Center for Studies in Demography and Ecology

Statistical Modeling of Social Networks: Practical Advances and Results

by

Steven M. Goodreau
University of Washington

David R. Hunter
The Pennsylvania State University

Martina Morris
University of Washington

UNIVERSITY OF WASHINGTON **CSDE Working Paper No. 05-01**

Statistical Modeling of Social Networks: Practical Advances and Results

Steven M. Goodreau

University of Washington

David R. Hunter

The Pennsylvania State University

Martina Morris

University of Washington

INTRODUCTION

Population scientists have become increasingly interested in the ways that the structure of social relations interrelate with traditional sociodemographic variables such age, sex, and race. A long line of research has examined the ways in which the structure of an individuals' social network, along with their sociodemographic profile, affects occupational mobility, probability of migration, or adoption of contraception. In recent years, the structure of the relations themselves have increasingly come to represent the object of analysis. The overall structure of social networks is of great importance for certain demographic processes such as infectious disease transmission; it also can provide us with a richer picture of the micro-level social processes that may generate the social structures in which individuals find themselves embedded. This latter connection requires an approach that is able not only to consider the ways in which micro-level phenomena lead to macro-level outcomes, but also to reverse the direction of analysis, and ascertain the underlying micro-level phenomena from observed macro-level data. This link has been hard to forge. Social relations typically exhibit complex dependence, not only on the individual sociodemographic characteristics of those involved in the relations, but on the other relations already present among the same set of actors. This feedback process among relationships has made it extremely difficult to conduct any kind of sound statistical inference on the structure of social networks. Recent advances in statistical network theory and in computation have finally begun to make such work possible.

In this paper we will apply these recent advances in the statistical analysis of social networks to examine specifically the nature of *assortativity* and *transitivity* in adolescent friendships. By assortativity, we refer to the tendency for social networks to contain, regardless of the reason, a disproportionate number of ties between actors who are similar according to some sociodemographic attribute. By transitivity, we refer to the tendency for social networks to contain more sets of triadic relationships (A links to B links to C links to A) than expected by chance. We use these terms in this paper specifically to refer to the manifest network configurations, regardless of the social process that generated them.

We select these phenomena to examine, partly because they are so common in networks of social relations, but also because they are jointly determined by (at least) two different social processes which are substantively quite distinct. One is through the simple process of *homophily*, which we define as the tendency for people to directly choose partners who are like themselves, regardless of the other relationships they are in. Homophily obviously creates assortativity; it also generally leads to an increase in transitivity, since three actors with the same attribute will be more likely under this scenario to all form joint relationships. When the attributes that actors match on are not influenced by the existing social relations, then the process of homophily is exogenous to the network; that is, the probability of two people forming a tie is not affected by the other ties in the network.

The second interacting social process is *balance*, or the tendency for actors with a mutual relational partner to become connected, regardless of their attributes; that is, if A and B are friends, and B and C are friends, then A and C are more likely to become friends, solely because they have a friend in common. This process is not synonymous with transitivity (as each is defined here), but is one means of creating it; homophily is another. Because with balance the probability of one relationship is directly dependent on the existence of other relationships in the network, we consider this an *endogenous* network process, in contrast to homophily.

Balance alone does not lead to an increase in assortativity, since any triangle is equally likely to form, regardless of the attributes of the actors. However, when homophily and balance are present simultaneously, balance can act to amplify the effects of homophily on assortativity (Moody 2001). That is, once A and B become friends because they share the same attribute, and B and C do the same, then A and C are extra likely to become friends: they share the same attribute, and they have a mutual friend in common. The two act in tandem both in increasing the number of triangle formations and the observed level of assortativity.

One may, then, when faced with observed levels of assortativity and transitivity, ask the question: what levels of homophily and balance might have jointly created such a network? Are both phenomena necessary to explain the observed patterns, or can one suffice? One can imagine that answering this lies in ascertaining the degree to which triangles appear regardless of the attributes of the actors, and within-group friendships

appear regardless of the actor's number of friends in common. Nevertheless, decomposing these effects fully is surprisingly complex, precisely because of the dependencies among ties inherent in these processes. Much of mathematical social network analysis has been concerned with determining whether more or fewer of certain configurations are observed than expected by chance; however, determining what should happen "by chance" has in the past usually required assuming that only one social process is underway at a time, since direct calculation of the expected network configurations in more complex and realistic scenarios is impossible. Moreover, these traditional mathematical methods do not measure the variation in outcome possible under different scenarios, and thus limit the ability to conduct inference and hypothesis testing on observed data. To conduct such inference generally, one needs to identify the range of social networks that emerge under a set of joint social processes of arbitrary magnitude. This is precisely the link that statistical modeling for social networks provides.

The modeling class we use, known originally as p^* modeling and more recently as exponential random graph modeling (ERGM), derives from the work of Besag (1973), Holland and Leinhardt (1981), Frank (1991), Frank and Strauss (x), Strauss and Ikeda (x), and Wasserman and Pattison (1996). Although the general theory has now been present for some time, applications have been limited for a number of interrelated technical and theoretical reasons. We will examine the major one of these, model degeneracy, with regards to how it effects the issue of model fitting. We then apply this a series of models from this class to the adolescent friendship data from the AddHealth study (xx) using the software package *statnet* (xx), which implements a number of recent

developments in this field. We will focus on three models: one in which only homophily is present, one in which transitivity is present, and one which combines both.

METHODS

The ERG model class

Social network data generally comprises a set of actors, their individual attributes, and the presence or absence of their pairwise relationships. One can think of these as forming a mathematical graph, represented in Figure 1. Here the nodes represent actors (the colors and shapes of which convey attribute information) and edges represent relational ties between actors.

We define Y_{ij} as the random variable for the relationship between actors i and j, with $Y_{ij} =$ 1 when i and j share a relationship, and 0 when they do not. These relations can be represented by an $n \times n$ matrix (where n is the number of actors in the network), which we call $\boldsymbol{Y}$. For most social relations (including our data below) the diagonal entries Y_{ii} are undefined, since one cannot share a partnership with oneself. For directed relationships (A sends a relationship to B distinct from B sending to A, e.g. the relation "gives money to") all of the off-diagonal entries are meaningful. For undirected relationships (e.g. "lives with"), the lower triangle and upper triangle are exactly symmetric. We focus on an undirected relationship in our review and example below.

The ERG modeling class specifies the probability of observing a set of relationships (the ***Y*** matrix) given a set of actors and their individual characteristics. Denoting the realization of the matrix of relations ***Y*** as ***y***, and the number of actors as *n*, the general ERGM class can be defined as:

$$P\left(Y = y \middle| n \text{ actors}\right) = \frac{\exp\left(\theta' z\left(x\right)\right)}{c}$$

The vector *z* represents a set of network statistics whose values are specific to *y* and that are posited to have affected the network's formation. Examples of possible *z* statistics are the total number of ties, the fraction of actors with no ties, or the number of triangles (A, B and C are all friends). The θ parameters represent the coefficients of these statistics for a specific set of data, and must be estimated. The denominator *c* represents the quantity from the numerator summed over all possible graphs with *n* actors; this constrains the probabilities of all of them to sum to 1.

The above formulation specified the probability of observing an entire set of relations, but it can also be interpreted in terms of a single relation. Reformulating the network this way yields:

$$\text{logit } P\left(Y_{ij} \middle| n \text{ nodes}, Y_{ij}^{c}\right) = \theta' \delta z\left(x\right)$$

where logit refers to the log odds of a tie, $\ln\left[P\left(Y_{ij}=1\right)/P\left(Y_{ij}=0\right)\right]$, Y_{ij}^{c} refers to the remainder of the graph other than Y_{ij}, and $\delta z\left(x\right)$ refers to the amount by which the *z* statistics change when Y_{ij} is toggled from 0 to 1. It is the presence of the "rest of the graph" in the conditional statement that makes the probabilities so difficult to calculate in the first place, since they are all mutually dependent. Nevertheless, this formulation allows for an easier interpretation of the parameter values: if the formation of a tie will increase a given *z* statistic by 1, then the log-odds of that tie being formed are increased by the θ parameter associated with that statistic.

The process of selecting *z* statistics is where social theory directly enters into the modeling process. In some cases the statistics that may be of interest to the researcher are intuitive; in cases where they are not, Frank and Strauss (199x) and Snijders et al. (2004) present methods and examples for deriving the set of statistics that correspond mathematically to various sets of theoretical approaches to the nature of dependence in social relations.

When only considering the total number of any of these configurations, it is implicitly assumed that all instances of a type of structure have the same probability, regardless of the individual actors within them. Thus, with a single triangle parameter in the model, the addition of one triangle to a graph would change its probability by an equal amount regardless of which three actors were involved in that triangle. This is known as the assumption of *uniformity*. With demographic data one commonly expects different patterns of social relations among actors with different attributes, e.g. race, age, sex. In

this case, attribute-specific versions of statistics may also be included in the model; e.g. the number of ties between two Hispanics. When counts of the attribute combinations in ties are the only statistics included in the *z* vector, these models become similar in form (but not exactly equivalent to) more traditional loglinear models for contingency tables (Koehly et al. 2004). Such models are said to exhibit "tie independence" because the probability of any tie does not depend on the presence or absence of other ties. Any model involving triangles would, on the other hand, involve tie dependence.

Given a proposed model containing a set of z statistics, one would like to identify the values for the θ parameters that maximize the likelihood of the model. Unfortunately, the denominator *c* makes the values of the probability distribution impossible to calculate. For instance, a graph with only 20 actors has $\binom{20}{2}$ = 190 actor pairs in it; each of these may or may not possess a relationship, yielding 2^{190}, or 1.6×10^{57} possible graphs over which the numerator must be summed. This inability to calculate the probabilities means that the maximum likelihood estimates (MLE) for θ cannot be determined directly. In the past, approximation approaches using logistic regression have been used (Besag 1974; Frank and Strauss, 1986; Strauss and Ikeda, 1990; Geyer and Thompson 1992), including an analysis of the same data consider here (Moody 2001). However, the resulting estimates, known as the maximum pseudolikelihood, can perform very badly in practice for tie-dependent models (Geyer and Thompson, 1992) and their theoretical properties are poorly understood (Handcock, 2003). This approach in fact relies on temporarily assuming that there is no dependence among ties, when it is precisely that dependence that is under investigation.

Alternately, approximation methods based on Markov chain Monte Carlo (MCMC) have now been developed (xxx), and have been proven to provide an unbiased estimate of the MLE of θ in the long run. These MCMC methods solve an additional problem induced by the presence of the normalizing constant c. Once the values of θ have been found (by whatever method), the z and θ vectors then jointly define the probability of any individual graph, although these still cannot be calculated directly because of c. The same MCMC method used in parameter estimation also provides a means for drawing samples of graphs with the appropriate probability. This may be useful to the researcher who wishes to simulate the flow of information or disease over structurally similar networks; it also provides a means for examining how well the model fits the data (Hunter et al. in press).

The issue of goodness of fit of model to data is of particular concern in networks. Unlike more familiar linear models, network models with tie dependence can exhibit complex forms of feedback which, when mis-specified, lead to results that are unexpected, difficult to interpret, and at worst clearly unsatisfactory. The extreme form of this problem is known as model degeneracy (Handcock xx), and it has been the main impediment to the wide application of these models since their introduction.

Degeneracy may be best understood through an example. Imagine an ERG model containing two statistics, one for the number of ties in a graph, and one for the number of triangles. On the surface, this may seem like an appropriate model to capture the magnitude of clustering in a network. In effect, the model proposes that there is some

overall tendency to form ties, which is then added to whenever the tie will create a triangle. The more triangles that will be completed by the addition of a tie, the more likely it is to occur. The behavior of the model in practice does not match intuition, however. For most data sets of more than a handful of actors, the MLE values of θ for this model imply a probability distribution in which the network is almost always full (every tie exists) or empty (no ties exist). The relatively probability of these two extremes is such that the average number of triangles present matches that observed in the data, but any realistic network that resembles the original data is virtually impossible.

In empirical networks the tie parameter is generally negative (since only a small fraction of possible ties actually form, even in the densest of social networks) and the triangle parameter positive. The negative tie parameter is the only "brake" placed on the formation of triangles; however, since the number of possible triangles is on the order of n^3 but the number of ties is only n^2, the negative tie effect is never large enough to contain the tendency to create more triangles. In other words, the statement about social processes implicit in the model itself (that an overall tendency to form ties and close triangles is an adequate description of relationship formation) is almost always wrong. Because the MCMC process used to estimate theta relied on drawing samples of the graph, these approaches often fail to converge on an estimate in the first place; this is not simply a function of the estimation procedure, but represents a fundamental lack of fit between the selected model and the data. Since most applications in the field have used pseudolikelihood estimation, the fact that many commonly used models were poorly

fitting has largely been hidden. For more a more technical analysis of model degeneracy for social networks, see Handcock (in press).

Since we are interested in transitivity in this paper, and since we know transitivity is important generally, this specific case of degeneracy is of particular importance. One solution that has proven successful is that of Snijders et al. (2004), who derive a new pair of statistics from assumptions about relational dependence, which they called alternating k-triangles and k-independent two-paths. Hunter and Handcock (2005) demonstrate that the former statistic is equivalent to what they refer to as a "weighted edgewise shared partner" (WESP). Under this fairly heavy but precise terminology lies a fairly straightforward idea: two actors have an increasing tendency to form ties as they have more partners in common, but that there is a "diminishing return" with increasing partners. That is, the chances of becoming friends increases more between having 1 vs. 2 friends in common than it does in having 6 vs. 7 friends in common. This approach is not only based on intuitive assumptions about relational dependence, but has been shown to work well in practice in overcoming model degeneracy and generating models that fit real data well, specifically the data used in this study (Hunter et al., in press).

In this paper, we will use MPLE estimation for tie-independence models (which are unbiased in this special case) and MCMC estimation for tie-dependence models. Both are implemented in the *statnet* package for R (http://csde.washington.edu/statnet), which implements numerous features that make the estimation both relatively fast and robust (Handcock et al. 2005b).

We next describe the data, and then describe the specific statistics we will use to fit it.

DATA

We consider the friendship data from the first wave of the National Longitudinal Study of Adolescent Health (AddHealth). AddHealth comprised a stratified sample of schools in the US containing students in grades 7 through 12; the first wave was conducted in 1994-1995. For the friendship networks data, AddHealth staff constructed a roster of all students in the school from school administrators; students were then provided with the roster and asked to select up to five close male friends and five close female friends. Students were allowed to nominate friends who were outside the school or not on the roster, or to stop before nominating five friends of either sex. Complete details of this and subsequent waves of the study can be found in Resnick et al. (1997) and Udry and Bearman (1998) and at http://www.cpc.unc.edu/projects/addhealth.

Each school community contains students in all six of the grades 7-12. In many cases, obtaining such a sample required multiple schools in the same community to be included. The initial sample consisted of schools containing an 11^{th} grade; if necessary, a feeder school was then randomly selected from among those schools sending students to the high school, with probability proportional to the size of the student body sent. Since most high schools do not have seventh graders, most school communities consist of more than one school; for simplicity we use the term "schools" to refer to such a school

community here. Our analysis includes 59 of the schools, ranging in size from 71 to 2209 surveyed students. We exclude schools with high amounts of missing data; this happened, among other reasons, for special education schools and for school districts that required explicit parental consent for student participation.

The original data consists of directed relations (e.g. A nominates B as a friend, but not vice versa). In this paper we consider the network of mutual friendships, those in which both actors in the pair nominate each other; this relationship is thus undirected. The limit on the number of allowed nominations means that the data are not complete, but we will assume for convenience that a lack of nomination in either direction between two individuals means that there is no mutual friendship.

The individual attributes include many measurements on each of the individuals in these networks, from grade and sex to drug use to club membership. These characteristics vary considerably in the degree to which they may change as a function of existing social relations, with sex at one extreme and alcohol and drug use at the other. Since these are cross-sectional data, it is extremely difficult to gain insight into the relative degree by which the individual characteristics affect the formation of friendship or the friendships shape individual characteristics. For this reason, we focus on three attributes which we consider relatively exogenous: sex, race and grade. (For an analysis of the same data that incorporates many other actor- and school-level variables, see Moody 2001). These are likely not perfectly exogenous; for instance, there may some influence of one's peers upon grade level because of failing and repeating. Likewise, self-identified race may be

affected by friendships, especially for students who are multiracial or from ethnic groups not neatly captured by the prevailing American racial typology. However, in the absence of any data we focus our attention on assuming these categories to be fixed, and predicting friendships as a result. What we term "race" is constructed from two questions on race and Hispanic origin; in our data, options for the race variable are "Hispanic (all races)", "Black (non-Hispanic)", "White (non-Hispanic)", "Asian (non-Hispanic)", "Native American (non-Hispanic)", and "Other (non- Hispanic)". We abbreviate these to "Hispanic", "Black", "White", "Asian", "Native American", and "Other". Note that different schools have different number of races present, and thus will possess different numbers of parameters in the models estimated here.

Multiple forms of missing data appear, including students who did not fill out the survey, students who filled out a survey but were not on the roster, and students who left individual questions blank. We exclude the first two types of students from our analysis; for the others, we allow an extra "NA" category for each of the attributes.

The statistics we will include in different models include:

TIES:

- s_1 = the total number of ties. This statistic acts as an intercept or grand mean. Its magnitude is directly affected by the density of the graph, and therefore by the number of actors. Its inclusion then allows the values of other parameters to be compared more directly across schools of different sizes.

ATTRIBUTE-SPECIFIC STASTISTICS:

- k_i = the total degree for all actors with attribute value i , where an actor's degree is defined as the number of ties they possess. These statistics serve as "main effects" for the attribute categories, allowing the different races, sexes and grades to have different overall levels of tie formation. (Each level of each attribute possesses its own statistic, although one value per attribute must be left out as a reference category, since the sum across all levels of the attribute equals $2s_1$).
- h_i = the total number of edges between actors who both possess attribute value i (differential homophily). There is one such statistic for each level of the attribute. Differential homophily statistics are included for race and for grade.
- h = the total number of edges between actors who both possess the same level of the attribute, regardless of which level (uniform homophily). Uniform homophily is used for sex, rather than differential homophily, since there are only two categories of interest for the variable; with edges and main effects already included, there is only one degree of freedom remaining (since with undirected data there are only three kinds of partnerships, MM, MF and FF).

BALANCE STATISTICS:

- $w = e\sum_{i=1}^{n-2}\left\{1-\left(1-e\right)^{i}\right\}p_i$, where p_i equals the number of pairs of actors who are friends, and who have exactly i friends in common. This is the "weighted edgewise shared partner" (WESP) statistic of Hunter and Handcock (200x), with τ set equal to 1. It is mathematically equal to the k-triangle statistic of Snijders et

al. (x). This statistic is included as a measure of transitivity, instead of a simple triangle count, because of the degeneracy issues discussed above.

We consider three models based upon this set of statistics: a homophily model (edges, main effects for attributes, attribute homophily), a balance model (edges, weighted shared partner), and a homophily/balance model (all statistics). We expect that the homophily model will yield comparatively high estimates for the homophily parameters, since the model attributes all of the observed levels of assortativity to homophily; likewise, the balance model will yield high estimates for the shared partner parameter, since it is attributing all transitivity effects to the balance process. The joint homophily/balance model should yield smaller estimates for both parameters than either of the single-effect model, since it is able to capture the joint interactive effects of the two processes. In the results below we consider whether this expectation holds up, and the magnitude to which the effect occurs in different settings.

RESULTS

Figure 2 shows the values for the parameters of the homophily model; boxplots represent the range that the parameter takes on across the 59 schools studied. Note that in any case where there were fewer than 5 students of a given race in a school, the parameter is excluded. This is because such small numbers mean the parameter values are capable of great changes based on small effects, and in extreme cases (i.e. with one student) are

undefined. Given the large number of parameters and schools, we do not show the individual values for each school, nor whether each term is significant (MCMC p-value < 0.05) for each school.

We see that the main effects for different grades are strong and consistent. The effects are almost always positive, meaning that each grade has on average more friends than students in Grade 7 (the reference category) do; the effect increases with grade, then declines for Grade 12. The last effect is almost certainly a result of Grade 12 students nominating more friends who are outside of the school and thus unable to reciprocate the tie in this data set. Homophily effects for grade are more consistently strong and positive across schools; this time, however, they are strongest in Grade 7, and decline with age. Again, the trend reverses with Grade 12, presumably for the same reason; with a portion of their out-of-grade friendships not captured, the remaining friendships are disproportionately within grade.

For race, the main effects are much less pronounced; that is, the overall number of friendships per person relative to whites does not differ systematically across all schools. Racial homophily is much clearer, however. Consistently large values are found among Blacks, followed by Asians, and Whites. (Recall that these parameters are ***not*** a function of the relative size of the group in the school). Hispanics, Native Americans and student of other race usually exhibit significant homophily, but each at times exhibits a significant level of the reverse trend, heterophily, with Hispanic students exhibit heterophily most frequently.

The main effect for sex is small, meaning females nominate slightly more friends than males. Homophily is moderate (but always significant); it would presumably be higher, but for the stratified nature of the friendship nomination question (name your male friends; name your female friends). What is remarkable is the incredibly tight consistency for both the main effects and the homophily across all 59 schools, which represent very different sociodemographic settings.

We now consider the balance model; the plot of the shared partner parameter across schools (plotted against school size) is found in Figure 3. This model converged for x out of the 59 schools, and lead to degeneracy in the remainder; those for which the model fit were typically among the smaller schools. Although the absolute value of this parameter is less straightforward to interpret than the attribute parameters, it is consistently positive, and tends to increases with school size.

What happens when we place the two models together? In concurrent work (Hunter et al. in press) we have found that the combined model fits significantly better than either sub-model for all schools thus far examined, both by standard likelihood measures and through comparison of simulated graphs to the original data. With regard to the parameter values, Figures 4 and 5 plot the values of the parameters from the first two models (x-axis) against those obtained in the joint homophily/balance model (y-axis). Grade, race and sex are covered in Figures 4a, 4b, and 4c, respectively, while the shared partner parameter is found in Figure 5. We see that our qualitative prediction largely

holds; the joint model yields lower parameter estimates than either of the sub-models in almost all cases. The magnitude of this decline is remarkably consistent within each type of attribute across a wide range of schools, but shows stronger variation across the four different parameter types. The confidence interval on the slope of the regression line fit through each of the four sets of points is shown in Table 1, with a consistently larger decline in the transitivity (shared partner) parameter than in any of the attribute parameters. The differences among different levels of each attribute (e.g. White vs. Black) are very small and not significant. Moreover, we examined the relationship between the level of decline and school size, as well as the fraction of the student body belonging to the relevant category (results not shown). No clear relations were found.

Discussion (To be developed)

- Both exogenous and endogenous social processes are needed to explain the observed sets of social relations among high schools.
- Adding balance to the model decreased the amount of homophily needed to explain observed levels of assortativity by about 15% across the board. This is not large, but it is remarkably consistent. One should not take observed levels of segregation as a direct indication of the degree to which actors directly choose partners like themselves; transitivity may be significantly amplifying this effect.
- These effects were remarkably consistent across different settings, with different racial/ethnic compositions.

- Examination of the circumstances under which Hispanics exhibited heterophily, and how balance affected this
- Discussion of attributes that can be affected by existing relations (e.g. smoking, self-esteem, etc.) In these cases, the clear boundary we are drawing between exogenous and endogenous effects becomes far more blurry.
- The promise of statistical models of complex social structure, now that many of the statistical and theoretical hurdles have been jumped.

FIGURE 1

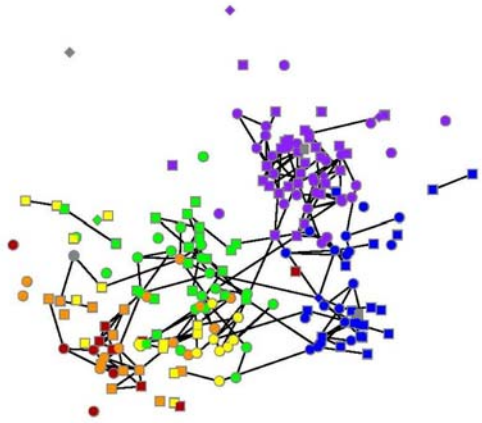

FIGURE 2

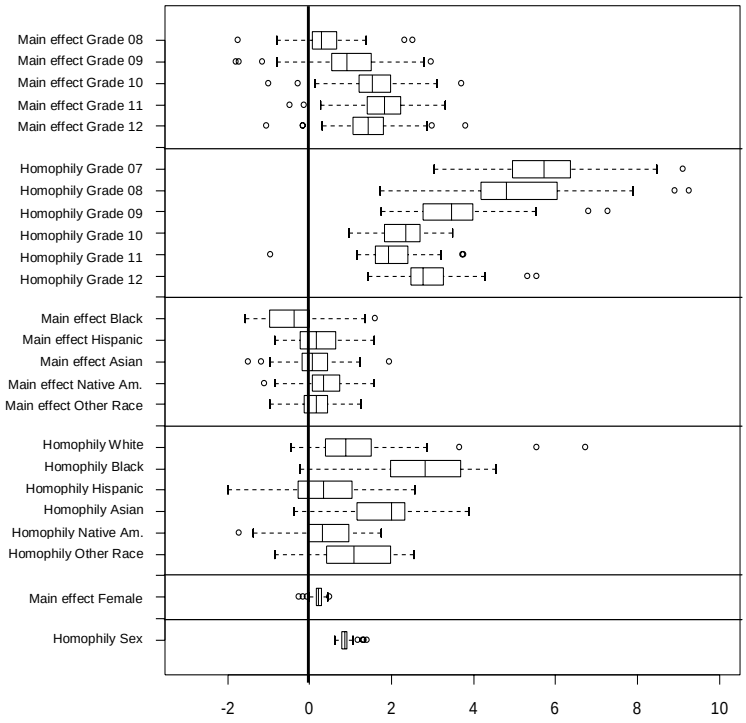
Main effect Grade 08
Main effect Grade 09
Main effect Grade 10
Main effect Grade 11
Main effect Grade 12
Homophily Grade 07
Homophily Grade 08
Homophily Grade 09
Homophily Grade 10
Homophily Grade 11
Homophily Grade 12
Main effect Black
Main effect Hispanic
Main effect Asian
Main effect Native Am.
Main effect Other Race
Homophily White
Homophily Black
Homophily Hispanic
Homophily Asian
Homophily Native Am.
Homophily Other Race
Main effect Female
Homophily Sex
-2
0
2
4
6
8
10

FIGURE 3

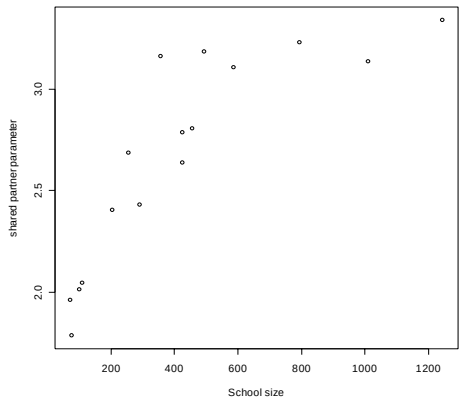
shared partner parameter
2.0
2.5
3.0
200
400
600
800
1000
1200
School size

FIGURE 4

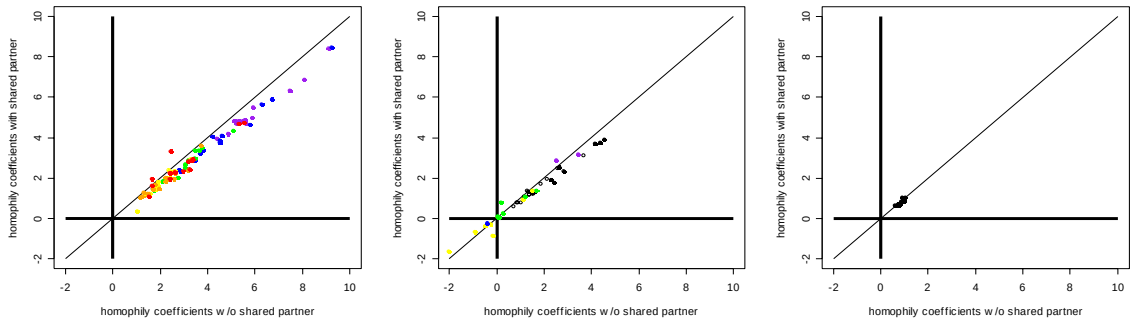

FIGURE 5

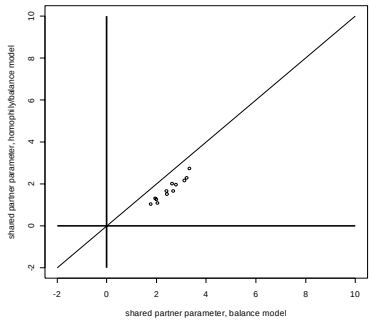

TABLE 1

Race	0.842 - 0.906
Grade	0.859 - 0.886
Sex	0.868 - 0.940
WESP	0.641 - 0.731