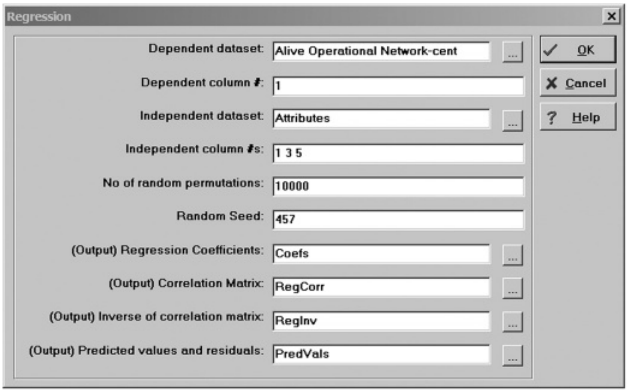
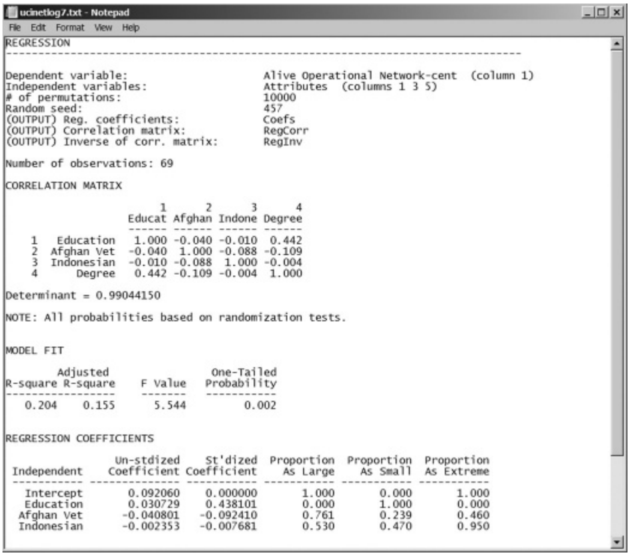


Figure 11.1. UCINET’s Attribute Regression Dialog Box

that the actor is Indonesian and “0’ indicates that the actor is not; and (3) Afghan veteran (column 5), which is also a binary variable where “1” indicates that the actor is an Afghan veteran and “0’ indicates that the actor is not. It is important that the attributes be either ordered or binary. Nominal (i.e., unordered) variables will produce nonsensical results. Indeed, the latter two attribute variables were recoded from the “Nationality” and “Military Training” variables included with the standard attribute data listed in Appendix 1. Note that the default number of random permutations is ten thousand; you can change this, but ten thousand is generally a large enough number to calculate a meaningful distribution of statistical measures.

After the datasets are loaded and the columns containing the dependent and independent variables are indicated (see Figure 11.1), click “OK,” and UCINET will generate a multivariate regression report (Figure 11.2). A correlation matrix appears at the top of the report (see Chapter 7 for discussion of correlation coefficients). Below this matrix are the measures of fit. The adjusted R^2, which is generally preferred over the unadjusted R^2 because it takes into account the number of variables included in the model, indicates that the independent variables account for 15.5 percent of the observed variation in the dependent variable (i.e., degree centrality). This is not huge, but given that the *p*-value of the F-statistic (F Value) equals 0.002, we can be reasonably confident that the model has some predictive power. Put differently, there is only a 0.2-percent chance that our independent variables provide no predictive power at all.

The regression coefficients and their associated levels of statistical significance are located at the bottom of the report. Two types of regression coefficients are reported: unstandardized and standardized. Unstandardized coefficients indicate the raw effect that each independent variable



```
ucinetlog7.txt - Notepad
File  Edit  Format  View  Help
REGRESSION
--------------------------------------------------------------------------------

Dependent variable:                   Alive Operational Network-cent  (column 1)
Independent variables:                Attributes  (columns 1 3 5)
# of permutations:                    10000
Random seed:                          457
(OUTPUT) Reg. coefficients:           Coefs
(OUTPUT) Correlation matrix:          RegCorr
(OUTPUT) Inverse of corr. matrix:     RegInv

Number of observations: 69

CORRELATION MATRIX

                         1      2      3      4
                    Educat Afghan Indone Degree
                    ------ ------ ------ ------
   1   Education     1.000 -0.040 -0.010  0.442
   2  Afghan Vet    -0.040  1.000 -0.088 -0.109
   3  Indonesian    -0.010 -0.088  1.000 -0.004
   4      Degree     0.442 -0.109 -0.004  1.000

Determinant = 0.99044150

NOTE: All probabilities based on randomization tests.


MODEL FIT

         Adjusted            One-Tailed
R-square R-square   F Value  Probability
-----------------   -------  -----------
   0.204    0.155     5.544        0.002


REGRESSION COEFFICIENTS

                    Un-stdized   St'dized  Proportion  Proportion  Proportion
 Independent   Coefficient Coefficient    As Large    As Small  As Extreme
 -----------   ----------- -----------  ----------  ----------  ----------
   Intercept      0.092060    0.000000       1.000       0.000       1.000
   Education      0.030729    0.438101       0.000       1.000       0.000
  Afghan Vet     -0.040801   -0.092410       0.761       0.239       0.460
  Indonesian     -0.002353   -0.007681       0.530       0.470       0.950
```

Figure 11.2. UCINET's Attribute Multivariate Regression Output

has on the dependent variable; in this case, the coefficients indicate the estimated amount by which normalized degree centrality changes for a one-unit change in each independent variable. Thus, a one-unit increase in education level leads to a 0.03 increase in normalized degree centrality, whereas being Indonesian or an Afghan veteran leads to decreases in normalized degree centrality of 0.04 and 0.002, respectively. By contrast, standardized coefficients indicate how many standard deviations a dependent variable will change per standard deviation increase in the independent variable. Standardized coefficients are helpful for comparing the effect of each independent variable on the dependent variable when the independent variables have different units of measure. Looking at the output we can see that the effect of education is far greater than that of being an Afghan vet or an Indonesian. Turning to the measures of statistical significance, we can see that UCINET provides three coefficients: the proportion of random trials that yielded a coefficient (1) as large or larger, (2) as small or smaller, and (3) as extreme as the observed value. Generally, the final column is where you will look to determine whether a coefficient should be regarded as statistically significant or not. The results tell us that, whereas we should pay attention to education coefficient, we should not put much stock in the Indonesian or Afghan veteran ones. Put differently, education level appears to have a positive effect on actors' centrality in the network, but being Indonesian or an Afghan veteran do not (i.e., the latter two variables

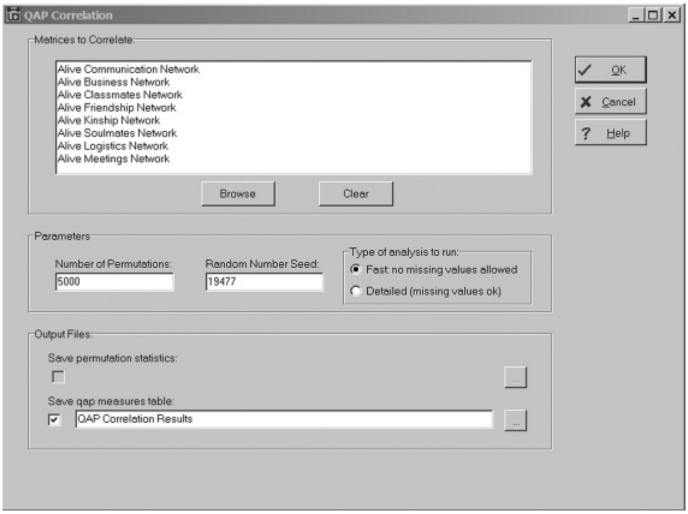


Figure 11.3. UCINET's QAP Correlation Dialog Box

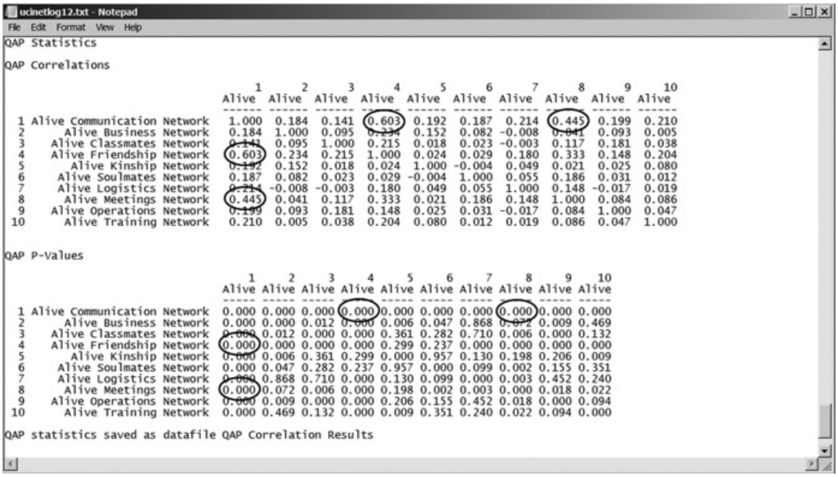
should not be seen as having either a positive or negative effect on actor centrality).

Multivariate Regression with Social Network Data. In the previous section we analyzed attribute data using correlation and regression techniques. The results from these types of analysis answer questions such as, "Are highly educated terrorists more likely to be central players in a terrorist network than are terrorists with very little education?" (The results suggest that they are.) By contrast, when we analyze social network data using correlation and regression methods, we are seeking to answer questions such as, "Are certain types of relationships (e.g., friendship ties) predictive of other types of relationships (e.g., operational ties) than are others?" We'll begin by first estimating the correlation between all of the Noordin alive networks before we regress a particular alive network (operations) on all of the other networks.

Tools>Testing Hypotheses> Dyadic (QAP)>QAP Correlation

To estimate the correlation between networks in UCINET, we use its *Tools>Testing Hypotheses>Dyadic (QAP)>QAP Correlation* command, which brings up a dialog box (Figure 11.3) that asks you to indicate (using the "Browse" button) which networks you want to include. For this exercise, we will examine the correlation between all ten of the alive networks,[7] most of which are visible in Figure 11.3 (note that the default number of permutations is five thousand).

7 Here we have unpacked the trust and operational networks and used the subnetworks contained therein.



ucinetlog12.txt - Notepad

File Edit Format View Help

QAP Statistics

QAP Correlations

		1 Alive	2 Alive	3 Alive	4 Alive	5 Alive	6 Alive	7 Alive	8 Alive	9 Alive	10 Alive
1	Alive Communication Network	1.000	0.184	0.141	0.603	0.192	0.187	0.214	0.445	0.199	0.210
2	Alive Business Network	0.184	1.000	0.095	~~0.234~~	0.152	0.082	-0.008	~~0.041~~	0.093	0.005
3	Alive Classmates Network	~~0.141~~	0.095	1.000	0.215	0.018	0.023	-0.003	0.117	0.181	0.038
4	Alive Friendship Network	0.603	0.234	0.215	1.000	0.024	0.029	0.180	0.333	0.148	0.204
5	Alive Kinship Network	~~0.192~~	0.152	0.018	0.024	1.000	-0.004	0.049	0.021	0.025	0.080
6	Alive Soulmates Network	0.187	0.082	0.023	0.029	-0.004	1.000	0.055	0.186	0.031	0.012
7	Alive Logistics Network	~~0.214~~	-0.008	-0.003	0.180	0.049	0.055	1.000	0.148	-0.017	0.019
8	Alive Meetings Network	0.445	0.041	0.117	0.333	0.021	0.186	0.148	1.000	0.084	0.086
9	Alive Operations Network	~~0.199~~	0.093	0.181	0.148	0.025	0.031	-0.017	0.084	1.000	0.047
10	Alive Training Network	0.210	0.005	0.038	0.204	0.080	0.012	0.019	0.086	0.047	1.000

QAP P-Values

		1 Alive	2 Alive	3 Alive	4 Alive	5 Alive	6 Alive	7 Alive	8 Alive	9 Alive	10 Alive
1	Alive Communication Network	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	Alive Business Network	0.000	0.000	0.012	~~0.000~~	0.006	0.047	0.868	~~0.072~~	0.009	0.469
3	Alive Classmates Network	~~0.000~~	0.012	0.000	0.000	0.361	0.282	0.710	0.006	0.000	0.132
4	Alive Friendship Network	0.000	0.000	0.000	0.000	0.299	0.237	0.000	0.000	0.000	0.000
5	Alive Kinship Network	~~0.000~~	0.006	0.361	0.299	0.000	0.957	0.130	0.198	0.206	0.009
6	Alive Soulmates Network	0.000	0.047	0.282	0.237	0.957	0.000	0.099	0.002	0.155	0.351
7	Alive Logistics Network	~~0.000~~	0.868	0.710	0.000	0.130	0.099	0.000	0.003	0.452	0.240
8	Alive Meetings Network	0.000	0.072	0.006	0.000	0.198	0.002	0.003	0.000	0.018	0.022
9	Alive Operations Network	~~0.000~~	0.009	0.000	0.000	0.206	0.155	0.452	0.018	0.000	0.094
10	Alive Training Network	0.000	0.469	0.132	0.000	0.009	0.351	0.240	0.022	0.094	0.000

QAP statistics saved as datafile QAP Correlation Results

Figure 11.4. UCINET's QAP Correlation Report

Once the files are loaded, click "OK," and UCINET will generate an extensive report, a portion of which is pictured in Figure 11.4. Most of the report contains a separate summary for the correlation between each pair of networks. Each summary reports the observed level of correlation, its associated *p*-value, the average random correlation and its associated standard deviation along with the minimum and maximum correlations, and the proportion of correlations that were larger and smaller than the observed level of correlation. Although the individual summaries are helpful because they provide a lot of useful information, most of the time you will want to scroll down to the bottom of the report where the correlations between the various networks and their associated *p*-values are succinctly summarized in two matrices (Figure 11.4). Looking at the top matrix we can see that there is a high level of correlation between the communication and friendship networks (0.603) and the communication and meetings networks (0.445); looking at the bottom matrix, we can see that both of these correlations are statistically significant ($p < .001$). In fact, the correlations between the communication network and all the other networks are statistically significant. Note also that the values above and below the diagonal are mirror images of each other. Thus, the correlation (and *p*-value) between two networks are actually listed twice, as illustrated by the correlations and *p*-values that are circled in Figure 11.4.

What about the operations network? Although it does not exhibit as high of a correlation with other networks as the communication network does, it does correlate relatively highly (and at statistically significant levels) with the communication (0.199, $p < 0.001$), classmates (0.181,

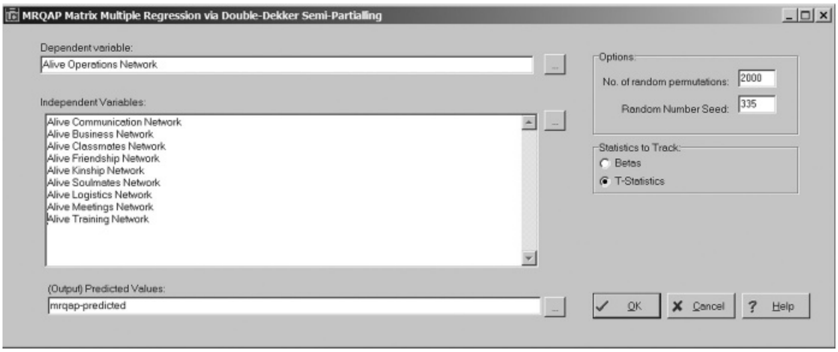


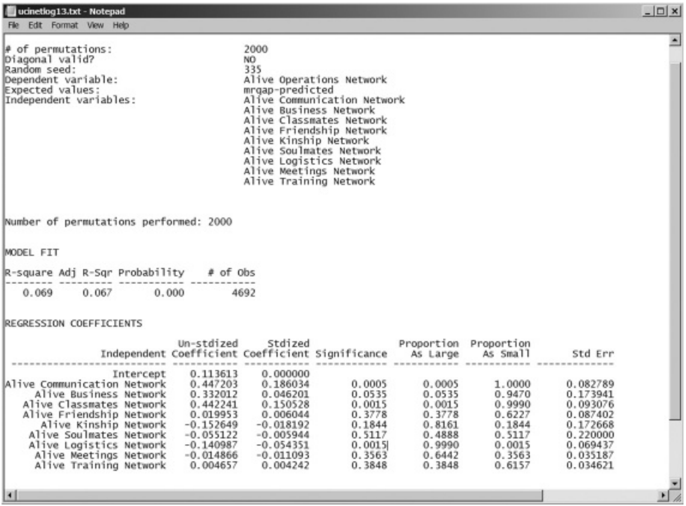
Figure 11.5. UCINET's MRQAP Network Regression Dialog Box

$p < 0.001$), and friendship (0.148, $p < 0.001$) networks. Its correlations with the business (0.093, $p < 0.01$) and meetings (0.084, $p < 0.05$) networks are also statistically significant, but their correlations are not as strong as the other three. Interestingly, there is a negative correlation between the operations and logistics networks, but it is not statistically significant, so we probably do not want to make too much of this, at least not at this point. What is also interesting is that the various networks that we have grouped together as the operational network (i.e., logistics, meetings, operations, and training) do not correlate too highly with one another, except for the meetings and logistics networks (0.148; $p < 0.00$), possibly suggesting that Noordin consciously attempted to keep various aspects of his operational network separate from one another. Whether all these correlations are genuine and whether an association that is not picked up by simply estimating correlations exists between two networks are not questions that quadratic assignment procedure (QAP) correlation can answer. Instead, we need to turn to multivariate regression, which is accessed with the *Tools>Testing Hypotheses>Dyadic (QAP)>QAP Regression>Double-Dekker Semi-Partialling MRQAP* command.[8] This calls up a dialog box (Figure 11.5) where we indicate the dependent (operations) and independent networks we will use in our analysis.

Tools>Testing Hypotheses> Dyadic (QAP)>QAP Regression> Double-Dekker Semi-Partialling MRQAP

It may take UCINET awhile to estimate the model because it recalculates each estimated coefficient numerous times (the default number of permutations equals two thousand). Once it finishes its calculations, it produces a report similar to the one presented in Figure 11.6, which reports the model's R^2, the adjusted R^2, the probability that the observed results could have occurred by chance, and the standardized and

8 MRQAP stands for Multiple Regression Quadratic Assignment Procedure. The Double-Dekker Semi-Partialling method was developed to adjust for network autocorrelation and mulitcollinearity (Dekker, Krackhardt, and Snijders 2007).



```
ucinetlog13.txt - Notepad
File Edit Format View Help

# of permutations:                     2000
Diagonal valid?                        NO
Random seed:                           335
Dependent variable:                    Alive Operations Network
Expected values:                       mrqap-predicted
Independent variables:                 Alive Communication Network
                                       Alive Business Network
                                       Alive Classmates Network
                                       Alive Friendship Network
                                       Alive Kinship Network
                                       Alive Soulmates Network
                                       Alive Logistics Network
                                       Alive Meetings Network
                                       Alive Training Network


Number of permutations performed: 2000

MODEL FIT

R-square Adj R-Sqr Probability    # of Obs
-------- --------- ----------- -----------
   0.069     0.067       0.000        4692

REGRESSION COEFFICIENTS

                          Un-stdized    Stdized             Proportion Proportion
              Independent Coefficient Coefficient Significance  As Large   As Small      Std Err
 ------------------------ ----------- ----------- ------------ ---------- ---------- ------------
                Intercept    0.113613    0.000000
Alive Communication Network  0.447203    0.186034       0.0005     0.0005     1.0000     0.082789
   Alive Business Network    0.332012    0.046201       0.0535     0.0535     0.9470     0.173941
  Alive Classmates Network   0.442241    0.150528       0.0015     0.0015     0.9990     0.093076
  Alive Friendship Network   0.019953    0.006044       0.3778     0.3778     0.6227     0.087402
    Alive Kinship Network   -0.152649   -0.018192       0.1844     0.8161     0.1844     0.172668
   Alive Soulmates Network  -0.055122   -0.005944       0.5117     0.4888     0.5117     0.220000
   Alive Logistics Network  -0.140987   -0.054351       0.0015     0.9990     0.0015     0.069437
    Alive Meetings Network  -0.014866   -0.011093       0.3563     0.6442     0.3563     0.035187
    Alive Training Network   0.004657    0.004242       0.3848     0.3848     0.6157     0.034621
```

Figure 11.6. UCINET's Network Regression Output

unstandardized coefficients along with their associated tests for statistical significance.

In this case the adjusted R^2 is not terribly high (6.7 percent), but because the p-value <0.001, we can conclude with relative confidence that the model does have some explanatory power. Clearly, though, a lot remains unexplained. The coefficients do tell something of an interesting story. Only the communication, classmates, and logistics networks exert a statistically significant effect on the operations network. More precisely, the presence of communication (0.447, $p < 0.001$) and classmate (0.442, $p < 0.01$) ties are positively associated with the presence of operational ties, while the presence of logistic ties (-0.141, $p < 0.01$) is negatively associated with presence of operational ties. Moreover, the standardized coefficients suggest that the effect of communication and classmate ties is about three times that of logistic ties.

What might these results tell us about Noordin's operations? As we noted previously, it appears that he consciously tried to keep his operations as separate from other aspects of his network (in particular, the logistic aspect of the network) as much as possible although he was unable to do so completely. The positive association of the communication with the operations network is not surprising given the important role that communications generally play in most dark network operations. This association suggests that targeting the communication network in some manner (e.g., through information operations or the removal of central players) might be an effective means for disrupting Noordin's operations. The positive association of classmate with operation ties is

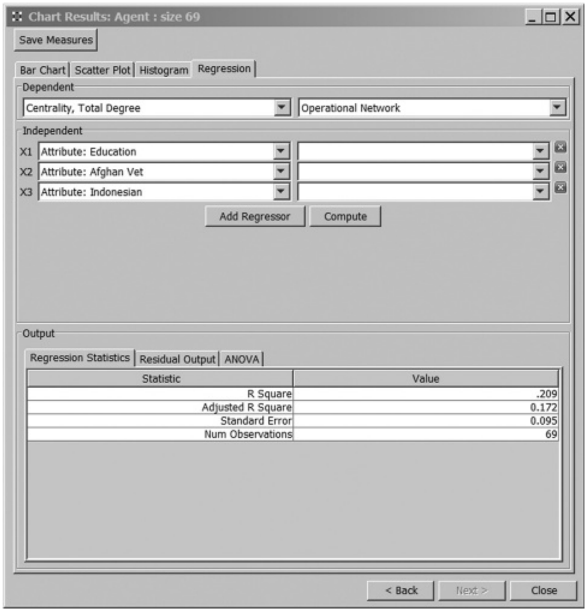


Figure 11.7. ORA's Attribute Regression (Overall Statistics)

perhaps not surprising either, given that we previously saw that classmate ties were central to Noordin's network. Nevertheless, this positive association once again highlights the important role that school ties have played in Noordin's network and indicates that the construction of alternative schools may provide a viable long-term strategy for disrupting not only Noordin's network but also other dark networks that recruit heavily from extremist schools. The construction of schools might also facilitate the process of reintegrating disaffected individuals back into Indonesian civil society.

Multivariate Regression with ORA

Multivariate Regression with Attribute Data. Let's begin by estimating the same multivariate regression model using attribute data that we previously estimated using UCINET. First, read the alive Noordin meta-network (Alive Network.xml) into ORA using its *File>Open Meta-Network* command. Then, select the *Visualizations>Measure Charts* command (or use the companion speed button). At the first dialog box (not shown) indicate the node class (i.e., agents, organizations, etc.) that you will be using. In this case, there is only one node class, so select it and click "Next." At the next dialog box (Figure 11.7), select the Regression tab, and in the "Dependent" box use the drop-down menu to select the

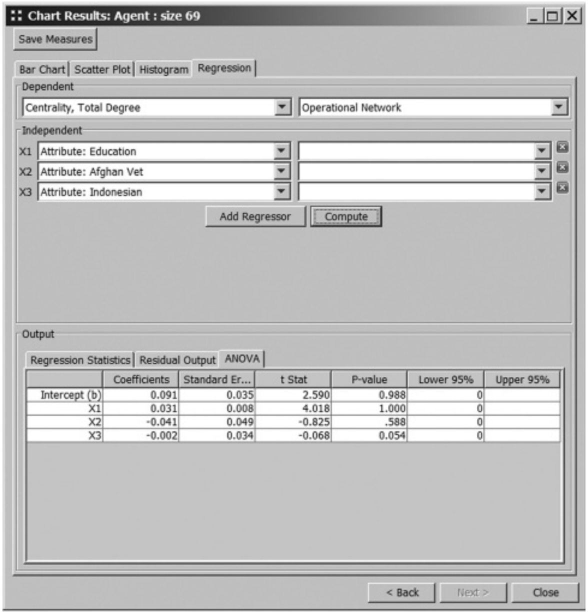


Figure 11.8. ORA's Attribute Regression (Coefficients)

dependent variable *Centrality, Total Degree*. To the right of this box, use the drop-down menu to indicate that you want total degree centrality to be calculated using the operational network. Next, we add our independent variables (i.e., regressors). Click on the "Add Regressor" button and add the three independent: (1) education level, (2) Afghan veteran, and (3) Indonesian. Finally, click "Compute" and the results will appear in the output portion of the dialog box (Figure 11.7). Note that the regression results are separated under three tabs. The first tab ("Regression Statistics") provides the overall results of the model. Comparing ORA's results with those of UCINET's (Figure 11.7) you will note that although they are similar, they are not identical. ORA's R^2 and adjusted R^2 are slightly higher than those estimated by UCINET, which is somewhat disconcerting, especially because an OLS regression estimated using the statistical package Stata confirmed UCINET's results.

Luckily, the coefficients for the independent variables agree with those estimated by UCINET. How do we know with ORA whether a coefficient is statistically significant? We look to the test statistic (t-stat), which indicates the number of standard deviations from the mean of the randomly generated results. If the absolute value of the test statistic is greater than 1.96, then the result is considered to be statistically significant.[9]

[9] If the observed results are greater than 1.96 standard deviations from the mean of the randomly generated results, this indicates that it lies beyond the 95th percentile. In other

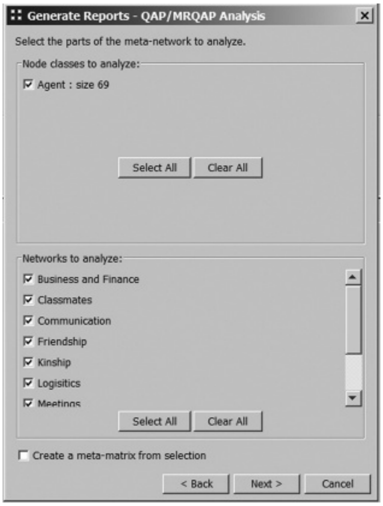


Figure 11.9. ORA's Network Regression Dialog Box

[ORA] File>Open Meta-Network

Analysis >Generate Reports >Statistical Procedures and Diagnostics >QAP/MRQAP Analysis

Multivariate Regression with Social Network Data. Finally, let us turn to see how to estimate a network regression equation in ORA. Open the alive Noordin meta-network (`Alive Noordin Network.xml`) in ORA using its *File>Open Meta-Network* command. Next, highlight the meta-network in the Meta-Network Manager and issue ORA's regression command: *Analysis>Generate Reports>Statistical Procedures and Diagnostics>QAP/MRQAP Analysis*. At the first dialog box (not shown) make sure that the meta-network is selected (it should be) and click "Next."

This calls up a dialog box similar to Figure 11.9 where you indicate the networks you want to analyze. In this case we want to analyze all ten, and by default all should be selected. However, it is possible to deselect some of the networks if you do not want to use them in your analysis. Clicking "Next" calls up the next dialog box (Figure 11.10) where, using the drop-down menus as well as the check boxes, we can indicate our dependent and independent networks. Note that the operations network has been identified as the dependent network and that boxes of all of the other networks have been checked to indicate that they are the independent networks. Be careful: The operations network is by default included as one of the independent networks, so make sure that its box is unchecked – see Figure 11.10). Also, the default number of permutations is set at 100, which is a bit small. Increase it to at least 1,000. Click "Next" and at

words, it indicates that there is less than a 5-percent chance that the observed result could have occurred by chance.

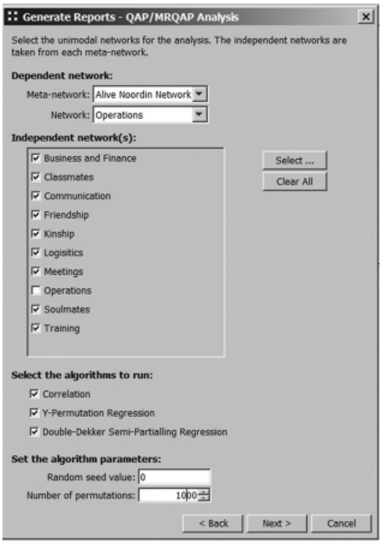


Figure 11.10. ORA's Network Regression Dialog Box

the next dialog box (not shown), provide an output file name and click "Finish."

Like UCINET, it will take ORA more than a few seconds to estimate the coefficients and their associated levels of significance. When it is done, it will produce a report similar to Figure 11.11. As you can see, ORA first reports the correlation between the operations network and all other networks and then the regression results. If you compare these results with those generated by UCINET, you will note that apart from rounding differences the correlation and regression coefficients are essentially the same as is the model R^2.[10] The measures of statistical significance do differ somewhat, but this is most likely due to the fact that we asked ORA to use only 1,000 permutations in its computations of statistical significance whereas in UCINET we used 2,000.[11] The differences do not change the story, however. Classmate and communication ties are positively associated with operation ties, while logistic ties are negatively associated, suggesting that targeting the communication network in some manner might be an effective short-term strategy for disrupting Noordin's operations, whereas building alternative schools might be a viable long-term strategy for either disrupting Noordin's operations or decreasing the likelihood that another dark network like Noordin will emerge in the future.

[10] Note that ORA does not report the unadjusted R^2.

[11] ORA includes both the Dekker semi-partialing method and the Y-permutation method for calculating statistical significance. The latter is the earliest form of MRQAP and is also implemented in UCINET.

Correlation Results

Network	Correlation	Significance	Hamming Distance	Euclidean Distance
Alive Noordin Network : Business and Finance	0.093	0.005	360	40.472
Alive Noordin Network : Classmates	0.181	0.000	476	39.925
Alive Noordin Network : Communication	0.199	0.000	552	39.950
Alive Noordin Network : Friendship	0.148	0.000	452	40.224
Alive Noordin Network : Kinship	0.025	0.192	366	40.768
Alive Noordin Network : Logisitics	-0.017	0.417	402	43.267
Alive Noordin Network : Meetings	0.084	0.020	406	47.497
Alive Noordin Network : Soulmates	0.031	0.169	362	40.719
Alive Noordin Network : Training	0.047	0.087	544	51.923

Regression Results

R-Squared: 0.068985

Variable	Coef	Std.Coef	Sig.Y-Perm	Sig.Dekker
Constant	0.114		0.000	
Alive Noordin Network : Business and Finance	0.332	0.046	0.051	0.044
Alive Noordin Network : Classmates	0.442	0.151	0.000	0.000
Alive Noordin Network : Communication	0.447	0.186	0.000	0.000
Alive Noordin Network : Friendship	0.020	0.006	0.398	0.385
Alive Noordin Network : Kinship	-0.153	-0.018	0.197	0.180
Alive Noordin Network : Logisitics	-0.141	-0.054	0.000	0.001
Alive Noordin Network : Meetings	-0.015	-0.011	0.391	0.380
Alive Noordin Network : Soulmates	-0.055	-0.006	0.437	0.490
Alive Noordin Network : Training	0.005	0.004	0.403	0.411

Figure 11.11. ORA's Network Correlation and Regression Report

11.4 Summary and Conclusion

In this chapter we have examined statistical models for attribute and network data that can help analysts disentangle genuine from spurious effects. As we saw, QAP correlation and multivariate regression models are invaluable to researchers when confronted with the problem of too many variables: that is, when they have identified several factors that could be associated with a particular outcome, it is impossible to distinguish which ones are truly from those that only appear to be. QAP regression is not the only statistical model available for cross-sectional network data. Also available are the p* models (also referred to as exponential random graph models or ERGMs), which differ from the QAP models in that whereas the latter are interested in the relationship between two or more attributes or networks, p* models attempt to draw inferences about the underlying structures of networks (Prell 2011:204). Unfortunately, like the actor-based models, p* models have yet to be implemented in UCINET, Pajek, or ORA (although ORA's Help function indicates that it is scheduled for the near future) and are difficult for the average user to implement in the available software.[12]

How have the approaches explored in this chapter improved our understanding of Noordin's network? The use of statistical models helped us to uncover the evidence that not only does education have a positive effect on the centrality of actors in Noordin's operational network but also ties with classmates are positively associated with operation ties. This is not the first time our analyses have highlighted the importance of schools, classmates, and education, but it does lend additional evidence to the argument that when crafting strategies for Noordin's network or others like it, the central role of education must be taken into account if interventions are to have any long-term effect.

12 Both actor-based and p* models can be implemented in the open-source statistical software, R (R Development Core Team 2011).