Title

Analysis of networks with missing data with application to the National Longitudinal Study of Adolescent Health

Journal

Journal of the Royal Statistical Society Series C (Applied Statistics), 66(3)

ISSN

0570-4936

Authors

Gile, Krista J
Handcock, Mark S

Publication Date

2017-04-01

DOI

10.1111/rssc.12184

Peer reviewed

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2017)
66, *Part* 3, *pp.* 501–519

Analysis of networks with missing data with application to the National Longitudinal Study of Adolescent Health

Krista J. Gile

University of Massachusetts at Amherst, USA

and Mark S. Handcock

University of California at Los Angeles, USA

[Received September 2015. Final revision July 2016]

Summary. It is common in the analysis of social network data to assume a census of the networked population of interest. Often the observations are subject to partial observation due to a known sampling or unknown missing data mechanism. However, most social network analysis ignores the problem of missing data by including only actors with complete observations. We address the modelling of networks with missing data, developing previous ideas in missing data, network modelling and network sampling. We use several methods including the mean value parameterization to show the quantitative and substantive differences between naive and principled modelling approaches. We also develop goodness-of-fit techniques to understand model fit better. The ideas are motivated by an analysis of a friendship network from the National Longitudinal Study of Adolescent Health.

Keywords: Dependent data; Exponential random-graph model; Missing data; Missingness not at random; Social networks

1. Introduction

Social network data typically consist of a set of n actors and a relational variable $Y_{i,j}$, measured on each ordered pair (i, j), $i, j = 1, \ldots, n$. We focus on binary relationships, for which $Y_{i,j}$ is a dichotomous variable indicating the presence or absence of some relationship of interest, such as communication or friendship. The data $Y = \{Y_{i,j}\}_{i \neq j}$ can be thought of as a graph in which the nodes are actors and the edge set is $\{(i, j) : Y_{i,j} = 1\}$. We consider a case where some $Y_{i,j}$ are unobserved because of out-of-design missingness. The analyses in this paper are concerned with characterizing the structure of friendships between high school students, measured as part of the National Longitudinal Study of Adolescent Health (Harris *et al.*, 2003).

1.1. The National Longitudinal Study of Adolescent Health

The National Longitudinal Study of Adolescent Health is a school-based longitudinal study of the health-related behaviours of adolescents and their outcomes in young adulthood. The study design sampled 80 high schools and 52 middle schools from the USA, representative with respect to region of country, urbanicity, school size, school type and ethnicity (Harris *et al.*,

Address for correspondence: Krista J. Gile, Department of Mathematics and Statistics, University of Massachusetts at Amherst, Amherst, MA 01003-9305, USA.
E-mail: gile@math.umass.edu

 0035–9254/17/66501

2003). In 1994–1995 an in-school questionnaire was administered to a nationally representative sample of students in grades 7–12. In addition to demographic and contextual information, each respondent was asked to nominate up to five boys and five girls within the school whom they regarded as their best friends. Thus each student could nominate up to 10 students within the school (Udry, 2003). This referral structure results in directed network data, where an *arc* or directed tie is said to exist from node i to node j if and only if i named j a friend. We define 'friend' to be one of a student's top five male or top five female friends, and we conduct all our analyses restricting networks and models to networks that are constrained by this definition. There is an extensive and growing literature describing and utilizing the survey—see the references of Resnick *et al*. (1997) and Udry and Bearman (1998) for a bibliography and more information.

We have selected one school, school 5, for our analysis. 70 students from this school completed the friendship nominations portion of the survey. From later waves of the survey, we could recover the sex and grade of 19 additional students who did not supply their friendship nominations in the original survey. In this section we consider the friendship nominations between 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 and known arcs sent by the 70 respondents to the 19 non-respondents. They do not include information on arcs between the 19 students who did not complete surveys or sent by the non-respondents to the respondents. Hence of the 7832 potential nominations $19 \times 88 = 1672$, or 21%, were unobserved. These missing arcs due to survey non-response constitute the missing data that we are concerned with.

The structure of the relations is usually dependent on the attributes of the actors. For example, for most social relations the likelihood of a relationship is a function of the age, gender, geography and race of the individuals. *Homophily* on attributes, or the tendency for like to share ties with like, is a common example (McPherson *et al*., 2001). In the adolescent friendships in our application, the social structure is highly dependent on class grade (grades 7–12) and sex. In addition to exogenous attributes of the actors, relationships are influenced by endogenous attributes such as their positions in the network (White *et al*., 1976). In the adolescent health data, we are particularly interested in examining the hierarchical or egalitarian structures of the friendship nominations, which we study by using endogenous structures.

1.2. Modelling networks with missing data

In this paper we consider the network over the set of actors to be the realization of a stochastic process and we model the process. The statistical modelling of such processes has a long history. Holland and Leinhardt (1981) appear to be the first to have proposed log-linear models for social networks. Their models resulted in each dyad—by which we mean each pair of actors—having edges independently of every other dyad. Frank and Strauss (1986) generalized to the case in which dyads exhibit a form of Markovian dependence: two dyads are dependent, conditionally on the rest of the graph, only when they share a node. Such exponentially parameterized random-graph models have connections to a broad array of literatures in many fields, such as spatial statistics, statistical exponential families and statistical physics (Geyer and Thompson, 1992). Since that time there have been many theoretical and applied developments (Lusher *et al*., 2012).

The analysis of sampled or missing data in networks is special for two reasons. First, we are often interested in models in which variables on all units of analysis are dependent. Thus, instead of inference from multiple independent observations of a given process, standard network modelling is based on a single observation of a dependent process. In this way, network modelling is similar to time series modelling. When the network is only partially observed, inference

must therefore be conducted on a single partially observed realization of the process. Second, networks consist of two fundamental units: nodes and dyads. In our framework, we consider the dyadic relations to be stochastic, and the simplest units of inference. Sampling and missing data processes, however, often act on the nodes, as much network data are collected through egocentric reporting processes (e.g. people reporting their own relations). Therefore, the units of inference reside between the units of observation. This is the pattern that is observed in our application, where missing friendships result from some students not completing the survey. Thus data are missing in highly dependent blocks, where all nominations of each non-respondent are unobserved.

Many practical settings result in missing network data. In this paper, we address what is perhaps the most common pattern: when dyads are observed through their incident nodes, and an attempted census of network nodes fails to reach some nodes, leaving some dyads unobserved. In human social networks including the National Longitudinal Study of Adolescent Health, such non-observation could be due to response refusal, absence or illness. Other forms of missing data in networks may result from non-observation of individual dyads or of nodal covariates. Despite the general acceptance that missing data are an important problem for social network analysis, there has been little work on inferential frameworks to treat social networks with missing data.

Some approaches to model-based treatment of missing data in social networks have been suggested but, because of the difficulty of the problem, they typically rely on special cases and assumptions. Stork and Richards (1992) advocated leveraging the strong effect of reciprocity in many networks to impute missing arcs, or directed edges, in directed networks by setting them equal to their opposite arcs, such that, if the relation from j to i is unobserved, it is set equal to that from i to j whenever the latter is observed. This approach is often more reasonable than treating the arc from j to i as a known non-arc, but it is not ideal for several reasons. First, as Stork and Richards pointed out, the approach is valid only for directed networks with very strong reciprocity. When reciprocity is not so strong (i nominating j does not strongly predict j nominating i), this approach may perform worse than pretending that the reciprocating arcs do not exist. This approach also treats the newly imputed arcs as true, rather than treating them probabilistically. 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.

The first model-based approach to networks with missing data was introduced by Robins *et al.* (2004), who used an exponential family model with the maximum pseudolikelihood estimates 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 allow for models that are applicable to the full populations of possible arcs. There is also evidence that the maximum pseudolikelihood estimator performs poorly for realistic network structures (van Duijn *et al.*, 2009).

The most systematic treatment of missing data in networks to date has been provided by Koskinen *et al.* (2013). They considered a Bayesian approach for missing tie variables and covariates, allowing for inference based on the full posterior of the parameters, as well as predictive inference for the unobserved parts of the network.

Our companion paper, Handcock and Gile (2010), building on Thompson and Frank (2000), developed a likelihood-based framework for the full network modelling of networks that are partially observed because of sampling. In this paper, we use the likelihood-based approach

of Handcock and Gile (2010) and Koskinen *et al*. (2013) while focusing on broader issues of data analysis, including goodness-of-fit diagnostics and leveraging the network data structure to address systematic patterns of missing data.

It is also worth mentioning a related area of research: techniques for sharing social network data that protect sensitive personal information privacy while retaining key statistical information. Karwa *et al*. (2014, 2015) have developed an approach to share synthetic networks with perturbed ties where the perturbation mechanism is carefully designed by the researcher to meet these differential privacy goals. Their statistical techniques are similar in approach to those developed by Handcock and Gile (2010) and this paper. The problem is substantially different, however, in that the perturbation mechanism is fully known whereas none of the data elements are known with certainty.

Following this introduction, in Section 2, we introduce several types of missing data in social networks. In Section 3, we review the general principles that are involved in fitting models to social networks with missing data. Section 4 introduces the widely used and powerful exponential family random-graph model class for networks and discusses the fitting of these models for networks with missing data. The approaches in this section are available in the `statnet` R package (Handcock *et al*., 2003).

Finally, we use these theoretical pieces to present an analysis of adolescent friendship data in Section 5, including the introduction of several descriptive and diagnostic approaches for partially observed network data. We finish with a discussion.

The code to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Missing data structures

We treat missing data as a special case of sampling in which the sampling mechanism is unknown and outside the control of the researcher, or an *out-of-design non-response mechanism*. As in Handcock and Gile (2010), we use the N-vector $\mathbf{S}$ to indicate the observed status of each node, and the $N \times N$ matrix D to represent the observed status of each directed dyad, such that

$$\begin{aligned} S_i &= \begin{cases} 1 & \text{node } i \text{ observed,} \\ 0 & \text{otherwise,} \end{cases} \\ D_{ij} &= \begin{cases} 1 & \text{relation } i, j \text{ observed,} \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{1}$$

We focus on the situation where the sampling design specifies $\mathbf{S} = \mathbf{1}$ and $D_{ij} = 1$ for all $i \neq j$, i.e. the researchers intended to observe all nodes and relations. In the case of missing data, however, the observed values of $\mathbf{S}$ and D are jointly determined by the sampling and missing data mechanisms, such that $D_{ij} = 0$ for some $i \neq j$. In this section, we explicitly describe several possible missing data mechanisms.

2.1. Non-responding nodes

Often a census of the network is attempted via a complete census of the nodes followed by the observation of all edges that are incident to each node. This is so, for example, in networks between people where each person is asked to report her relations to all others in the network: a sampling design corresponding to a nodal census. However, if some nodes do not respond, many ties variables will be missing. In an undirected network in which we observe all dyads that are incident to each observed node,

$$D = \mathbf{S}\mathbf{1}^{\mathrm{T}} + \mathbf{1}\mathbf{S}^{\mathrm{T}} - \mathbf{S}\mathbf{S}^{\mathrm{T}}.$$

For a directed network in which we observe all dyads originating at an observed node, $D = \mathbf{S}\mathbf{1}^{\mathrm{T}}$. Note that both of these mappings are identical to the case of sampling that was considered in Handcock and Gile (2010). Unlike in the case of sampling, the missing data mechanism ϕ is typically unknown, even up to a model class. In the simplest case, we might consider a missing data mechanism corresponding to a simple random sample of nodes whereby

$$P(D = d|Y, \psi) = \psi^{\mathbf{1}^{\mathrm{T}}\mathbf{s}}(1-\psi)^{n-\mathbf{1}^{\mathrm{T}}\mathbf{s}}.$$

More complex mechanisms include functions depending on nodal characteristics or observed network features, or those depending on unobserved features. Such missing data structures can also result from *link tracing* designs, where the intention is to sample the contacts of all previously sampled nodes, but the referral and contact process does not reach each contact. When this structure is by design, as in partial wave link tracing samples, the data are not missing.

2.2. Unobserved (directed) dyads

Particular dyads or directed pairs may also be unobserved, even if their incident nodes are observed. A particularly difficult form of this pattern is when some edges are observed, but few or no non-edges are observed. This is often so in protein interaction networks, where references tend to report observed protein interactions, but not tests for interaction with negative results.

2.3. Partially observed nodal attributes

Often, observing a node implies observation of associated nodal covariate information, as in self-report surveys. There is also sometimes non-response on individual items, even for observed nodes. It is also possible that nodal covariate information is available even for nodes that are not sampled, as in administrative databases on well-defined populations.

2.4. Boundary specification

All the examples thus far assume that the set of actors is well defined and the number of actors in the population is known. Often it is unclear which nodes should be considered part of the population of interest. Such cases are beyond the realm of models that are currently in standard use, and also beyond the scope of this paper. In our analyses, we assume that we know the exact set of nodes in the network, but that some of the dyads are unobserved.

2.5. Frameworks for analysis

We focus on two strategies for inference, which we refer to as *complete-case* (CC) and *all observations* (AO) analysis. In CC analysis, only nodes with fully observed data are considered. This approach is also referred to as *subnet* analysis, as, in this case, only the subnet that is induced by the nodes with full available information is analysed. The advantage of this approach is that is does not require any special software for missing data. However, it ignores both the larger size of the full network to which we wish to apply a model, as well as any additional information that is available on the cases that are not complete. Shalizi and Rinaldo (2013) showed that analyses based on subnetworks cannot be consistently applied to the true larger network, so this naive approach is not statistically principled. Therefore, researchers who are interested in finding a principled model fit for the true full network, and interested in using all available data,

should prefer the more principled approach, which we call the AO approach. In this approach, all available data are used in the analysis, including all observed relations, the population size and all known nodal characteristics. The majority of our paper treats these two inferential approaches and compares the resulting model fits, as these are the approaches that are most likely to be employed in practice.

We also include two more specialized model fits. First, for comparison, we also consider an *incomplete-case* (IC) approach, which fits a model over the full known size and nodal covariates of the network, but treating all dyads involving non-respondents as unobserved. This fit helps us to distinguish the separate effects of the additional data and the larger network size on the differences between CC and AO fits. Finally, we apply a *differential popularity* (DP) analysis to model observed systematic differences between respondents and non-respondents in our data directly, in partial adjustment for irregularities in the missing data process.

3. Principled likelihood inference for partially observed networks

Our development here follows the development in Handcock and Gile (2010), which followed Little and Rubin (2002) and Thompson and Frank (2000). For most of this treatment, we consider the case of fully observed covariates information X. Consider a parametric model for the random relational matrix Y, depending on a parameter p-vector $\boldsymbol{\eta}$ and an $N \times q$ matrix of nodal covariates X:

$$P_{\boldsymbol{\eta}}(Y = y|X = x), \qquad \boldsymbol{\eta} \in \Xi, \quad y \in \mathcal{Y}(x), \tag{2}$$

where Ξ is the space of possible parameter values $\boldsymbol{\eta}$ and $\mathcal{Y}(x)$ is the set of possible networks on the n actors with covariates $X = x$. In the model-based framework, if Y and X are completely observed, inference for $\boldsymbol{\eta}$ can be based on the likelihood

$$L[\boldsymbol{\eta}|Y_{\text{obs}}, X] \propto P_{\boldsymbol{\eta}}(Y = Y_{\text{obs}}|X = x).$$

This situation has been considered in detail in Hunter and Handcock (2006) and the references therein. In the general case where Y may be only partially observed, we denote the observed part of Y by $Y_{\text{obs}} = \{Y_{ij} : D_{ij} = 1\}$ and the unobserved part by $Y_{\text{mis}} = \{Y_{ij} : D_{ij} = 0\}$; then $Y = \{Y_{\text{obs}}, Y_{\text{mis}}\}$. The *complete data*; $\{Y_{\text{obs}}, Y_{\text{mis}}, D\}$, are not fully observed, and the *observed data* are $\{Y_{\text{obs}}, D\}$. Following Handcock and Gile (2010), we make the convention that undefined numbers act as identity elements in addition and multiplication, such that $Y = Y_{\text{obs}} + Y_{\text{mis}}$. Letting lower-case symbols represent the observed values of random variables, we let $\mathcal{Y}(y_{\text{obs}}, x) = \{v : y_{\text{obs}} + v \in \mathcal{Y}(x)\}$ represent the set of possible values of Y_{mis}, consistent with y_{obs}. Then $y_{\text{obs}} + \mathcal{Y}(y_{\text{obs}}, x)$ is the subset of $\mathcal{Y}(x)$ that is consistent with y_{obs}.

If the missing data mechanism is missingness at random (MAR) (Rubin, 1976), in the sense that

$$P(D = d|Y = y, X = x; \psi) = P(D = d|Y_{\text{obs}} = y_{\text{obs}}, X = x, \psi) \qquad \text{for all } y \in y_{\text{obs}} + \mathcal{Y}(y_{\text{obs}}, x), \tag{3}$$

and the parameters ψ and $\boldsymbol{\eta}$ are distinct, then the likelihood for $\boldsymbol{\eta}$ and ψ is

$$L[\boldsymbol{\eta}, \psi|Y_{\text{obs}} = y_{\text{obs}}, D = d_{\text{obs}}, X = x] \propto L[\psi|D = d_{\text{obs}}, Y_{\text{obs}} = y_{\text{obs}}, X = x]\, L[\boldsymbol{\eta}|Y_{\text{obs}} = y_{\text{obs}}, X = x]$$

Thus likelihood-based inference for $\boldsymbol{\eta}$ from $L[\boldsymbol{\eta}, \psi|Y_{\text{obs}}, D, X = x]$ will be the same as likelihood-based inference for $\boldsymbol{\eta}$ based on $L[\boldsymbol{\eta}|Y_{\text{obs}}, X = x]$. The latter is typically easier to compute:

$$L[\boldsymbol{\eta}|Y_{\text{obs}} = y_{\text{obs}}, X = x] \propto P(Y_{\text{obs}} = y_{\text{obs}}|\boldsymbol{\eta}, X = x) = \sum_{v \in \mathcal{Y}(y_{\text{obs}}, x)} P_{\boldsymbol{\eta}}(Y = y_{\text{obs}} + v|X = x).$$

Hence we can evaluate the likelihood by just enumerating the full data likelihood over all possible values for the missing data.

4. Exponential family random-graph models

We model the random behaviour of Y by using an exponential family random-graph model. The standard exponential family model form is

$$P_{\boldsymbol{\eta}}(Y = y|X = x) = \exp\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z}(y|X = x) - \kappa(\boldsymbol{\eta}, x)\} \qquad y \in \mathcal{Y}(x), \tag{4}$$

where $\mathbf{Z}(Y)$ is a p-vector of statistics, $\boldsymbol{\eta} \in R^p$ is a parameter vector and

$$\exp\{\kappa(\boldsymbol{\eta}, x)\} = \sum_{u \in \mathcal{Y}(x)} \exp\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z}(u|X = x)\}$$

is the normalizing constant (Barndorff-Nielsen, 1978).

A wide range of network statistics could be included in $\mathbf{Z}(y|X = x)$ (Lusher *et al.*, 2012). In the network modelling literature these are referred to as exponential family random-graph models (Hunter and Handcock, 2006). We allow the vector $\mathbf{Z}(y|X = x)$ to include covariate information about nodes or dyads in the network in addition to information that is derived directly from the matrix y itself.

In a sampling-focused companion paper, we addressed the fitting of exponential family random-graph models for partially observed network data with an MAR structure (Handcock and Gile, 2010) using a natural extension of this method. Consider that

$$L[\boldsymbol{\eta}|Y_{\mathrm{obs}} = y_{\mathrm{obs}}|X = x] \propto \exp\{\kappa(\boldsymbol{\eta}|y_{\mathrm{obs}}, x) - \kappa(\boldsymbol{\eta}, x)\}, \tag{5}$$

where

$$\exp\{\kappa(\boldsymbol{\eta}|y_{\mathrm{obs}}, x)\} = \sum_{u: u + y_{\mathrm{obs}} \in \mathcal{Y}(x)} \exp\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z}(u + y_{\mathrm{obs}}|X = x)\}$$

is the normalizing constant of the conditional distribution of $Y_{\mathrm{mis}}|Y_{\mathrm{obs}}, X = x$:

$$P_{\boldsymbol{\eta}}(Y_{\mathrm{mis}} = y|Y_{\mathrm{obs}} = y_{\mathrm{obs}}, X = x) = \exp\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z}(y + y_{\mathrm{obs}}|X = x) - \kappa(\boldsymbol{\eta}|y_{\mathrm{obs}}, x)\}, \qquad y \in Y(y_{\mathrm{obs}}, x). \tag{6}$$

To find the maximum likelihood estimate (MLE), we therefore maximize an estimate of expression (5) computed as the ratio of the two normalizing constants. The $\exp\{\kappa(\boldsymbol{\eta}, x)\}$ term is estimated by using unconditional samples of Y, as in standard exponential family random-graph model fits, whereas $\exp\{\kappa(\boldsymbol{\eta}|y_{\mathrm{obs}}, x)\}$ is estimated by conditionally sampling from $Y_{\mathrm{mis}}|Y_{\mathrm{obs}}$ and $X = x$ according to equation (6) (Geyer and Thompson, 1992; Hunter and Handcock, 2006). This is implemented in the `statnet` R package (Handcock *et al.*, 2003).

5. Analysing adolescent friendship networks

We consider four modelling approaches for the National Longitudinal Study on Adolescent Health adolescent friendship network with missing data. We begin by describing the pattern of missing data. We then introduce the common model that is used in all the approaches. We next compare the approach that was presented in Section 4 with a naive approach, modelling the subnetwork consisting of respondents only (a CC approach), and we use several methods to estimate the magnitude of the difference between the two approaches. We illuminate both numerical and substantive differences. We illustrate some diagnostic procedures for partially testing the MAR assumption and introduce a third modelling approach partially correcting for

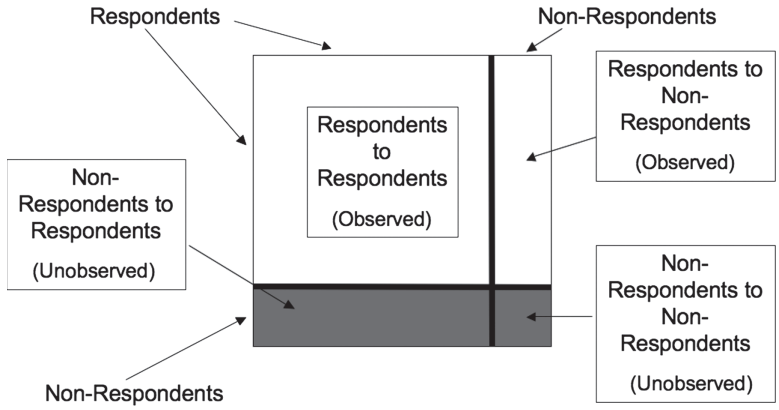

Fig. 1. Schematic depiction of observed and unobserved arc data

the failure of the MAR assumption. We also include a fourth modelling approach to explain some differences between the first two.

5.1. Missing data pattern

The data pattern is shown in Fig. 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 consist of the full sociomatrix. The first two blocks contain the observed data (the arcs that are sent by respondents), and the second two blocks contain the unobserved data (those sent by non-respondents).

Almost all analysis of the adolescent health network data uses the *CC* approach, treating the network among the respondents only, excluding those who did not complete the survey (Bearman *et al.*, 2004; Harris *et al.*, 2003), and corresponding to considering only the upper left-hand block of Fig. 1. We can also visualize the excluded data by plotting the network both including and excluding the non-respondents, and then plotting only the arcs to non-respondents as in Fig. 2. For clarity, the positions of the nodes are the same in each plot.

5.2. Model specification

We specify an exponential random-graph model for the social process in which $\mathbf{g}(y, X)$, the set of network statistics, has 21 terms. The first term, named *density*, captures the overall tendency for edges in the network. The corresponding sufficient statistic is the total number of arcs: $\Sigma_{i \neq j} y_{ij}$. In an exponential family random-graph model, this term has a role that is similar to the intercept in a regression model. The next term, *mutuality*, captures the tendency for arcs to be reciprocated and has sufficient statistic $\Sigma_{i<j} y_{ij} y_{ji}$. The next seven terms capture the differential tendency for nodes of different classes to receive arcs. Grade 7 females serve as the reference category. The additional tendency for grade 8 females to receive arcs is given by the *grade 8 popularity* term with sufficient statistic

$$\sum_{i \neq j} y_{ij} \mathbb{I}(\text{grade}_j = 8), \tag{7}$$

where $\mathbb{I}(k)$ is the indicator function taking the value 1 when k is true, and 0 otherwise. The remainder of the grade popularity terms are defined similarly, and the *male popularity* term has sufficient statistic

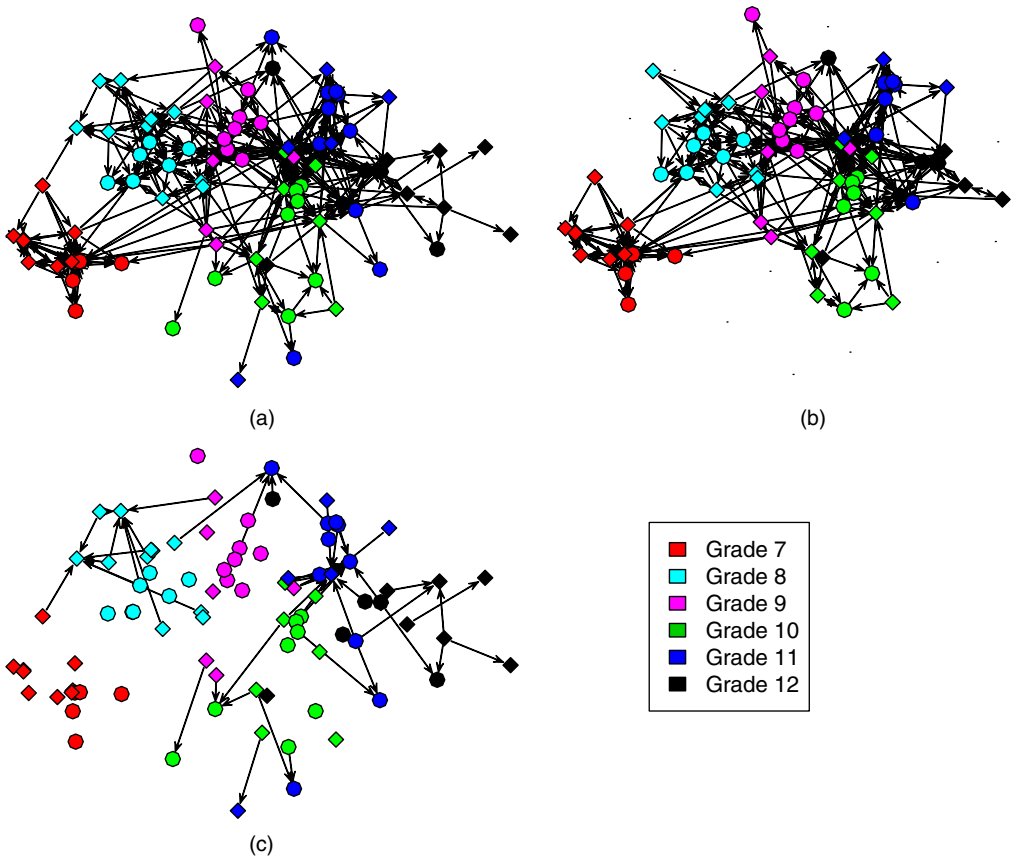

Fig. 2. Depiction of the data excluded by the CC analysis (the positions of the nodes are the same in each subplot; node colour represents grade and shape represents sex (◇, female): (a) all the observed data; (b) only the relations between respondents (the data considered in a CC analysis); (c) difference between (a) and (b)

$$\sum_{i \neq j} y_{ij} \mathbb{I}(\text{sex}_j = \text{'male'}). \tag{8}$$

We leave the definition of the *non-respondent popularity* term to Section 5.5.

The *sex* and *grade mixing* terms capture the differential tendencies for arcs across sex and grade classes, and respect the potentially asymmetrical patterns of these relations. The *girl to same grade boy* term has sufficient statistic

$$\sum_{i \neq j} y_{ij} \mathbb{I}(\text{grade}_i = \text{grade}_j, \text{sex}_i = \text{'female'}, \text{sex}_j = \text{'male'}) \tag{9}$$

and captures the differential tendency for female students to send arcs to males in the same grade, against the reference of females in the same grade. The *boy to same grade girl* term is similarly defined.

The remaining terms in this set capture linear functions of grade differences. For example, the *girl to older girl* term has sufficient statistic

$$\sum_{i \neq j} y_{ij} |\text{grade}_j - \text{grade}_i| \mathbb{I}(\text{grade}_i < \text{grade}_j, \text{sex}_i = \text{sex}_i = \text{'female'}), \tag{10}$$

such that the corresponding parameter indicates the change in tendency for arcs corresponding

to each one-grade difference between two female students. This parameterization assumes that each 1-year difference in grade has the same effect on friendship formation. The remainder of the terms in this section are similarly defined.

The *transitivity* terms are a measure of hierarchy or equality in social networks. Having captured the hierarchical tendencies across sex and grade in the sex and grade mixing terms, here we consider the transitivity effects within sex and grade only. The *transitive same sex and grade* term is based on *transitive triad* structures. If arcs are sent in a hierarchical manner, then if B names A (making A above B), and C names B (making B above C), then the triad is likely to be completed in a transitive manner with C naming A, since by transitivity A would be above C. The sufficient statistic capturing this structure is

$$\sum_{i\neq j\neq k} y_{ij}y_{jk}y_{ik}\mathbb{I}(\text{grade}_i=\text{grade}_j=\text{grade}_k,\text{sex}_i=\text{sex}_j=\text{sex}_k). \tag{11}$$

Similarly, a *cyclical triad* structure is indicative of an egalitarian relational structure. Here, B $\rightarrow$ A and C $\rightarrow$ B make A, B and C about equal, making cyclical triadic completion (A $\rightarrow$ C) likely. These structures are captured with sufficient statistic

$$\sum_{i\neq j\neq k} y_{ij}y_{jk}y_{ki}\mathbb{I}(\text{grade}_i=\text{grade}_j=\text{grade}_k,\text{sex}_i=\text{sex}_j=\text{sex}_k). \tag{12}$$

The final term, *isolation*, captures the tendency for some nodes to receive no friendship nominations, beyond what would be expected given the rest of the terms in the model. This term is based on the sufficient statistic

$$\sum_i \mathbb{I}\Big(\sum_j y_{ji}=0\Big). \tag{13}$$

We estimate the MLE of the parameters of this model conditionally on the restriction in the data that no node may nominate more than five female friends or five male friends.

5.3. Model fit

The parameter estimates under the AO approach are summarized in the second and sixth columns of Table 1. All terms are nominally significant at the 0.01-level except the terms capturing the differential popularity by grade and sex, and the terms comparing cross-sex and within-sex popularity within the same grade. Ninth and 10th graders and males do show nominally significant differences in popularity at the 0.05-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, ninth and 10th graders receive significantly fewer friendship nominations than the reference seventh 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 class years. 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. 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 hierarchical manner, rather than in

Table 1. Estimated coefficients and standard errors for the parameters of the model fits under the AO, CC, DP and IC approaches

	Results for the following approaches:							
	AO	*CC*	*DP*	*IC*	*AO standard error*	*CC standard error*	*DP standard error*	*IC standard error*
Density	−1.929	−1.557	−1.901	−1.923	0.19†	0.19†	0.19†	0.20†
Mutuality	1.728	1.963	1.726	1.854	0.19†	0.20†	0.19†	0.21†
Sex and grade factors								
Grade 8 popularity	−0.161	−0.218	−0.144	−0.402	0.12	0.12	0.12	0.15‡
Grade 9 popularity	−0.301	−0.330	−0.324	−0.353	0.14§	0.14§	0.14§	0.16§
Grade 10 popularity	−0.318	−0.277	−0.303	−0.374	0.14§	0.14§	0.14§	0.16§
Grade 11 popularity	−0.043	−0.033	0.027	−0.042	0.17	0.18	0.18	0.24
Grade 12 popularity	−0.095	0.062	−0.010	−0.175	0.16	0.17	0.16	0.20
Male popularity	0.407	0.461	0.452	0.504	0.16§	0.16‡	0.16‡	0.21§
Sex and grade mixing								
Non-respondent popularity	—	—	−0.313	—	—	—	0.12‡	—
Girl to same grade boy	0.193	0.001	0.175	0.074	0.21	0.23	0.22	0.29
Boy to same grade girl	−0.217	−0.155	−0.231	−0.078	0.22	0.23	0.21	0.25
Girl to older girl	−0.956	−0.959	−0.962	−1.115	0.16†	0.18†	0.16†	0.23†
Girl to younger girl	−1.318	−1.308	−1.334	−1.340	0.21†	0.21†	0.21†	0.25†
Girl to older boy	−0.901	−1.066	−0.906	−1.069	0.14†	0.17†	0.14†	0.20†
Girl to younger boy	−1.326	−1.375	−1.339	−1.894	0.22†	0.23†	0.22†	0.38†
Boy to older boy	−0.876	−1.137	−0.885	−0.943	0.15†	0.21†	0.15†	0.22†
Boy to younger boy	−1.789	−2.082	−1.807	−2.696	0.33†	0.40†	0.33†	0.71†
Boy to older girl	−0.680	−0.521	−0.683	−0.533	0.14†	0.14†	0.14†	0.16†
Boy to younger girl	−1.114	−1.048	−1.125	−0.959	0.17†	0.17†	0.17†	0.17†
Transitivity								
Transitive same sex and grade	0.502	0.502	0.497	0.477	0.05†	0.06†	0.05†	0.05†
Cyclical same sex and grade	−0.913	−0.995	−0.891	−0.865	0.18†	0.20†	0.18†	0.21†
Isolation	2.664	3.059	2.355	3.617	0.90†	0.62†	0.94†	0.71†

†$p < 0.001$.
‡$p < 0.01$.
§$p < 0.05$.

an egalitarian regime. This finding is probably the most scientifically interesting of the processes supported by this model.

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

The third and seventh columns of Table 1 present the corresponding CC model fit. It is of interest to compare the AO and CC models. The question here is the effect of erroneously using the CC model when the full network model is of interest. A natural way to compare the models is the Kullback–Leibler divergence of the CC model from the AO model when both are used to model the CC subnetwork. Specifically, we can consider the probability distribution that the AO model for the full network places over the CC network dyads, $P_{\eta_{\mathrm{AO}}}(Y^{\mathrm{CC}}|X)$, and use it to compute the Kullback–Leibler divergence:

$$\mathbb{E}_{\eta_{\mathrm{AO}}}\left[\log\left\{\frac{P_{\eta_{\mathrm{AO}}}(Y^{\mathrm{CC}}|X)}{P_{\eta_{\mathrm{CC}}}(Y^{\mathrm{CC}}|X)}\right\}\right] \tag{14}$$

where $P_{\eta_{\mathrm{AO}}}$ is the AO model for the full network, $P_{\eta_{\mathrm{CC}}}$ is the CC model for the CC subnetwork and Y^{CC} is the set of dyads from the full network in the CC subnetwork. The method to compute this divergence is given in Appendix A. The value of the divergence is 159. The large magnitude of this divergence indicates that the AO and CC models are substantially different representations of the CC subnetwork (Cover and Thomas, 2006).

Because the AO and CC approaches fit models to different networks, with different sets of nodes, we know from Shalizi and Rinaldo (2013) that a direct comparison of the natural parameters of the two approaches is not valid: the interpretations of coefficients are different in different node set contexts. Nonetheless, researchers often draw substantive conclusions based on the magnitudes and significances of model coefficients, so it is of interest to compare the conclusions that might be drawn by researchers using one approach or the other. At first glance, a comparison of the model fits in Table 1 reveals striking similarities between the natural parameters for these two approaches. The fits find nearly identical patterns of statistical significance. A researcher basing conclusions on the sign and significance of individual model terms would draw nearly the same conclusions from either of these fits. That said, there are also notable differences in the magnitudes of coefficients. In particular, the CC fit reflects a greater popularity of 12th graders, and a greater tendency for students to receive no arcs. It also suggests a tendency for girls to send fewer arcs to same grade boys and fewer arcs to older boys. The CC fit suggests that 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 CC fit reflects higher overall popularity of 12th graders, do lesser estimates for terms for arcs sent to older students merely reflect that this 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 2.

The mean value parameterization provides an alternative to the natural parameterization of the exponential family random-graph model. The mean value parameters are given by

$$\boldsymbol{\mu}(\boldsymbol{\eta}) = \mathbb{E}_{\boldsymbol{\eta}}[\mathbf{g}(y, X)] \tag{15}$$

(Handcock, 2003).

This parameterization puts the coefficients on the scale of the network statistics rather than on the conditional log-odds scale of the natural parameters. Looking at the mean value parameters provides a sense of the implications of the model fit for the network statistics. Although we assume that both models are intended to model the structure of the full network of 89 nodes, there is no principled way to apply the CC fit directly to the larger network (Shalizi and Rinaldo, 2013). We therefore compare the models on the basis of their mean value parameterizations applied to the portion of the network for which they each provide valid probability models: the subnetwork of 70 respondents. For the AO fit, these values are determined by marginalizing over the rest of the network. This puts both fits on the same scale to allow meaningful comparisons.

Table 2 shows the MLEs of the mean value parameters. To begin with, the expected number of arcs demonstrates that the CC fit implies about 7% more arcs (394) than the AO fit (367), and 21% more reciprocated arcs (94 *versus* 77). The mean value parameters of other model terms support conclusions that are suggested by the natural parameters. Under the CC fit, 12th graders receive more arcs (7% more), and more students receive no friendship nominations (almost twice as many). Differences in rates of cross-sex nominations within grade are not large. The weighted sum of arcs from girls to older boys is lower (9%). The weighted sums of arcs from boys to older and younger boys are reduced (15% and 4% respectively), and those to older girls are increased

Table 2. Estimated mean value parameters and standard errors for the model fits under the AO, CC, DP and IC approaches

	Results for the following approaches:							
	AO	*CC*	*DP*	*IC*	*AO standard error*	*CC standard error*	*DP standard error*	*IC standard error*
Density†	7.606	8.158	8.134	7.245	0.34	0.41	0.34	0.36
Mutuality	77.453	93.990	86.485	77.168	7.84	9.00	8.19	7.95
Sex and grade factors								
Grade 8 popularity	90.895	92.070	96.585	72.062	9.19	10.44	9.36	9.81
Grade 9 popularity	67.230	71.994	69.218	66.624	8.71	9.67	8.62	9.24
Grade 10 popularity	55.296	61.975	59.456	53.841	8.08	9.78	8.20	8.64
Grade 11 popularity	47.571	57.913	55.296	47.946	6.44	7.24	6.79	6.46
Grade 12 popularity	33.789	36.097	38.175	32.159	5.67	7.81	5.91	5.92
Male popularity	182.424	200.999	198.345	172.386	10.41	12.06	10.64	10.72
Sex and grade mixing								
Girl to same grade boy	57.373	59.055	60.687	57.111	5.88	6.27	5.95	6.00
Boy to same grade girl	39.111	42.027	41.133	40.263	5.25	5.47	5.27	5.30
Girl to older girl	21.021	22.017	22.358	16.073	5.99	6.25	6.15	5.08
Girl to younger girl	15.069	16.040	15.501	14.223	4.64	4.81	4.72	4.45
Girl to older boy	38.211	34.954	41.820	32.316	7.75	7.17	8.03	6.94
Girl to younger boy	16.595	19.992	17.771	9.316	4.53	4.99	4.63	3.23
Boy to older boy	21.301	18.046	23.371	19.534	6.07	5.16	6.30	5.72
Boy to younger boy	6.212	5.992	6.704	2.347	2.73	2.60	2.81	1.58
Boy to older girl	29.558	40.000	31.683	33.661	6.95	8.62	7.17	7.84
Boy to younger girl	22.542	24.019	23.258	24.964	5.77	6.05	5.85	6.26
Transitivity								
Transitive same sex and grade	153.144	216.549	186.802	153.131	40.72	50.47	45.75	39.45
Cyclical same sex and grade	35.247	54.847	45.497	36.800	11.09	14.93	13.16	11.10
Isolation	2.065	3.991	1.341	4.184	1.36	1.88	1.11	1.88

†The density coefficient is in the percentages of possible ties.

(35%). Unexpectedly, the number of transitive and cyclical triads within sex and grade are substantially higher in the CC fit (41% and 56% 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 to be due to greater concentration of arcs within sex and grade for the CC fit. This phenomenon is consistent with the relatively higher rate of sex–grade homophilous arcs from respondents to respondents, as opposed to from respondents to non-respondents. Fig. 3 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 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 the decreased rate of most other arc types across sex and grade under the CC fit.

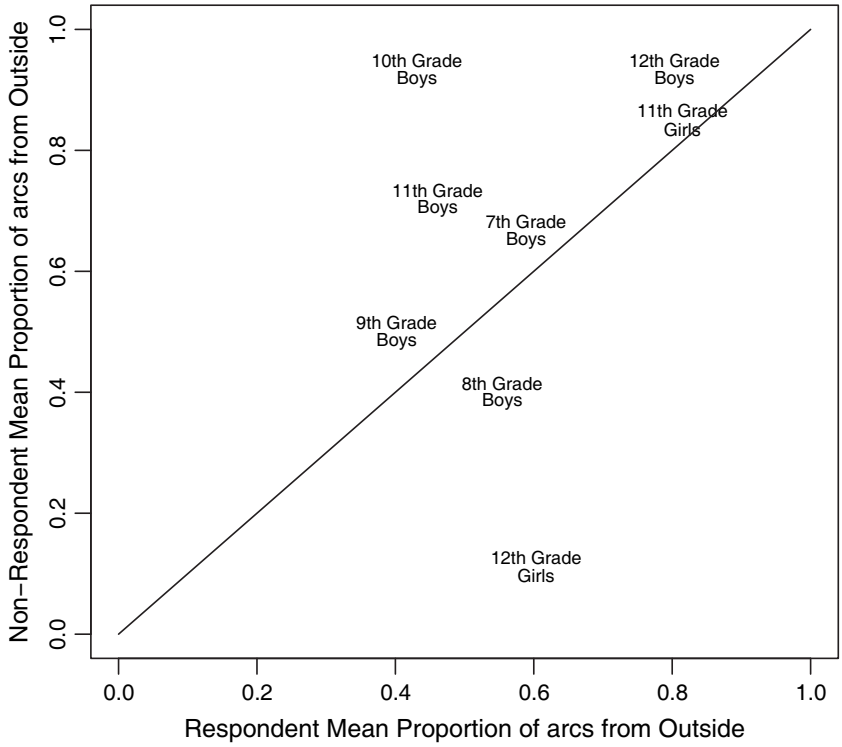

Fig. 3. Mean proportion of nominations received from outside sex and grade, by sex and grade

The AO approach relies on two types of information that are not used in the CC approach: the full size of the network, and the additional data in the arcs sent to non-respondents. To help to 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. The fifth and ninth columns of Table 1 present the resulting fit, which we refer to as the IC fit. A naive reading of Shalizi and Rinaldo (2013) may suggest that the parameters for the full network cannot be estimated by applying the model to the subnetwork data alone. As we see, the parameter estimates are close to the AO case (that uses the full observed data and the same model). The same is true for the mean value parameters given in the fifth column of Table 2. These are useful as they indicate that the uncertainty in the mean value parameter estimates for the isolates is large. The Kullback–Leibler divergence of the IC model from the AO model when both are used to model the CC subnetwork is 8.4. As the corresponding divergence for the CC model is 159, this indicates that the IC fit is much closer to the AO fit than the CC fit for the CC subnetwork. This suggests that much of the difference between the AO and CC fits is due to the assumed size of the full network.

5.4. Goodness of fit

Hunter *et al*. (2008) presented a method for evaluating the fit of network models, based on network statistics that are not modelled directly. They proposed comparing the distribution of selected statistics of substantive interest (e.g. the degree distribution and shortest path length distribution) with their observed values. They drew a sample of networks from the model that is specified by the MLE and compared the observed with the sampled distribution of statistics

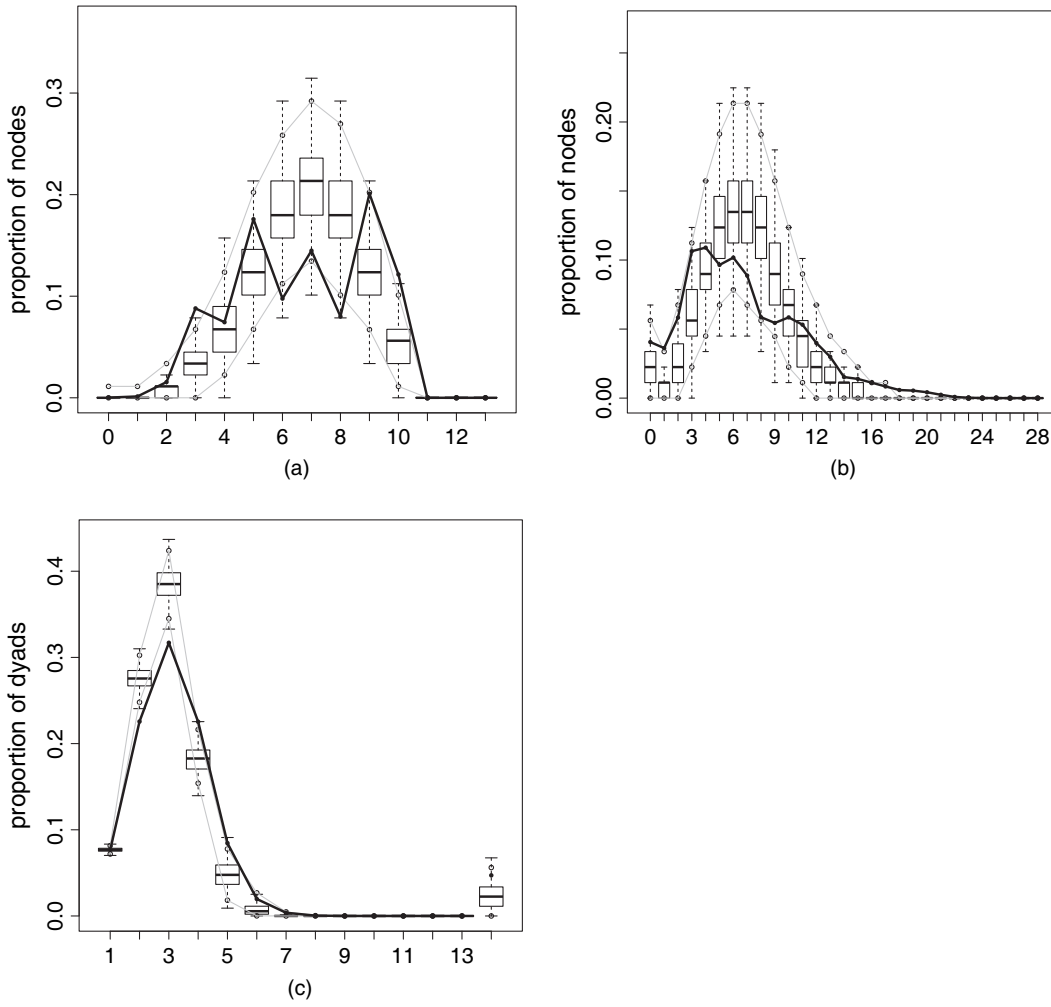

Fig. 4. Goodness-of-fit diagnostic plots for the AO model fit (———, mean proportions from conditional simulations under the model fit; □, distributions of proportions across unconditional simulations; · · · · · ·, middle 95% of the simulated distributions): (a) out-degree; (b) in-degree; (c) geodesic distance

via boxplots. The closer the observed statistics are to the middle of the sample distributions, the better the fit of the model. We extend this approach to networks that are modelled with missing data.

When the model includes missing data, we are still interested in the features of the full network, but it is only partially observed. For this reason, we estimate the reference distribution as the network statistics based on simulations of the unobserved data conditional on the observed data under the MLE. The means of these values are taken as the reference distribution, depicted with a full line in Fig. 4. Boxplots representing the distribution under the model are then added on the basis of unconditional simulations under the MLE, as in the fully observed data case. Fig. 4 depicts three such plots for the AO model fit. This model reproduces the in-degree distribution quite well. It smooths some jaggedness in the out-degree distribution. Also, it recovers the distribution of mean *minimum geodesic distances*, or the minimum number of arcs between each

pair of nodes, fairly well, although it slightly underestimates these distances. Corresponding plots for the CC and DP (to be introduced in the next section) fits are very similar to these.

5.5. Addressing the missingness at random assumption

Such goodness-of-fit analyses can also be conducted on other statistics. In particular, we may be interested in systematic differences between respondents and non-respondents, as related to the MAR assumption.

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 Fig. 1. We have observed the first two blocks, the arcs sent by respondents, and these observations provide a basis for comparing the respondents and non-respondents.

Each model implies expected densities in each of the four blocks, which can be estimated by drawing unconditional samples from the model and averaging the resulting densities. 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. The block densities are different in the two observed blocks. 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. In theory, it is possible that this is due to the different compositions of nodal covariates among respondents and non-respondents. If these nodal covariates are the only difference between respondents and non-respondents (i.e. a grade 12 boy respondent behaves the same as a grade 12 boy non-respondent), and, if we have accounted for the network features that are related to these nodal covariates, then this constitutes data *missing at random*, as in equation (3), and the AO modelling approach is valid.

The expected block densities resulting from the AO fit are represented in Fig. 5(b). These densities do not reflect the DP of non-respondents in the observed network. Thus, this result constitutes the failure of the MAR assumption. There are systematic differences between the respondents and non-respondents, beyond what can be explained by the observed data.

Note that testing the MAR assumption typically requires outside information, such as an expensive follow-up study of non-respondents. Because the primary units of inference (directed dyads) are nested between the primary units of observation (nodes), however, often the available data include information about non-respondents, such as the in-arcs that they receive from respondents. In this way, the missing data structure is similar to that of longitudinal data with partial non-response. In such a case, we may first measure any systematic differences between respondents and non-respondents. A common approach to improving inference is then to include additional parameters capturing differences between respondents and non-respondents, as per the mixture model approach that was advocated by Little, Rubin and others (Little, 1995; Little and Rubin, 2002) and, in many cases, requiring the collection of additional data. Robins *et al*. (2004) applied a variant of this approach when they used separate model terms for arcs sent to respondents and to non-respondents.

Our approach here is less extreme. Unlike Robins *et al*. (2004), we use a network model with most terms applying to the full $N \times N$ relational matrix Y, thereby leveraging the information in the observed portion to infer features of the unobserved portion. However, we also introduce a term capturing the observed systematic difference between respondents and non-respondents: their tendency to receive friendship nominations. We refer to this term as *non-respondent popularity* and use the sufficient statistic

$$\sum_{i \neq j} y_{ij} \mathbb{I}(S_j = 0). \tag{16}$$

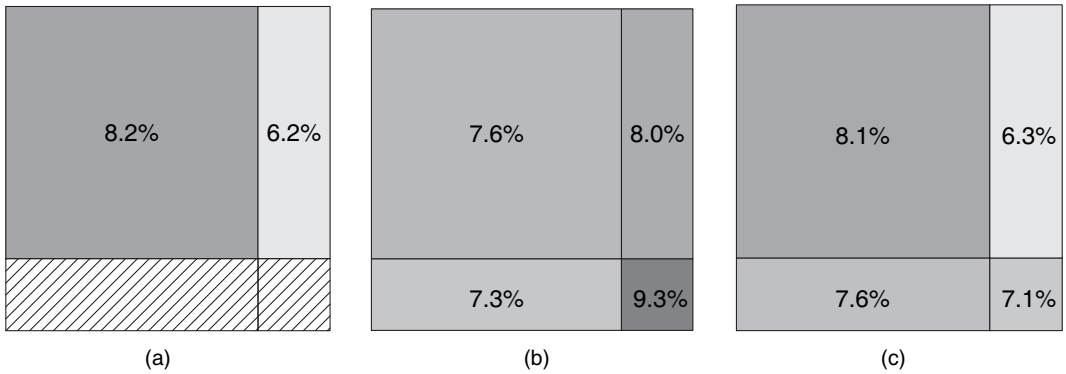

Fig. 5. Percentage of possible arcs in each of four blocks in (a) observed data and expected under (b) AO and (c) DP model fits

The resulting model fit is in the fourth and eighth columns of Table 1 (natural parameterization) and Table 2 (mean value parameterization). We see that the term is negative and significant, indicating a significant difference between the two subgroups. The other parameter estimates are not significantly different in this new model fit, although Fig. 5(c) illustrates that the observed densities of ties to respondents and non-respondents are reproduced almost exactly. Thus we have successfully accounted for one feature of data that are not missing at random, although it is clear that there may be other non-MAR features that we have not addressed.

6. Discussion

In this paper we provide an analysis of the mechanisms governing friendship formation between students in a US high school. We find that friendship nominations are often mutual and more likely to occur between students of the same grade and sex. We find that friendships within sex and grade show patterns of hierarchical structure, and also that there is a tendency for some students to receive no friendship nominations, at a higher rate than we would expect at random.

Through this analysis, we offer an exposition of methodology for the modelling of networks with missing data, expanding on previous work in missing data, network modelling and network sampling. We primarily treat the MAR case but also introduce a framework for treatment of some limited instances of data not missing at random. We show that, under these conditions, available software can be used to analyse networks that are partially observed because of out-of-design missing data mechanisms.

The analysis also illustrates some specific points. The first is that only analysing the CCs can lead to different conclusions from those by analysing all the observations. Comparing the CC fit treating respondents only and the IC fit, using the same data but respecting the true network size and nodal composition illustrates the practical implications of Shalizi and Rinaldo (2013), showing that the subnetwork model fit differs considerably from the full network fit, even using identical dyadic data. The further differences between the IC and AO fits illustrate the effect of ignoring the information in observed ties to non-respondents. We also show that the overall fit to the data is improved by extending the model to represent differences between respondents and non-respondents.

We illustrate extensions of existing network analysis techniques to the missing data setting. In particular, we apply the mean value parameterization to study differences between modelling

approaches to the same data. We also extend the goodness-of-fit techniques of Hunter *et al.* (2008) to understand models fitted to partially observed data better.

We find that it is typically worthwhile to retain as much information as possible from the data. This is unsurprising but often not obvious in the network setting. The CC approach, discarding all information about nodes with only partially available information, is straightforward to implement and seems an attractive alternative. However, we have shown that, in principle and in practice, it is possible and natural to work with models for the full network, using all observed data, even when some data might be missing. As with any missing data situation, it is helpful wherever possible to retain any information that is available on the full sampling frame, including non-respondents. In this paper, we have retained two types of data on non-respondents: exogenously available covariate data and friendship nominations received.

It is also sometimes possible to improve model fit by capturing observable differences between respondents and non-respondents. We have illustrated one such effect in our DP model fit. It is important to remember, however, that missing data are, by definition, beyond the control of researchers and often follow unpredictable patterns. In many cases, valid inference may require further study of non-response patterns, or sensitivity analysis.

Acknowledgements

This material is based on work supported by the National Science Foundation (MMS-085155, SES-1357619, IIS-1546259 and SES-1230081, including support from the National Agricultural Statistics Service), National Institute of Child Health and Human Development (R21HD063000, R21HD075714 and R24-HD041022) and the Office of Naval Research (N00014-08-1-1015).

Appendix A: Computational procedure for Kullback–Leibler divergence

This appendix details the estimation procedure of Section 5. A natural way to compare the network models is the Kullback–Leibler divergence. In Section 5 we use it to compare the CC model with the AO model when both are used to model the CC subnetwork. Specifically, we can consider the probability distribution that the AO model for the full network places over the CC network dyads, $P_{\eta_{\mathrm{AO}}}(Y^{\mathrm{CC}}|X)$, and use it to compute the Kullback–Leibler divergence:

$$\mathbb{E}_{\eta_{\mathrm{AO}}}\left[\log\left\{\frac{P_{\eta_{\mathrm{AO}}}(Y^{\mathrm{CC}}|X)}{P_{\eta_{\mathrm{CC}}}(Y^{\mathrm{CC}}|X)}\right\}\right]$$

where $P_{\eta_{\mathrm{AO}}}$ is the AO model for the full network, $P_{\eta_{\mathrm{CC}}}$ is the CC model for the CC subnetwork and Y^{CC} is the set of dyads from the full network in the CC subnetwork. From equations (4) and (5):

$$\log\{P_{\eta_{\mathrm{AO}}}(Y^{\mathrm{CC}}|x)\} = \kappa(\boldsymbol{\eta}_{\mathrm{AO}}|Y^{\mathrm{CC}}, x) - \kappa(\boldsymbol{\eta}_{\mathrm{AO}}, x),$$
$$\log\{P_{\eta_{\mathrm{CC}}}(Y^{\mathrm{CC}}|x)\} = \boldsymbol{\eta}_{\mathrm{CC}}^{\mathrm{T}}\mathbf{Z}(Y^{\mathrm{CC}}|x) - \kappa_{\mathrm{CC}}(\boldsymbol{\eta}_{\mathrm{CC}}, x)$$

so the Kullback–Leibler divergence is

$$\mathbb{E}_{\eta_{\mathrm{AO}}}[\kappa(\boldsymbol{\eta}_{\mathrm{AO}}|Y^{\mathrm{CC}}, x) - \boldsymbol{\eta}_{\mathrm{CC}}^{\mathrm{T}}\mathbf{Z}(Y^{\mathrm{CC}}|x)] + \kappa_{\mathrm{CC}}(\boldsymbol{\eta}_{\mathrm{CC}}, x) - \kappa(\boldsymbol{\eta}_{\mathrm{AO}}, x).$$

The first term is computed by generating full networks from the AO model and then the conditional normalizing constants for each of their CC subnetworks. We provide the `statnet` code for this computation as it is of general interest for modelling networks with missing data.

References

Barndorff-Nielsen, O. E. (1978) *Information and Exponential Families in Statistical Theory*. New York: Wiley.

Bearman, P. S., Moody, J. and Stovel, K. (2004) Chains of affection: the structure of adolescent romantic and sexual networks. *Am. J. Sociol.*, **110**, 44–91.
Cover, T. M. and Thomas, J. A. (2006) *Elements of Information Theory*, 2nd edn. New York: Wiley.
van Duijn, M. A. J., Handcock, M. S. and Gile, K. J. (2009) A framework for the comparison of maximum pseudo likelihood and maximum likelihood estimation of exponential family random graph models. *Socl Netwrks*, **31**, 52–62.
Frank, O. and Strauss, D. (1986) Markov graphs. *J. Am. Statist. Ass.*, **81**, 832–842.
Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc.* B, **54**, 657–699.
Handcock, M. S. (2003) Assessing degeneracy in statistical models of social networks. *Working Paper 39*. Center for Statistics and the Social Sciences, University of Washington, Seattle.
Handcock, M. S. and Gile, K. J. (2010) Modeling networks from sampled data. *Ann. Appl. Statist.*, **4**, 5–25.
Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M. and Morris, M. (2003) statnet: software tools for the statistical modeling of network data. *R Package Version 2.0*. University of Washington, Seattle.
Harris, K. M., Florey, F., Tabor, J., Bearman, P. S., Jones, J. and Udry, J. R. (2003) The National Longitudinal Study of Adolescent Health: research design. *Technical Report*. Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill. (Available from `http://www.cpc.unc.edu/projects/addhealth/design`.)
Holland, P. W. and Leinhardt, S. (1981) An exponential family of probability distributions for directed graphs (with comments by Ronald L. Breiger, Stephen E. Fienberg, Stanley S. Wasserman, Ove Frank and Shelby J. Haberman and a reply by the authors). *J. Am. Statist. Ass.*, **76**, 33–65.
Hunter, D. R., Goodreau, S. M. and Handcock, M. S. (2008) Goodness of fit for social network models. *J. Am. Statist. Ass.*, **103**, 248–258.
Hunter, D. R. and Handcock, M. S. (2006) Inference in curved exponential family models for networks. *J. Computnl Graph. Statist.*, **15**, 565–583.
Karwa, V., Krivitsky, P. N. and Slavković, A. B. (2015) Sharing social network data: differentially private estimation of exponential-family random graph models. *Preprint arXiv 1511.02930*.
Karwa, V., Slavković, A. B. and Krivitsky, P. (2014) Differentially private exponential random graphs. *Lect. Notes Comput. Sci.*, **8744**, 143–155.
Koskinen, J. H., Robins, G. L., Wang, P. and Pattison, P. E. (2013) Bayesian analysis for partially observed network data, missing ties, attributes and actors. *Socl Netwrks*, **35**, 514–527.
Little, R. J. (1995) Modeling the drop-out mechanism in repeated-measures studies. *J. Am. Statist. Ass.*, **90**, 1112–1121.
Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, 2nd edn. Hoboken: Wiley.
Lusher, D., Koskinen, J. and Robins, G. (2012) *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge: Cambridge University Press.
McPherson, M., Smith-Lovin, L. and Cook, J. M. (2001) Birds of a feather: homophily in social networks. *A. Rev. Sociol.*, **27**, 415–444.
Resnick, M. D., Bearman, P. S., Blum, R. W., Bauman, K. E., Harris, K. M., Jones, J., Tabor, J., Beuhring, T., Sieving, R., Shew, M., Ireland, M., Bearinger, L. H. and Udry, J. R. (1997) Protecting adolescents from harm: findings from the National Longitudinal Study of Adolescent Health. *J. Am. Med. Ass.*, **278**, 823–832.
Robins, G., Pattision, P. and Woolcock, J. (2004) Missing data in networks: exponential random graph (p^*) models for networks with non-respondents. *Socl Netwrks*, **26**, 257–283.
Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
Shalizi, C. R. and Rinaldo, A. (2013) Consistency under sampling of exponential random graph models. *Ann. Statist.*, **41**, 508–535.
Stork, D. and Richards, W. D. (1992) Nonrespondents in communication network studies: problems and possibilities. *Grp Organizn Mangmnt*, **17**, 193–209.
Thompson, S. K. and Frank, O. (2000) Model-based estimation with link-tracing sampling designs. *Surv. Methodol.*, **26**, 87–98.
Udry, J. R. (2003) The National Longitudinal Study of Adolescent Health: (Add Health), waves I and II, 1994-1996; wave III, 2001-2002. *Technical Report*. Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill.
Udry, J. R. and Bearman, P. S. (1998) New methods for new research on Adolescent sexual behavior. In *New Perspectives on Adolescent Risk Behavior* (ed. R. Jessor), pp. 241–269. Cambridge: Cambridge University Press.
White, H. C., Boorman, S. A. and Breiger, R. L. (1976) Social-structure from multiple networks I: Blockmodels of roles and positions. *Am. J. Sociol.*, **81**, 730–780.