Model-based Assessment of the Impact of Missing Data on Inference for Networks[1]

Krista Gile
University of Washington, Seattle

Mark S. Handcock
University of Washington, Seattle

Working Paper no. 66
Center for Statistics and the Social Sciences
University of Washington

September 7, 2006

[1]Krista Gile, Department of Statistics, University of Washington, Box 354322, Seattle WA 98195-4322. E-mail: kgile@stat.washington.edu and; Mark S. Handcock is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle WA 98195-4322. E-mail: handcock@stat.washington.edu; Web: www.stat.washington.edu/handcock. The authors are grateful to Martina Morris for numerous helpful suggestions. This research is supported by Grant DA012831 from NIDA and Grant HD041877 from NICHD.

Abstract

Most inference using social network models assumes that the presence or absence of all relations is known. This is rarely the case. Most social network analysis ignores the problem of missing data by including only actors with complete observations.

In this paper we use a statistical model for the underlying social network to demonstrate that the computationally parsimonious complete case approach can lead to different conclusions from an approach utilizing all observations. We also show that the overall fit to the data is improved by extending the model to represent differences between respondents and non-respondents.

The ideas are motivated and illustrated by an analysis of a friendship network from the National Longitudinal Study of Adolescent Health.

Key Words: AddHealth, exponential random graph model, maximum likelihood estimation, non-response, sample survey, statnet

1 Introduction

Social network analysis can be used to characterize the patterns of arcs between actors. For example, in a sexual network related to disease spread, social network analysis can address questions such as: Who tends to have sex with whom? Do people form sexual partnerships entirely at random, or do they tend to choose partners from among their friends and friends' friends, leading to a clustering of sexual relations? Social networks in other contexts present questions such as: Do consumers make purchases completely at random, or do they tend to purchase some specific items together? Do the patterns of telephone conversations between suspected terrorists become more clustered, with more calls within a specific small group, just before a terrorist attack?

Statistical modeling of social networks allows researchers to statistically compare the patterns of arcs observed in the network to the patterns that we might have observed if the arcs had been formed completely at random. In this way, we can statistically formalize our inference about which processes are likely to have generated the network we've observed. Models based on statistical exponential families have a long history in social network analysis (Holland and Leinhardt 1981, Frank and Strauss 1986). These models allow complex heterogeneity in arc formation to be represented in an interpretable manner. In particular they can capture social structures such as clustering, transitivity and hierarchy in a parsimonious manner.

Statistical exponential family models can also quantify the strengths of the various social processes most likely to have given rise to the network observed. An important property of the family is that they can represent a dependence between the arcs within a dyad — by which we mean each pair of actors — and those of other dyads.

An exponential random graph model (ERGM) with a transitive triad term is an example of dyad dependent model. An ERGM fit with a significant positive estimate for a transitive triad coefficient, for example, suggests that there are more hierarchal clusters of arcs than we would expect at random. This implies that the network was unlikely to have been generated by a process where each arc forms at random, with regard only to the characteristics of the two actors forming the arc. Instead, this term implies that three arcs that form a transitive triad are more likely than three arcs between random pairs.

In many cases where social network data are available, the full network information is not observed. The network is comprised of the actor's identities, attributes, and arcs with other actors. In many measurement settings some of this information may be known to be missing. Any missing information can influence inference about the processes responsible for generating the social network. For example, consider a directed network where we survey each actor to determine her or his out-arcs. Suppose we are unable to survey one quarter of the actors so that the presence or absence of all their out-arcs is unknown. Suppose we do observe the presence or absence of all arcs sent by actors who are surveyed. If we consider only the network among the subset of actors completing the survey, we have excluded seven sixteenths of the possible arcs in the network, and about one quarter of the out-arcs of each observed actor. When estimating the propensity for friends to have friends in common, this exclusion may strongly influence our conclusions about the importance of this factor in generating the network.

The problem of missing data in social networks has important differences from most missing data problems. To begin with, the unit of analysis in social network data is usually different from the unit of sampling. In general, we sample actors and obtain arc information from them. But network analyses are based on arcs, for which the fundamental unit is a pair of actors. This

sharing of units of analysis across units of sampling, together with most any interesting network model induces a complex dependence structure between the arcs in the network and, in particular, between the arcs that are observed and the arcs that are not observed. In this respect, missing data in social networks is similar to missing data in time series analysis (Little and Rubin 1987, Section 11.6). Also similar to the time series problem is the fact that in social network analysis we generally have only one sample from the process about which we wish to make inference. In time series analysis we wish to characterize the underlying time series process. Similarly, in social network analysis we are often interested in the population-level characteristics of the full network. When there is missing data, we are not merely missing replicates where each sample is an independent observation of the process of interest. We are missing parts of a single realization of a dependent process.

Despite the general acceptance that missing data is an important problem for social network analysis, there has been little work on appropriate systematic model-based frameworks to treat social networks with missing data.

Thompson and Frank (2000) review the closely related literature on network sampling designs based on sequentially sampling individuals nominated by individuals already sampled. They consider estimation from such data when the dyads in the social network are modeled as independent given the characteristics of the individuals.

Kossinets (2006) conducted a descriptive simulation study of the effect of missing data on network statistics. Starting with an existing network, he randomly removed data according to several different patterns and observed a set of network characteristics of interest. With respect to actor non-response, he treats a smaller amount of missing data than we do here, as he considers an undirected network where an arc is counted if reported by either party. Kossinets found that omission of arcs between non-respondents led to underestimation of degree associativity estimates and underestimation of the clustering coefficient.

Some approaches to model-based treatment of missing data in social networks have been suggested, but due to the difficulty of the problem, they are partial and ad-hoc. Stork and Richards (1992) advocate leveraging the strong effect of reciprocity in many networks to impute missing arcs in directed networks by setting them equal to their opposite arcs. For example, if actor i has reported an arc to actor j, but j's reported arcs are unavailable, we assume that j would have reciprocated i's report of an arc. This approach is often more reasonable than treating the arc from j to i as a known non-arc, but is not ideal for several reasons. First, as Stork and Richards point out, the approach is only valid for networks with very strong reciprocity. When reciprocity is not so strong (i nominating j does not strongly predict j nominating i), then this approach may perform worse than pretending the reciprocating arcs do not exist. This approach also treats the newly imputed arcs as true, rather than treating them probabilistically, potentially biasing the estimates. In addition, this approach does not address the arcs that may originate from the missing actors which are not reciprocated, or any arcs between missing actors. Finally, this approach is not applicable to undirected networks.

Robins, Pattison, and Woolcock (2005) use an exponential family model with the maximum pseudo-likelihood estimates (MPLE) of the parameters based on treating arcs between respondents and other respondents separately from arcs from respondents to non-respondents. This approach is most helpful if it is known that the arc-related characteristics of non-respondents are different from those of respondents in ways that are not captured in the terms in the model. However, it does not allow for the consideration of network structures which span the boundary between observed and

unobserved parts of the network, or address the full-network implications of arcs that may have been sent by non-respondents. In addition there is evidence that the MPLE is poor for realistic network structures (van Duijn, Gile and Handcock 2006).

Handcock (2002) developed a likelihood-based framework for the treatment of missing network arcs based on standard ideas of missing data. This extended the work of Thompson and Frank (2000), and we apply his framework here.

The approach of Handcock (2002) addresses the problems with these approaches. He uses an ERGM framework to find the maximum likelihood estimates based on the observed arcs (and non-arcs), through sampling of the missing arcs conditional on the observed arcs. This approach treats the missing arcs probabilistically based on the precise estimates of mutuality and other network characteristics fit from the observed data. Furthermore, it allows for network statistics reflecting probabilistic representations of the full network. By including model terms representing differences between respondents and non-respondents, this approach can be extended to allow for different characteristics of arcs involving respondents and non-respondents.

In this paper we examine the implications of ignoring the actors who did not supply arc information in the modeling of social network data. In Section 2 we describe the National Longitudinal Study of Adolescent Health. This survey has been and will continue to be a primary resource for researchers interested in the health-related behaviors of adolescents. In this paper we develop methodology to adjust for the characteristics of this survey. In Section 3 we begin with exponential family models for networks, we then introduce the particular model developed in the paper, then describe the observed data likelihood approach to fitting an exponential random graph model. Section 4 presents our results, beginning with an analysis of the model fit to all observations through the observed data likelihood.

Then we consider inference for the same model applied to the network of respondents only. We compare the implications of these two approaches in four ways. First, in terms of model coefficient estimates and nominal statistical significance. Next, by comparing the mean value parameterizations of each model fit. Then by simulating complete networks under each model fit and comparing network parameters not in the model to those observed. And finally, by comparing samples of the unobserved arcs conditional on the observed arcs. We end with a discussion of our overall findings, and some extensions of this methodology in Section 5.

2 Introduction to the AddHealth Survey

The social network example we analyze here is from the first wave (1994-5) of the National Longitudinal Study of Adolescent Health (AddHealth). AddHealth includes a stratified sample of US schools including grades 7-12 (Harris et al., 2003). In each school, students were asked many individual questions including, for our purposes, grade and sex. Each student was also presented with a roster of all students in the school and asked to identify up to five of her or his best female friends and up to five of her or his best male friends. There is an extensive and growing literature describing and utilizing the AddHealth survey - see Resnick et al. (1997) and Udry and Bearman (1998) for a bibliography and more information.

These friendship arcs define the friendship network we are modeling. This network includes a directed arc between actor i and actor j if and only if i named j a friend. We represent these arcs as an $n \times n$ matrix, Y where a 1 in the (i, j) cell indicates that i considers j a friend, and n is the

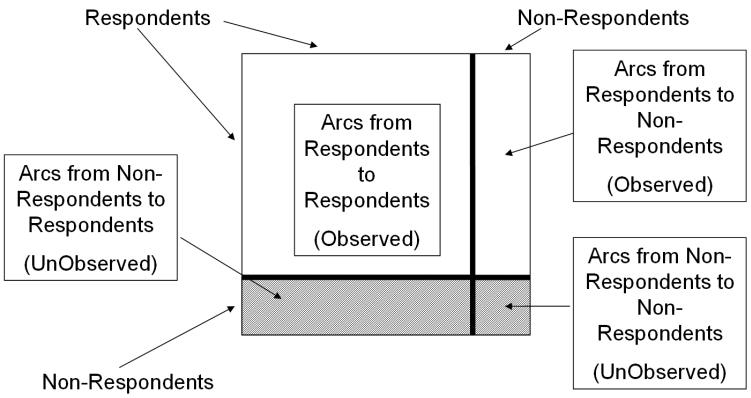

Figure 1: Schematic depiction of observed and unobserved arc data.

number of students in the school. We also consider an $n \times q$ matrix of covariates on the nodes, X. In this example, X includes the grade and sex of each student at the time of the survey.

We have selected one school, School 5, for our analysis. Seventy students from this school completed the friendship nominations portion of the survey. From later waves of the survey, we were able to recover the sex and grade of 19 additional students who did not supply their friendship nominations in the original survey. The students were asked to name up to five of their best male and up to five of their best female friends. The relationship we study here is that of being named one of these friends.

In this paper we consider the friendship nominations among these 89 students to be the focus of scientific interest. In particular we are interested in inferring the social process that generated the observed set of friendship arcs among the 89 students. Of these, 70 reported arcs and 19 did not report arcs. Thus our data contain known arcs and non-arcs between the 70 students who completed surveys, known arcs sent by the 70 respondents to the 19 non-respondents, and do not contain arcs among the 19 students who did not complete surveys and sent by the non-respondents to the respondents. These missing arcs due to survey non-response constitute the missing data we are concerned with.

The data pattern is shown in Figure 1. Consider a partition of respondents from non-respondents and the corresponding 2×2 blocking of the sociomatrix, with the four blocks representing arcs from respondents and non-respondents to respondents and non-respondents. The complete data consists of the full sociomatrix. The first two blocks contain the observed data, the arcs sent by respondents, and the second two blocks contain the unobserved data, those sent by non-respondents.

Almost all network analysis of the AddHealth survey models or describes the network among the respondents only, excluding those individuals who are did not complete the survey (Bearman, Moody, and Stovel 2004; Harris et al., 2003)

3 The Exponential Random Graph Model

Exponential Random Graph Models (ERGM) are a powerful and flexible tool for modeling the behavior of a matrix of social arcs conditional on a matrix of covariates.

An ERGM is an exponential family model in which the data to be modeled is the $n \times n$ matrix of social arcs, and the sufficient statistics are a set of user-defined functions $g(Y, X)$ of the sociomatrix Y and the matrix of covariates X. Models take the form:

$$P_{\boldsymbol{\eta}}(Y = y|X) = c^{-1} \exp\{\boldsymbol{\eta}^T g(y, X)\} \tag{1}$$

where the normalizing constant $c \equiv c(\boldsymbol{\eta})$ is defined by

$$c = \sum_{w} \exp\{\boldsymbol{\eta}^T g(w, X)\} \tag{2}$$

and the sum (2) is taken over the whole sample space of allowable graphs. The statistics $g(Y, X)$ are chosen to capture the hypothesized social structure of the network.

For this study, we fit an ERGM of the form (1) where the set of allowable graphs is restricted to those having no more than five male out-arcs and five female out-arcs for each student. This reflects the nature of the relationship modeled and ensures the distribution is over the appropriate sample space.

3.1 Specification of the Model

We specify a model for the social process in which, $g(y, X)$, the set of network statistics has twenty-one terms. These are summarized in Table 1.

This first term represents the overall tendency for students to nominate friends. This captures the overall density of arcs - that is the number of arcs divided by the number of possible arcs in the network. This characteristic is found in the simplest network model, the Bernoulli or Rényi-Erdős graph. We use the number of arcs in the network as the sufficient statistic for this feature. Note that although this term is based on an arc count, its estimation, like that of the other model terms, implicitly reflects the number of potential arcs in the network, so is comparable across networks of different sizes.

The second term represents the propensity for arcs to be reciprocated. The corresponding sufficient statistic is the number of dyads with arcs in both directions. A positive parameter for this term suggests that students are more likely to nominate friends who nominate them.

The third through seventh terms capture the differential tendencies for students in different grades to be named as friends. The reference category is 7th grade, so the 8th grade popularity term, for example, captures the degree to which the tendency for 8th graders to receive friendship arcs exceeds that of 7th graders. The sufficient statistics for these terms are the counts of in-arcs of students in each grade.

The eighth term captures the tendency for boys to receive arcs, beyond the tendency of girls. A non-zero parameter estimate here suggests that boys and girls have different tendencies to be named as friends.

The following two terms, “girl to same grade boy” and “boy to same grade girl,” capture the relative propensity for same-grade arcs to be sent to a same-sex or opposite sex nominee. These effects are allowed to be different for males and females.

The next six terms address the relative tendencies of students to choose friends who are older or younger, closer or farther from their grade, and same or opposite sex. These tendencies are

operationalized as dyadic covariates. As an example, consider the "girl to older girl" term. This term applies only to dyads from a female in a lower grade to a female in a higher grade. The coefficient is scaled by the number of years between the two grades. If the "girl to older girl" parameter estimate is α, this means that the log odds of an arc from a girl to a girl one grade older is α more than the log odds of an arc from a girl to a same grade girl. The log odds of an arc to a girl two grades older is an additional α more. Using the number of years older or younger as a linear covariate serves as a way to address the decaying likelihood of friendship across increasing grade differences. The covariates measured here are all measured in terms of "absolute number of years difference."

If these terms were the only ones in the ERGM then each dyad would have arcs independently of every other dyad. Together, they capture the propensities for arcs to be formed between senders and receivers of various characteristics, as well the overall rate at which each group tends to receive arcs. The remaining terms address the correlation or clustering behavior among arcs. We know that most arcs are among actors of the same grade and sex, so this is where we focus our attention on the patterns of clustering.

The two triad-based dependence terms address the propensity for arcs to form transitive and cyclical triads respectively among actors of the same grade and sex. These terms are valuable in describing the clustering behavior in the network. A positive parameter for the transitive triad term suggests that friendships within students of the same sex and grade are likely to form in hierarchal patterns, whereby if Anne nominates Betty (giving greater attractiveness, and greater prestige to Betty), and Betty nominates Carol (even greater prestige), then Anne is more likely to nominate Carol as well. Cyclical triads, on the other hand, can be interpreted as an indication of friendships forming on an egalitarian basis. If Anne nominates Betty (making Anne and Betty close and equal), and Betty nominates Carol, then Carol is more likely to nominate Anne. Together, these two terms capture the flavor of the clustering behavior in the observed network. In practice, the transitive term is often positive and the cyclical term often negative (as in this case). This characterization of the clustering process is often a point of great scientific interest in social network research.

The final term captures one additional facet of clustering. Not all nodes have in arcs. None of the other terms have captured this tendency for some people to simply not be nominated as friends. Therefore, we have included a term to explicitly account for the tendency of the network to contain nodes with no in arcs.

Note that all of these terms measure tendencies with respect to the set of realizations that are possible given the network covariates. The "boy to older boy" term, for example, captures the propensity for arcs from a younger to an older boy, with respect to the total older-younger boy dyads in the network of interest. For this reason, it is reasonable to apply the model fit on the smaller network of respondents only to the larger network of all 89 students.

3.2 Inference with Incomplete Observation of the Network

Typically in network analysis, it is assumed we have complete data on the arcs between all actors in the network. Without a way to address missing observations of the arcs, a common approach is to fit a model to the sub-network of actors for whom the out-arcs are observed. We refer to this as the *respondents only* (RO) approach. In our situation this would mean fitting the network model to the complete case social network, consisting of only the 70 students who completed the survey. In this paper we use likelihood-based inference for the model and report maximum likelihood estimates

(MLEs). We fit the model to the respondents only network using a Markov Chain Monte Carlo (MCMC) algorithm to estimate the log-likelihood corresponding to the model (1). We compute approximations to the maximum likelihood estimates and standard errors of the model parameters based on the MCMC log-likelihood as a surrogate for the log-likelihood. Our approach is described in detail in Handcock et al. (2003) and Hunter and Handcock (2006).

An alternative to the respondents only approach is to focus on all actors and base inference on all the observed arcs and non-arcs. To fit the model with all the observations, accounting for the missing data, we use the observed data likelihood. Denote the subset of the sociomatrix corresponding to the out-arcs that are observed by Y_{obs} and the unobserved out-arcs by Y_{miss}. The distribution of the observed data is:

$$P_{\boldsymbol{\eta}}(Y_{obs} = y_{obs}|X) = \sum_{s} P_{\boldsymbol{\eta}}((Y_{obs}, Y_{miss}) = (y_{obs}, s)|X). \tag{3}$$

where s ranges over all values of Y_{miss} consistent with allowable graphs. Handcock (2002) describes the mathematics involved in maximizing the observed data likelihood based on (3). In principle, the approach is based on enumerating the probabilities of all networks that are consistent with the data we've observed. The observed data likelihood is formed by summing these probabilities. Computationally, there are far too many conditional networks to fully enumerate, so we use an MCMC algorithm to sample from the conditional and full networks, and thereby estimate the observed data likelihood corresponding to the model (3). We compute approximations to the maximum likelihood estimates and standard errors of the model parameters based on the MCMC log-likelihood as a surrogate for the log-likelihood. Details are given in Handcock (2002, 2003).

Under a missing at random (MAR) assumption the observed data likelihood is the statistically appropriate likelihood to use based on the full data we have observed. This approach includes the information in all the observations.

We note that standard errors based on the curvature of the estimated log-likelihood and approximations to the sampling distribution of the MLE based on asymptotic arguments require non-standard justifications. In the results below the standard approximation to the sampling distribution is supported by a parametric bootstrap exercise. The analyses here were performed with `statnet` (Handcock et al., 2003).

4 Results

We fit the model to the friendship network using the two approaches outlined in the previous section: the respondents only (RO) approach, the all observations (AO) approach. In both cases we use the same model terms. Mean value parameters for both fits are on the scale of the full set of 89 students. The fits for these models in their natural parameterizations are summarized in Table 2.

In the next sub-sections we describe the interpretation of the All Observations model fit, then explore the implications of the differences between the fits.

4.1 Interpretation of All Observations Fit

In the fit based on the AO approach, all terms are nominally significant at the .01 level except the terms capturing the differential activity by grade and sex, and the terms comparing cross-sex and within-sex attractiveness within the same grade. Tenth graders and males do show nominal significant differences in popularity at the .05 level. This fit supports several scientific hypotheses about the social mechanisms giving rise to this observed network.

First, friendship arcs are reciprocated at a higher rate than we would expect at random given the other terms in the model. With regard to grade, 10th graders seem to receive significantly less friendship nominations than the reference 7th graders, although this finding is weaker than the others.

Males receive within-sex nominations at a nominally higher rate than females. Both males and females seem less likely to nominate friends outside their grades, with the chance of nomination decreasing with the number of years between the two. Looking at the effect sizes for the Sex and Grade Mixing terms together, we note that, although not significant, boys show a stronger aversion to sending cross-sex nominations within grade, while girls show a stronger aversion to sending cross-sex nominations out of grade (to both older and younger boys). We also see that both sexes appear more likely to nominate older (higher grade) rather than younger (lower grade) friends. This effect is stronger in males, with a particularly strong prohibition against males nominating younger males as friends.

The positive significant transitive triad, and negative significant cyclical triad terms suggest that friendship arcs within sex and grade tend to form in a hierarchal manner, rather than in an egalitarian regime. This finding is likely the most scientifically interesting of the processes supported by this model.

Finally, arcs are clustered so as to produce more nodes receiving no friendship nominations than we would expect from the rest of the terms in the model.

4.2 Comparison of the Respondent Only to the All Observation Model Fit

A likelihood ratio can be used to make a heuristic overall comparison between the AO and RO model fits (Hunter and Handcock 2006). The appropriate likelihood for this purpose is the observed data likelihood, including all the nodes and integrating over the unobserved arcs. Thus the appropriate ratio is:

$$\frac{P_{\boldsymbol{\eta}_{AO}}(Y_{obs} = y_{obs}|X)}{P_{\boldsymbol{\eta}_{RO}}(Y_{obs} = y_{obs}|X)} \tag{4}$$

The logarithm of this ratio for the models in Table 2 is about 7. Because we are comparing two fits to the same model, one the MLE, the other arrived at by another method (namely the MLE after throwing away a substantial portion of the data), we cannot apply any standard testing criteria to this ratio. Therefore, this ratio tells us only that the data we have observed is about $\exp(7) \approx 1000$ times more likely under the AO fit than under the RO fit. This seems a substantial difference. If we imagine a Bayesian analysis applying any reasonable prior, we would conclude that the posterior probability of the AO fit is approximately 1000 times greater than the posterior probability of the RO fit. This result suggests that there are overall differences between the RO and AO fits.

At first glance, comparison of the model fits in Table 2 reveals striking similarities among the natural parameters. The fits find nearly identical patterns of statistical significance. With the exception of 9th and 10th grade popularity, a researcher basing conclusions on hypotheses concerning individual model terms would draw the same conclusions from either of these fits. That said, there are also notable differences in the magnitudes of coefficients. In particular, the RO fit reflects a greater popularity of 12th graders, and a lesser tendency for students to receive no arcs. It also suggests a tendency for girls to send less arcs to same-grade boys and less arcs to older boys. The RO fit suggests boys are more likely to send arcs to same-grade girls, less likely to send arcs to older or younger boys, and more likely to send arcs to older girls. The interpretation of these effects is complicated by the many terms in the model. If the RO fit reflects higher overall popularity of 12th graders, do lesser estimates for terms for arcs sent to older students merely reflect that this phenomenon has already been captured by the 12th grade popularity term?

We can better compare the marginal effects of the two fits by comparing the mean value parameterizations of the two fits, as presented in Table 3.

4.3 Comparison of Model Fits under the Mean-Value Parameterization

The mean value parameterization provides an alternative to the natural parametrization of the ERGM model. The mean-value parametrization is (Handcock 2003):

$$\boldsymbol{\mu}(\boldsymbol{\eta}) = \mathrm{E}_{\boldsymbol{\eta}}\left[g(y, X)\right] \tag{5}$$

This parameterization puts the coefficients on the scale of the network statistics rather on the conditional log-odds scale on which the natural parameters are based. Looking at the mean value parameters provides a sense of the implications of the model fit for the network statistics implied by the model fit. Although the RO model is fit on a smaller network, we assume this analysis is in the interest of making inference regarding the "whole school", and report mean value parameters implied by applying the natural parameter estimates to interpolate the network of size 89. This puts both fits on the same scale to allow for meaningful comparisons.

Table 3 shows the MLE of the mean value parameters and their standard errors. To begin with, the expected number of arcs demonstrate that the RO fit implies about 1% more arcs (607) than the AO fit (600), and 6% more reciprocated arcs (144 vs 135). The mean value parameters of other model terms support conclusions suggested by the natural parameters. Under the RO fit, 12th graders receive more arcs (17 %), and fewer students receive no friendship nominations (32% less). Differences in rates of cross-sex nominations within grade are not large. The weighted sum of arcs from girls to older boys is lower (19%). The weighted sums of arcs from boys to older and younger boys are reduced (35% and 31% respectively), and those to older girls are increased (29%). Unexpectedly, the number of transitive and cyclical triads within sex and grade are substantially higher in the RO fit (16% and 21% respectively), although the natural parameter estimates for these terms were nearly identical. Since these terms are focused on arcs within sex and grade, the observed differences are likely due to greater concentration of arcs within sex and grade for the RO fit. This phenomenon is consistent with the relatively higher rate of sex-grade homophilious arcs from respondents to respondents, as opposed to from respondents to non-respondents. Figure 2 compares the proportion of observed in-arcs received from outside one's own sex and grade for respondents and non-respondents of the same sex and grade. Note that for six of the eight

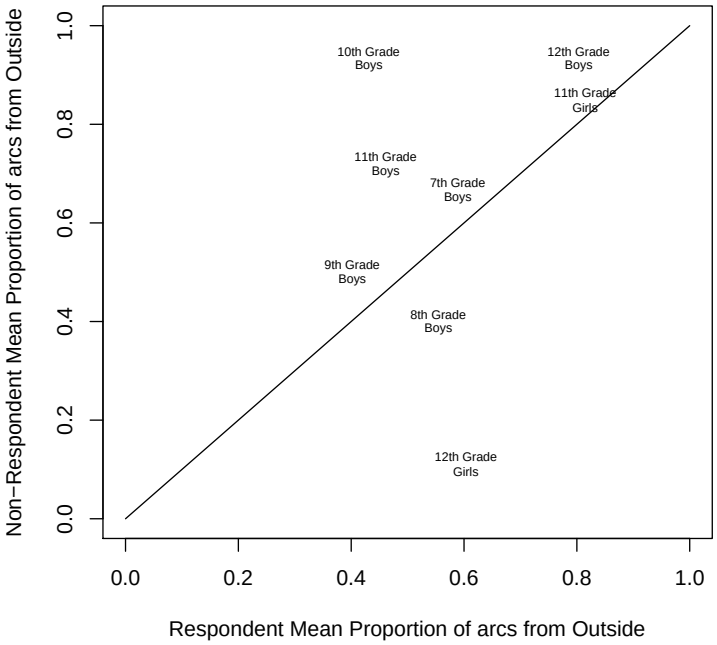

Figure 2: Mean proportion of nominations received from outside sex and grade, by sex and grade.

sex-grades with non-respondents, non-respondents received a higher proportion of nominations from outside their own sex and grade. The greatest exception to this pattern is 12th grade girls, for whom non-respondents receive a lower proportion of nominations from outside their sex and grade than their respondent counterparts. This is consistent with the increased rate of "boy to older girl" nominations, and decreased rate of most other arc types across sex and grade under the RO fit.

The AO approach relies on two types of information not used in the RO approach: the full size of the network, and the additional data in the arcs sent to non-respondents. To help distinguish the effects of these two differences, we fit the same model to a network of size 89×89 with only the respondents to respondents block observed. This fit resulted in mean value parameter estimates almost identical to those of the RO approach. In particular, the mean value parameter estimates for the triad terms are nearly identical to those of the RO fit. In addition, by strategically removing several influential non-respondents, we are able to produce mean values for cross-grade arcs that are nearly identical in the RO and AO approaches. This suggests that most of the differences between the RO and AO fits can be attributed to the additional data available in the AO approach.

4.4 Goodness-of-Fit to Other Structural Properties

Hunter, Goodreau, and Handcock (2005) present a method for evaluating the fit of network models, based on network statistics not modeled directly. They propose comparing the distribution of selected statistics of substantive interest (e.g, degree distribution and shortest path length distribution) to their observed values. They then draw a sample of networks from the model specified by

the MLE, and compare the observed to the sampled distribution of statistics via box plots. The closer the observed statistics are to the middle of the sample distributions, the better the fit of the model. Plots of this sort representing in degree, out degree, and minimum geodesics, and adjusting for missing data did not demonstrate obvious differences between the performance of these two model fits (Figures not shown).

4.5 Goodness-of-Fit to Sub-Group Densities

Consider the partition of respondents from non-respondents and the corresponding four blocks representing arcs from respondents and non-respondents to respondents and non-respondents given in Figure 1. We have observed the first two blocks, the arcs sent by respondents, and these observations provide a basis for comparing the densities of the unknown arcs, sent by non-respondents.

We can use each model to estimate the probabilities of an arc for each entry in the last two blocks conditional on the observed arcs. Conceptually this can be done by simulating graphs from the model conditional on the observed data and averaging over the sociomatrices.

If the non-respondents were equally likely to be any of the 89 students, the expected densities of all four blocks would be the same. To begin with, the block densities are different in the observed portion of the network. Respondents nominate other respondents with density 0.082 and non-respondents with density only 0.062, reflecting different in degrees between respondents and non-respondents. We extend this discussion of differences between respondents and non-respondents in the discussion.

Table 4 summarizes these interpolated densities under the RO and AO fits. The first thing we notice is that both model fits imply a higher density of arcs between non-respondents than in any other block. Some of this effect is due to the greater homogeneity of sex and grade within non-respondents as compared to within the other blocks. To assess the magnitude of this effect, we independently sample 19 students with sex and grade matching those of the non-respondents from each of the 500 conditional samples from each model, and compute the densities of arcs among those students. The non-respondents to non-respondents block is at the 69th percentile for the RO fit, and the 56th percentile for the AO approach. While this does not make them significantly different from others of their sex and grades, they are both slightly elevated. This is consistent with the many terms in the model regulating in degrees, and the strong mutuality parameter in all the models. If a 12th grade boy who is a non-respondent has many fewer arcs than a 12th grade boy who is a respondent, a strong mutuality parameter suggests that the deficit might be accounted for by unobserved mutual arcs between him and other non-respondents.

We also note that the AO fit implies slightly fewer arcs in the unobserved blocks than the RO approach. This is consistent with the slightly higher overall density implied by the RO fit.

5 Discussion

Any treatment of missing data requires researchers to make assumptions. Unfortunately, those assumptions are often made due to methodological limitations rather than scientific conviction. Here, we have demonstrated the implications of the computationally parsimonious respondents only approach to treating social networks with missing data. This example demonstrates that the RO approach can lead to different conclusions from the all observations approach, to a large degree

due to the additional information available in the arcs sent from respondents to non-respondents. Using the method introduced by Handcock (2002), we can do a better job of estimating the model fit for all observed data. By introducing additional terms for any differences between respondents and non-respondents, we can extend this approach to produce even better model fits.

Faced with incomplete observation of a network of interest, a researcher may choose to redefine the network of scientific interest to that formed among the respondents only (RO). Such a reduction is usually the result of a lack of methodology rather than based on scientific principal. Given the choice, few researchers would choose to define the boundary of the network of interest as an artifact of the measurement process rather than based on the phenomena itself.

The RO approach is most defensible in the case of a model assuming independence among all arcs, with non-respondents selected with respect to attributes in the model only (sex and grade). In this case, the RO likelihood is a true likelihood for the portion of the data it considers. Even in this best case, however, the RO approach ignores much of the observed data (28% in this example), resulting in a loss of efficiency.

In this network, we have reason to suspect the non-respondents may differ systematically from the respondents. We have observed the number of in arcs, or in degrees of both groups. The average in degree of respondents is 5.6, while the average in degree for non-respondents is only 4.4. We might imagine this difference is due to the sex and grade composition of the non-respondents. But repeated random sampling of respondents with sex and grade matching those of non-respondents yields average in degree above that of non-respondents in 95% of samples (mean 5.5). This result supports the hypothesis of a systematic difference between respondents and non-respondents.

The AO approach relaxes the assumptions of the RO approach, by appropriately accounting for any missing at random (MAR) patterns captured by any terms in the model. If, for example, the non-respondents differ systematically from respondents in that they have lower prestige, and therefore fewer in arcs, the AO approach will appropriately account for this difference. This approach simultaneously allows us to relax our assumptions about the pattern of missing data, and allows us to make use of all the data we have actually observed. An important remaining assumption of this approach is that the same statistical model applies to arcs involving both respondents and non-respondents.

To examine this assumption, we can sample networks from the AO model fit and examine whether they reproduce the observed data patterns, in particular the differential popularity between respondents and non-respondents. Table 5 summarizes the mean densities of the respondent and non-respondent blocks implied by the AO fit. These densities do not reflect the differential popularity of non-respondents in the observed network.

To capture differences in the models applying to respondents and non-respondents, we can include any estimable term in the AO model. Since we have observed differential popularity of respondents and non-respondents, we might consider including a term for the differential popularity of non-respondents in our model. The natural and mean value parameter estimates under this model are nearly identical to the AO fit, and the term for differential popularity is small, negative, and significant (-0.28, s.e. 0.11), supporting the observation that non-respondents receive systematically less in-arcs. This approach reduces the implied conditionally simulated density of arcs between non-respondents from 0.087 (AO) to 0.072, but does not reduce the implied conditionally simulated density of arcs from non-respondents to respondents (0.070 AO, 0.073 Differential Popularity). Table 5 shows that this model does adequately capture the differential popularity of non-respondents, as unconditional samples reproduce the densities of the observed blocks almost

exactly.

This approach could be extended to the broader class of models including any other terms capturing differences between respondents and non-respondents. This represents the least restrictive set of assumptions for treating these data. Note, however, that additional terms are limited to terms we can estimate based on the observed data. We could not, for example, estimate a term for differential activity level of non-respondents from these data.

In this paper, we have found deficiencies in the respondents only approach to treating missing arc data. Surprisingly, we did not see these deficiencies in all areas of comparison. In particular, the natural parameter estimates were qualitatively very close to the AO approach. For a researcher interested only in these estimates, the RO fit would not negatively impact their analysis in this case. We must stress that we have no reason to expect that this phenomenon would extend to a broader set of cases. In our example, the RO fit performs less well on the mean value parameterization, with some parameters mis-estimated by as much as a third. Furthermore, the RO approach does not allow for extensions to model any differences between respondents and non-respondents. Historically, the RO approach was widely used because there was no methodology or software available to implement the AO approach. With the current availability of software to implement the AO methodology, there is no longer a need to resort to the RO approach.

An `R` package called `statnet` implementing the procedures in this paper will be publicly available at `http://csde.washington.edu/statnet`.

References

van Duijn, M. A. J, K. Gile, M. S. Handcock (2006), Comparison of Maximum Pseudo Likelihood and Maximum Likelihood Estimation of Exponential Random Graph Models, Working Paper, Center for Statistics and the Social Sciences, University of Washington. Available from `http://www.csss.washington.edu/Papers/`

Bearman, P., Moody, J., and Stovel, K. (2004), Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks. *American Journal of Sociology*, 110(1), 44-91.

Frank, O. and D. Strauss (1986), Markov graphs, *Journal of the American Statistical Association*, **81**: 832–842.

Handcock, M. S. (2002), Missing Data for of social networks, Manuscript, Center for Statistics and the Social Sciences, University of Washington.

Handcock, M. S. (2003), Assessing degeneracy in statistical models of social networks, Working Paper no. 39, Center for Statistics and the Social Sciences, University of Washington. Available from `http://www.csss.washington.edu/Papers/`

Handcock, M. S., D. R. Hunter, C. T. Butts, S. M. Goodreau, and M. Morris (2003) `statnet`: An `R` package for the Statistical Modeling of Social Networks
`URL: http://www.csde.washington.edu/statnet`.

Harris, K. M., F. Florey, J. Tabor, P. S. Bearman, J. Jones, and R. J. Udry (2003). The National Longitudinal Study of Adolescent Health: Research design [www document]. Technical report, Carolina Population Center, University of North Carolina at Chapel Hill, Available at: `http://www.cpc.unc.edu/projects/addhealth`.

Holland, P. W. and S. Leinhardt (1981), An exponential family of probability distributions for directed graphs, *Journal of the American Statistical Association*, **76**: 33-50.

Hunter, D. R., S. M. Goodreau, and M. S. Handcock (2005), Goodness of Fit of Social Network Models, Working Paper no. 47, Center for Statistics and the Social Sciences, University of Washington. Available from `http://www.csss.washington.edu/Papers/`

Hunter, D. R. and M. S. Handcock (2006), Inference in curved exponential family models for networks, *Journal of Computational and Graphical Statistics*, in press.

Kossinets, G. (2006), Effects of missing data in social networks, *Social Networks*, **28**, 3, 247–268.

Little, R. J. A., and D. B. Rubin (1987), *Statistical Analysis with Missing Data*, New York: Wiley.

Morris, M. (2003), Local rules and global properties: Modeling the emergence of network structure, pp. 174 – 186 in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, edited by Ronald Breiger, Kathleen Carley, and Philippa E. Pattison. Washington, DC: National Academy Press.

Resnick, M. D., P. S. Bearman, R. W. Blum, et al. (1997), Protecting adolescents from harm. Findings from the National Longitudinal Study on Adolescent Health, *Journal of the American Medical Association*, **278**: 823–832.

Robins, G., P. Pattison, and J. Woolcock (2005), Missing data in networks: exponential random graph (p*) models for networks with non-respondents, *Social Networks*, **26**: 257–283.

Stork, D., W.D. Richards (1992), Nonrespondents in communication network studies: Problems and possibilities, *Group & Organization Management*, **17**: 193–209.

Thompson, S. K. and O. Frank, (2000), Model-based estimation with link-tracing sampling designs, *Survey Methodology*, **26**: 87–98.

Udry, J. R. and P. S. Bearman (1998), New methods for new research on adolescent sexual behavior, in *New Perspectives on Adolescent Risk Behavior*, R. Jessor, ed. New York: Cambridge University Press, pp. 241–269.

Model Term	Description
Density	Overall rate of arc formation, similar to an intercept
Mutuality	Increased propensity for arcs that are reciprocated
Sex and Grade Factors	
Grade 8 Popularity	Propensity for 8th graders to receive arcs, beyond that of 7th graders
Grade 9 Popularity	Propensity for 9th graders to receive arcs, beyond that of 7th graders
Grade10 Popularity	Propensity for 10th graders to receive arcs, beyond that of 7th graders
Grade 11 Popularity	Propensity for 11th graders to receive arcs, beyond that of 7th graders
Grade 12 Popularity	Propensity for 12th graders to receive arcs, beyond that of 7th graders
Male Popularity	Propensity for males to receive arcs, beyond that of females
Sex and Grade Mixing	
Girl to Same Grade Boy	Propensity for girls to send arcs to boys over girls in the same grade
Boy to Same Grade Girl	Propensity for boys to send arcs to girls over boys in the same grade
Girl to Older Girl	From a younger girl to an older girl as the grade difference increases
Girl to Younger Girl	From an older girl to a younger girl as the grade difference increases
Girl to Older Boy	From a younger girl to an older boy as the grade difference increases
Girl to Younger Boy	From an older girl to a younger girl as the grade difference increases
Boy to Older Boy	From a younger boy to an older boy as the grade difference increases
Boy to Younger Boy	From an older boy to a younger boy as the grade difference increases
Boy to Older Girl	From a younger boy to an older girl as the grade difference increases
Boy to Younger Girl	From an older boy to a younger girl as the grade difference increases
Transitivity	
Transitive Same Sex and Grade	Propensity for arcs within students of the same sex and grade to form transitive triads
Cyclical Same Sex and Grade	Propensity for arcs within students of the same sex and grade to form cyclical triads
Isolation	Propensity for students to receive no arcs

Table 1: Description of Model Terms. The “Sex and Grade Mixing” terms for students of different ages are interpreted as the rate of change of log-odds of an arc from an individual in the first group to an individual in the second group as the grade difference increases.

	Respondents Only (RO)	All Observations (AO)	RO s.e.	AO s.e.
Density	−1.555	−1.508	0.19***	0.19***
Mutuality	1.963	1.951	0.21***	0.22***
Sex and Grade Factors				
Grade 8 Popularity	−0.217	−0.171	0.15	0.14
Grade 9 Popularity	−0.330	−0.297	0.16*	0.16
Grade10 Popularity	−0.278	−0.346	0.17	0.16*
Grade 11 Popularity	−0.031	−0.052	0.20	0.19
Grade 12 Popularity	0.061	−0.147	0.20	0.18
Male Popularity	0.461	0.400	0.16**	0.16*
Sex and Grade Mixing				
Girl to Same Grade Boy	0.001	0.172	0.24	0.23
Boy to Same Grade Girl	−0.156	−0.255	0.24	0.23
Girl to Older Girl	−0.959	−0.928	0.19***	0.17***
Girl to Younger Girl	−1.310	−1.300	0.23***	0.22***
Girl to Older Boy	−1.067	−0.882	0.17***	0.14***
Girl to Younger Boy	−1.374	−1.358	0.24***	0.23***
Boy to Older Boy	−1.140	−0.859	0.22***	0.16***
Boy to Younger Boy	−2.081	−1.825	0.41***	0.35***
Boy to Older Girl	−0.520	−0.641	0.14***	0.14***
Boy to Younger Girl	−1.050	−1.102	0.19***	0.19***
Transitivity				
Transitive Same Sex and Grade	0.501	0.505	0.06***	0.05***
Cyclical Same Sex and Grade	−0.994	−1.002	0.20***	0.20***
Isolation	3.051	3.613	0.64***	0.68***

Table 2: Estimated coefficients and standard errors for the parameters of the model fits under the Respondents Only (RO) and All Observations (AO) approaches. $* = p < .05$, $** = p < .01$, $*** = p < .001$

	Respondents Only (RO)	All Observations (AO)	RO s.e.	AO s.e.
Density	607.30	600.56	19.74	20.48
Mutuality	143.79	135.24	10.56	9.79
Sex and Grade Factors				
Grade 8 Popularity	122.84	129.09	11.22	10.67
Grade 9 Popularity	81.14	84.01	9.05	11.10
Grade10 Popularity	95.10	87.34	9.52	10.65
Grade 11 Popularity	119.30	124.93	8.62	9.54
Grade 12 Popularity	102.40	87.54	8.86	9.96
Male Popularity	319.08	324.37	11.89	12.04
Non-Resp Popularity	141.39	137.96	10.86	11.41
Sex and Grade Mixing				
Girl to Same Grade Boy	88.01	90.96	6.64	7.05
Boy to Same Grade Girl	66.80	63.60	6.73	6.75
Girl to Older Girl	31.68	29.87	7.77	6.86
Girl to Younger Girl	19.99	20.49	5.35	5.45
Girl to Older Boy	49.15	60.78	8.01	10.03
Girl to Younger Boy	29.62	28.48	5.89	5.97
Boy to Older Boy	27.18	41.62	6.52	8.82
Boy to Younger Boy	8.06	11.68	3.26	3.89
Boy to Older Girl	68.70	53.17	11.12	9.09
Boy to Younger Girl	34.88	35.57	6.80	7.27
Transitivity				
Transitive Same Sex and Grade	401.29	344.87	68.79	62.09
Cyclical Same Sex and Grade	96.31	79.47	18.73	16.19
Isolation	2.58	3.78	1.52	1.78

Table 3: Estimated mean value parameters and standard errors for the model fits under the Respondents Only (RO) and All Observations (AO) approaches.

	Respondents Only (RO)	All Observations (AO)
Respondents to Respondents	0.082	0.082
Respondents to Non-Respondents	0.062	0.062
Non-Respondents to Respondents	0.071	0.070
Non-Respondents to Non-Respondents	0.092	0.087

Table 4: Interpolated densities of blocks of arcs under each fit, conditional on observed arcs.

	Respondents Only (RO)	All Observations (AO)	Diff Pop. (DP)
Respondents to Respondents	0.076	0.076	0.081
Respondents to Non-Respondents	0.081	0.080	0.062
Non-Respondents to Respondents	0.075	0.072	0.074
Non-Respondents to Non-Respondents	0.099	0.093	0.071

Table 5: Estimated densities of blocks of arcs under each fit unconditional on the observed arcs.