Statistica Sinica

Dynamic Network Analysis with Missing Data: Theory and Methods

Zack W. Almquist
University of Minnesota

Carter T. Butts
University of California, Irvine

Abstract: Statistical methods for dynamic network analysis have advanced greatly in the past decade. This article extends current estimation methods for dynamic network logistic regression (DNR) models, a subfamily of the Temporal Exponential-family Random Graph Models, to network panel data which contain missing data in the edge and/or vertex sets. We begin by reviewing DNR inference in the complete data case. We then provide a missing data framework for DNR families akin to that of Little and Rubin (2002) or Gile and Handcock (2010a). We discuss several methods for dealing with missing data, including multiple imputation (MI). We consider the computational complexity of the MI methods in the DNR case and propose a scalable, design-based approach that exploits the simplifying assumptions of DNR. We dub this technique the "complete-case" method. Finally, we examine the performance of this method via a simulation study of induced missingness in two classic network data sets.

Introduction

Interest in the collection and analysis of dynamic network data has increased dramatically over the past decade (e.g., Snijders (2005); Almquist and Butts (2014b); Krivitsky (2012); Hanneke et al. (2010)). This growth stems primarily from advances in computational resources and statistical theory (particularly developments in the simulation and analysis of relational data), with additional impetus stemming from the rapid growth of Internet-based data collection (e.g., Leskovec (2011)). The scientific study of dynamic networks is pervasive in the social sciences, arising in the context of problems such as the evolution of friendship ties (e.g., Newcomb (1961)),

communication and face-to-face interaction over time (e.g., van de Rijt (2011)), the dynamics of disease transmission networks (e.g., e.g., sexual contact networks or needle sharing networks; Entwisle et al. (2007)), emergent organizational networks during disaster response (e.g., Carley (1999)), and organizational collaboration dynamics (Powell et al. (1996)). Beyond the social sciences, dynamic networks have been explored in computer science (e.g., online networks, see Leskovec (2008)), physics (e.g., coevolution of dynamical states and interactions, see Zimmermann et al. (2004)) and engineering (e.g., human, cyber and physical traffic engineering, see Wang et al. (2006)), among other fields.

Here we consider a *network* to be any system that can be represented as a *graph*, where a graph is defined by two sets: a vertex set (V), and a edge set (E) consisting of ordered or unordered pairs from V (reflecting undirected and directed relations, respectively). In a network context, edges pertain to relations (e.g. friendship or chemical bonds) and vertices generally represent entities (e.g. Mike or Microsoft). Modern data collection through sensors (e.g. cell phones), surveys, and online social network systems (OSNs) have allowed for larger and more detailed network data collection efforts, especially in the area of dynamic networks; however, even with improved measurement tools there still exists the persistent problem of *missing data*, either by design (e.g. sampled data) or out of design (e.g. machine failure). Thus, the collection of large dynamic networks often results in various types of missingness, which can complicate analysis (Hipp

et al. (2015)). Particular complexities arise with missing data in a dynamic context because most plausible temporal network models (e.g., Hanneke et al. (2010); Cranmer and Desmarais (2011)) rely on conditioning on the past. Conditioning on the past can yield missing data on both the dependent and independent variables, and accounting for missingness raises both computational and theoretical challenges.

Here, we consider the case of so-called network panel data, a series of network snapshots over time. The framework we employ builds on the *exponential-family random graph models* (ERGMs), positing that each network in the time series is drawn from a discrete exponential family conditional on past draws and exogenous covariates. This class of models is often referred to as temporal ERGMs or TERGMs (Hanneke et al. (2010)). Further, under certain conditions (primarily, conditional independence of edge and/or vertex states in the present given the past) these models have a conditional Bernoulli structure closely resembling logistic regression; members of this class are referred to as *dynamic network logistic regression* (DNR) models (Almquist and Butts (2014b)), and are of particular interest because of their simplicity, interpretability, and computational scalability. While generally employed for networks with fixed or exogenously changing vertex sets, TERGM (and DNR) can be further extended to model endogenous vertex dynamics (Almquist and Butts (2014b)).

To date, work on TERGMs has assumed complete network data (i.e., no missing edges or vertices); however, given that each network cross-section

contains $\mathcal{O}(N^2)$ edge variables, there is considerable opportunity for missingness to occur. Here, we propose a general framework for conceptualizing missingness in a network panel data context. In this we build on the work on cross-sectional ERG inference with missing data introduced by Gile and Handcock (2010a) and Koskinen et al. (2010), as well as the broader statistical literature on missing data (for a review see e.g. Little and Rubin (2002)). Further, we discuss some specific implications of *missing at random* (MAR) data for DNR, and provide a computationally scalable approach to parameter estimation for DNR families with ignorably missing data. When N is large, complete enumeration of an entire network often becomes infeasible, resulting in the omission of nodes and/or edges either unintentionally or by design. With respect to the latter, there are now numerous methods for acquiring probability samples of network data; these include: uniform or weighted independence sampling of nodes (for a detailed review see, Kolaczyk (2009)), respondent-driven sampling methods (RDS) (for a review see Gile and Handcock (2010b)), and random-walk methods (e.g., Gjoka et al., 2010). While it is well known that discrete exponential-family models for cross-sectional networks can be quite challenging to model, both theoretically (see e.g. Schweinberger and Handcock (2015)) and computationally (see e.g., van Duijn et al. (2009)), it has been shown that in certain contexts it may be easier to model dynamic networks conditioned on the past (e.g., Desmarais and Cranmer (2012)). This effect appears to be largely due to the ability to leverage past information to reduce the strong depen-

dence that cross-sectional designs with very little covariate information are forced to model (a special case of the "strong covariate" effects shown by (Butts (2011), p332). Dynamic networks – when available – not only allow more direct investigation of social mechanisms, but can thus also be easier to work with Almquist and Butts (2014b).

Dynamic Network Logistic Regression

A social or other network on vertex set V and edge set E is often represented as a graph ($G = (V, E)$), where the *size* of the graph is defined by the number of vertices ($N = |V|$). This framework can be extended to incorporate a temporal dimension by indexing each set by time (i.e., $G_t = (V_t, E_t)$ is the state of G at time t). A dynamic graph may also be represented as a series of *adjacency matrices*, $\ldots, Y_t, \ldots$ with $Y_t \in \{0,1\}^{N_t \times N_t}$, such that Y_{ijt} is an indicator for the presence of an edge from vertex i to vertex j at time t, and the size of the network is the row or column dimension of Y_t (N_t). Lower case y_{ijt} will represent the observed edge value. TERGMs are generally specified in a manner similar to VAR processes via a kth order temporal Markov assumption. Specifically, let Y_t be the state of the network at time t, given vertex set V_t. We then assume that, for all times t, $Y_t|Y_{t-1}, \ldots, Y_{t-k}$ is independent of $Y_{t-k-1}, \ldots$. The general TERGM framework (for full details see the Core Concepts Section in the online supplement), like ERGMs, can parametrize an extremely broad class of models, not all of which are statistically or computationally tractable. While well-

specified ERGMs have been successfully used to study a wide range of social phenomena, poorly chosen ERGMs can have issues of instability, sensitivity, degeneracy, and scalability (challenges that have spawned a literature in their own right; see e.g., Handcock (2003); Butts (2011); Schweinberger and Handcock (2015)). Hanneke and Xing (2007) and Hanneke et al. (2010) have shown that the general TERGM case includes model families with similar properties, but also that under certain conditions the inclusion of temporal structure can improve model behavior; this work naturally leads to the assumptions underlying dynamic network regression, the most important of which is that edge variables are independent in the present conditional on the past (and any covariates). Here we follow the approach of Almquist and Butts (2014b), who modeled the vertex and edge set coevolution as separable DNR processes with vertex (2) and edge (1) likelihoods:

$$\Pr(V_t \mid Z_{t-k}^{t-1}, X_t) = \prod_{i=1}^{n} B\left(\mathbb{I}(v_i \in V_t) \left| \text{logit}^{-1}\left(\psi^T w(i, Z_{i-k}^{i-1}, X_t)\right)\right.\right), \tag{1}$$

$$\Pr(Y_t \mid V_t, Z_{t-k}^{t-1}, X_t) = \prod_{(i,j)\in V_t \times V_t} B\left(Y_{ijt} \left| \text{logit}^{-1}\left(\theta^T u(i, j, V_t, Z_{i-k}^{i-1}, X_t)\right)\right.\right), \tag{2}$$

where B is understood to be the Bernoulli pmf, $\mathbb{I}$ is the indicator function, X_t is a covariate set (potentially including dynamic latent variables, see supplement for discussion), $Y_{t-k}^{t-1} = Y_{t-1}, \ldots, Y_{t-k}$ is the graph structure given the vertex set from time $t-k$ to $t-1$, $Z_t = (Y_t, V_t)$ is the joint vertex and edge set structure, $Z_{t-k}^{t-1} = Z_{t-k}, \ldots, Z_{t-1}$ is the joint edge/vertex set

structure from time $t-k$ to $t-1$, u and w are sufficient statistics for the edge and vertex models (respectively), and θ and ψ are the respective edge and vertex parameter vectors. It can often be assumed that V_t is fixed, in which case the above reduces to a DNR family on Y alone. As the form of Eq 1-2 suggests, inference for θ and ψ reduces to logistic regression (hence the term "DNR"). Missing data introduces some complications, however, as we discuss below.

Ignorable Missingness and DNR

Rubin (1976) introduced a typology for typical forms of missingness in social science data, that has served as the basis for a widely used framework for modeling missing data (Rubin (1976); Little and Rubin (2002)). This framework allows for arbitrarily complicated forms of missingness if the mechanism of missingness is known; however, under typical situations the mechanism of missingness is not known and is often assumed to be either Missing Completely at Random (MCAR) or Missing at Random (MAR). These may be defined as follows. Let R be an indicator function such that $R = 1$ if random variable Y is observed and $R = 0$ if Y is missing. This naturally defines a model for the missing data process, $Pr(R = r \mid Y = y) = f_{R|Y}(r \mid y, \xi)$ with parameters $\xi \in \Xi$ governing the missing data mechanism. With this notation we can define MCAR to be the case when $Pr(R = r \mid Y = y) = Pr(R = r)$, equivalently $f_{R|Y}(r|y, \xi) = f_R(r, \xi)$, and MAR to be the case when $Pr(R = r \mid Y = y) = Pr(R = r \mid Y_{obs})$, equivalently $f_{R|Y}(r|y, \xi) = f_{R|Y_{obs}}(r|y_{obs}, \xi)$. MAR is typically interpreted

as implying that knowledge about Y_{mis} does not provide any additional information about R if Y_{obs} is already known. Of these assumptions, we generally prefer the weaker MAR condition.

We begin by extending our notation to the missing data case (for more details see the supplement). Following the development of Little and Rubin (2002), we decompose the complete data into an *observed* part (Y^o) and a *missing* part (Y^m). Under the adjacency matrix characterization of a graph the missing data are definitionally the edge variables (i.e., y_{ij}^m) when not considering vertex dynamics. Similarly, in the vertex dynamics case we take V^o to be the set of vertices whose presence or absence is observed, and V^m the vertices that are missing (not observed). Since an edge can only be present if its endpoints are present, it further follows that $v_i \in V^m$ implies that all y_{ij}, y_{ji} are missing for all j. Focusing on the fixed-vertex case, it is convenient to express the likelihood function for the model of (2) as

$$L(\theta \mid Y_t) = \prod_{t=k}^{T} \prod_{i,j=1}^{N_t \times N_t} f_Y(y_{ijt} \mid Y_{t-k}^{t-1}, X_t, S(Y_{t-k}^{t-1})), \tag{3}$$

where S is a function of sufficient statistics of the graph or graph sequence (e.g. degree or the triad census), the X_t are exogenous covariates (possibly time varying) and T is the length of the dynamic data. We can then define the *observed data likelihood* as the integral of the joint likelihood over the possible states of the missing data (weighted by the probability of the specific pattern of observations obtained, R). Under the assumption that R is ignorable, it follows that

$$L(\theta, \xi | Y_t^o, R) = \int f_{R|Y_t}(R|Y_t^o, y_t^m|\xi) f_{Y_t}(Y_t^o, y_t^m|\theta) dy_t^m \propto f_{R|Y_t}(R|Y_t^o, \xi) l(\theta|Y_t^o), \tag{4}$$

where ξ is a vector of parameters related to the inclusion pattern. (Additional discussion of the assumptions involved is contained in the Section on Missing at Random in the online supplement.) Although this is in principle straightforward, in practice it may be very difficult to compute (particularly when the number of missing variables is large). We thus propose a simplified approach, based on what we call the "complete-case likelihood," which permits scalable inference at the expense of some loss in statistical efficiency.

The Complete-case Likelihood

We focus here on the common use case of DNR in which the vertex set is fixed, leaving us with a model on the set of edge variables. Our development begins by allocating these to three sets, based on observability: (1) $O_t := \{(x, y, t) \,|\, x, y, t \in \{R = 1\}\}$; (2) $C_t := \{(x, y, t) \mid (x, y, t), \ldots, (x, y, t-k) \in \{R = 1\}\}$; and (3) $M_t^c := \{(x, y, t) \mid \{(x, y, t), \ldots, (x, y, t-k)\} \cap \{R = 0\} \neq \varnothing\}$. We take $N_t^o = |O_t|$ and $N_t^c = |C_t|$. This, together with the DNR and ignorability assumptions, allows us to write the likelihood of the observed data in terms of a collection of edge variables within each time point;

$$L(\theta \mid Y_t^o) = \prod_{t=k}^{T} \prod_{i,j \in O_t} f_Y(y_{ijt}^o \mid Y_{t-k}^{t-1}, X_t, S(Y_{t-k}^{t-1})), \tag{5}$$

bearing in mind that some values on which we are conditioning may not be observed (an issue to which we return below). Notice that MAR and distinctness guarantees that the maximizer of the observed-data likelihood is the MLE (i.e., $\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta, \xi | Y_t^o, R) \iff \hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta | Y_t^o)$). Little and Rubin (2002) point out that MAR is typically regarded as the more important condition in ignorability, in the sense that if the data are MAR but separability does not hold, inference based on the ignorable likelihood is still valid from the frequentist perspective, but not fully efficient. Thus, the face value likelihood of y^o contains less information than the true observed data likelihood (of y^o and R, jointly), but can still lead to estimators with other good properties. An alternative strategy is to transform this problem into one that is relatively easy and has comparable asymptotics to the "non-missing" version of DNR. In particular, we can reframe our problem as one arising from sampling theory. We began our consideration of this problem from the perspective of a logistic regression on the complete data with missingness; however, we can alternatively think of the observed data as a *random sample of dyads* arising from a population. Specifically, we can view the observed edge variables as a sample of n dyads from the set of all dyads, rather than as a complete dyad set from which some members are missing. This allows us to exploit the conditional independence of DNR and the MAR assumptions to derive a likelihood from the observed data that is

computationally facile and has good asymptotic properties, at the cost of some loss of efficiency. Here, we define this *complete-case likelihood* to be the likelihood of the data that is observed in both Y_t and Y_{t-k}^{t-1}. Following from the DNR and MAR assumptions, this likelihood can be written as

$$L(\theta \mid Y_t^o, (Y_{t-k}^{t-1})^o) = \prod_{t=k}^{T} \prod_{i,j \in C_t} f_Y(y_{ijt}^o \mid (Y_{t-k}^{t-1})_{ij}^o, X_t, S_{ij}(Y_{t-k}^{t-1})). \qquad (6)$$

We take $\hat{\theta}_{cc} = \arg\max_{\theta \in \Theta} L(\theta | Y_t^o, (Y_{t-k}^{t-1})^o)$ to be an estimator of θ_0, some of whose asymptotic properties we prove below. As a convenience for this purpose, we define N_t^o to be the cardinality of the set of observed edge variables and N_t^m the cardinality of the set of missing edge variables.

We provide some observations and a theorem regarding this estimator. For the moment, imagine that we can always observe the needed elements from the past so that we can calculate $\Pr(y_t^o | y_{<t})$. While this is clearly not realistic in many settings of interest, the property itself will be informative for our later development (where we will weaken this condition). In particular, to the extent that we can approximate this condition, we will be able to closely approximate the desired likelihood. Under this assumption, the complete-case likelihood trivially converges to the full likelihood as $N_t \to \infty$ such that $N_t^m / N_t \to 0$. It follows that $\hat{\theta}_{cc}$ approaches $\hat{\theta}$ as the fraction of missing data goes to zero with $\hat{\theta} \to N(\theta_0, [I(\hat{\theta})]^{-1})$, where $I(\hat{\theta})$ is the expected (Fisher's) information. These observations hold under weak regularity conditions on f by noting that if $N_t^m = 0$ and $N_t^o > 0$, then $y_t^o = y_y$, and hence $f(y_t^o) = f(y_t)$ for all f, y_t. We summarize the key

results as follows:

Theorem 1. *Let $\hat{\theta}_{cc}$ be the maximizer of the complete case likelihood for $y^o \sim Y, R$ with finite parameter θ_0, where Y is a DNR family of finite order k with finite, affinely independent statistics, and ignorable measurement process R. Then*

(i) If $\hat{\theta}$ is the MLE of θ_0 under the complete data, y and $\left(\sum_{t=1}^{T} |M_t^c|\right) / \left(\sum_{t=1}^{T} N_t^c\right) \to 0$, $\hat{\theta}_{cc} \to \hat{\theta}$

(ii) As $\left(\sum_{t=1}^{T} N_t^c\right) \to \infty$, the sampling distribution of $\hat{\theta}_{cc}$ converges to $N(\theta_0, [I(\theta_0)]^{-1})$, and $[I(\hat{\theta}_{cc})]^{1/2}(\hat{\theta}_{cc} - \theta_0) \to N(0, I)$.

Proof. We define the CC observation process, $R_t^{'}$, in terms of R_t (the time-indexed missing data mechanism) as $R_t^{'} = R_t \cap R_{t-1}^{t-k}$. We can then write down our CC likelihood in terms of $Y^{cc} = \{y_{ijt}^o \mid y_{ij,t}^o \in Y_t^o \text{ and } y_{ij,t-k}^o \in (Y_{t-k}^{t-1})^o\}$ and $R_t^{'}$:

$$L(\theta, \xi | Y_t^o, (Y_{t-k}^{t-1})^o) = \prod_{ij \in C_t} \prod_{t=k}^{T} \Pr(Y_{ijt} | Y_{t-k}^{t-1}, \theta) \Pr(R_t^{'} | Y_{t-k}^{t-1}, \xi). \tag{7}$$

By MAR, we may factor $R_t^{'}$ from (7). MAR and separability of parameters further imply that $R_t^{'}|Y^{cc}$ is constant with respect to θ. We can then write our CC likelihood as

$$L(\theta, \xi | Y_t^o, (Y_{t-k}^{t-1})^o) \propto \prod_{ij \in C_t} \prod_{t=k}^{T} \Pr(Y_{ijt} | Y_{t-k}^{t-1}, \theta) \propto L(\theta | Y_t^o, (Y_{t-k}^{t-1})^o). \tag{8}$$

By the definition of the DNR family, this likelihood is equivalent to that of a logistic regression with fixed parameter θ and data degrees of freedom

equal to the size of the complete case set, and $\hat{\theta}_{cc}$ is equivalent to the MLE of θ in the corresponding problem. It is then a standard result (e.g. McCullagh and Nelder, 1999) that, for true, finite parameter θ_0 and affinely independent statistics, $\hat{\theta}_{cc}$ will converge to $N(\theta_0, [I(\theta_0)]^{-1})$ in distribution as $\left(\sum_{t=1}^{T} N_t^c\right) \to \infty$, and $[I(\hat{\theta}_{cc})]^{1/2}(\hat{\theta}_{cc} - \theta_0) \to N(0, I)$. This establishes (ii). For (i), we observe that, under the assumed conditions, $L(\theta|Y_t^o, (Y_{t-k}^{t-1})^o) \to L(\theta|Y_t, (Y_{t-k}^{t-1}))$ as $\left(\sum_{t=1}^{T} |M_t^c|\right) / \left(\sum_{t=1}^{T} N_t^c\right) \to 0$, and hence a limiting maximizer of the former must also be a maximizer of the latter. (i) follows immediately from the definitions of $\hat{\theta}$ and $\hat{\theta}_{cc}$.

We do not treat the derivation of the information matrix here in detail, but note that the equivalence of the complete-case likelihood to a standard logistic regression problem implies that conventional approaches (e.g., approximation via the inverse Hessian of the log-likelihood) from the latter case apply here as well. The equivalence of the complete case likelihood and the logistic regression likelihood also implies Gaussian posterior asymptotics under standard Bayesian theory (Gelman et al. (2004)) precisely as in the case of the logistic regression for ignorably sampled data, and with the same caveats. For a further discussion of Bayesian analysis of DNR with/without vertex dynamics see Almquist and Butts (2014a).

Approximation of the Complete-case Likelihood

These results seem to suggest that for DNR we can avoid the entire issue of imputation; however, the assumption that we can exactly calculate

$S_{ij}(Y_{t-k}^{t-1})$ (for all $i, j, t-1, \ldots, t-k$ of interest) completely from the observed data is not always true. In special cases it is possible to use the complete-case likelihood additionally constrained to only the cases $S(Y_{t-k}^{t-1})^o$ are also observed; this most often occurs at low levels of missingness when the graph statistics of interest are local (e.g., degree). In general, however, missing edge variables have effects that propagate into S_{ij} in ways that make subsetting alone insufficient. To illustrate this point, consider a simple example. Let S be a statistic on Y_t that outputs the average degree for the endpoints of a given Y_{ij} edge variable. S then depends on the values of all Y_{ik}, Y_{kj} for $k \in \{V_t \setminus i, j\}$, and cannot be exactly computed if any of these are missing. Since each missing edge variable here interferes with the statistic on $2(|V_t| - 2) + 1$ edge variables in each time slice, it is apparent that only a small number of missing edges (here $\mathcal{O}(|V_t|)$ per slice) are needed to prevent exact calculation of the complete-case likelihood in the worst case. Thus, we often cannot calculate S (and hence the complete-case likelihood) exactly from the observed data. We can, however, *approximate* S with a reasonable estimator in many contexts of interest, substituting $\hat{S}_{ij}$ for S_{ij} in the likelihood calculation. We then work with the resulting approximation to the complete-case likelihood,

$$l(\theta \mid Y_t^o, (Y_{t-k}^{t-1})^o, \hat{S}) = \prod_{t=k}^{T} \prod_{i,j \in C_t} f_Y(y_{ijt}^o \mid (Y_{t-k}^{t-1})_{ij}^o, X_t, \hat{S}((Y_{t-k}^{t-1})_{ij}^o)). \qquad (9)$$

Clearly, these results for the CC likelihood hold in the limit when $\hat{S} \to S$. Often it is possible to choose an $\hat{S}$ that approximates S well, though this

cannot be guaranteed. Thus this process transforms a problem of TERGM estimation with missing data to a problem of logistic regression with measurement error, which can be dealt with in standard ways. Different strategies for handling the $\hat{S}$ function can be employed, allowing the researcher to exploit properties of the model statistics and/or missingness process on a case-by-case basis.

The $\hat{S}$-function

Given that observations of the graph Y_t contain missingness, it follows that $S(Y_{t-k}^{t-1})$ must typically be approximated by $\hat{S}((Y_{t-k}^{t-1})^o)$; in general, this will introduce some level of error into our measurement of a given graph statistic of interest. To illustrate this point, consider a simple graph statistic such as the indegree of actor j for a the case with a single lag Y_{t-1} and no missing vertices. In this case, $S_j(Y_{t-1}) = \sum_{i=1}^{n_{t-1}}(Y_{t-1})_{ij}$. Where there is missingness in Y_{t-1}, we can decompose this function into its observed and missing components, e.g. $S_j(Y_{t-1}) = \sum(Y_{t-1}^o)_{ij} + \sum(Y_{t-1}^m)_{ij}$ (for the fixed vertex case), where the latter term is unobserved in our setting. Thus, introducing an estimate for the unknown term leads to an error in the associated statistic, which in the case of DNR is equivalent to the introduction of error to a regression covariate.

To understand the extent of this error and to provide intuition as to how effective simple heuristic imputation schemes might be in practice, we begin by proposing three naive approximation schemes for S that represent

various limiting cases. The first estimator we refer to as the "0 estimator," $\hat{S}^0$, which treats all missing edge variables as if they had values of 0 for purposes of calculating sufficient statistics, each $(Y_{t-k}^{t-1})_{ij}^m = 0$. (Huisman and Steglich (2008) used a related idea in one stage of a model-based imputation scheme for actor-level missingness in SAB models.) The second is the "1 estimator," $\hat{S}^1$, which treats all missing edge variables as if they had a value of 1 for calculative purposes, $(Y_{t-k}^{t-1})_{ij}^m = 1$. The third estimator is the "density estimator," which can be thought of as a simple "grand mean" imputation strategy.

The density estimator has the potential for further variation based on whether we compute a density over the entire time period for imputation, or if we compute the density at each individual time period for a separate imputation estimator at each time period. In the first case, we take all missing edge variables to be independent Bernoulli trials with success probability equal to the fraction of 1's in the observed data, while in the second we instead employ the fraction of 1's at each respective time step. Further elaboration by subsetting density between covariate-defined classes of vertices is possible, though we do not pursue it here.

We also considered a fourth class of imputation schemes based on *local prediction*, where an estimator for missing values in $\hat{S}$ is to be found through a naive statistical model fit to the observed data (R-imputation). Here we considered simple linear regression. This amounts to approximating $\hat{S} = Z^T\beta + \epsilon$. We recommend assessing any proposed training data model for

prediction validation. Here, separate out the calculable ("observed") and incalculable ("missing") S values S_{ij}^o and S_{ij}^m, selecting a training fraction α (in the following Sections we take $\alpha = 3/4$th) of the Y_{ij}^o edge variables, fitting a linear model to the training data, and evaluating its prediction accuracy with the remaining $1 - \alpha$ of the observed data. The model with the highest prediction accuracy is taken to be the best model for imputing the missing S_{ij}^m values, we replace S_{ij}^m with S_{ij}^{pred}.

Simulation Analysis

Having introduced a set of simple heuristics for complete-case likelihood approximation, we explored the performance of these approaches when applying DNR to data with missingness under MAR assumptions. To do this we worked with two data-sets. First, we employed a dynamic inter- and intra-group blog citation network with a fixed vertex set; second, we used a month of daily interpersonal communication data on windsurfers congregating on a beach in Southern California (a data set that varies with respect to both edges and vertices). For full details on the computational methods employed, see the Computation Subsection in the online supplement.

Our choice of missing data percentages were based on the empirical literature for static and dynamic networks. Missing data rates in social networks has been observed to be up to 20% for survey data at a single time point (Brewer and Webster (2000)); in the AddHealth survey up to 75% of edges over the three waves of data can be missing (Hipp et al. (2015));

in the context of political blogs, Adamic and Glance (2005) estimated 15% missing for the vertex set of interest and 32% edges over 624 days; and in a dynamic Twitter network of US Emergency Management-related Organizations was missing due to computer failure (Almquist et al. (2016)). Here we employed the following statistics for describing missingness levels: the fraction of vertices missing, the fraction of edges missing, and the fraction of edge and vertices missing for a given number of time points, e.g. a vector (p_v^m, p_e^m, t^m).

Data: Blog Citation Network: Our first data set involves observations of a dynamic inter- and intra-group blog citation network collected by Butts and Cross (2009) and analyzed with DNR in Almquist and Butts (2013). This temporal network consists of interactions among all blogs credentialed by the U.S. Democratic National Committee (DNC) or Republican National Committee (RNC) for their respective 2004 conventions. (For full details see Section Data: Blog Citation Network in the online supplement).

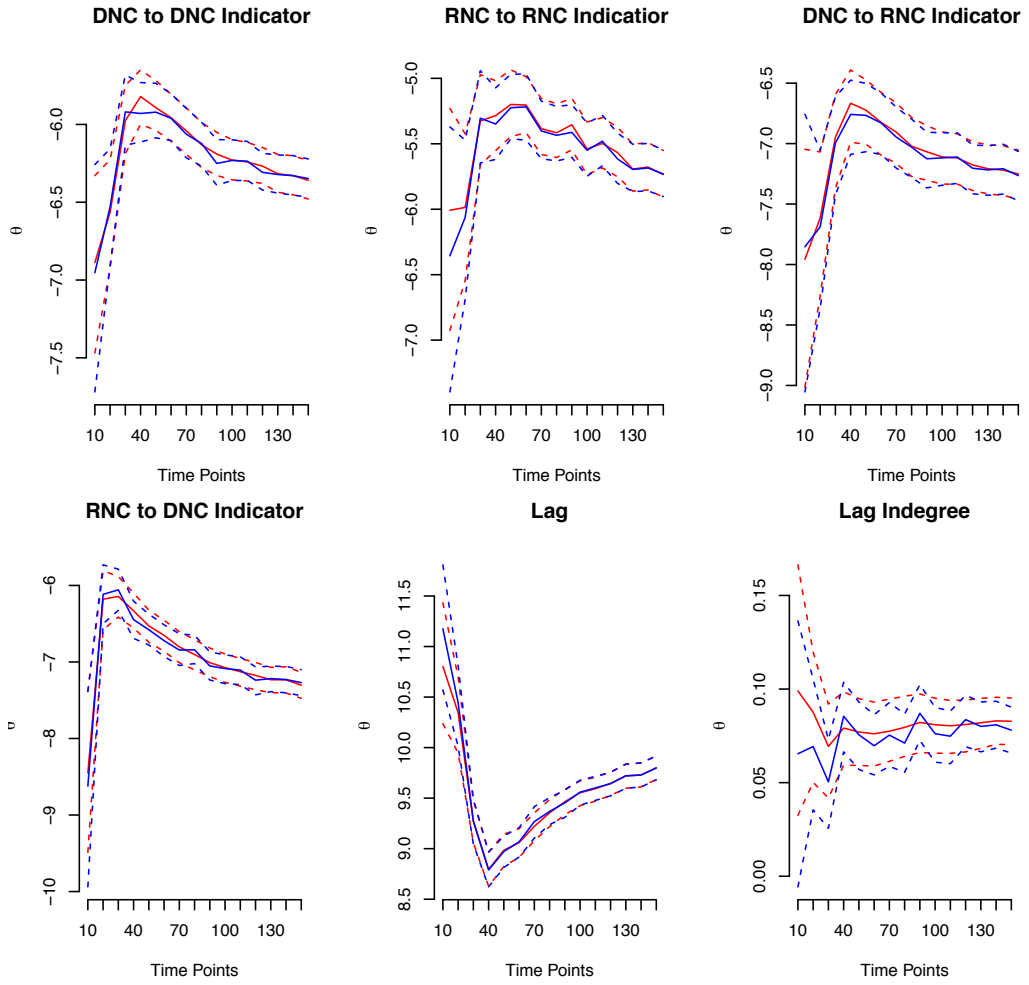
Simulation Study of the Effects of Missingness: To assess efficiency and accuracy we first examined parameter estimates for θ – in this case Bayesian parameter estimates under diffuse t priors (the ML estimate results are similar) – for the case with missingness in the edges at given time points (e.g. 10% in the number of zeros and 10% in the number of 1's at time t). We fit a simple model to the blog data. The model contained three statistics: a mixing statistic which is constituted by four dummy variables (DNC to DNC indicator, RNC to RNC indicator, DNC to RNC indicator,

and RNC to DNC indicator), a lag statistic (in this case Y_{t-1}), and an in-degree statistic ($S(Y_{t-1}) = indegree(Y_{t-1})$). For these simulations we broke down the missingness into three different parameters: missingness in the 0's, missingness in the 1's, and the number of time points which contain missingness. When it is obvious in context, we will express these parameters as a vector (a, b, c). We chose to look at two estimators: the MAP, and 95% Bayesian posterior interval (PI) (the ML estimate of the parameter and 95% confidence interval are similar in this context).

We performed a simple trace through various levels of missingness. We considered the parameter estimates under missingness versus the model without any missing data, where missingness was simulated in every time point and contained every combination from 0.05 to 0.5 by 0.05 in both the 0's and 1's for $\hat{S}$ with 0-imputation, 1-imputation, and δ-imputation models. (See Figures 1 - 3 available in Supplement). In all three cases the lag parameter and associated Bayesian PI were estimated surprisingly well. As the missingness in the 1's increased, the graph density terms (the mixing effects) grew more biased. To an extent, the fact that this model performed well in the lag-term is unsurprising given that our scheme selects precisely those cases where both pre- and post- states were observed, which should provide a fairly good basis for estimating the lag. The surprise here was how well approximate inference using $\hat{S}$ performed under all of the imputation schemes. Our next exploration into missingness was to simulate something akin to missingness caused by machine failure (e.g.

a fixed amount of missingness) with a steadily growing number of time points. We began with fixed (0.10, 0.30, 5) amount of missingness in the data. We varied the available number of time points from 10 to 150 for all four imputation schemes (the regression imputation (R) was performed using predictive calibration on a set of simple regression models to predict indegree, see Figures 4 - 7 available in Supplement). At this level of missingness, the parameter estimates and Bayesian PIs settled to approximately the true values after about 15-20 time points for all the parameters. As before, the lag parameter was almost always well-estimated. We then repeated the procedure starting with (0.15, 0.30), but increasing the number of time points which contain missingness at 1 to 5 ratio (see Figures 8-11 available in Supplement). Again, the parameter estimates typically improve as the amount of data increased. The 1-imputation case performed very poorly for the parameter of $\hat{S}$ in this case. We show in the main text the results of the density-imputation heuristic (Figure 1) because it serves as the simplest baseline model and beginning point for any MCMC-based scheme, as explored in the Discussion Section. Finally, we performed a closer inspection on the effect of fixed-rate missingness on performance. We simulated missingness rates of (.15, .30, 25) on 50 time points and obtained the resulting parameter estimates over 100 simulations (see Figures 8-12 available in Supplement). The R-imputation method performed best in the sense that the parameter estimate was consistently close to the true value, and the PI was always wider than the "true" Bayes PI (reflecting the expected uncer-

Figure 1: Missingness maintained at 5 random time-points with 10 percent missing in the zeros and 30 percent in the 1s. Parameter estimates were generated for DNR under complete case with d-imputation. Red is the "true" parameter and PI (i.e., MAP estimate from the complete data) and the blue line is the estimated parameter and PI from the data with simulated missingness.



tainty expansion due to missing data). The 0-imputation and δ-imputation schemes were consistently biased downwards in both the Bayesian PI and the parameter estimate. This is consistent with the literature in mean im-

putation methods (see Little (1992)), noting that the 0-imputation method is similar to the mean in this context. The 1-imputation scheme was biased downwards for mixing effects, but biased upwards for the lag and biased downwards for the indegree effect. This suggests that over-estimating the graph density is likely to cause much more trouble for sparse-graph models than does biasing the density downward.

Interpretation and Suggestions: We summarize our findings as follows. The R-imputation method, while powerful, has some obvious drawbacks. If there is too much missingness in the data, there may not be enough observed data to fit a regression model with good predictive performance. The blog data is particularly sparse, so we see that the δ-imputation and 0-imputation methods are very similar in this case. We suspect in a less sparse case the δ-imputation method would outperform the 0-imputation case. Given a sufficiently dense network we would expect the 1-imputation case to perform much better. Finally, we point out that the "intercept" or graph density terms can be very biased depending on choice of imputation scheme for $\hat{S}$, and that the lag term is surprisingly robust.

Data: Windsurfers (DNR with Vertex Dynamics case): Freeman et al. (1988) collected a dynamic interpersonal communication network of windsurfers in the late 1980's that he subsequently analyzed only statically. The network was originally collected daily (aggregated over a morning and an afternoon observation period) for 31 days (August 28, 1986 to September 27, 1986). For full details see Section Data: Windsurfers in the online

supplement.

Simulation Study of the Effects of Missingness: To test performance, we considered the impact on parameter estimates of missingness in the vertices (e.g. 5 percent of the vertices being missing at time t). We fit a simple model to the fully observed beach data. The edge model contains a graph density term, a density scaling effect ($\log(N_t + 1)$) (Butts and Almquist, 2015), a lag term (Y_{t-1}), and a degree statistic ($S(Y_{t-1}) = degree(Y_{t-1})$). The vertex model contains an intercept term, a lag term (V_{t-1}), and a degree sum statistic ($S_V(Y_{t-1})_i = \sum_j degree(Y_{ijt-1})$). We express our missingness rates as a vector (a, c) where a is the percent missingness in the vertex and edge sets (both) and c is the number of time points which contain missingness. We evaluated the impact of missingness on MAP and 95% Bayesian PI estimates (the ML estimate of the parameter and associated CI show similar behavior).

We simulated the missing data at rates similar to the earlier section, with complete details in the supplement. This case study focuses on the (.50,12) case over 24 time points, and the resulting parameter estimates (see Figures 13-15 available in Supplement). We primarily found bias in the intercept and $\hat{S}_V$ parameters, and, noticed that the vertex model suffers the most (particularly under the 0-imputation method). We observed (.10, 24) on 24 time points and show the resulting parameter estimates (see Figures 13-15 available in Supplement). The results are consistent with similar examples from the literature on imputation (see Little (1992)), where we

see that, as the amount of missing data grows, we see downward bias in our δ-imputation scheme (and in the 0-imputation scheme). The 1-imputation scheme again appears to behave better in this context, but again less predictably. Overall, the vertex model was more sensitive than the edge model with the 1-imputation performing the worst.

Interpretation and Suggestions: The R-imputation method is much more complex in the case of vertex dynamics and not always feasible if there is missingness at every time point. Efficiency or power is a much bigger concern in the case with vertex dynamics: nodes may not be observed very often and, if missing, can bias the model substantially. However, if the missingness rate is relatively low the basic heuristics perform quite well.

Discussion

To provide a point of comparison, we consider a few alternative estimation schemes for the DNR case with missing data. For this purpose we focus on DNR in the Blog Citation Network with (.2, .4) missingness in all time points, a relatively high level of missingness, but not unreasonable in a practical setting due to limitations in data collection, coding, or software error. For our case, we consider a series with 100 time points.

To examine the quality of our heuristics versus alternative approaches, we employed a local approximation to a full MCMC algorithm that directly approximates the full observed data likelihood of (4) by integrating across the full set of possible missing edge values. To obtain an approximate

MCMC algorithm we employed both past and future states as covariates for a local ERGM with Left and Right averaging; here the issue arises that we have missingness in the predictors. We alleviated this issue by imputing the first time step in the usual way, performing model-based imputation on each step given the previous one, and then estimating our model from the full data generated by the imputation procedure. We then repeated the whole process $K = 10$ times to obtain multiple imputations. One can then follow the Little and Rubin (2002) strategy for computing the parameters and SE, or one could fit to the marginalized likelihood over the realizations. Here, we fit to the marginalized likelihood over the imputed data. We followed the same basic procedure as Koskinen et al. (2010), providing again a local approximation to the full MCMC algorithm for the model-based missing data procedure discussed earlier. (We employed the software for the cross-sectional missing data approach developed by Koskinen et al. (2010). Because this code was implemented using interpreted R functions it did not scale to our full case, and we thus chose to limit analysis to a much smaller set of time points. We note that all other procedures used here were implemented in the same fashion, but employed algorithms that by nature are several orders of magnitude faster in typical settings.) We found that the local ERGM with Left and Right averaging approach in this case study results in parameters that are typically biased downward or to zero by small amount (e.g., Table 1). Parameter variance on $\hat{S}$ is typically lower (e.g., Table 2), though sometimes to the point of overconfidence. The

Table 1: Comparison of bias for imputation strategies and local MCMC approximation.

Bias Comparison

Par	1-Imp	0-Imp	δ-Imp	LA MCMC	Par	1-Imp	0-Imp	δ-Imp	LA MCMC
RNC	-0.16	-0.59	-0.21	0.45	Y_{t-1}	-0.02	-0.14	0.01	-0.90
RNC	-0.04	-0.49	-0.10	0.40	IDeg(Y_{t-1})	0.02	-0.02	0.01	0.00
DNCtoRNC	-0.28	-0.71	-0.34	0.42	ODeg(Y_{t-1})	-0.01	-0.02	-0.00	-0.00
RNCtoDNC	-0.19	-0.58	-0.23	0.47	3Cycle(Y_{t-1})	0.06	0.01	0.01	0.05

Table 2: Comparison of 95% BCI for imputation strategies and local MCMC approximation.

Bayesian Credible Intervals

Par	0-Imput		1-Imput		δ-Imput		LA MCMC		True Values	
	2.5 %	97.5 %	2.5 %	97.5 %	2.5 %	97.5 %	2.5 %	97.5 %	2.5 %	97.5 %
RNC	-7.12	-6.60	-7.12	-6.60	-7.12	-6.60	-6.49	-6.17	-6.88	-6.50
RNC	-6.44	-5.75	-6.44	-5.75	-6.44	-5.75	-5.94	-5.52	-6.29	-5.79
DNCtoRNC	-8.07	-7.32	-8.07	-7.32	-8.07	-7.32	-7.26	-6.82	-7.67	-7.14
RNCtoDNC	-8.54	-7.62	-8.54	-7.62	-8.54	-7.62	-7.74	-7.21	-8.19	-7.56
Y_{t-1}	10.05	10.57	10.05	10.57	10.05	10.57	9.42	9.73	10.14	10.51
IDeg(Y_{t-1})	0.07	0.15	0.07	0.15	0.07	0.15	0.07	0.11	0.07	0.11
ODeg(Y_{t-1})	-0.05	0.03	-0.05	0.03	-0.05	0.03	-0.02	0.01	-0.02	0.02
3Cyc(Y_{t-1})	-0.09	0.30	-0.09	0.30	-0.09	0.30	0.04	0.14	-0.01	0.09

improvement due to MCMC use is greatest for complex terms (e.g. degree and cycle terms) that benefit from the more refined imputation strategy. We expect this might be due to MCMC approach overstating tie formation, improving the ability to estimate complex terms, but pushing the model to think the overall density is higher than actually observed.

We also tried a local version of the Bayesian data augmentation procedure implemented by Koskinen et al. (2010). (This procedure was developed for static rather than temporal imputation, so we have extended it in the same manner as the MCMC local averaging procedure. Due to computational cost, we only perform this procedure on 20 time points.) We found that this local approximation to Koskinen et al. (2010) tends to bias the

covariate effects by a large margin: biases in Homophily terms are (2.76, 0.65, 2.938, 2.95), and the bias in the lag term is large, (-5.7). However, the bias in more complex terms such as Indegree, Outdegree and Three Cycle (0.005, -0.0002, -0.038) again resulted in an improvement, suggesting that a local MCMC procedure for $\hat{S}$ imputation may be preferable for such terms. This procedure appears to decouple the lag term to a much larger degree than the local ERGM with Left and Right averaging approach. Finally, we found that the Bayesian PI was shifted downward, e.g., the Outdegree term interval for the Koskinen et al. (2010) approximation was (-0.030, -0.006) compared to the true interval of (-.018, 0.006). Overall, we find that in interpreted R code the simple $\hat{S}$-heuristics are an order of magnitude or more faster than MCMC-based procedures, and generally perform well. Nevertheless, we can in some cases obtain improved performance from MCMC-based imputation procedures for $\hat{S}$ statistics if the time series is sufficiently short or the network size is sufficiently small, particularly for models that depend on complex structural statistics. In such settings, it may be worth employing a local approximation to the full MCMC to obtain improved $\hat{S}$ estimates. However, we find that the performance of the complete-case method seems to be quite good even with fairly simple heuristics, and it appears that the cost/performance tradeoff of the heuristic methods will prove difficult to beat as data size increases. While we think that there is considerable merit in continuing to consider MCMC-based imputation strategies, our findings suggest that complete-case estimation with density

or regression imputation (as feasible) is a reasonable default strategy for researchers working with large data sets. Clearly, there is room for future exploration on improved model estimation based on multiple imputation (Little and Rubin (2002); Gile and Handcock (2010a); Wang et al. (2016)), whether by MCMC or by other simulation techniques. We expect that more sophisticated $\hat{S}$ approximation methods could further improve the performance of the complete case approach. We find it interesting however, that for DNR TERGMs with missing data, simple methods can often yield favorable results.

Supplementary Material. Extended details may be found online.

Acknowledgements. This work was supported in part by ONR award N00014-08-1-1015, ARO awards W911NF-14-1-0577 (YIP) and W911NF-14-1-0552, NSF award IIS-1526736, and NIH/NICHD award 1R01HD068395-01.

References

Adamic, L. A. and N. Glance (2005). The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, pp. 36–43. ACM.

Almquist, Z. W. and C. T. Butts (2013). Dynamic network logistic regression: A logistic choice analysis of inter and intra-group blog citation dynamics in the 2004 us presidential election. *Political Analysis 21*(4), 430–448.

Almquist, Z. W. and C. T. Butts (2014a). Bayesian analysis of dynamic network regression with joint edge/vertex dynamics. In I. Jeliazkov and X.-S. Yang (Eds.), *Bayesian inference in the social and natural sciences*. New York City, NY: John Wiley & Sons.

Almquist, Z. W. and C. T. Butts (2014b). Logistic network regression for scalable analysis of networks with joint edge/vertex dynamics. *Sociological Methodology 44*(1), 273–321.

Almquist, Z. W., E. S. Spiro, and C. T. Butts (2016). Shifting attention: Modeling follower relation-

ship dynamics among us emergency management-related organizations during a colorado wildfire. In A. Faas and E. Jones (Eds.), *Social Network Analysis of Disaster Response, Recovery, and Adaptation*. Elsevier.

Brewer, D. D. and C. M. Webster (2000). Forgetting of friends and its effects on measuring friendship networks. *Social Networks 21*(4), 361–373.

Butts, C. T. (2011). Bernoulli graph bounds for general random graphs. *Sociological Methodology 41*, 299–345.

Butts, C. T. and Z. W. Almquist (2015). A flexible parameterization for baseline mean degree in multiple-network ergms. *The Journal of mathematical sociology 39*(3), 163–167.

Butts, C. T. and B. R. Cross (2009). Change and external events in computer-mediated citation networks: English language weblogs and the 2004 u.s. electoral cycle. *The Journal of Social Structure 10*(3), 1–29.

Carley, K. (1999). On the evolution of social and organizational networks. *Research in the Sociology of Organizations 16*, 3–30.

Cranmer, S. J. and B. A. Desmarais (2011). Inferential network analysis with exponential random graph models. *Political Analysis 19*(1), 66–86.

Desmarais, B. and S. Cranmer (2012). Statistical mechanics of networks: Estimation and uncertainty. *Physica A: Statistical Mechanics and its Applications 391*(4), 1865 – 1876.

Entwisle, B., K. Faust, R. R. Rindfuss, and T. Kaneda (2007). Networks and contexts: Variation in the structure of social ties. *American Journal of Sociology 112*(5), 1495–1533.

Freeman, L. C., S. C. Freeman, and A. G. Michaelson (1988). On human social intelligence. *Journal of Social Biological Structure 11*, 415–425.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian Data Analysis* (2nd ed.). New York, NY: Chapman & Hall/CRC.

Gile, K. J. and M. S. Handcock (2010a). Modeling networks from sampled data. *Annals of Applied Statistics 4*(1), 5–25.

Gile, K. J. and M. S. Handcock (2010b). Respondent-driven sampling: An assessment of current methodology. *Sociological Methodology 40*, 285–327.

Gjoka, M., M. Kurant, C. T. Butts, and A. Markopoulou (2010). Walking in facebook: A case study of unbiased sampling of osns. In *Proceedings of IEEE INFOCOM 2010*.

Handcock, M. S. (2003). Statistical models for social networks: Inference and degeneracy. In R. Breiger, K. M. Carley, and P. Pattison (Eds.), *Dynamic Social Network Modeling and Analysis*, pp. 229–240. Washington, DC: National Academies Press.

Hanneke, S., W. Fu, and E. P. Xing (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics 4*, 585–605.

Hanneke, S. and E. P. Xing (2007). *Statistical Network Analysis: Models, Issues, and New Directions: ICML 2006 Workshop on Statistical Network Analysis, Pittsburgh, PA, USA, June 29, 2006, Revised Selected Papers*, Volume 4503 of *Lecture Notes in Computer Science*, Chapter Discrete Temporal Models of Social Networks, pp. 115–125. Springer-Verlag.

Hipp, J. R., C. Wang, C. T. Butts, R. Jose, and C. M. Lakon (2015). Research note: The consequences of different methods for handling missing network data in stochastic actor based models. *Social networks 41*, 56–71.

Huisman, M. and C. Steglich (2008). Treatment of non-response in longitudinal network studies. *Social networks 30*(4), 297–308.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Series in Statistics. New York, NY: Springer.

Koskinen, J. H., G. L. Robins, and P. E. Pattison (2010). Analysing exponential random graph (p-star) models with missing data using bayesian data augmentation. *Statistical Methodology 7*(3), 366–384.

Krivitsky, P. N. (2012). Modeling of dynamic networks based on egocentric data with durational information. Technical Report Series 12-01, The Pennsylvania State University, University Park, PA 16802.

Leskovec, J. (2008). *Dynamics of large networks*. ProQuest.

Leskovec, J. (2011). Stanford large network dataset collection. *http://snap.stanford.edu/data/index.html*.

Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association 87*(420), 1227–1237.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis With Missing Data* (2nd ed.). Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons, Inc.

McCullagh, P. and J. A. Nelder (1999). *Generalized Linear Models* (2nd ed.). Chapman & Hall/CRC.

Newcomb, T. M. (1961). *The Acquaintance Process*. New York, NY: Holt, Reinhard & Winston.

Powell, W. W., K. W. Koput, and L. Smith-Doerr (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly 41*(1), 116–145.

Rubin, D. B. (1976). Inference and mising data. *Biometrika 63*(3), 581–592.

Schweinberger, M. and M. S. Handcock (2015). Local dependence in random graph models: characterization, properties and statistical inference. *JRSS: B (Statistical Methodology) 77*(3), 647–676.

Snijders, T. (2005). Models for longitudinal network data. In P. Carrington, J. Scott, and S. Wasserman (Eds.), *Models and methods in social network analysis*. New York: Cambridge University Press.

van de Rijt, A. (2011). The micro-macro link for the theory of structural balance. *Journal of Mathe-*

matical Sociology 35(1), 94–113.

van Duijn, M., K. Gile, and M. S. Handcock (2009). Comparison of maximum pseudo likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks 31*(1), 52–62.

Wang, C., C. T. Butts, J. R. Hipp, R. Jose, and C. M. Lakon (2016). Multiple imputation for missing edge data: A predictive evaluation method with application to Add Health. *Social Networks 45*, 89–98.

Wang, H., H. Xie, L. Qiu, Y. R. Yang, Y. Zhang, and A. Greenberg (2006). Cope: traffic engineering in dynamic networks. *ACM SIGCOMM Computer Communication Review 36*(4), 99–110.

Zimmermann, M. G., V. M. Eguìluz, and M. San Miguel (2004). Coevolution of dynamical states and interactions in dynamic networks. *Phys. Rev. E 69*, 065–102.