Using retrospective sampling to estimate models of relationship status in large longitudinal social networks

A. James O'Malley, *Dartmouth College*
Sudeshna Paul, *Emory University*

Journal Title: Computational Statistics and Data Analysis
Volume: Volume 82
Publisher: Elsevier | 2015-02-01, Pages 35-46
Type of Work: Article | Post-print: After Peer Review
Publisher DOI: 10.1016/j.csda.2014.08.001
Permanent URL: https://pid.emory.edu/ark:/25593/rmfp3

Final published version: http://dx.doi.org/10.1016/j.csda.2014.08.001

Copyright information:

Published in final edited form as:

Comput Stat Data Anal. 2015 February 1; 82: 35–46. doi:10.1016/j.csda.2014.08.001.

Using Retrospective Sampling to Estimate Models of Relationship Status in Large Longitudinal Social Networks

A. James O'Malley[a,*] and **Sudeshna Paul**[b]

A. James O'Malley: James.OMalley@Dartmouth.edu; Sudeshna Paul: sudeshna.paul@emory.edu

[a]The Dartmouth Institute for Healh Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, Lebanon, NH 03766, USA

[b]Nell Hodgson Woodruff School of Nursing, Emory University, Atlanta, GA 30322, USA

Abstract

Estimation of longitudinal models of relationship status between all pairs of individuals (dyads) in social networks is challenging due to the complex inter-dependencies among observations and lengthy computation times. To reduce the computational burden of model estimation, a method is developed that subsamples the "always-null" dyads in which no relationships develop throughout the period of observation. The informative sampling process is accounted for by weighting the likelihood contributions of the observations by the inverses of the sampling probabilities. This weighted-likelihood estimation method is implemented using Bayesian computation and evaluated in terms of its bias, efficiency, and speed of computation under various settings. Comparisons are also made to a full information likelihood-based procedure that is only feasible to compute when limited follow-up observations are available. Calculations are performed on two real social networks of very different sizes. The easily computed weighted-likelihood procedure closely approximates the corresponding estimates for the full network, even when using low sub-sampling fractions. The fast computation times make the weighted-likelihood approach practical and able to be applied to networks of any size.

Keywords

Conditional independence; Longitudinal; Retrospective sampling; Social network; Sociocentric design; Sparse data; Weighting

1. Introduction

In this paper we develop, apply, and evaluate a new method of estimating a dynamic model of the relationship status of all *dyads* (pairs of individuals) in a social network, where both the number of individuals (N) and the number of observation times (T) can be large.

*Correspondence to: A. James O'Malley, The Dartmouth Institute for Healh Policy and Clinical Practice, Geisel School of Medicine, Dartmouth College, Lebanon, NH 03766, USA. Ph: 603-653-0854; Fax: 603-653-0896; James.OMalley@Dartmouth.edu.

Analyses of complete lattices of dyadic data (referred to as *sociocentric network data*) in general seek to identify the important determinants of dyadic relationships and gain insights into properties or determinants of the network. For example, one phenomena that is often thought responsible for the formation of relationships is *homophily* – commonly described as "birds of a feather flock together" – whereby individuals with similar attributes are more likely to form or maintain relationships, leading to clusters of individuals with similar traits within the network. However, the primary objective of this paper is demonstrating that the new estimation method is feasible to implement on networks of any N and T, overcoming the practical limitations of existing methods. The following two publicly-available social network data sets, judiciously chosen due to the difference in N and T between then, will be analyzed and used to appraise our method of computing estimates.

The smaller network is from the excerpt of 50 schoolgirls in the Teenage Friends and Lifestyle Study (TFLS) described in Snijders (2014). Students in the study named up to 12 close friends at three surveys conducted during 1995–1997 (Michell and Amos, 1997; West and Sweeting, 1995). After dropping the two girls who did not nominate and were not nominated by anyone, the final network comprised $N = 48$ girls (1,128 dyads) observed on $T = 3$ occasions (two relationship change opportunities). The number of friends named by each schoolgirl (*out-degree*) could range from 0 to 12 while the number of times a girl could be named by others as their friend (*in-degree*) had a range from 0 to 47. The students were also asked about substance use and adolescent behavior associated with lifestyle, sporting behavior and tobacco, alcohol and cannabis consumption. A particular question of importance is whether homophily of smoking behavior exists; were girls who were both smokers or both non-smokers more likely to become friends.

The second and larger longitudinal friendship network is from the offspring cohort of the Framingham Heart Study (FHS). Since the offspring cohort's inception in 1971, its members have been followed from 1971–2008 through eight periodic health exams, at which an extensive array of personal and medical information (e.g., height, weight, age, smoking status) was collected. Friendship ties at each exam were ingeniously obtained from the nomination of close-friends who might be in a position to know where the study member would be in two to four years (Christakis and Fowler, 2007, 2008). Subjects were not restricted from naming multiple friends but on most occasions only named a single friend, resulting in a sparsely-connected network. Out-degrees were typically 0 or 1 while in-degrees were more widespread with values ≥ 2 relatively common. Emulating Paul and O'Malley (2013), all FHS offspring members who named or were named by another offspring cohort member over any two consecutive exams were included in the analysis, yielding $N = 831$ individuals observed at up to $T = 8$ exams (7 relationship change opportunities). A plethora of personal characteristics (gender, age, BMI, smoking status, various medical quantities) are available although herein we focus on age. More details of both the FHS and TFLS networks appear in Paul and O'Malley (2013).

In these networks relationship status (close friendships between schoolgirls or between study members) is presumed known for all $N(N-1)/2$ dyads, yielding complete sociocentric data. Close friendship is represented as a binary random variable (1 = yes, 0 = no) with the presence thereof referred to as a *tie*. Because there is no constraint that a tie from one

individual to another implies that a tie exists in the reverse direction the networks are *directional*.

To identify the presence of homophily or some other relationship feature (e.g., reciprocity) in the network, other possible explanations for the formation and dissolution of ties need to be statistically adjusted for or controlled. Finding the important determinants of a network is aided by longitudinal data. However, such data has historically been elusive. Not surprisingly, methods for longitudinal analysis of sociocentric data are scarce and those that do exist are confronted by computational challenges. For example, we previously developed a novel model for a longitudinally-observed sociocentric network that allowed homophily effects and other network phenomena to be estimated. Although the methodology was sound, implementation was restricted to small- to mid-sized networks by CPU and time constraints (Paul and O'Malley, 2013). One of the reasons for the challenging computations is that the number of dyads in a sociocentric dataset has order N^2 as opposed to the order N number of observations in individual level analyses. Because large networks with $N \geq 1000$ are becoming commonplace, the development of methods of estimating models of networks for any N and T is timely.

The method proposed herein adapts ideas from survey sampling methodology to accurately approximate estimates of the full network in minimal computational time. The genesis of the method is the observation that as N increases the number of dyads that remain null (no ties) over time increases. Therefore, as long as the sampling design is accounted for in the analysis, in large networks only a small fraction of the always-null dyads may be needed to accurately approximate the estimates computed on the full network. To account for the dependencies introduced by sampling, we develop a novel *weighted likelihood* (WL) estimation procedure that weights the observations for each dyad by the inverse of the probability of sampling that dyad. The proposal to subsample null-dyads is not without precedent (Raftery et al., 2012; Kleinbaum, 2012). However, to our knowledge we are the first to consider subsampling in the context of longitudinal sociocentric networks.

In Section 2 we define notation and specify models for longitudinal analysis of sociocentric data. In Section 3 we describe our proposed sampling design and develop associated WL estimation and implementation procedures. To evaluate the efficacy of the WL estimation procedure, we compare it to a full information *observed data likelihood* (ODL) procedure on the smaller TFLS network data for which estimation of the ODL procedure is feasible and discuss the limitations of ODL methods on larger or more intensely observed networks. The estimation methods are applied to the two longitudinal sociocentric network data sets described above in Section 4 with comparisons between the methods and other results reported in Section 5. Section 6 reviews the primary findings and discusses limitations.

2. Notation, Nertwork Phenomena, and Model Specification

Let Y_{ijt} denote the presence of a tie (1 = friend, 0 = not a friend) from individual i to individual j ($i, j \in \{1, \ldots, N\}$) at time $t \in \{1, \ldots, T\}$. The bivariate random variable $D_{ijt} = (Y_{ijt}, Y_{jit})$, the status of dyad ij at time t, is the primary unit of analysis and the subject of our statistical model. Clearly, D_{ijt} contains 0 (null friendship), 1 (directional friendship), or 2

(mutual friendship) ties. For notational convenience, the sequence of states held by a dyad is collated as $D_{ij} = (D_{ij1}, \ldots, D_{ijT})$, the whole network at a given time as $D_t = \{D_{ijt}\}_{i<j}$, and the sequence of networks beyond baseline ($t = 1$) as $D = (D_2, \ldots, D_N)$.

In both the TFLS and FHS networks several network phenomena are of interest. If the prevalence of mutual dyads is greater than expected (i.e., if knowing *i* named *j* as a friend makes it more likely than otherwise that *j* named *i* as a friend) all else equal then *reciprocity* is present. A distinct phenomena from reciprocity is the propensity of an individual to name others as friends being correlated with the propensity of them being named by others as a friend. A positive correlation suggests that expansive individuals are also popular while a negative correlation might suggest the presence of powerful individuals who keep few close friends despite many others wanting to be their friend. *Transitivity*, the phenomena commonly referred to as a "friend of a friend is a friend," is the most well-known form of between-dyad dependence. In addition, various forms of homophily may be present and as in any longitudinal study, observations may be serially dependent (O'Malley and Marsden, 2008). A model for longitudinal sociocentric network data is needed to distinguish the effects of each of these and other terms (Handcock et al., 2003; Robins et al., 2007; Lewis et al., 2008). The statistical model considered in this paper was developed in our own prior work (Paul and O'Malley, 2013), which built off or was motivated by the work of several others (van Duijn et al., 2004; Zijlstra et al., 2006; Hoff, 2005, 2008).

A vector of covariates x_{ijt} includes the homophily (or similarity) measures (e.g., difference in age, both smokers or both non-smokers) for actors *i* and *j* at time *t*, any observed predictors specific to individuals *i* and *j* at time *t*, and any network-based covariates capturing transitivity or other forms of triadic dependence determined from $D_{t'}$ for $t' < t$ as elements. An example of the latter is the *common source* covariate, given by $I(\sum_{k\neq i,j} Y_{ki(t-1)}Y_{kj(t-1)} > 0)$ where I(event) = 1 if event is true and 0 otherwise, allowing an individual $k \neq i, j$ naming both *i* and *j* as a friend at $t-1$ to have an effect on the likelihood of ties between *i* and *j* at *t*. Time-lagged versions of such triadic-type covariates are used in place of contemporaneous predictors in order to ensure that the resulting model is well-defined and self-consistent (Paul and O'Malley, 2013). To allow different effects on tie-formation and tie-retention, two versions of the common source covariate are created through multiplication by $1- Y_{ij(t-1)}$ and $Y_{ij(t-1)}$, respectively.

The bivariate D_{ijt} may be represented as a four-category multinomial random variable and $D_{ij(t-1)} \rightarrow D_{ijt}$ is represented by a 4×4 transition matrix. We focus on dyadic models with assumed Markov dependence across time so that the transition matrix is a sufficient representation of the serial dependence of a dyad on its prior states. However, this assumption could be relaxed to allow dependence on earlier states of the dyad (and more generally the network). We further assume that transitions in dyad status follow a generalized mixed effect logistic regression equation (an alternative link function such as the inverse of the standard normal cumulative distribution function as for probit regression could instead be considered). Let $\theta_{ij} = (a_i, b_i, a_j, b_j)$ denote the individual-specific propensities of *i* and *j* to form (*a*) and receive (*b*) friendships. The probabilities of the four possible states of D_{ijt} are represented in the form:

$$\Pr(D_{ijt}=d_{ijt},|D_{t-1}=d_{t-1},x_{ijt},\theta_{ij})=k_{ijt}^{-1}\exp(\mu_{ijt}y_{ijt}+\mu_{jit}y_{jit}+\rho_{ijt}y_{ijt}y_{jit}),\quad(1)$$

where

$$k_{ijt}=1+\exp(\mu_{ijt})+\exp(\mu_{jit})+\exp(\mu_{ijt}+\mu_{jit}+\rho_{ijt}),\quad(2)$$

$$\mu_{ijt}=\beta_0+\beta_1y_{ij(t-1)}+\beta_2y_{ji(t-1)}+\beta_3y_{ij(t-1)}y_{ji(t-1)}+\beta_x^Tx_{ijt}+a_i+b_j,\quad(3)$$

and

$$\rho_{ijt}=\lambda_0+\lambda_1(y_{ij(t-1)}+y_{ji(t-1)})+\lambda_2y_{ij(t-1)}y_{ji(t-1)}.\quad(4)$$

In (1) the terms μ_{ijt} and ρ_{ijt} are linear predictors that relate the systematic components of (3) and (4) to the four state probabilities of dyad *ij* at *t*. The term μ_{ijt} includes factors associated with the likelihood that $Y_{ijt} = 1$ but not necessarily with the likelihood that $Y_{jit} = 1$. Dependence between Y_{ijt} and Y_{jit} not attributed to observed characteristics of the individuals is quantified by the extent that $\rho_{ijt} = \rho_{jit}$ and is known as reciprocity. The further ρ_{ijt} is from 0 the greater the difference between $\Pr(Y_{ijt} = Y_{jit} = 1 \mid D_{t-1} = d_{t-1}, x_{ijt}, \theta_{ij})$ and $\exp(\mu_{ijt} + \mu_{jit})/k_{ijt}$, its value under statistical independence of Y_{ijt} and Y_{jit}.

The density parameter β_0 reflects the rate of tie-formation in dyads whose current state is null. The reciprocity parameter λ_0 allows the rate of formation of a tie from *i* to *j* to be correlated with that of a tie from *j* to *i* in dyads whose current state is null. The parameters β_1, β_2, and β_3; and λ_1 and λ_2 are modifications of β_0 and λ_0 when the dyad is in the various non-null states at $t - 1$. For example, λ_1 measures the increased propensity of a bidirectional tie at *t* if it is asymmetric as opposed to null at $t - 1$.

We assume that the sender and receiver effects in θ_{ij} only impact μ_{ijt} and are time invariant although these assumptions could be relaxed. To complete the model we assume (a_i, b_i) are random effects from a bivariate normal distribution having mean 0 and covariance matrix Σ (standard deviations σ_a and σ_b; correlation coefficient ν). The model in (1–4) is then a longitudinal extension of the P_2 model (Duijn et al., 2004).

The joint likelihood function of the observed data and the random effects is given by

$$p(D=d,\theta|D_1=d_1,x)=\left(\prod_{i<j}\prod_{t=2}^{T}p(d_{ijt}|\theta_{ij},d_{t-1},x_{ijt};\Omega)\right)\prod_i p(a_i,b_i;\sum),\quad(5)$$

where $\Omega = (\beta, \lambda, \theta)$ and Σ denote the transition model and random effect parameters, respectively. The factorization evident in (5) follows from the conditional independence of the dyadic observations given (a, b) and the independence of the random effects.

2.1. Bayesian Estimation

Due to the hierarchical structure of the model, Bayesian methods are natural to use for estimation. As in Paul and O'Malley (2013) we complete a Bayesian specification of the model by assuming diffuse normal priors (mean 0 and variance 10^6) for the regression coefficients (β, λ). Instead of specifying priors directly for Σ, for additional flexibility we express the joint distribution of the random effects (a_i, b_i) in the product normal form (Spiegelhalter, 1998; Cooper et al., 2007). The product-normal form is advantageous compared to assuming (e.g.) the conjugate inverse-Wishart prior for Σ as it allows the prior for the variance and correlation components of Σ to be specified with different levels of precision. In the analyses of the TFLS and FHS networks we assume $a_i \sim N(0, \sigma_a^2)$ and $b_i | a_i \sim N(\phi a_i, \tau_b^2)$, where σ_a^2 and τ_b^2 in-turn have inverse-gamma priors with mean 1 and variance 10^3, and $\varphi \sim N(0, 10^6)$.

3. Sampling Always-Null Dyads

In previous work we found that model estimation rapidly became too CPU intensive and time-consuming due to the $O(N^2)$ number of dyads. However, because the proportion of dyads that remain in their null state throughout ("always-null dyads") naturally increases with N (Dunbar, 1992; Gladwell, 2000), it is reasonable to expect that sampling such dyads will result in minimal loss of precision. Therefore, we propose the sampling-design that samples non always-null dyads with probability 1 and always-null dyads with probability π_0. Let $r_{ij} = \prod_{t=1}^{T} (1 - y_{ijt})(1 - y_{jit})$ so that $r_{ij} = 1$ if dyad ij is an always-null dyad and $r_{ij} = 0$ otherwise, and $S_{ij} = 1$ if dyad ij is sampled and 0 otherwise. Then the associated sampling probability is given by $\Pr(S_{ij} = 1 | D_{ij}, \theta_{ij}, x_{ij}) = \Pr(S_{ij} = 1 | r_{ij}) = \pi_0^{r_{ij}}$.

The above sampling scheme is easily extended by allowing the sampling probabilities to depend on x_{ij}. For example, setting $\Pr(S_{ij} = 1 \mid D_{ij}) = 1$ if individuals i and j are both friends with an individual $k \geq i, j$ at some t and 0 otherwise would ensure that the dyads that contain the most information about triadic effects are assured of being sampled. Such sampling designs are highly desirable when certain levels of a predictor occur infrequently to ensure the model is estimable on the sample data.

Because the sample inclusion indicators depend on D_{ij} they are informative and ignoring them is likely to lead to sample selection bias. The joint likelihood function of D and θ given the sample inclusion indicators S has the form $p(D, \theta | S, X, D_1)$, where X contains the matrix of covariates from across the observations and D_1 is the initial state of the network. Because $p(D, \theta | S, X, D_1) = p(D | S, \theta, X, D_1) p(\theta | S, X, D_1)$ it is convenient to consider $p(D | S, \theta, X, D_1)$ and $p(\theta | S, X, D_1)$ separately. Although the likelihood function is well-defined and maximum-likelihood or Bayesian estimators inherit the associated optimality properties, conditioning on S leads to non-standard expressions. As illustrated in the Appendix, $p(D | S, \theta, X, D_1)$ is available in closed-form but is laborious to evaluate when T is large, while $p(\theta | S, X, D_1)$ is not available in closed-form and so exact computation would entail numerical evaluation of an unresolved integral. Therefore, while traditional estimators exist

theoretically, computation of them would be onerous. To avoid such bias we propose to use a WL procedure to estimate the model parameters.

3.1. WL Point Estimation

In the design-based approach to the analysis of survey data, weighted estimators are often used to account for informative sampling designs. The general procedure is to weight observations by the inverse of their sample inclusion probability, thereby ensuring that inferences pertain to the individuals in the sampled population. The analogy for the sociocentric network is to weight each sampled dyadic observation by the inverse of its sampling probability.

Because we are interested in estimating the parameters of the model in (1–4), we propose to weight the contribution of each sampled observation to the likelihood function. This procedure emulates an approach to estimating hierarchical models on survey data given informative sampling weights (Browne et al., 2002; Pfeffermann et al., 1998). We first ignore the presence of random effects and focus on the weighted-likelihood alternative to $p(D \mid S, \theta, X, D_1)$. For our sampling scheme, the weighted log-likelihood function conditional on θ is

$$\mathcal{L} = \sum_{i<j} w_{ij} \mathcal{L}_{ij}, \quad (6)$$

where $\mathcal{L}_{ij} = \sum_{t=2}^{T} \{\mu_{ij2} y_{ijt} + \mu_{ji2} y_{jit} + \rho_{ijt} y_{ijt} y_{jit} - \log(k_{ijt})\}$ and $w_{ij} = \Pr(S_{ij} = 1 | r_{ij})^{-1} = \pi_0^{-r_{ij}}$ are the likelihood contribution and sampling weight, respectively, for dyad *ij*. The WL in (6) is not a true likelihood function as the weights do not represent the frequency of observations. However, from the perspective of approximating the estimates that would have obtained had the full sociocentric network been analyzed, theory suggests maximizing (6) is a more desirable procedure than the naïve unweighted alternative.

To accommodate the random effects θ, we extend (6) to

$$\mathcal{L} = \sum_{i<j} w_{ij} \mathcal{L}_{ij} + \sum_{i} \log(p(a_i, b_i; \sum)). \quad (7)$$

The unweighted second term on the right-hand-side of (7) arises because the sample design does not directly depend on θ_{ij}; therefore, the information contained in S about θ is likely to be minimal. Therefore, we make the approximation $p(\theta \mid S, X, D_1; \Omega) \simeq p(\theta; \Sigma) = \Pi_i p(a_i, b_i; \Sigma)$. Raftery et al. (2012) similarly argues that information in the sample inclusion indicators about an individual's position in the latent "social space" (the random parameters in this context) need not be used to construct individual-level weights.

3.2. Bootstrap interval estimation procedure

An advantage of using (7) (or (6) if no covariates) for point estimation is that it is easily maximized. However, because it is not a true likelihood function, the WL is not calibrated to the information in the sample and so standard MCMC-derived variance estimates cannot be assumed to adequately represent the precision of knowledge about the true parameters. Therefore, we use a non-parametric bootstrap to compute interval estimates for the WL procedure. The resulting $100(1-\alpha)$-level intervals are used to approximate the exact $100(1-\alpha)\%$ equal-tailed Bayesian credible intervals. For the analyses conducted herein we assume $\alpha = 0.05$.

The following pseudo-code describes the bootstrap procedure:

1. For $k = 1 : n_{\text{boot}}$:
    a. Re-sample with replacement $N(N-1)/2$ dyads from the network, keeping the set of observations attached to each dyad intact. Denote the re-sampled data by D^k.
    b. Sample the always-null dyads in D^k with probability π_0 and augment these with the non always-null dyads (sampled with probability 1). Denote the sampled data by D^k_{sub}.
    c. Use the WL procedure to fit the model to D^k_{sub}. Denote the posterior mean estimates of the model parameters by $\Omega^k = (\beta^k, \lambda^k, \theta^k)$ and Σ^k.
2. Take the $100\alpha/2$ and $100(1-\alpha/2)$ percentiles of each element of $\{\Omega^k, \Sigma^k\}_{k=1:n_{\text{boot}}}$ as the 95% interval estimate.

Because the bootstrap fits the model on each re-sampled dataset (in our case $n_{\text{boot}} = 100$) in step 1c, we recommended that the bootstrap calculations be conducted in parallel.

The key assumption underlying the bootstrap is that the units of observations being re-sampled are independent. In the case of the model in (1–4), dyadic independence does not hold as multiple dyadic observations depend on the same elements of θ_{ij}. Furthermore, a two-stage bootstrap re-sampling scheme that re-samples clusters and observations within clusters does not solve the problem as the sender (a) and receiver (b) random effects are cross-classified between dyads. Therefore, the proposed bootstrap procedure is in general an approximation that warrants evaluation.

3.3. Validation of WL Estimation Procedure

A theoretically supported estimation procedure is to evaluate the likelihood function of the observed data conditional on $S = \{(i, j) : S_{ij} = 1\}$, the set of the sample inclusion indicators. Maximization of the resulting observed data likelihood (ODL) function is a statistically efficient procedure as all the information in the sample is utilized, including the information contained in the sample inclusion indicators. The ODL procedure emulates estimation of models for case-control studies, an approach that has been previously considered for cross-sectional network data (Raftery et al., 2012; Kleinbaum, 2012) but not for longitudinal network data.

Because $p(D \mid S, \theta, X, D_1)$ has a closed-form expression for all T (see Appendix), exact full information likelihood-based estimates can be computed for the case when (a_i, b_i) are constant across i. In general, because $p(\theta \mid S, X, D_1; \Omega)$ does not separate into disjoint components due to the involvement of a_i and b_i in multiple dyads (see Equation 10 in the Appendix), the ODL function involves a $2N$-dimensional integral that does not have a closed-form. Therefore, exact likelihood-optimization or Bayesian evaluation is computationally challenging. However, because S only depends on θ_{ij} indirectly through D_{ij}, it is reasonable to assume S is only weakly informative about θ. Therefore, we approximate $p(\theta \mid S, X, D_1; \Omega)$ by $p(\theta, \Sigma) = \Pi_i\, p(a_i, b_i; \Sigma)$ and use the resulting "pseudo ODL,"

$$L = \prod_{i<j} p(D_{ij} \mid S_{ij} = 1, x_{ij}, D_1; \Omega) \prod_i p(a_i, b_i; \sum), \quad (8)$$

as a "working" likelihood function for computing estimates against which to evaluate the performance of the WL procedure.

In those cases when ODL is available and able to be evaluated exactly, it provides a gold standard against which to evaluate the WL procedure. However, the above approximations notwithstanding, application of the ODL procedure is limited to situations where T is small (see Appendix for details).

4. Analyses of Network Data

In analyzing the TFLS and FHS networks, we assume (a_i, b_i) is a bivariate normal random variable with unknown covariance. When using the ODL procedure to assess the WL procedure, we also consider the special case where the variances are 0 (i.e., the random effects are constant across dyads). Because the objective of the WL procedure is to recover the estimates obtained from the full sample, the $\pi_0 = 1$ case yields an upper bound on the performance of the WL procedure.

4.1. TFLS Network

In the TFLS analysis, x_{ijt} consists of indicators of whether individuals i and j either both smoked or both did not smoke ("same smoking status") at t and whether a third individual named both individuals as a friend at $t - 1$ ("lagged common source"). The common source covariate allows a test of whether a new tie is more likely to form or an existing tie is more likely to remain intact if the presence of that tie would result in a transitive triad.

The smaller ($N = 50$, $T = 3$) TFLS network allows the impact of sampling always-null dyads to be evaluated efficiently for each of: (1) naïve analysis of the sampled data (ignoring the fact that sampling has occurred), (2) the WL approach, and (3) the ODL procedure to assess the efficacy of the WL procedure. We evaluate results when $\pi_0 = 0.005$, 0.01, 0.02, 0.05, 0.10, 0.25, 0.5, and 1 to assess the extent to which WL recovers the estimates for the full network as $\pi_0 \to 0$. To determine which parameters are the most sensitive to π_0, results are compared between the elements of (β, λ, Σ).

4.2. FHS Network

The model for analysis of the FHS network is analogous to that for the TFLS network except that x_{ijt} includes the absolute difference of age for actors i and j as a homophily covariate in place of the same smoking status covariate. Because $T = 8$, the denominator of each contribution of the ODL function evaluates and sums $4^7 = 16,384$ terms, a laborious computation that needs to be frequently performed (see Appendix). Therefore, we only apply the WL procedure to the FHS Network.

To fully evaluate the utility of WL on the FHS network we evaluate its performance with π_0 as small as 0.001. For the purpose of the bootstrap we treat dyads as independent units – a reasonable assumption in the FHS due to the fact that individuals seldom name more than one close-friend at a single wave. Computations were performed using computer code written in the C programming language and, for preliminary testing, the R statistical language. The bootstrap was implemented by running analyses in parallel on a machine with eight dual-core processors. CPU times are reported for the evaluation of point estimates of the parameters.

5. Results

5.1. TFLS Network: Recovery of Estimates for Full Network from Sampled Data

Approximately 5% of possible ties were present at any given wave. The rate of friendship transitions (formation or dissolution) across consecutive waves was about 7% (173 changes in friendship occurred out of 2256 opportunities). The observed number of triads across the three waves was 86, 88 and 137 – substantially greater than would be expected by chance and implying that accounting for triadic dependence is important (Paul and O'Malley, 2013).

It is clear from the results for the naïve method that ignoring sampling leads to substantial bias (Figure 1). The dependence of the point estimates on π_0 exhibits asymmetry for pairs of parameters in which one parameter is a modification of another (e.g., β_0 and β_1, θ_0 and θ_1) reflecting the high collinearity between them. Sensitivity of the WL and ODL estimates to π_0 is confined to the range [0, 0.05], suggesting that $\pi_0 = 0.05$ is a sufficiently large sampling probability. The fact that the results for β are the least sensitive to π_0 is a consequence of the invariant sampling probabilities. If the sampling design was covariate-specific (e.g., dyads with two or more individuals naming them as friends were sampled with high probability), the estimates of β would likely be more sensitive to the sampling probabilities.

When the analysis was repeated without random effects, the results were similar suggesting that the approximations in the random effect component of the likelihood function have minimal impact. In addition, both the WL and ODL procedures successfully recovered the full sample estimates for $\pi_0 \geq 0.02$. However, WL was inferior to ODL for $\pi_0 < 0.02$, an observation consistent with the general decline in performance of weighted estimators as the variability of the weights increases.

The variances associated with WL and ODL are expressed in Figure 2 as ratios relative to the posterior variance under the naïve method. In general, the posterior variances under WL are larger than the ODL variance estimates. The parameters for whom the variance ratios are most sensitive to π_0 are β_0 and λ_0. These parameters utilize information from the always-null dyads while estimates of the remaining parameters (β_k, λ_k for $k > 0$) are identified solely from non-null dyads, which are included in the sample with probability 1 and thus are less sensitive to π_0.

Table 1 shows the fitted model parameters for the TFLS data when $\pi_0 = 1$ and when $\pi_0 = 0.05$ for WL and ODL. The existence of an individual in the role of a common source is associated with an increase in the likelihood of tie formation whereas same smoking status was non-significant in all scenarios, particularly as π_0 decreased.

5.2. FHS Network: Comparison of CPU times

Whereas the TFLS network was based on individuals naming up to 12 friends, the FHS network was based on individuals most commonly naming a single close friend. Together with the large N, this led to fewer than 0.1% of dyads being non-null at any given wave. Due to sample attrition (e.g., due to death), the number of dyads in the sample declined from 306,687 at exam 1 to 181,329 at exam 8.

Under WL and when $\pi_0 = 0.005$, the estimates of all terms other than the elements of Σ are similar to those when $\pi_0 = 1$, implying that 0.5% is a sufficient sampling fraction under the WL procedure (Table 2). Therefore, WL appears to be an efficient means of analyzing large networks.

The age-difference homophily covariate has a significant negative effect under both the weighted and full data analyses implying that individuals with more disparate ages were less likely to form or maintain friendships. Because the design of the FHS only required that a single close-friend be named at each wave, the variability of the out-degree distribution was much smaller than for the TFLS network leading to imprecise estimates of ν, the correlation of the latent propensities of individuals to name friends (*expansiveness*) and be named as friends (*popularity*).

The reduction in CPU time as a function of π_0 was more profound for the FHS than the TFLS analyses (Figure 3). For example, the CPU time used by WL when $\pi_0 = 0.005$ is 61 times less than that of the full sample analysis on the FHS network while the analogous ratio for the TFLS is 2.7. The substantial difference of the ratios reflects the increased benefit from sampling always null dyads in large sparse networks due to the lower information content of each always-null dyad. Clearly, the larger the network the greater the utility of sampling the always-null dyads and, from a practical perspective, the more likely a computation that was unfeasible (years to complete) is to become feasible (days or hours to complete).

6. Conclusion

The analysis of large sociocentric networks encounters various methodological and computational problems. For example, in the exponential random graph (ERGM) or pstar (p^*) family of models (Frank and Strauss, 1986; Wasserman and Pattison, 1996), exact computations become unfeasible when $N > 20$ (Hunter and Handcock, 2006). Although approximate numerical estimates for ERGMs have been obtained using Markov-chain Monte Carlo (MCMC) for larger N (Goodreau, 2007), problems due to insufficient memory and computational time are commonly encountered. In the *latent space* family of models, which the models considered herein are a close derivative of, computations have been reported as becoming troublesome when $N > 1000$ (Raftery et al., 2012).

To overcome the above concerns we used informative dyadic sampling combined with a weighted-likelihood (WL) estimation method to estimate a longitudinal socio-centric network model. We validated the procedure using the small TFLS network for which likelihood-based estimates (or approximations thereof) could be computed and used as a bound to measure the performance of the WL procedure. We then showed that the WL procedure permitted rapid model estimation on the large FHS network with accurate results when using always-null dyad sampling fractions as low as 0.001.

Furthermore, because the density of ties in networks tends to decrease with N, smaller π_0 may be used thereby allowing WL to be applied to very large networks. A challenge facing WL is that bootstrap interval estimates must be computed for each re-sampled dataset. However, the bootstrap can be implemented such that each sample is analyzed in parallel, alleviating this concern. Therefore, the WL procedure has general applicability for any (N, T). Although the ODL procedure might be feasible when T is small (≤ 3), its feasibility will decline relative to that of analyzing the full network as T increases.

The dyadic sampling plan described herein is applicable irrespective of the complexity of the statistical model for the network. The only requirement is that the probability of sampling each dyad is known or is able to be accurately approximated. The always-null dyadic sampling scheme considered in this paper can be extended to allow dyads embedded in network positions of particular interest (e.g., if the individuals comprising them share certain traits or connections to common others) to be sampled with high probability or even with certainty. The mixed-effects generalized logistic regression model in (1–4) can also be extended in various ways, including allowing the addition of latent variables representing individuals' positions in a "social space" to account for contemporaneous triadic dependence (Hoff et al., 2002; Hoff, 2005, 2008). A sampling scheme that over-samples triads would make estimation of such a model less computationally burdensome.

The evaluation of standard errors for the WL procedure warrants further research. An alternative to the proposed heuristic would be to form blocks of individuals in the network such that relationship statuses in different networks are as close to independent as possible. One strategy is to use a *community detection algorithm* that partitions individuals into clusters such that the ratio of ties within clusters to ties between clusters is maximized (Newman and Girvan, 2004). A bootstrap would then re-sample the clusters without

replacement as opposed to re-sampling individual dyads. The closer the "communities" are to having no ties between them, the more accurately the bootstrap would be expected to perform. However, the low density of ties in the FHS network makes it unlikely that this more complex procedure would have obtained meaningfully different results.

Acknowledgments

Research for the paper was supported by NIH grant P01 AG031093. We thank Joel Hoff for expert programming and Nicholas Christakis and Alan Zaslavsky for helpful comments.

References

Browne WJ, Draper D, Goldstein H, Rasbash J. Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. Computational Statistics & Data Analysis. 2002; 39(2):203–225.

Christakis N, Fowler J. The spread of obesity in a large social network over 32 years. New England Journal of Medicine. 2007; 357:370–379. [PubMed: 17652652]

Christakis N, Fowler J. Dynamics of smoking behavior in a large social network. New England Journal of Medicine. 2008; 358:2249–2258. [PubMed: 18499567]

Cooper NJ, Lambert PC, Abrams KR, Sutton AJ. Predicting costs over time using bayesian markov chain monte carlo methods: an application to early inflammatory polyarthritus. Health Economics. 2007; 16:37–56. [PubMed: 16981192]

Duijn MV, Snijders TAB, Zijlstra B. P2: A random effects model with covariates for directed graphs. Statistica Neerlandica. 2004; 58:234–254.

Dunbar R. Neocortex size as a constraint on group size in primates. Journal of Human Evolution. 1992; 22(6):469–493.

Frank O, Strauss D. Markov graphs. Journal of American Statistical Association. 1986; 81:832–842.

Gladwell, M. The Tipping Point: How Little Things Make a Big Difference. Little, Brown and Company; 2000.

Goodreau S. Advances in exponential random graph (p*) models applied to a large social network. Social Networks. 2007; 29:231–248. [PubMed: 18449326]

Handcock MS, Robins GL, Snijders TAB, Moody J, Besag J. Assessing degeneracy in statistical models of social networks. Journal of American Statistical Association. 2003; 76:33–50.

Hoff, P. Advances in Neural Information Processing Systems. Vol. 20. MIT Press; 2008. Modeling homophily and stochastic equivalence in symmetric relational data; p. 657-664.

Hoff PD. Bilinear mixed effects models for dyadic data. Journal of American Statistical Association. 2005; 100:286–295.

Hoff PD, Raftery AE, Handcock MS. Latent space models for social networks analysis. Journal of American Statistical Association. 2002; 97:1090–1098.

Hunter DR, Handcock MS. Inference in curved exponential family models for networks. Journal of Computational and Graphical Statistics. 2006; 15:565–583.

Kleinbaum AM. Organizational misfits and the origins of brokerage in intrafirm networks. Administrative Science Quarterly. 2012; 57:407–452.

Langholz B, Goldstein L. Conditional logistic analysis of case-control studies with complex sampling. Biostatistics (Oxford). 2001; 2(1):63–84.

Lewis K, Kaufman J, Gonzalez M, Wimmer A, Christakis N. Tastes, ties, and time: A new social network dataset using facebook.com. Social Networks. 2008; 30:330–342.

Michell L, Amos A. Girls, pecking order and smoking. Social Science and Medicine. 1997; 44:1861–1869. [PubMed: 9194247]

Neuhaus JM, Jewell NP. The effect of retrospective sampling on binary regression models for clustered data. Biometrics. 1990; 46:977–990. [PubMed: 2085642]

Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E. 2004; 6910.1103/PhysRevE.69.026113

O'Malley AJ, Marsden PV. The analysis of social networks. Health Services & Outcomes Research Methodology. 2008; 8(4):222–269. [PubMed: 20046802]

Paul S, O'Malley AJ. Hierarchical longitudinal models of relationships in social networks. Journal of the Royal Statistical Society, Series C (Applied Statistics). 2013; 62(5):705–722.

Pfeffermann D, Skinner CJ, Holmes DJ, Goldstein H, Rasbash J. Weighting for unequal selection probabilities in multilevel models (Disc: P41-56). Journal of the Royal Statistical Society, Series B: Statistical Methodology. 1998; 60:23–40.

Raftery A, Niu X, Hoff P, Yeung K. Fast inference for the latent space network model using a case-control approximate likelihood. To appear: Journal of Computational and Graphical Statistics. 2012

Robins GL, Snijders TAB, Wang P, Handcock MS, Pattison PE. Recent developments in exponential random graph (p^*) models for social networks. Social Networks. 2007; 29(2):192–215.

Snijders, TAB. [Accessed: July 29, 2014] Description excerpt of 50 girls from "teenage friends and lifestyle study" data. 2014. http://www.stats.ox.ac.uk/~snijders/siena/s50data.htm

Spiegelhalter DJ. Bayesian graphical modelling: A case-study in monitoring health outcomes. Journal of the Royal Statistical Society, Series C: Applied Statistics. 1998; 47:115–133.

van Duijn M, Snijders TAB, Zijlstra B. P2: A random effects model with covariates for directed graphs. Statistica Neerlandica. 2004; 58:234–254.

Wasserman S, Pattison P. Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p^*. Psychometrika. 1996; 61:401–425.

West, P.; Sweeting, H. In Working Paper no. 52. Glasgow: MRC Medical Sociology Unit; 1995. Background rationale and design of the west of scotland 11–16 study.

Zijlstra BJH, van Duijn M, Snijders TAB. The multilevel p2 model: A random effects model for the analysis of multiple social networks. Methodology. 2006; 2:42–47.

Appendix

Observed Data Likelihood (ODL)

We seek the observed data likelihood (ODL) given the set of sampled dyads, denoted $p(D, \theta \mid S, X, D_1)$, where $S = \{(i, j) : S_{ij} = 1\}$ is the set of the sample inclusion indicators and X contains the covariates of all observations. Because $p(D, \theta \mid S, X, D_1) = p(D \mid S, \theta, X, D_1)p(\theta \mid S, X, D_1)$ we derive the joint posterior distribution of (D, θ) by computing $p(D \mid S, \theta, X, D_1)$ and $p(\theta \mid S, X, D_1)$ separately.

Under the assumed model, the elements of $\{D_{ij}\}_{i<j}$ are conditionally independent given θ, implying $p(D \mid \theta, X) = \Pi_{i<j}\, p(D_{ij} \mid \theta_{ij}, x_{ij})$. Therefore,

$$p(D|S,\theta,X,D_1) = \frac{p(S|D,\theta,X,D_1)p(D|\theta,X,D_1)}{p(S|\theta,X,D_1)}, \\ = \prod_{i<j:S_{ij}=1} \frac{\pi_0^{r_{ij}} p(D_{ij}|\theta_{ij},x_{ij},D_{ij1})}{\Pr(S_{ij}=1|\theta_{ij},x_{ij},D_{ij1})}, \quad (9)$$

where $\Pr(S_{ij}=1|\theta_{ij}, x_{ij}, D_{1ij}) = \sum_{\mathscr{D}_{ij}} \pi_0^{r_{ij}} p(D_{ij}|\theta_{ij}, x_{ij}, D_{ij1})$ and $\mathscr{D}_{ij}$ denotes the set of 4^{T-1} possible values of $(D_{ij2}, \ldots, D_{ijT})$. Similarly,

$$
\begin{aligned}
&p(\theta|S,X,D_1)\\
&\quad=\frac{p(S|\theta,X,D_1)p(\theta)}{P(S|X,D_1)}\\
&\quad=\frac{\prod_{i<j:S_{ij}=1}\sum_{\mathscr{D}_{ij}}\pi_0^{r_{ij}}p(D_{ij}|a_i,b_i,a_j,b_j,x_{ij},D_{ij1})\prod_i p(a_i,b_i;\sum)}{\int_{a_1,\ldots,a_N,b_1,\ldots,b_N}\prod_{i<k:S_{ij}=1}\sum_{\mathscr{D}_{ij}}\pi_0^{r_{ij}}p(D_{ij}|a_i,b_i,a_j,b_j,x_{ij},D_{ij1})\prod_i p(a_i,b_i;\sum)da_i db_i},
\end{aligned}
\tag{10}
$$

where (a_i, b_i, a_j, b_j) is substituted for θ_{ij} to make the form of the dependence on the random effects clear.

6.1. Relationship of ODL procedure to complete data maximum likelihood

We temporarily ignore the presence of θ in the model in order to (i) establish a connection between maximization of (9) and maximum likelihood estimation and (ii) show that ODL becomes computationally unfeasible as T increases, even in the absence of θ. The former provides a useful characterization of the impact of retrospective sampling and makes a connection to the general literature on retrospective sampling (Neuhaus and Jewell, 1990; Langholz and Goldstein, 2001). The latter justifies the use of WL estimation on large networks.

Connection of ODL to Maximum Likelihood—When $T = 2$ it follows that $r_{ij1} = (1 - y_{ij1})(1 - y_{ji1})$ and

$$
\begin{aligned}
\Pr(D_{ij2}|S_{ij}=1,x_{ij},D_1;\Omega)&=\frac{\pi_0^{\prod_{t=1}^{2}(1-y_{ijt})(1-y_{jit})}\exp(\mu_{ij2}y_{ij2}+\mu_{ji2}y_{ji2}+\rho_{ij2}y_{ij2}y_{ji2})}{\sum_{y_1,y_2\in\{0,1\}}\pi_0^{(1-y_{ij1})(1-y_{ji1})(1-y_1)(1-y_2)}\exp(\mu_{ijt}y_1+\mu_{jit}y_2+\rho_{ijt}y_1y_2)},\\
&=\frac{\pi_0^{r_{ij1}(1-y_{ij2})(1-y_{ji2})}\exp(\mu_{ij2}y_{ij2}+\mu_{ji2}y_{ji2}+\rho_{ij2}y_{ij2}y_{ji2})}{\pi_0^{r_{ij1}}+\exp(\mu_{ijt})+\exp(\mu_{jit})+\exp(\mu_{ijt}+\mu_{jit}+\rho_{ijt})},\\
&=\frac{\exp((\mu_{ij2}-r_{ij1}\log(\pi_0))y_{ij2}+(\mu_{ji2}-r_{ij1}\log(\pi_0))y_{ji2}+(\rho_{ij2}+r_{ij1}\log(\pi_0))y_{ij2}y_{ji2})}{1+\exp(\mu_{ijt}-r_{ij1}\log(\pi_0))+\exp(\mu_{jit}-r_{ij1}\log(\pi_0))+\exp(\mu_{ijt}+\mu_{jit}+\rho_{ijt}-r_{ij1}\log(\pi_0))},\\
&=\tilde{k}_{ijt}^{-1}\exp(\tilde{\mu}_{ij2}y_{ij2}+\tilde{\mu}_{ji2}y_{ji2}+\tilde{\rho}_{ij2}y_{ij2}y_{ji2}),\\
&=\widetilde{\Pr}(D_{ij2}|d_1,x_{ij};\tilde{\Omega}),
\end{aligned}
$$

where $\widetilde{\Pr}(D_{ij2}|x_{ij},D_1;\tilde{\Omega})$ equals $\Pr(D_{ij2}|, x_{ij}, D_1; \Omega)$ when $\tilde{\beta_k} = \beta_k - \log(\pi_0)$ for $k = 0, 3$; $\tilde{\beta_k} = \beta_k + \log(\pi_0)$ for $k = 1, 2$; $\tilde{\theta_k} = \theta_k + \log(\pi_0)$ for $k = 0, 2$; and $\tilde{\theta_1} = \theta_1 - \log(\pi_0)$ are substituted in (3) and (4). If the sampling design is generalized such that $\Pr(S_{ij} = 1 \mid D_{ij}) = \pi_1$ when $r_{ij} = 0$ then $\log(\pi_0/\pi_1)$ is substituted for $\log(\pi_0)$.

The above derivation reveals that after accounting for retrospective sampling of dyads when $T = 2$, the ODL function has the same form as the likelihood function for the full network. Therefore, estimates of the model parameters are obtained by fitting the model given by (1) – (4) to obtain estimates of $\tilde{\beta}$ and $\tilde{\theta}$ and then adding or subtracting $\log(\pi_0)$ as appropriate.

Impracticality of ODL when T is large—Let $\mathscr{D}^+$ denote the set of non-null values of D_{ij}. For general T,

$$\Pr(D_{ij}|S_{ij}=1,x_{ij};\Omega)=\frac{\pi_0^{r_{ij1}\prod_{t=2}^{T}(1-y_{ijt})(1-y_{jit})}\exp\{\sum_{t=2}^{T}\mu_{ijt}y_{ijt}+\mu_{jit}y_{jit}+\rho_{ijt}y_{ijt}y_{jit}\}}{\pi_0^{r_{ij1}}+\sum_{\mathscr{D}^+}\exp\{\sum_{t=2}^{T}\mu_{ijt}y_{ijt}+\mu_{jit}y_{jit}+\rho_{ijt}y_{ijt}y_{jit}\}}. \quad (11)$$

Because the model does not contain predictors of lag two or greater, the term $r_{ij1}\prod_{t=2}^{T}(1-y_{ijt})(1-y_{jit})$ contains indicator variables that are not inside the exponential in the numerator of (11), preventing reduction to the same form of likelihood function as for the complete data case. Therefore, specialized methods of optimizing (11) are needed.

Because the likelihood function has a closed-form expression, MCMC and other likelihood-based methods of estimation can be directly applied. However, summing the 4^{T-1} terms in the denominator of (11) for each of the $N(N-1)/2$ dyads and each time the likelihood function is evaluated, rapidly becomes infeasible as T increases. For example, if $T = 8$ (as in the FHS network), 16,384 terms are summed each time the likelihood contribution of a single dyad is evaluated. In general, ODL is impractical when T is large.

Accessing Data Used in Paper

The two longitudinal network datasets and additional individual specific covariate information are available from:

1. Excerpt of 50 students from the Teenage Friends and lifestyle study (TFL): Publicly available at: http://www.stats.ox.ac.uk/~snijders/siena/Glasgowdata.htm
2. Offspring cohort of the Framingham Heart study (FHS): Network Data can be accessed through the application process at: http://www.ncbi.nlm.nih.gov/gap. Search for "framingham social network" to locate the data. Instructions for applying for data access are on the site.

In both cases, the relational data are in the form of adjacency matrices (i.e. an $N \times N$ matrix).

Highlights

1. Estimation of statistical models for social networks is challenging
2. Dyads with no relationship ("null-dyads") are common in large social networks
3. Proposal to subsample the "always-null" dyads proposed
4. Develop weighted likelihood Bayesian estimation method
5. Method enables large social networks to be analyzed feasibly and accurately

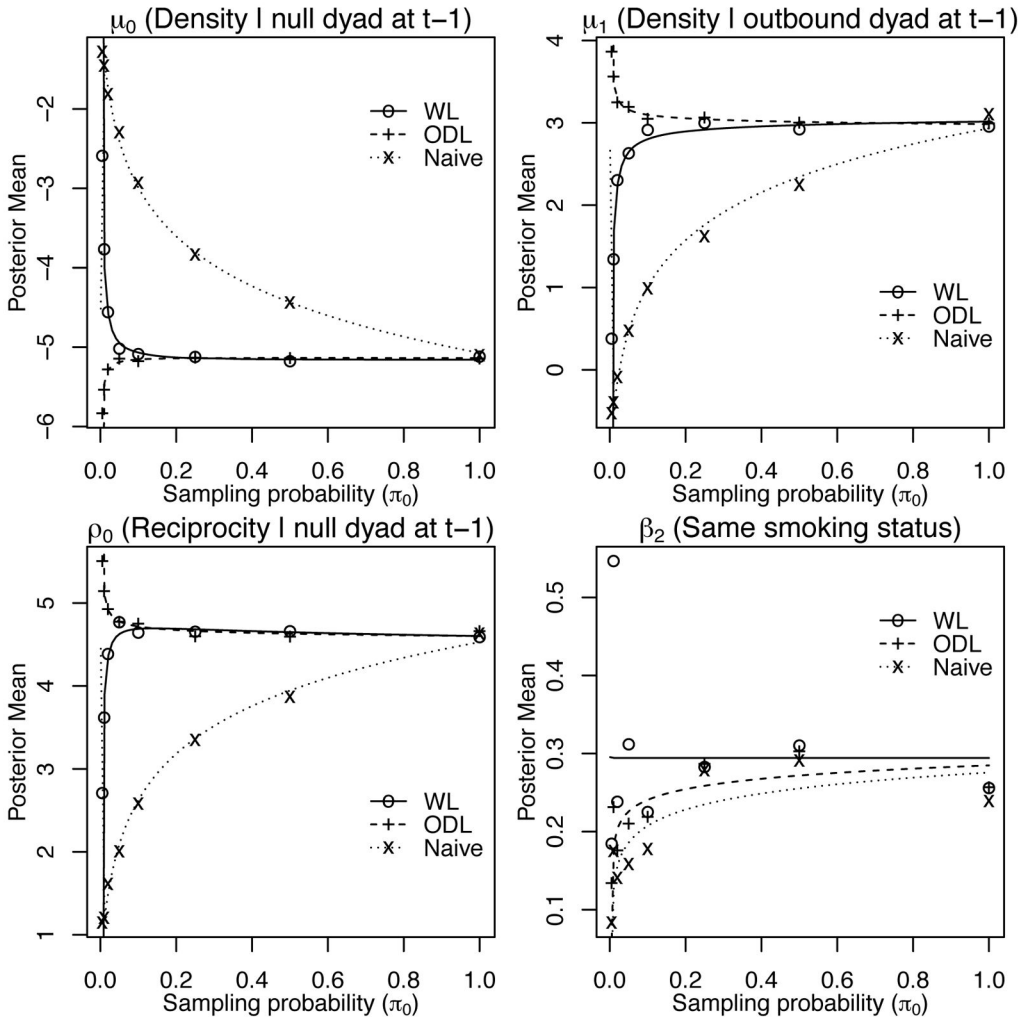

Figure 1.

Trajectories of point estimates of selected density, reciprocity, and covariate effects for the TFLS data. Fitted curves are used to help identify the point where the estimator fails to accurately approximate the estimate for the full network.

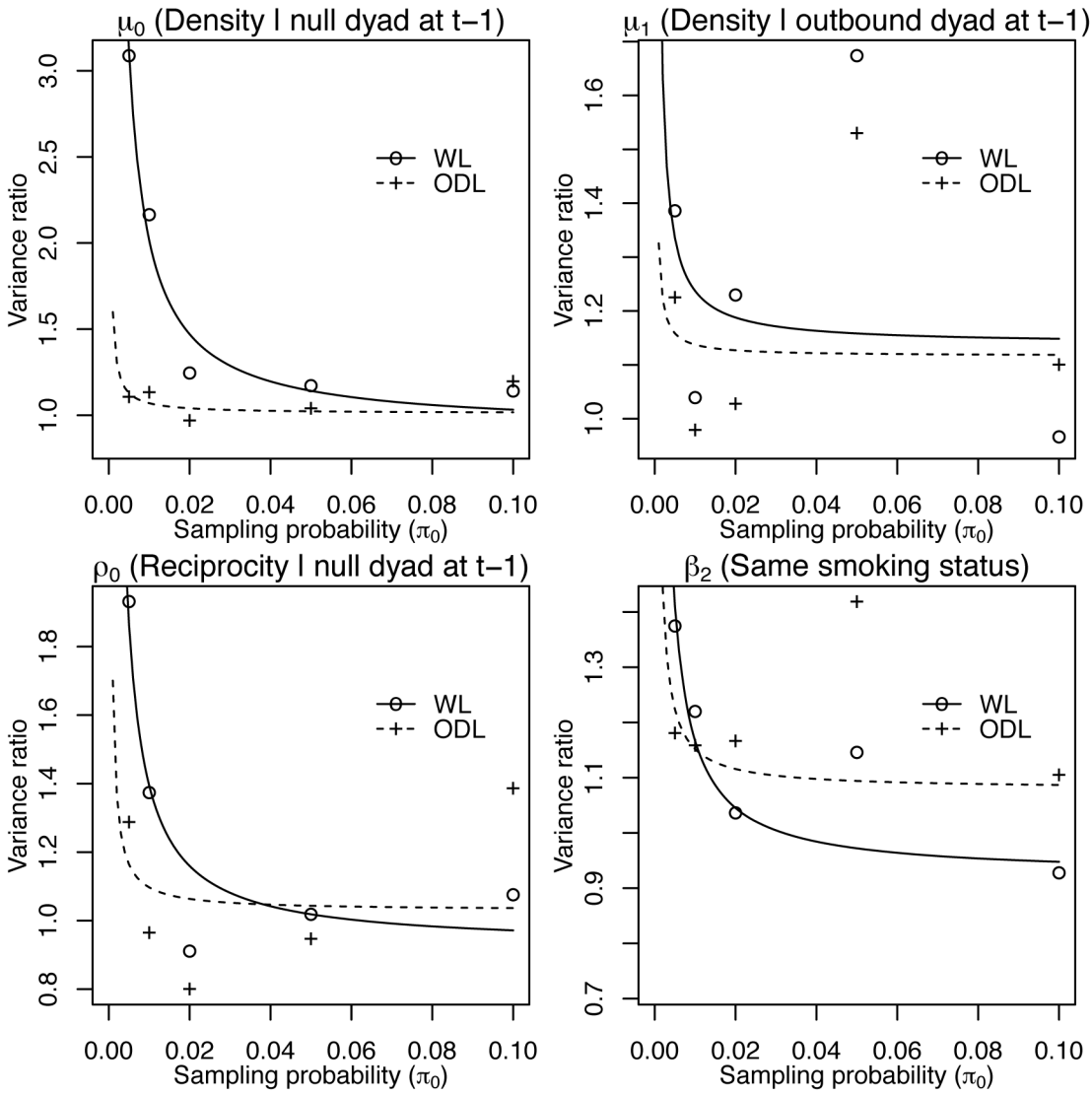

Figure 2.

Ratio of variance estimates under WL and ODL with respect to the naïve method on TFLS data. To make the point where the precision of estimation rapidly deteriorates clear, a smooth curve overlays the simulated values.

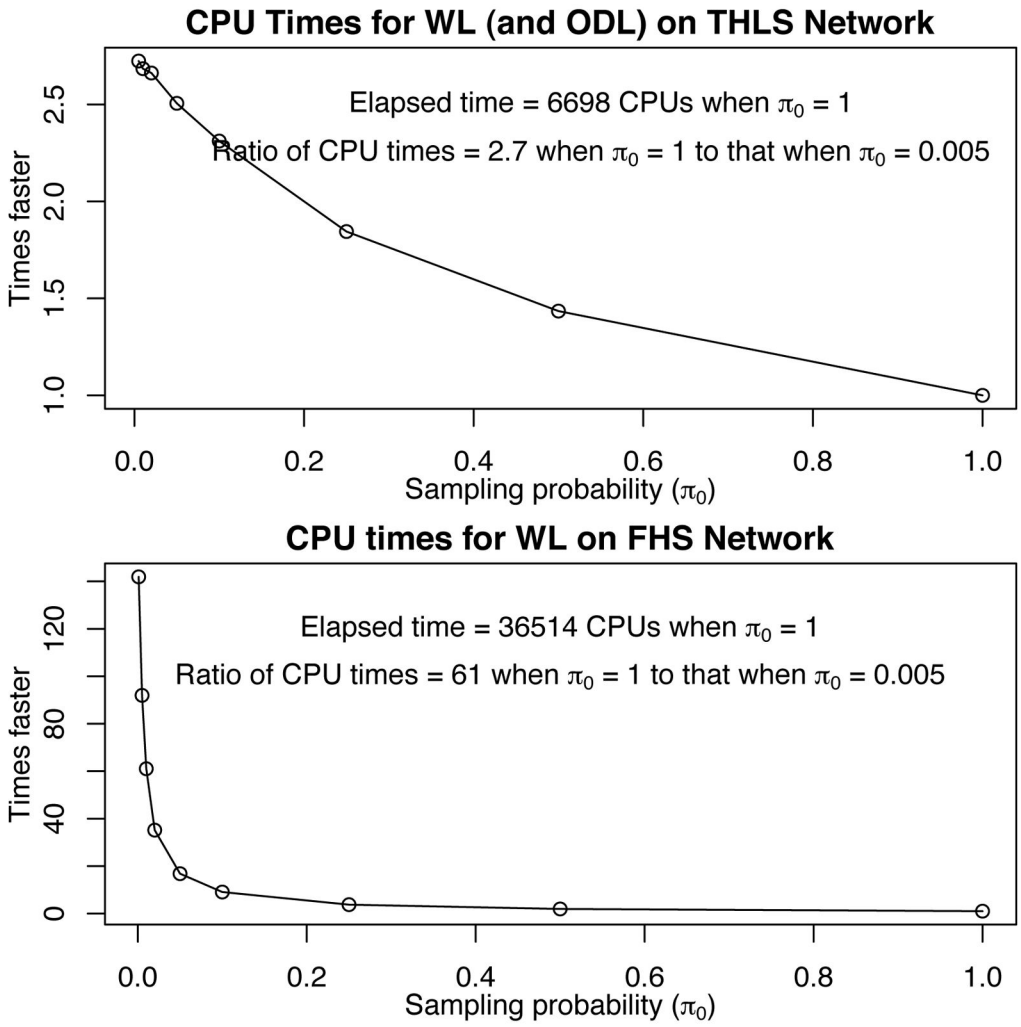

Figure 3.

CPU time ratios for WL on the TFLS network (with $T = 3$, ODL estimates are nearly indistinguishable to WL estimates) and in the FHS network (with $T = 8$, ODL is unfeasible) as a function of π_0. The plotted values are the ratios of the CPU time of the full network analysis to the CPU time of the analysis of the sampled data when always-null dyads are retained with probability π_0.

Table 1

Posterior mean estimates of the model parameters for the mixed effects model fit to the TFLS network

	Full Network		WL (5% sample)		ODL (5% sample)	
Term	**Mean**	**(2.5, 97.5)**	**Mean**	**(2.5, 97.5)**	**Mean**	**(2.5, 97.5)**
Density terms						
β_0 (baseline)	−5.144	(−5.566, −4.724)	−5.020	(−5.701, −4.588)	−5.144	(−5.622, −4.687)
β_1 (lagged 1-way)	2.994	(2.196, 3.837)	2.630	(2.204, 4.185)	3.194	(2.284, 4.028)
β_2 (lagged other way)	1.946	(0.932, 2.992)	1.882	(−0.403, 4.721)	2.159	(0.849, 3.180)
β_3 (lagged 2-way)	−1.397	(−2.977, 0.018)	−1.421	(−4.415, 0.787)	−1.814	(−3.376, −0.168)
β_x (lagged common nominator)	1.453	(1.073, 1.905)	1.637	(0.435, 2.260)	1.348	(0.857, 1.882)
β_x (same smoking)	0.257	(−0.055, 0.562)	0.312	(−0.383, 0.803)	0.210	(−0.151, 0.633)
Reciprocity						
θ_0 (baseline)	4.660	(3.908, 5.478)	4.768	(3.874, 5.673)	4.774	(4.020, 5.602)
θ_1 (lagged 1-way)	−1.131	(−2.531, 0.137)	−1.092	(−3.955, 1.024)	−1.466	(−2.927, 0.201)
θ_2 (lagged 2-way)	0.985	(−1.552, 3.697)	1.284	(−3.269, 6.350)	1.541	(−1.696, 4.449)
Covariance matrix						
σ_a^2 (sender Var)	0.456	(0.240, 0.779)	0.949	(−0.252, 1.184)	0.466	(0.223, 0.830)
σ_b^2 (receiver Var)	0.634	(0.312, 1.156)	1.494	(0.090, 1.341)	0.716	(0.313, 1.348)
ν (correlation)	−0.679	(−0.874, −0.285)	−0.603	(−0.967, 0.035)	−0.547	(−0.847, −0.032)

The weighted likelihood (WL) and observed data likelihood (ODL) estimates are for a 5% sample of always-null dyads; 95% interval estimates are enclosed in parentheses.

Table 2

Posterior mean estimates of the model parameters for the mixed effects model fit to the FHS network

	Full Network		WL (0.5% sample)	
Term	**Mean**	**(2.5, 97.5)**	**Mean**	**(2.5, 97.5)**
Density terms				
β_0 (baseline)	−8.365	(−8.525, −8.204)	−8.078	(−9.017, −8.122)
β_1 (lagged 1-way)	10.103	(9.923, 10.276)	9.998	(10.010, 11.238)
β_2 (lagged other way)	4.953	(4.044, 5.651)	4.797	(1.567, 6.108)
β_3 (lagged 2-way)	−5.286	(−6.204, −4.098)	−5.006	(−6.468, −1.734)
β_x (lagged common nominator)	2.532	(−0.448, 4.320)	2.175	(−0.582, 2.206)
β_x (difference in age)	−0.051	(−0.063, −0.038)	−0.064	(−0.076, −0.019)
Reciprocity				
ρ_0 (baseline)	5.441	(4.490, 6.101)	5.005	(1.883, 6.295)
ρ_1 (lagged 1-way)	−4.558	(−5.491, −3.161)	−4.417	(−5.919, 0.958)
ρ_2 (lagged 2-way)	4.633	(2.357, 6.237)	4.504	(−3.550, 6.845)
Covariance matrix				
σ_a^2 (sender Var)	0.104	(0.068, 0.143)	0.138	(0.007, 0.011)
σ_b^2 (receiver Var)	0.211	(0.110, 0.323)	0.600	(0.161, 1.703)
ν (correlation)	−0.309	(−0.740, 0.225)	0.654	(−0.686, 0.843)

The weighted likelihood (WL) estimates are for a 0.5% sample of always-null dyads; 95% interval estimates are enclosed in parentheses. Observed data likelihood (ODL) estimates are not available as the calculations are too laborious when T = 8.