University of California
Los Angeles

Exponential Family Random Network Models

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Ian Edward Fellows

2012

Abstract of the Dissertation

Exponential Family Random Network Models

by

Ian Edward Fellows

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2012

Professor Mark S. Handcock, Chair

Random graphs, where the presence of connections between nodes are considered random variables, have wide applicability in the social sciences. Exponential-family Random Graph Models (ERGM) have shown themselves to be a useful class of models for representing complex social phenomena. We generalize ERGM by also modeling nodal attributes as random variates, thus creating a random model of the full network, which we call Exponential-family Random Network Models (ERNM). We demonstrate how this framework allows a new formulation for logistic regression in network data. We develop likelihood-based inference for the model in the case of a fully observed network and an MCMC algorithm to implement it.

We then develop a theory of inference for ERNM when only part of the network is observed, as well as specific methodology for missing data, including non-ignorable mechanisms for network-based sampling designs and for latent class models. We also consider contact tracing sampling designs which are of considerable importance to infectious disease epidemiology and public health. This culminates in a treatment of respondent driven sampling (RDS), which is a widely used link tracing design.

The dissertation of Ian Edward Fellows is approved.

Robert E. Weiss

Qing Zhou

Frederick P. Schoenberg

Mark S. Handcock, Committee Chair

University of California, Los Angeles

2012

Table of Contents

List of Figures

List of Tables

Vita

2009-2012	President: Fellows Statistics Inc.
2006-2009	Statistician: University of California, San Diego.
2005-2006	Data Analyst: San Diego State University Research Foundation.
2006	M.S. (Biostatistics), SDSU, San Diego, California.
2003	B.A. (Mathematics and Philosophy): University of Redlands, Redlands, California.

Publications

Fellows I., (2012). Deducer: A data analysis GUI for R, Journal of Statistical Software, 49(8)

Fellows I., (2010). The Minimaxity of the Mid P-value under the Estimated Truth Framework, Communications in Statistics: Theory and Methods, 40(2):244-54

Fellows I., (2008). Pseudo-Optimal Solutions to Texas Holdem Poker with Improved Chance Node Abstraction. (Technical Report UCSD)

Vahia I., Depp C., Palmer B., Fellows I., Golshan S., Thompson W., Allison,M., Jeste,D. Correlates of spirituality in older women Aging & Mental Health Vol. 15, Iss. 1, 2010

Fang D., Young C., Golshan S., Fellows I., Moutier C., Zisook S., Depression in Premedical Undergraduates: A Cross-Sectional Survey Prim Care Companion, J Clin Psychiatry. 2010; 12(6)

Vahia, I. V., Palmer, B. W., Depp, C., Fellows, I., Golshan, S., Kraemer, H. C. and Jeste, D. V. (2010), Is late-onset schizophrenia a subtype of schizophrenia?. Acta Psychiatrica Scandinavica, 122: 414426

Kasckow J., Fellows I., Golshan S., Solorzano E., Meeks T., Zisook S., (2010). Treatment of Subsyndromal Depressive Symptoms in Middle-Age and Older Patients With Schizophrenia: Effect of Age on Response. American Journal of Geriatric Psychiatry, Am J Geriatr Psychiatry. 2010 September; 18(9): 853857

Kasckow J., Lanouette N., Patterson T., Fellows I., Golshan S., Solorzano E., Zisook S. (2010) Treatment of subsyndromal depressive symptoms in middle-aged and older adults with schizophrenia: effect on functioning. International Journal of Geriatric Psychiatry International Journal of Geriatric Psychiatry, Volume 25, Issue 2, pages 183190

Nyer M., Kasckow J., Fellows I., Lawrence E., Golshan S., Solorzano E., Zisook S., (2010). The relationship of marital status and clinical characteristics in middle aged and older patients with schizophrenia and depressive symptoms. Clinical Schizophrenia and Related Psychoses Annals of Clinical Psychiatry 2010;22(3):172-179

Zisook S, Kasckow JW, Lanouette NM, Golshan S, Fellows I, Vahia I, Mohamed S, Rao S. (2010) Augmentation with citalopram for suicidal ideation in middle-aged and older outpatients with schizophrenia and schizoaffective disorder who have subthreshold depressive symptoms: a randomized controlled trial. J Clin Psychiatry. 2010 Jul;71(7):915-22. Epub 2010 Mar 9.

Palmer B., Savla G., Fellows I., Twamley E., Jeste D., Lacro J., (2010). Do people with schizophrenia have differential impairment in episodic memory and/or working memory relative to other cognitive abilities? Schizophrenia Research, Volume 116, Issue 2, Pages 259-265

Jeste D., Palmer B., Golshan S., Eyler L., Dunn L., Meeks T., Glorioso D., Fellows I., Kraemer H., Appelbaum P., (2009). Multimedia Consent for Research in People with Schizophrenia and Normal Subjects: A Randomized Controlled Trial, Schizophrenia Bulletin 35(4) 719-729

Zisook S., Kasckow J., Golshan S., Fellows I., Solorzano E., Lehman D., Mohamed S., Jeste D., (2009). Citalopram Augmentation for Subsyndromal Depressive Symptoms in Middle-Aged and Older Patients with Schizophrenia and Schizoaffective Disorder: A Randomized Controlled Trial, Journal of Clinical Psychiatry, 70(4):562-71

Jeste D., Jin H., Golshan S., Mudaliar S., Glorioso D., Fellows I., Kraemer H., Arndt S. (2009). Discontinuation of Quetiapine From an NIMH-Funded Trial Due to Serious Adverse Events Am J Psychiatry 2009 166: 937-938

Folsom D., Depp C., Palmer B., Mausbach B., Golshan S., Fellows I., Cardenas V., Patterson T., Kraemer H., Jeste D., (2008) Physical and mental health-related quality of life among older people with schizophrenia, Schizophrenia Research, 108(1) 207-213

Dunn L., Kim D., Fellows I., Palmer B., (2008). Worth the Risk? Relationship of Incentives to Risk and Benefit Perceptions and Willingness to Participate in Schizophrenia Research, Schizophrenia Bulletin. Schizophr Bull, 35 (4): 730-737

Roseman A., Kasckow J., Fellows I., Osatuke K., Patterson T., Mohamed S., Zisook S., (2008). Insight, Quality of Life, and Functional Capacity in Middle-Aged and Older Adults with Schizophrenia, International Journal of Geriatric Psychiatry, 23: 760-765

Zisook S., Montross L., Kasckow J., Mohamed S., Palmer B., Patterson T., Golshan S., Fellows I., Lehman D., Solorzano E., (2008). Subsyndromal Depressive Symptoms in Middle aged and Older Persons with Schizophrenia, American Journal of Geriatric Psychiatry 15:1005-1014

Kasckow J., Patterson T., Fellows I., Golshan S., Solorzano E., Mohamed S., Zisook S., (2008). Functioning in Middle Aged and Older Patients with Schizophrenia and Depressive Symptoms: Relationship to Psychopathology, American Journal of Geriatric Psychiatry, 16: 660-663

CHAPTER 1

Introducing Exponential Family Random Graph Models

A graph is a collection of nodes, each of which may either be connected or not connected to each other node. For the purposes of this document, nodes may not be connected to themselves. In addition to the graph connections, each node may have characteristics which are of interest to the researcher. A network is defined as the union of a graph and the nodal characteristics. Random graphs, where connections between nodes are random but nodal characteristics are either fixed or missing, have a long history in the mathematical literature starting with the simple Erdős-Rényi model (Erdos and Renyi, 1959), and including the more general exponential-family random graph models (ERGM) for which inference requires modern Markov Chain Monte Carlo (MCMC) methods (Frank and Strauss, 1986, Hunter and Handcock, 2006). On the other hand we have Gibbs/Markov random field models where nodal attributes are random but interconnections between nodes are fixed. A simple example is the Ising model of ferromagnetism (Ising, 1925) from the statistical physics literature which is exactly solvable under certain network configurations (Baxter, 1982); however, most field models require more complex methodologies for inference (Zhu and Liu, 2002).

In the social network literature, these two classes of models are conceptually defined as "social selection" and "social influence" models. In social selection models, the probability of social ties between individuals are determined by nodal characteristics such as age or sex (see Robins et al. (2001a) and references therein).

In social influence models, individuals' nodal characteristics are determined by social ties (see Robins et al. (2001b) and references therein). Leenders (1997) argues that the processes of tie selection and nodal variate influence are co-occurring phenomena, with ties affecting nodal variates and visa versa, and should therefore be considered together. This chapter presents a joint exponential-family model of connections between nodes (dyads), and nodal attributes, thus representing a unification of social selection and influence. We will refer to this model as an exponential-family random network model (ERNM).

We note that we are not developing a model for the coevolution of the tie and nodal variables. We are modeling the joint relation between the processes of tie selection and nodal variate influence in a cross-sectional network. As such our model explicitly represents the endogenous nature of the relational ties and nodal variables. If network-behavior panel data is available then it may be possible to statistically separate the effects of selection from those of influence. For a discussion of these issues for dynamic and longitudinal data, see Steglich et al. (2010).

In this chapter we introduce the ERNM class and give simple examples, then develop aspects of the class that are important for statistical modeling. We then apply the modeling approach to the study of substance abuse in adolescent peer networks and compares it to standard approaches.

Chapter 2 extends the ERNM class to handle cases where only part of the network is observed, providing a basis for inference in link tracing designs, which we apply to a simulated dataset. Chapter 3 generalizes the computation methods used to find the maximum likelihood estimates outlined in chapters 1 and 2 to all exponential-family distributions, and provides computational heuristics for the algorithm. Chapter 4 applies the methods in chapter 2 to the respondent driven sampling survey design, and explores the properties of the resulting estimator with simulations. Chapter 5 proposes a new, more rigorous, link tracing design

based on respondent driven sampling. Finally, chapter 6 concludes with a broader discussion.

1.1 ERNM Specification

Let the graph Y be an n by n matrix whose entries $Y_{i,j}$ indicate whether subject i and j are connected, where n is the size of the population. Further let X be an $n \times q$ matrix of nodal variates. We define the network to be the random variable (Y, X). Let $\mathcal{N}$ be the set of possible networks of interest (the sample space of the model). For example, $\mathcal{N} \subseteq 2^{\mathbb{Y}} \times \mathcal{X}^n$, the power set of the dyads in the network times the power set of the sample space of the nodal variates. A joint exponential family model for the network may be written as

$$P(X = x, Y = y|\eta) = \frac{1}{c(\eta, \mathcal{N})} e^{\eta \cdot g(y,x)}, \qquad (y, x) \in \mathcal{N} \tag{1.1}$$

where η is a vector of parameters, g is a vector valued function, and $c(\eta, \mathcal{N})$ is a normalizing constant such that the integral of P over the sample space of X and Y is 1 (See equation (1.2)). The model parameter space is $\eta \in H \subseteq \mathbb{R}^q$. This functional form is the familiar exponential family form, and is extremely general depending on the choice of g (see Barndorff-Nielsen (1978) and Krivitsky (2011)). Formally, let $(N, \mathcal{N}, P_0)$ be a $\sigma-$finite measure space with reference measure P_0. A probability measure $P(X = x, Y = y|\eta)$ is an ERNM with respect to this space if it is dominated by P_0 and the Radon-Nikodym derivative of $P(X = x, Y = y|\eta)$ with respect to P_0 is expressible as

$$\frac{dP(X = x, Y = y|\eta)}{dP_0} = \frac{1}{c(\eta, \mathcal{N})} e^{\eta \cdot g(y,x)}, \qquad (y, x) \in \mathcal{N}$$

where

$$c(\eta, \mathcal{N}) = \int_{(y,x) \in \mathcal{N}} e^{\eta \cdot g(y,x)} dP_0(y, x) \tag{1.2}$$

and $H \subseteq \{\eta \in \mathbb{R}^q : c(\eta, \mathcal{N}) < \infty\}$. See Barndorff-Nielsen (1978) for further properties of the exponential-family class of probability distributions.

1.1.1 Relationship with ERGM and Random Fields

Let $\mathcal{N}(x) = \{y : (y, x) \in \mathcal{N}\}$ and $\mathcal{N}(y) = \{y : (y, x) \in \mathcal{N}\}$ then

$$\begin{aligned} P(Y = y|X = x; \eta) &= \frac{1}{c(\eta; \mathcal{N}(x), x)} e^{\eta \cdot g(y,x)} \quad y \in \mathcal{N}(x) \\ P(X = x|Y = y; \eta) &= \frac{1}{c(\eta; \mathcal{N}(y), y)} e^{\eta \cdot g(y,x)} \quad x \in \mathcal{N}(y) \end{aligned}$$

The first model is the ERGM for the network conditional on the nodal attributes. Analysis of models of this kind have been the staple of ERGM (Frank and Strauss, 1986, Hunter and Handcock, 2006, Goodreau et al., 2009). The second model is an exponential-family for the field of nodal attributes conditional on the network. This will be a Gibbs/Markov field when the process satisfies the pairwise Markov property (i.e., If $Y_{ij} = 0$ then X_i and X_j are conditionally independent given all other X) (Besag, 1974). However the model is more general than this as $g(y, x)$ can be arbitrary. We will refer to it as a Gibbs measure (Georgii, 1988).

The model (1.1) can be expressed as

$$P(X = x, Y = y|\eta) = P(Y = y|X = x, \eta)P(X = x|\eta) \tag{1.3}$$

where

$$P(X = x|\eta) = \frac{c(\eta; \mathcal{N}(x), x)}{c(\eta, \mathcal{N})} \quad x \in \mathcal{X}.$$

$P(X = x|\eta)$ is the marginal representation of the nodal attributes and is not necessarily an exponential-family with canonical parameter η. These decompositions demonstrate why the joint modeling of Y and X via ERNM (as proposed here) is different and novel compared to the conditional modeling of Y given X via ERGM.

1.1.2 Interesting Model-Classes of ERNM

1.1.2.1 Example: Separable ERGM and Field Models

Suppose that g is composed such that the model can be expressed as

$$P(X = x, Y = y|\eta_1, \eta_2) = \frac{1}{c(\eta_1, \eta_2, \mathcal{N})} e^{\eta_1 \cdot h(x) + \eta_2 \cdot g(y)} \qquad (y, x) \in \mathcal{N} \tag{1.4}$$

where $\mathcal{N}$ is the product space $\mathcal{Y} \times \mathcal{X}$ with $\mathcal{Y}$ is the space of Y and $\mathcal{X}$ is the space of X. x and y in this model are separable and therefore may be considered independently. The model (1.4) can be decomposed as the product of

$$\begin{aligned} P(X = x|\eta_1) &= \frac{1}{c_1(\eta_1, \mathcal{X})} e^{\eta_1 \cdot h(x)} \\ P(Y = y|\eta_2) &= \frac{1}{c_2(\eta_2, \mathcal{Y})} e^{\eta_2 \cdot g(y)}. \end{aligned}$$

This type of model is particularly simple because of the separation of the two components. The first term is a general exponential-family model for the attributes (e.g., generalized linear models McCullagh and Nelder (1989)). The second term is a separate ERGM for the relations that has no dependence on the nodal attributes. Such separable models are usually not applicable as the phenomena that we are interested in studying is precisely the relationship between X and Y, thus independence is typically an unrealistic assumption.

1.1.2.2 Example: Joint Ising Models

An important aspect of social networks is the increased likelihood of a connection existing between nodes charing a characteristic. This property is generally referred to as homophily. If X is univariate and binary $x_i \in \{-1, 1\}$, previous social selection models (Goodreau et al., 2009) have used the following statistic to model

homophily

$$\text{homophily}(y, x) = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i y_{i,j} x_j. \tag{1.5}$$

This statistic is a count of the number of ties between nodes homophilous in the nodal covariate. Such a statistic is useful as a basis for a joint model. A simple example would include a term for homophily and a term graph density, explicitly

$$P(X = x, Y = y|\eta_1, \eta_2) \propto e^{\eta_1 density(y) + \eta_2 homophily(y,x)} \qquad (y, x) \in \mathcal{N}.$$

where density(y) $= \frac{1}{n} \sum_i \sum_j y_{i,j}$ and $\mathcal{N} = \mathcal{Y} \times \mathcal{X} = \{0, 1\}^{(2^n)} \times \{-1, 1\}^n$. If we look at the conditional distribution of Y given X we get

$$P(Y_{i,j} = y_{i,j}|X = x, \eta_1, \eta_2) \propto e^{\eta_1 \frac{1}{n} y_{i,j} + \eta_2 x_i y_{i,j} x_j} \quad y \in \{0, 1\},\ x \in \mathcal{X}.$$

The dyadic variables $y_{i,j}$ are independent of each other, and thus is a dyad-independent model for Y. We can recognize the functional form of the conditional distribution of Y given X as identical to logistic regression, and thus the conditional likelihood could be maximized using familiar generalized linear model (GLM) algorithms (McCullagh and Nelder, 1989). Conditioning X on Y we arrive at

$$P(X = x|Y = y, \eta_2) \propto e^{\eta_2 \sum_i \sum_j x_i y_{i,j} x_j} \qquad (y, x) \in \mathcal{N},$$

which we can recognize as the familiar Ising model (Ising, 1925) for the field over X with its lattice defined by Y.

This joint Ising model has the advantage of being mathematically parsimonious. Unfortunately, the results in section 1.2.1 indicate that it displays unrealistic statistical characteristics, which may rule it out as a reasonable representation of typical social networks.

1.2 Development of ERNM

In this section we develop ERNM, including issues of model degeneracy, the specification of network statistics and likelihood-based inference. In particular, we specify a class of logistic regression models for ERNMs.

An important consideration when modeling with the ERNM class is the specification of the statistics $g(y, x)$. As each choice of $g(y, x)$ leads to a valid model for the network process, there is much flexibility in this for modeling. The particular choices are very application dependent. However, as for ERGM, a stable of statistics can be created to capture primary the features of networks (Morris et al., 2008).

It is important to note that the ERNM class is quite different from the ERGM class (despite the formal similarity in equation (1)). ERNM require the specification of stochastic models for the nodal attributes (which ERGM do not permit). Further statistics which are meaningless for ERGM, for example, any statistic of X alone, play a prominent role in ERNM.

1.2.1 Model Degeneracy

Exponential family models for networks have been known to suffer from model degeneracy (Strauss, 1986, Handcock, 2003, Schweinberger, 2011), and even simple Markov models have similarly been shown to have degenerate states (sometimes called phase transitions in the statistical physics literature (Dyson, 1969)). Degeneracy is loosely defined as a set of model parameters where a small change in the parameters yield a massive change in the types of networks produced, usually this change is between two stable states. For example, one set of parameter values may indicate a low density graph, but a small change in the values leads to high density graphs. Parameter values at the tipping point of this change often yield bimodal network statistics. Because ERNM models represent the unification of

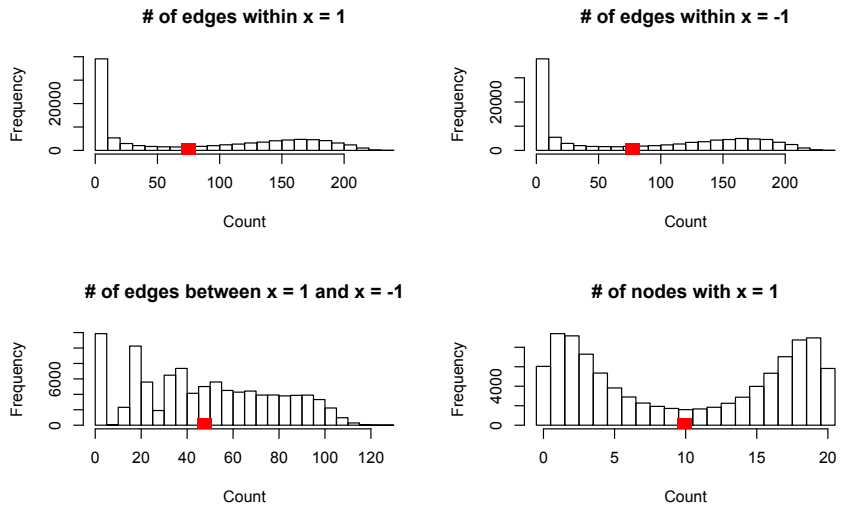


Figure 1.1: 100,000 draws from an Ising Joint Model with $\eta_1 = 0$ and $\eta_2 = 0.13$. Mean values are marked in red.

these two classes of models, a consideration of degeneracy must be undertaken. For example, while the joint Ising model of Section 1.1.2.2 is pleasing in its parsimonious simplicity, it unfortunately displays pathological degeneracy under mild homophily. Consider a joint Ising model of a 20 node network, with $\eta_1 = 0$ and $\eta_2 = 0.13$. In this model, 76% of edges are between nodes with matching x values, whereas 24% are between miss-matched nodes. Figure 1 shows the marginal statistics of 100,000 draws from this model.

Despite the fact that the homophily is not particularly severe, Figure 1 displays a great deal of degeneracy. The counts of edges are highly skewed. By symmetry we know that the expected number of nodes with $x = 1$ is 10, however, when inspecting the marginal histogram, we see that it is bimodal and puts very low probability on the value of 10. This severe degeneracy greatly reduces the usefulness of this model for practical networks.

We note that this phenomena will likely be as prevalent for ERNM models as for ERGM, and will have similar solutions. We recommend that model degeneracy be assessed for all proposed ERNM models.

1.2.2 Non-Degenerate Representation of Homophily within ERNM

Specification of the network's statistics via g is fundamental to ERNM. A natural source are analogues of those terms developed for ERGM (Morris et al., 2008). However, the degeneracy of the homophily specification in Section 1.1.2.2 suggests that careful thought is required in considering some network statistics. Suppose x is categorical with category labels $1, \ldots, K$. To define homophily we start by defining fundamental statistics of the network. Let $d_i(y)$ be the degree of node $i = 1, \ldots, n$ and $n_k(x) = \sum_i I(x_i = k)$ be the category counts, that is, the number of nodes in category $k = 1, \ldots, K$. Here I is the indicator function. Let $d_{i,k}(y, x) = \sum_{i<j} y_{ij} I(x_j = k)$ be the number of edges connecting node i to nodes in category k. We can generalize Equation (1.5) as

$$\text{homophily}_{k,l}(y, x) = \sum_{i=1}^{n} \sum_{j=1}^{n} I(x_i = k) y_{i,j} I(x_j = l).$$

As with Equation (1.5), this term has the nice property that it is dyad independent, meaning that conditional upon X, the marginal distribution of each dyad is independent of all others. Unfortunately, it displays the same degeneracy we saw in Section 1.1.2.2. We propose an alternate regularized homophily statistic which can be expressed as

$$\text{rhomophily}_{k,l}(y, x) = \sum_{i:x_i=k} [\sqrt{d_{i,l}(y, x)} - E_{\perp\!\!\!\perp}(\sqrt{d_{i,l}(Y, X)} | Y = y, n(X) = n(x))],$$

where $E_{\perp\!\!\!\perp}(g(Y, X)|Y = y, n(X) = n(x))$ is the expectation of the statistic $g(Y, X)$ conditional upon the graph Y and number of nodes in each category of x ($n(x) = \{n_k(x)\}_{k=1}^{K}$), under the assumption that X and Y are independent. Specifically, this distribution is

$$P(X = x|Y = y, n(X) = n(x)) \propto 1 \qquad\qquad (y, x) \in \mathcal{N},$$

There are many possible definitions of homophily, and this is one of many ways to formulate the relationship and in some applications, there may be a superior form. The justification for this particular formula is primarily empirical in that it captures the relationship between nodal variates and dyads well, and does not display the degeneracy issues that plague other forms of homophily. There are, however, some features of the statistic which provide justification for its form. The statistic $d_{i,l}(y,x)$ is transformed by a square root to roughly stabilize the variance based on the Poisson count model. This is important as nodes with high degree should not have qualitatively larger influence than nodes with low degree. Subtracting off the expectation based on the uniform independence model is essential in avoiding degeneracy because degenerate networks where all, or almost all, nodes belong to the same category should have homophily near zero.

1.2.3 Logistic Regression for Network Data

Let us consider a specific form of Equation (1.1) where X is partitioned into a binary nodal variate of particular interest $Z \in \{0,1\}$ (i.e. an outcome variable), and a matrix of regressors X.

$$P(Z=z, X=x, Y=y|\eta,\beta,\lambda) = \frac{1}{c(\beta,\eta,\lambda)} e^{z\cdot x\beta+\eta\cdot g(y,x)+\lambda\cdot h(y,z)}. \tag{1.6}$$

While most relationships in this model are left in a general formulation, it implies that the relationship of X to Z is described by $z\cdot x\beta$.

We can then write the distribution of z_i conditional upon all other variables as

$$P(z_i=1|z_{-i},x_i,Y=y,\beta,\lambda) = \frac{e^{x_i\beta}}{e^{\lambda\cdot[h(y,z^-)-h(y,z^+)]}+e^{x_i\beta}}. \tag{1.7}$$

where z_{-i} represents the set of z not including z_i, z^+ represents the variant of z

where $z_i = 1$, z^- is the variant of z where $z_i = 0$, and x_i represents the ith row of X. Suppose all variables remain fixed at their value except for x_i, which changes to x_i^*, then using equation (1.7), we can write the log odds ratio as

$$\text{logodds}(z_i = 1|z_{-i}, x_i, Y = y, \beta, \lambda) - \text{logodds}(z_i = 1|z_{-i}, x_i^*, Y = y, \beta, \lambda) = \beta(x_i - x_i^*).$$

Thus, the coefficients β may be interpreted as a conditional logistic regression model (i.e. conditional upon the rest of the network, a unit change in x_i leads to a β change in the log odds). Though the interpretation of the coefficients is familiar, the usual algorithms for estimating a logistic regression can not be used because the distribution of z_i depends on z_{-i} and thus the independence assumption does not hold.

1.2.4 Likelihood-based Inference for ERNM

The likelihood in equation (1.1) can be maximized using the methods of Geyer and Thompson (1992) and Hunter and Handcock (2006). Let y_{obs} and x_{obs} be the observed network, and ℓ be the log likelihood function. The log likelihood ratio for parameter η relative to η_0 can be written as

$$\ell(\eta) - \ell(\eta_0) = (\eta - \eta_0) \cdot g(y_{obs}, y_{obs}) - \log[E_{\eta_0}(e^{(\eta - \eta_0) \cdot g(y,x)})].$$

Given a sample of m networks (y_i, x_i) generated from $P(X = x, Y = y|\eta_0)$ the log likelihood can be approximated by

$$\ell(\eta) - \ell(\eta_0) \approx (\eta - \eta_0) \cdot g(y_{obs}, x_{obs}) - \log(\frac{1}{m} \sum_{i=1}^{m} e^{(\eta - \eta_0) \cdot g(y_i, x_i)})). \tag{1.8}$$

Appendix B provides the details of the Metropolis-Hastings algorithm used to sample from $P(X = x, Y = y|\eta_0)$ when the normalizing constant c is intractable

(which is usually the case). The approximation in equation (1.8) degrades as η diverges from η_0, motivating the following algorithm for estimating the maximum likelihood parameter estimates

1. Choose initial parameter values η_0.
2. Use Markov Chain Monte Carlo to generate m samples (y_i, x_i) from $P(X = x, Y = y|\eta_0)$.
3. With the sample from step 2, find η_1 maximizing a Häjek estimator (Thompson, 2002) of Equation (1.8) subject to $abs(\eta_1 - \eta_0) < \epsilon$.
4. If convergence is not met, let $\eta_0 = \eta_1$ and go to step 2.

This approximation to the log-likelihood can then be used to derive the Fisher information matrix and other quantities used for inference. The usual asymptotic approximations based on $n \to \infty$ may not apply to this situation as n is often endogenous to the social process.

1.3 Application to Substance Use in Adolescent Peer Networks

In addition to collecting data on health related behaviors, the National Longitudinal Study of Adolescent Health (Add Health) also collected information on the social networks of subjects (Harris et al., 2003a).

The network data we study in this chapter was collected during the first wave of the study. The Add Health data came from a stratified sample of schools in the US containing students in grades 7 through 12; the first wave was conducted in 1994-1995. For the friendship networks data, Add Health staff constructed a roster of all students in the school from school administrators. Students were then provided with the roster and asked to select up to five close male friends and five

close female friends. Complete details of this and subsequent waves of the study can be found in Resnick et al. (1997) and Udry and Bearman (1998).

Previous studies have investigated the social network structure of Add Health schools (Bearman et al., 2004), including Hunter et al. (2008), Goodreau et al. (2009), Handcock and Gile (2007) who used ERGM models to investigate network structure.

Here we analyze one of these schools; the high school had 98 students, of which 74 completed surveys. Students who did not complete the survey were excluded from analysis. The data contains many measurements on each of the individuals in these networks with some measurements, like sex, not influenced by network structure in any way, termed *exogenous*. Other covariates may exhibit strong non-exogeneity (e.g., substance use may be influenced through friendships).

1.3.1 A Super-population Model for an Add Health High School

Using the MCMC-MLE algorithm in Section 1.2.4, we fit an ERNM model to the high school data. The model has six terms modeling the degree structure of the network, three modeling the counts of students in each grade, and two representing the homophily within and between grades. Table 1.1 defines each of the terms, and explicit formulas are listed in Appendix A. Note that many terms could be added to this model to make it a more complex representation of the social structure, including terms similar to those in Handcock and Gile (2007), however, here we prefer a simple parsimonious model of the network, with particular focus on the relationship between X and Y.

Table 1.2 shows the fitted model along with standard errors and $p-$values based upon the Fisher information matrix. We can see that students in the same grade are much more likely to be friends, as the Within Grade Homophily term is positive, and is nominally highly significant. The positive coefficient for '+1

Table 1.1: ERNM Model Terms: The terms in the first block are graph statistics (ERGM-type), those in the second block model nodal attributes, and the last are joint. Terms in the the last two blocks can not be represented in an ERGM.

Form	Name	Definition
Y	Mean Degree	Average degree of students
Y	Log Variance of Degree	The log of the variance of the student degrees
Y	In Degree = 0	# of students with in degree 0
Y	In Degree = 1	# of students with in degree 1
Y	Out Degree = 0	# of students with out degree 0
Y	Out Degree = 1	# of students with out degree 1
Y	Reciprocity	# of reciprocated ties
X	Grade = 9	# of freshmen
X	Grade = 10	# of sophomores
X	Grade = 11	# of juniors
X, Y	Within Grade Homophily	Pooled homophily within grade
X, Y	+1 Grade Homophily and the grade above it	Pooled homophily between each grade

Grade Homophily' indicates that students also tend to form connections to the grades just below or just above them.

We can evaluate the fit of the model in two ways. The first is to simulate networks from the fitted model, and visually compare them to the observed network (Hunter et al., 2008). Figure 1.2 shows one such simulation. The observed network and simulated network look similar, giving some support that the fitted model is reasonable. Next we can simulate network statistics from the model and compare them to the observed network. The box plots in Figure 1.3 represent network statistics from 1000 draws from the fitted model, and the red dots are the statistics of the observed network. The degree structure matches well. Looking at the number of edges between grades, we see that the two homophily terms capture the 16 mixing statistics quite well. If desired, we could have added additional terms for each of the 16 mixing categories, but our interest was in a reasonable parsimonious representation of the network. The number of students within each grade are perfectly centered around the observed statistics. This is

Table 1.2: ERNM Model with Standard Errors Based on the Fisher Information

Term	$\hat{\eta}$	Std. Error	Z	$p-$value
Mean Degree	-217.02	7.81	-27.80	<0.001
Log Variance of degree	25.07	9.06	2.77	0.006
In-Degree 0	2.62	0.50	5.20	<0.001
In-Degree 1	1.05	0.40	2.62	0.009
Out-Degree 0	4.09	0.52	7.91	<0.001
Out-Degree 1	1.93	0.45	4.25	<0.001
Reciprocity	2.71	0.23	11.77	<0.001
Grade = 9	1.46	0.62	2.37	0.018
Grade = 10	1.93	0.71	2.72	0.007
Grade = 11	2.08	0.59	3.54	<0.001
Grade Homophily	4.34	0.46	9.41	<0.001
+1 Grade Homophily	0.63	0.21	2.98	0.003

expected, as the number of students in each grade are explicitly included in the model, and thus the mean counts from the model match the observed counts in the high school.

1.3.2 Logistic Regression on Substance Use

One aspect of the Add Health data that is of particular interest is the degree to which students use, or have used, tobacco and alcohol. In this section we will investigate the relationship between substance use and sex. We define substance use as either current use of tobacco or having used alcohol at least 3 times. Overall 19 students reported having used substances. A naive logistic regression model with X as an indicator that the sex of the adolescent is male shows a significant effect of sex (Table 1.3). That this model implies separability between the distribution of the network and the distribution of the outcome as in Section 1.1.2.1. This is an unreasonable assumption if friends tend to influence each other's substance abuse patterns, which we expect to be the case.

We extend the model in Section 1.3.1 with terms for substance and gender homophily, as well as terms for the logistic regression of sex on substance use.

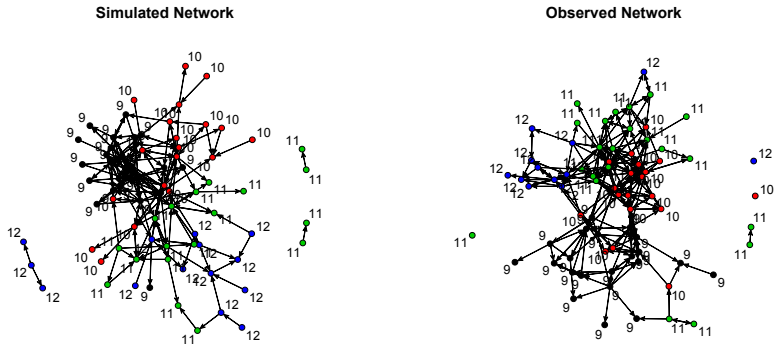


Figure 1.2: Model-Based Simulated High School

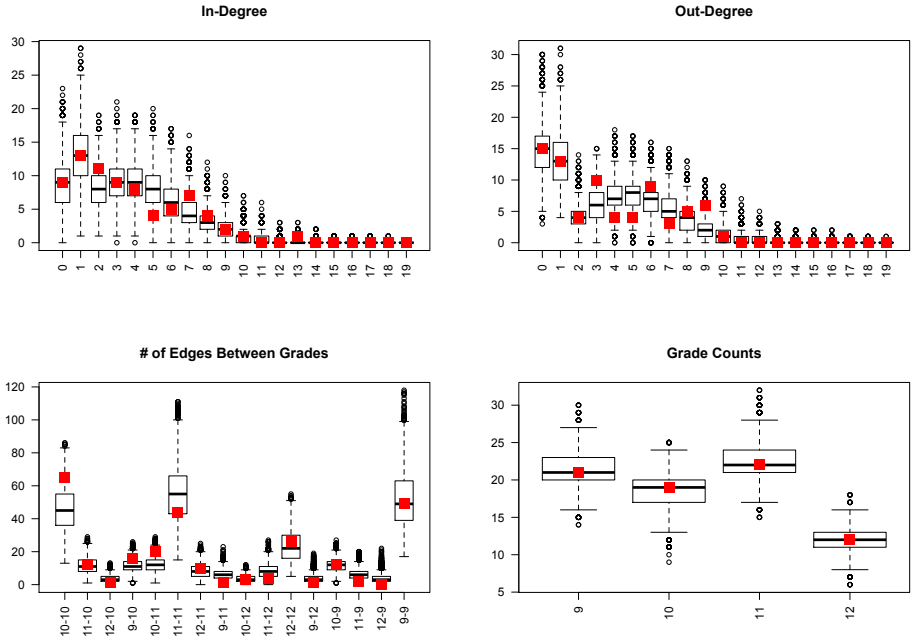


Figure 1.3: Model Diagnostics

Table 1.3: Simple Logistic Regression Model Ignoring Network Structure. This is the standard approach to regression in network data that ignores social influence and selection.

	β	Std. Error	Z	$p-$value
Intercept	-1.70	0.44	-3.84	<0.001
Gender	1.18	0.57	2.09	0.037

Table 1.4: Network Logistic Regression Parameter Estimates: These are based on the ERNM which models social influence and selection. The effect of gender on substance abuse is different than that in simple model (Table 3).

	η	Bootstrap Std. Error	Asymptotic Std. Error	Z	$p-$value
Mean Degree	-215.50	8.32	8.15	-26.44	<0.001
Log Variance of degree	24.46	8.80	8.91	2.75	0.006
In-Degree 0	2.68	0.55	0.48	5.55	<0.001
In-Degree 1	1.07	0.43	0.41	2.60	0.009
Out-Degree 0	4.15	0.54	0.52	8.03	<0.001
Out-Degree 1	1.94	0.50	0.45	4.31	<0.001
Reciprocity	2.71	0.25	0.23	11.96	<0.001
Grade Homophily	4.28	0.44	0.47	9.18	<0.001
+1 Grade Homophily	0.62	0.21	0.21	2.99	0.003
Gender Homophily	0.78	0.24	0.24	3.27	0.001
Substance Homophily	0.76	0.25	0.25	3.02	0.003
Intercept	-1.72	0.50	0.44	-3.91	<0.001
Gender	0.92	0.55	0.51	1.79	0.073

Whereas, Grade was considered random in the model in Section 1.3.1, because substance use is of primary interest in this model, all covariates are fixed except for Substance use. Table 1.4 displays the parameter estimates as well as p-values based on the Fisher information. Under regularity and asymptotic conditions, the MLE parameter estimates are distributed normally with covariance equal to the inverse of the negative Fisher information $(-cov_{\hat{\eta}}(g(T)))^{-1}$. Because inferences using Fisher information are typically justified using asymptotic arguments which don't apply here, we also ran a parametric bootstrap procedure with 1000 bootstraps, and bootstrap standard errors are included in Table 1.4. To perform parametric bootstrap we simulated independent networks from the MLE model, then found the MLE parameter estimates for the simulated networks. The standard deviations of the parameters among these estimate are reported in table 1.4. There is very close agreement between the bootstrap standard errors and the asymptotic ones, indicating that the Fisher information is a reliable measure for this model.

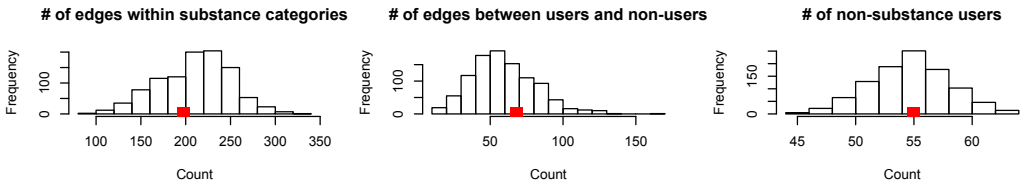


Figure 1.4: Substance Use Homophily Diagnostics. The values of the observed statistics are marked in red.

We see that the first 9 terms in the model are similar to their counterparts in Table 1.2. Two additional homophily terms are added, one for gender, and one for substance use. Both of these are highly significant, lending support to the position that it is unwise to simply perform a logistic regression ignoring network structure. The last two terms in Table 1.4 represent the network aware logistic regression of gender of substance use, and are analogous to the terms in Table 1.3. The parameter for sex is 22% smaller than in Table 1.3 leading to a non-significant $p-$value.

Similarly to the model in Section 1.1.2.2, in the fitted model, 73% of edges occur between students with the same substance abuse classification, whereas 27% are between users and non-users. Figure 1.4 shows model diagnostics for the homophily on substance abuse. Each marginal histogram puts high probability on the observed statistics (marked in red) and are not highly skewed, indicating that our model both captures the homophily relation, and is a reasonable model of that relation.

1.4 Discussion

We have developed a new class of joint relational and attribute models for the analysis of network data. These models represent a generalization of both ERGM and Gibbs random field models with each expressible as a special case of the new class. The new model provides a principled way to draw inferences about not only

the graph structure, but also the nodal characteristics of the network.

A ramification of the joint class is a natural way to specify conditional logistic regression on nodal variables. Previous models for network regression have struggled with the specification due to the ambiguity induced by endogenous nodal variable. The ERNM framework clarifies the model formulation and the interpretation of the parameters.

Further work on specifying model statistics is necessary to unlock the power of the ERNM class. The regularized homophily statistic of Section 3.2 is a good illustration of the issues involves. It is a good way to represent homophily on nodal characteristics. However, alternatives need to be developed for other features such as transitivity.

As could be expected based on presence of degeneracy in many ERGM models, we found that there exist degenerate states in even simple ERNM models. In particular, we found that the usual statistic used to represent homophily (the major relation of interest in a joint model) displayed significant degeneracy issues, and proposed an alternative that does not.

An R package implementing the methods developed in this chapter will be made available on CRAN (R Development Core Team, 2012).

1.5 Appendix A: Specifics of ERNM Terms

Here we explicitly define the network terms in (1.1). Let n be then number of nodes in the network, $d^x_{i,j} = \sum_k y_{i,k} I(x_k = j) + \sum_k y_{k,i} I(x_k = j)$ be the degree of node i to category j of x, and $d^+_i = \sum_k y_{k,i}$, $d^-_i \sum_k y_{i,k}$, $d_i = d^+_i + d^-_i$ be the in,

out and overall degree respectively. Then the model terms can be expressed as:

$$
\begin{aligned}
\text{mean degree} &= \frac{\sum_i^n d_i}{n} \\
\text{log variance of degree} &= \log(\frac{\sum_i^n (\text{mean degree} - d_i)^2}{n}) \\
\text{in-degree k} &= \sum_i^n I(d_i^- = k) \\
\text{out-degree k} &= \sum_i^n I(d_i^+ = k) \\
\text{reciprocity} &= \sum_i^n \sum_j^n y_{i,j} y_{j,i} \\
\text{within grade homophily} &= \sum_{k \in \{9,10,11,12\}} \sum_{i:grade=k} \sqrt{d_{i,k}} - E_{\perp\!\!\!\perp}(\sqrt{d_{i,k}}) \\
\text{+1 grade homophily} &= \sum_{k \in \{9,10,11\}} \sum_{i:grade=k} \sqrt{d_{i,k+1}} - E_{\perp\!\!\!\perp}(\sqrt{d_{i,k+1}}) + \\
&\quad \sum_{k \in \{10,11,12\}} \sum_{i:grade=k} \sqrt{d_{i,k-1}} - E_{\perp\!\!\!\perp}(\sqrt{d_{i,k-1}})
\end{aligned}
$$

For large networks some computational efficiency can be obtained by approximating the the expectations $E_{\perp\!\!\!\perp}(\sqrt{d_{i,k}})$ by that of the square root of a binomial variable, with probability equal to the proportion of nodes in category l, and size equal to the out-degree of node i. Each term of the sum is then the square root of the number of connections to category l, from node i, minus what would be expected by chance. Note that the expectation would more accurately be a hypergeometric distribution, due to the fact that only one edge can connect two nodes, however, the binomial approximation is much faster to compute and is asymptotically correct for sparse graphs. This approach was used in the application of Section 1.3.

1.6 Appendix B: An MCMC algorithm for ERNM

We use a Metropolis-Hastings algorithm to sample networks from an ERNM (Gilks et al., 1996). The algorithm alternates between proposing a change to a dyad with probability p_{dyad} and proposing a change to a nodal variable. Because the graphs for social networks are usually sparse, when proposing a dyad change the algorithm selects an edge to remove with probability p_{edge} and a random dyad to toggle with probability $1 - p_{edge}$. We found that this leads to better mixing than simply toggling a random dyad (Morris et al., 2008). When proposing a change to the nodal attributes, an attribute is picked at random. If it is categorical, a random new category is chosen. If it is continuous, it is perturbed by adding a small constant ϵ.

The following algorithm can be used to generate a random draw from an ERNM probability distribution (1.1) with an intractable normalizing constant:

Require: Arbitrary $(y^0, x^0) \in nets(Y, X)$, $p_{dyad} \in [0, 1]$, $p_{edge} \in [0, 1]$ and S sufficiently large

for $s \leftarrow 1$ to S **do**
  $y^* \leftarrow y^{(s-1)}$
  $x^* \leftarrow x^{(s-1)}$
  $u_{dyad} \leftarrow \text{Uniform}(0, 1)$
  if $u_{dyad} < p_{dyad}$ **then**
    $u_{edge} \leftarrow \text{Uniform}(0, 1)$
    if $u_{edge} < p_{edge}$ **then**
      $(i, j) \leftarrow$ RandomEdge(y^*)
      $y^*_{i,j} \leftarrow 0$
      $q \leftarrow \frac{\text{NumberOfEdges}_{(y*)}}{\text{NumberOfEdges}_{(y*)+}\ \text{NumberOfDyads}_{(y*)}}$
    else
      $(i, j) \leftarrow$ RandomDyad(y^*)
      if $y^*_{i,j} = 0$ **then**
        $y^*_{i,j} \leftarrow 1$

15: $q \leftarrow \frac{\text{NumberOfEdges}(y*)}{\text{NumberOfEdges}(y*)+ \text{NumberOfDyads}(y*)}$

16: **else**

17: $y^*_{i,j} \leftarrow 0$

18: $q \leftarrow 1 + \frac{\text{NumberOfDyads}(y*)}{\text{NumberOfEdges}(y*)+1}$

19: **else**

20: $(k, l) \leftarrow$ RandomAttribute(x^*)

21: **if** IsContinuous$(x^*_{*,l})$ **then**

22: $\epsilon \leftarrow$ Normal$(0, \sigma)$

23: $x^*_{k,l} \leftarrow x^*_{k,l} + \epsilon$

24: $q \leftarrow 1$

25: **else**

26: $x^*_{k,l} \leftarrow$ RandomCategory$(x^*_{*,l})$

27: $q \leftarrow 1$

28: $r \leftarrow qe^{\eta(g(x^*,y^*)-g(x^{(s-1)},y^{(s-1)}))}$

29: $u \leftarrow \text{Uniform}(0, 1)$

30: **if** $u < r$ **then**

31: $(y^s, x^s) \leftarrow (y^*, x^*)$

32: **else**

33: $(y^s, x^s) \leftarrow (y^{s-1}, x^{s-1})$

34: **return** (y^S, x^S)

An adjustment to the calculation of q must be made when toggling the graph when less than two edges are present in the network. If we are removing the last edge, then $q \leftarrow 1/(\text{ NumberOfDyads}(y*) + .5)$, and if we are adding an edge to an empty graph, then $q \leftarrow 0.5(\text{ NumberOfDyads}(y*) + 1)$.

For this algorithm to be fast, we must calculate the likelihood ratio $e^{\eta \cdot (g(x^*,y^*)-g(x^{(s-1)},y^{(s-1)}))}$ quickly, preferably in constant time relative to the size of the network. We do this by calculating the change in our statistics for a hypothetical change in the network (Morris et al., 2008). we can usually calculate the differences in the g statistics given small changes to the graph y or nodal attributes x in constant time. Morris et al. (2008) review change statistics for

commonly used ERGM terms and these can be reused here for changes in the graph (i.e. $g(x^{(s-1)}, y^*) - g(x^{(s-1)}, y^{(s-1)})$). ERNM require additional terms, such as those specified in Section 1.2.2, and also require that all change statistics be generalized to allow for changes in nodal attributes (i.e. $g(x^*, y^{(s-1)}) - g(x^{(s-1)}, y^{(s-1)})$).

CHAPTER 2

Analysis of Partially Observed Networks via Exponential-family Random Network Models

It is not uncommon for researchers to collect data on a subset of a single network rather than observing the full network. This partially observed case has been studied within the framework of exponential-family random graph models (ERGM) by Handcock and Gile (2010), however their formulation suffers from the limitation that any nodal attributes included in the model must be fully observed, and only dyads may be missing. This assumption is not met in most sampling designs, where only some of the nodes are surveyed by the researcher, and reduces the practical usage of ERGMs in the missing data setting.

By including nodal attributes as variates rather than fixed quantities, exponential-family random network models (ERNM) were shown in Chapter 1 to provide a convenient basis for inference in cases where the data is partially unobserved, either due to design, or out-of-design (e.g., non-response) mechanisms. While our framework is applicable to all partial observation mechanisms we consider three common mechanisms for partial observations in more detail, specifically:

Missing Data: If the population is comprised of a large number of units, or the number of edges is large, it is relatively common to find that the resources to observe a full network are not available. Often units or dyads are unavailable for sampling or do not provide complete responses to a survey instrument. In this case, only some of the dyads and nodal characteristics

are collected. We treat missing data as a form of sampling in which the sampling mechanism is unknown and outside the control of the researcher, or an out-of-design missing data mechanism. A good example of this is the National Longitudinal Study of Adolescent Health (Add Health), a school-based, longitudinal study of the health-related behaviors of adolescents and their outcomes in young adulthood. The study design sampled 80 high schools and 52 middle schools from the U.S., representative with respect to region of country, urbanicity, school size, school type, and ethnicity (Harris et al., 2003b). In 1994-95 an in-school questionnaire was administered to a nationally representative sample of students in grades 7 through 12. In addition to demographic and contextual information, each respondent was asked to nominate up to five boys and five girls within the school whom they regarded as their best friends. Thus each student could nominate up to ten students within the school (Udry, 2003). The nominations and contextual information were not available for some of the adolescents, either due to absence from school while the survey was being conducted, or refusal to participate. Thus, both the graph and nodal variates contained missing values.

Network sampling designs: Many studies in hard to reach populations use study designs that trace the linkages of an underlying social network. In these designs, the network is partially observed, however it is not of primary interest to the researcher. Such sampling designs have been exploited to estimate population disease rates (Gile and Handcock, 2010, Gile, 2011a, Gile and Handcock, 2011).

Latent variables: Some quantities of the network may be in principle unobservable. The probability model for a network may posit the existence of unknown variables which do not correspond to any observable quantity. For example, stochastic block models (Nowicki and Snijders, 2001) posit the

existence of classes of nodes, conditional upon which the dyads are independent. These classes are unobserveable nodal characteristics and must be inferred from the relational data. Similarly, latent position cluster models (Handcock et al., 2006) posit the existence of unobservable continuous nodal quantities that provide a spatial geometry for the network structure.

In this chapter we develop approaches for each of these scenarios in the context of ERNMs. Sections 2.1 through 2.3 introduce ERNM and extend the theory to incorporate partially observed populations. Section 2.4 develops methodology for each of the scenarios. Sub-section 2.4.1 looks at the effect of random non-response, and sub-section 2.4.2 applies a latent class model to extract unknown clusters from a real dataset. Sub-section 2.4.3 develops estimates based on contact tracing designs, which is of vital importance to the public health community. To our knowledge, the methods outlined in this chapter represent the first statistically justifiable approach to inference in contract tracing data.

2.1 Exponential Random Network Models

As we saw in Chapter 1, exponential random network models are a generalization of the exponential-family random graph model (Frank and Strauss, 1986, Hunter and Handcock, 2006), where both dyads and nodal characteristics are treated as random variates. Formally, in a population of n units, let $Y_{i,j}$ indicate that unit i has a tie to unit j. Let Y be an $n \times n$ matrix $[Y_{i,j}]$ and X be a an $n \times K$ matrix $[X_{ik}]$ of unit covariates. We define a network T as the union of the nodal covariates and the graph structure (i.e. $T = \{X, Y\}$). An exponential family model of T is expressed as

$$P(T = t|\eta) \quad = \quad \frac{1}{c(\eta, \mathcal{T})} e^{\eta \cdot g(t)} \qquad t \in \mathcal{T}, \tag{2.1}$$

where $\eta \in R^q$ is a vector of parameters, g is a $q-$vector valued function defining a set of sufficient statistics, $\mathcal{T}$ is the sample space of networks and $c(\eta, \mathcal{T}) = \sum_{t \in \mathcal{T}} e^{\eta \cdot g(t)}$ is the normalizing constant.

2.1.1 The Simple Homophily Model

Though any set of network statistics can be represented by g in equation (2.1), the examples in this paper will focus on a relatively simple network model. Suppose that $X = (X_1, \ldots, X_n)$ is a univariate categorical variable with m levels, labeled $0, \ldots, m-1$. If $X_i = l$ we say that unit i is in group l. A simple yet interesting joint model for X and Y is

$$P(T = (y, x)|\eta) = \frac{1}{c(\eta, \mathcal{T})} e^{\eta_0 \sum_{i,j} y_{i,j} + \eta_2 h(y,x) + \sum_{l=0}^{m-2} \eta_{j+3} \sum_{i=1}^{n} I(x_i = l)} \qquad (y, x) \in \mathcal{T}.$$

The first term of this model is the number of edges, and controls the density of the graph. The last term represents the number of nodes in each category of x ,except for the last level, which is dropped to maintain identifiability of the model. The second term h is the regularized sample homophily of x, as introduced in Chapter 1, and is defined as

$$h(y, x) = \sum_{k=0}^{m-1} \sum_{i:x_i=k} \sqrt{d_{i,k}(y, x)} - E_{\perp\!\!\!\perp}(\sqrt{d_{i,k}(y, x)}),$$

where $d_{i,k}(y, x)$ is the number of edges between node i and nodes in group k, and $E_{\perp\!\!\!\perp}(f(Y, X))$ is the expectation of the statistic $f(Y, X)$, conditional upon $Y = y$ and the category counts (that is, the number of nodes in each category of x, $n(x) = \{n_k(x)\}_{k=1}^{K}$), assuming that X and Y are independent.

2.2 Partially Observed Networks

Handcock and Gile (2007) developed a theory of missing data for ERG models, and the specification for ERN models proceeds similarly, though our formulation supports a more general class of missingness processes known as missing not at random (MNAR; see Rubin (1976)). We define T_{obs} and T_{miss} to be the observed and unobserved part of T respectively. We write $T = (T_{obs}, T_{miss})$, with realizations $t = (t_{obs}, t_{miss})$. Let W be a random variable representing the sampling process with realization w. The probabilistic distribution of W is the sampling mechanism, and must fully specify the sample selection process, including the partition of T into T_{obs} and T_{miss}. Typically, W will consist of an n by n matrix indicating whether the dyad was sampled, and an n by K matrix indicating which nodal attributes are missing; however, W may contain additional information about the sampling, such as the order of sampling.

Let us define the full data likelihood as

$$p(T = t, W = w|\eta, \theta) = p(W = w|T = t, \theta)\frac{1}{c(\eta, \mathcal{T})}e^{\eta \cdot g(t)}.$$

We wish to draw inferences about η from the observed data likelihood, defined as

$$p(T_{obs} = t_{obs}, W = w|\eta, \theta) = \sum_{t_{miss}} p(W = w|t = (t_{obs}, t_{miss}), \theta)\frac{1}{c(\eta, \mathcal{T})}e^{\eta \cdot g((t_{obs}, t_{miss}))}. \tag{2.2}$$

In the case where the sampling probabilities only depend on the observed data, then the sampling design is amenable to the model Handcock and Gile (2010), and is ignorable in the sense of Rubin (1976). In this case, the likelihood simplifies

to

$$\begin{aligned} p(T_{obs} = t_{obs}, W = w|\eta, \theta) &= \sum_{t_{miss}} p(W = w|T_{obs} = t_{obs}, \theta) \frac{1}{c(\eta, \mathcal{T})} e^{\eta \cdot g((t_{obs}, t_{miss}))} \\ &= p(W = w|T_{obs} = t_{obs}, \theta) \sum_{t_{miss}} \frac{1}{c(\eta, \mathcal{T})} e^{\eta \cdot g((t_{obs}, t_{miss}))} \\ &\propto \sum_{t_{miss}} \frac{1}{c(\eta, \mathcal{T})} e^{\eta \cdot g((t_{obs}, t_{miss}))}. \end{aligned} \tag{2.3}$$

Thus, when the sampling process is ignorable, inferences on η are not affected by $p(W = w|T_{obs} = t_{obs}, \theta)$, and so knowledge of the sampling process is not essential for the process of inference.

Having defined the full and observed likelihood, it is also useful to define the missing data likelihood:

$$p(T_{miss} = t_{miss}|W = w, T_{obs} = t_{obs}, \eta, \theta) = \frac{1}{c(t_{obs}, w, \eta, \theta)} p(W = w|T = (t_{obs}, t_{miss}), \theta) e^{\eta \cdot g((t_{obs}, t_{miss}))}$$

where

$$c(t_{obs}, w, \eta, \theta) = \sum_{t_{miss}} p(W = w|T = (t_{obs}, t_{miss}), \theta) e^{\eta \cdot g((t_{obs}, t_{miss}))}.$$

The (observed data) likelihood can then be rewritten as the ratio of two normalizing constants

$$\begin{aligned} p(T_{obs} = t_{obs}, W = w|\eta, \theta) &= \frac{1}{c(\eta, \mathcal{T})} \sum_{t_{miss}} p(W = w|T = t, \theta) e^{\eta \cdot g(t)} \\ &= \frac{c(t_{obs}, w, \eta, \theta)}{c(\eta, \mathcal{T})}, \end{aligned}$$

and using this, we may write the observed data log likelihood ratio of (η, θ) versus

(η_0, θ_0) as

$$\begin{aligned}
\ell(\eta,\theta) - \ell(\eta_0,\theta_0) &= \log(\frac{c(t_{obs},w,\eta,\theta)}{c(t_{obs},w,\eta_0,\theta_0)}) - \log(\frac{c(\eta,\mathcal{T})}{c(\eta_0,\mathcal{T})}) \\
&= \log(\sum_{t_{miss}} \frac{p(W=w|T=t,\theta)}{p(W=w|T=t,\theta_0)} e^{(\eta-\eta_0)\cdot g(t)} \frac{p(W=w|T=t,\theta_0)e^{\eta_0\cdot g(t)}}{c(t_{obs},w,\eta_0,\theta_0)}) \\
&\quad - \log(\sum_{t_{miss}} e^{(\eta-\eta_0)\cdot g(t)} \frac{e^{\eta_0\cdot g(t)}}{c(\eta,\mathcal{T})}) \\
&= \log(E_{\eta_0,\theta_0}(\frac{p(W=w|T,\theta)}{p(W=w|T,\theta_0)} e^{(\eta-\eta_0)\cdot g(T)})|W=w, T_{obs}=t_{obs}) \qquad (2.4) \\
&\quad - \log(E_{\eta_0}(e^{(\eta-\eta_0)\cdot g(T)})) \\
&= \log(E_{\eta_0,\theta_0}(e^{(\eta-\eta_0)\cdot g(T)}|T_{obs}=t_{obs})) - \log(E_{\eta_0}(e^{(\eta-\eta_0)\cdot g(T)})) \\
&\quad + \log(\frac{E_{\eta,\theta}\ (P(W=w|T,\theta\ )|T_{obs}=t_{obs})}{E_{\eta_0,\theta_0}(P(W=w|T,\theta_0)|T_{obs}=t_{obs})}). \qquad (2.5)
\end{aligned}$$

Both equation 2.4 and equation 2.5 motivate algorithms to draw inferences about η and θ. Section 2.3 describes the algorithm motivated by equation 2.4, and Appendix A.1 outlines an algorithm using equation 2.5.

2.3 Calculating the MLE with MCMC

For most models, equation 2.4 is not analytically solvable. However we may approximate it by Markov Chain Monte Carlo (MCMC). Let $t^{(i)}$ and $t_m^{(i)}$ where $i \in (1,\ldots,M)$ be samples from the full likelihood and missing data likelihood respectively with parameters η_0, θ_0. Then equation 2.4 may be approximated by

$$\ell(\eta,\theta) - \ell(\eta_0,\theta_0) \approx \log(\frac{1}{M}\sum_i^M \frac{p(w|t_m^{(i)},\theta)}{p(w|t_m^{(i)},\theta_0)} e^{(\eta-\eta_0)\cdot g(t_m^{(i)})}) - \log(\frac{1}{M}\sum_i^M e^{(\eta-\eta_0)\cdot g(t^{(i)})}) \qquad (2.6)$$

As η, θ move away from η_0, θ_0 the quality of this approximation degrades. Because we will be optimizing equation 2.4, it is useful to have both the first and second derivatives of the log likelihood, which are

$$\frac{\delta \ell}{\delta \eta} = E_{\eta,\theta}(g_i(t)|T_{obs} = t_{obs}, W = w) - E_{\eta,\theta}(g_i(T))$$

$$\frac{\delta^2 \ell}{\delta \eta_i \delta \eta_j} = -cov(g_i(T), g_j(T)) + cov(g_i(T), g_j(T)|T_{obs} = t_{obs}, W = w).$$

The expectations and covariances in these derivatives can be approximated using the conditional and unconditional MCMC samples and thus we can then use the following algorithm to approximate the MLE.

1. Let $k = 0$ and choose initial parameter values $\eta^{(0)}$, θ_0.
2. Use MCMC to generate k samples, $t^{(i)}_{miss}$ from $P(T_{miss} = t_{miss}|\eta^k, T_{obs} = t_{obs}, W = w)$.
3. Use MCMC to generate m samples $t^{(i)}$ from $P(T = t|\eta^k)$.
4. Using the samples from step 2 and 3 in equation (2.6), find η^{k+1}, θ^{k+1} maximizing the likelihood ratio, subject to $||\eta^{k+1} - \eta^k|| < \epsilon$ and $||\theta^{k+1} - \theta^k|| < \epsilon$.
5. If the likelihood has not converged, set $k = k + 1$ and go to step 2.
6. Let the MLE estimate be $\hat{\eta} = \eta^{k+1}$ and $\hat{\theta} = \theta^{k+1}$

Asymptotic standard errors for $\hat{\eta}$ may be obtained using an MCMC approximation to the Fisher information (i.e. the second derivative of the log likelihood). While asymptotics of the Fisher information are not assured with respect to ERNM (or ERGM) models, Fellows and Handcock (2012) show strong empirical agreement between the Fisher information standard errors and parametric bootstrap simulations. Standard errors for the mean value parameters $\hat{\mu} = E(g(T)|\eta = \hat{\eta})$ can be approximated by MCMC sampling.

2.4 Specific Forms of Partial Observation

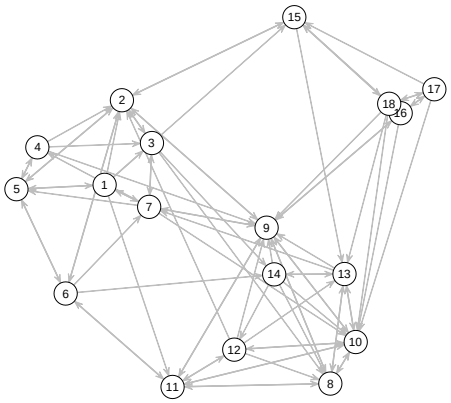
In this section we consider the three common forms of partial observation considered in the introduction, each corresponding to a different mechanism of partial observation or conceptualization of that mechanism.

2.4.1 Missing Data: Unobserved Relational Information

It is common when surveying networked populations that there are insufficient resources to conduct a census of the population and their relations. For efficiency reasons, a sampling based survey is undertaken, or the full network is partially observed due to non-response. In this sub-section, we give an illustration of the effect of non-response where the dyad information is missing completely at random. We consider the relations of "liking" among 18 monks in a monastery (Sampson, 1969). The network analyzed has a directed edge between two monks if the sender monk ranked the receiver monk in the top three monks for positive affection in any of the three interviews given over a twelve month period (Hoff et al., 2002). The sociogram of this dataset is shown in Figure 2.1. One nodal attribute of interest is an indicator of attendance at the minor "Cloisterville" seminary before coming to the monastery.

We fit a simple homophily model on Cloisterville status using the full data. We then ran simulations on the effect of missingness by selecting dyads, and Cloisterville status variates, completely at random and setting them to missing. Figure 2.2 shows one simulated missingness pattern with 15% missing. We ran 100 simulations at each missingness percentage. Means and standard deviations of the ERNM models fit to these simulated missingness patterns are displayed in Figure 2.3.

We see that the standard deviations of the estimates increase as the amount of missingness increases. At the higher missingness levels some bias is apparent



1	Ramauld (L)	10	Gregory (T)
2	Bonaventure (L)	11	Hugh (T)
3	Ambrose (L)	12	Boniface (T)
4	Berthold (L)	13	Mark (T)
5	Peter (L)	14	Albert (T)
6	Louis (L)	15	Amand (O)
7	Victor (L)	16	Basil (O)
8	Winfred (T)	17	Elias (O)
9	John (T)	18	Simplicius (O)

Figure 2.1: Relationships among monks within a monastery and their affiliations as identified by Sampson: Young (T)urks, (L)oyal Opposition, and (O)utcasts.

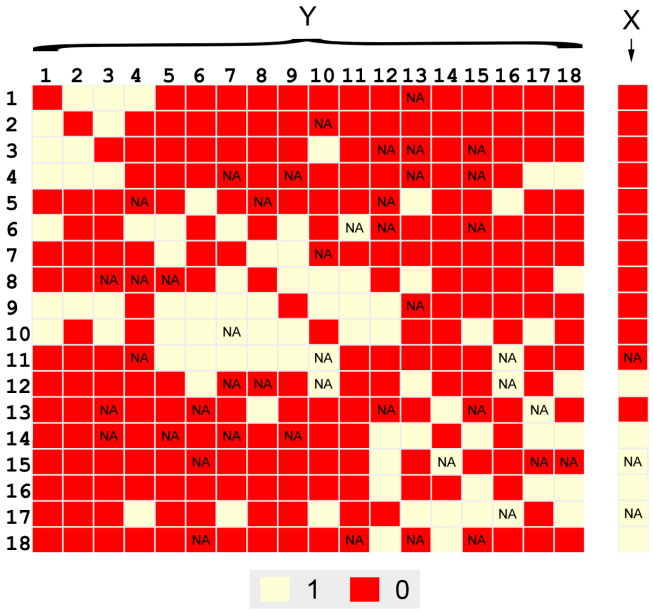


Figure 2.2: Sampson's monks with 15% missingness. Cloisterville status marked on the right hand side.

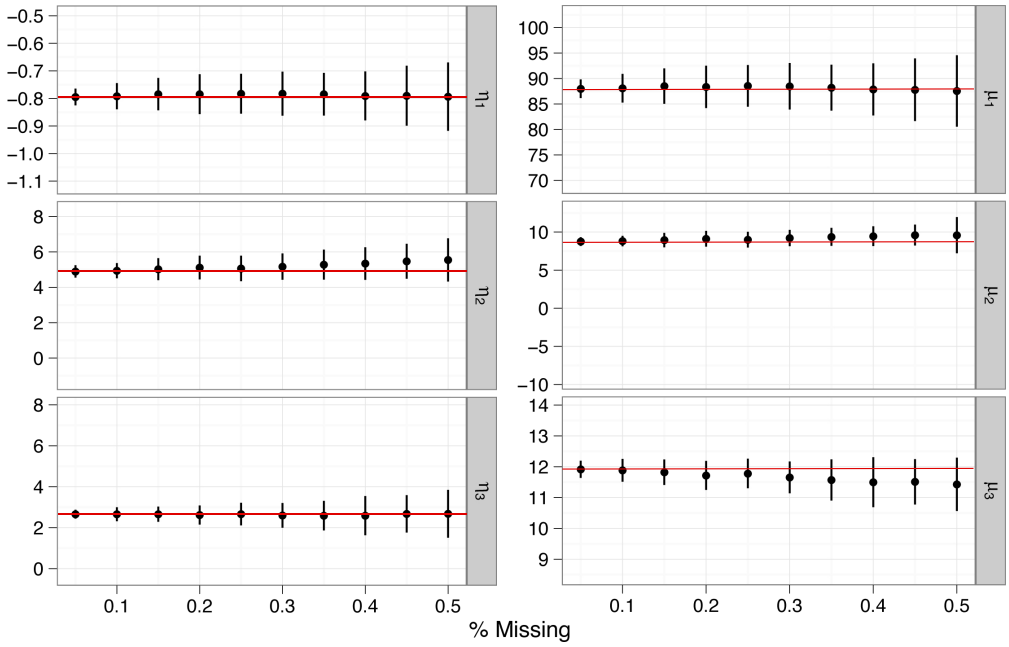


Figure 2.3: Means and standard deviations of model estimates. Red lines indicate fully observed MLE

relative to the full data MLE, but not more than one standard deviation. One possible explanation for this bias is that there were only six monks who attended Cloisterville, and so at 50% missingness, a significant number of samples will include no (or perhaps a single) Cloisterville monk.

2.4.2 Latent Variables: Stochastic Block Models

In this sub-section we consider the situation where some characteristics of the network are posited but unobserved. Specifically, we consider the case where each node of the network belongs to a latent class, and the structure of the network depends on that latent class. The traditional approach to this has been stochastic block models Nowicki and Snijders (2001), and here we show how these models fall naturally out of our general formulation.

It is apparent from Figure 2.1 that the pattern of "liking" between the monks

may exhibit clustering. Through close sociological study, Sampson (1969) identified three clusters which he dubbed the Turks, Loyal Opposition and the Outcasts (see: Figure 2.1). Here we will attempt to identify clusters by inferring class membership from the graph. We fit the simple homophily model of Section 2.1.1 to this data, assuming a class covariate, X, with three levels, and that all of the monks are "missing" their class covariate. The simple homophily model treated this way represents a novel latent block model in the spirit of Nowicki and Snijders (2001). Note that the missingness process here is ignorable because it does not depend on unobserved quantities as all of the x values are missing regardless of the Y values. We fit the model using the algorithm in Section 2.3. Table 2.1 shows the maximum likelihood parameter estimates, along with standard errors of the estimators based on the Fisher information.

Term	$\hat{\eta}$	$\hat{\mu}$	$s.e.(\hat{\eta})$	$s.e.(\hat{\mu})$
# of edges	-0.58	88.23	0.14	7.48
Homophily	7.28	15.30	0.91	1.33
# in group 0	-2.50	3.95	1.44	1.08
# in group 1	-0.02	6.95	1.31	0.99

Table 2.1: Latent Class model for Sampson's monks.

The natural parameter estimates indicate significant homophily in tie formation based on the class. It also indicates that the number of monks in the third class is significantly more than those of the other two classes, which are not statistically significantly different in size. The mean value parameters indicate that the expected number of ties is about 88, and the expected numbers in the three groups are 4, 7 and 7.

An advantage of this approach is that we can investigate the probability of class membership, which is well defined through our framework as $p(X = x|Y = y_{obs}, \eta)$. To compute $p(X = x|Y = y_{obs}, \eta)$ we simulated a large number of samples from $p(X = x|Y = y_{obs}, \hat{\eta})$ using MCMC to show the probability of the monks being in

the classes displayed in Figure 2.1 to be above 0.9999. These clusters were also identical to those chosen by Sampson (1969) and verified by later research Breiger et al. (1975), Handcock et al. (2006).

In addition to assuming a set number of latent classes for the model, we can also use the MLE procedure to select an appropriate number of clusters for the data. We fit the simple homophily model with a latent variable X able to take a potentially large number of values (e.g., the number of monks). In this case $p(X = x|Y = y_{obs}, \hat{\eta})$ places zero mass for all but three of the groups. This is evidence that the three groups we have identified are a good classification for these data. More sophisticated model selection approaches for choosing the number of clusters are possible (Handcock et al., 2006), and are left for future work.

Our form of the stochastic block model is conceptually very clean with the ability to naturally incorporate additional covariates, multiple membership variables, and extensions to an unbounded numbers of classes. Inference is straightforward, and quantities such as the probability of class membership are well defined and interpretable. We leave a full exploration of these for latter work.

2.4.3 Network Sampling: Biased Seed Link-Tracing

Handcock and Gile (2010) explored the idea of sampling networks by tracing the edges. As a general concept, link tracing involves selecting one or more seed nodes, and then observing the edges connected to those seeds. One or more of these edges are then followed to the neighboring node, whose ties are observed, and the process is continued. Each iteration of this process is known as a wave.

Provided that the seed nodes are chosen at random, and the method by which edges are chosen to be followed depends only on the observed data, this missingness process is ignorable. To be explicit, consider a link tracing process with k waves. Let w_i be the ordered set of nodes and edges sampled in

the ith wave in the order in which they were sampled, $w = \{w_0, ..., w_k\}$, and $w_{-i} = \{w_0, ..., w_{i-1}, w_{i+1}, ..., w_k\}$. If the seeds are chosen at random, and the edges followed by the sampling process are also chosen at random, then at each step in the sampling process, the nodes sampled in the next step depend only on the ties connected to an observed node, which are observed as well. Thus, $p(W = w|T = t, \theta) = p(W = w|T_{obs} = t_{obs}, \theta)$, implying that the missingness is ignorable.

In many cases, however, the seeds are not chosen at random from the population, but are some form of convenience sample. For example, in a population where some people have an infection and others do not, we may start with a sample of s_i seeds picked at random from among the infected individuals, and s_{-i} seeds picked from the non-infected individuals. These seeds are then used as a starting point for standard link tracing. We may then write the sampling probability as

$$\begin{aligned} p(w|t,\theta) &= p(w_0|t,\theta)p(w_{-0}|t_{obs}, w_0, \theta) \\ &= \frac{(n_i - s_i)!}{n_i!}\frac{(n_{-i} - s_{-i})!}{n_{-i}!}p(w_{-0}|t_{obs}, w_0, \theta), \end{aligned}$$

where n_i and n_{-i} are the number of infected and non-infected in the population, respectively. Note that $p(w_{-0}|t_{obs}, w_0, \theta)$ does not depend on t_{miss} and may be factored out of the likelihood in equation (2.2). Thus there is no need to calculate $p(w_{-0}|t_{obs}, w_0, \theta)$ explicitly, as it makes no impact on the likelihood. Hence, in this case, we can compute a likelihood without knowing the specific mechanism of seed selection.

2.4.4 Network Sampling: Positive Contact Tracing

As emerging epidemics develop, control measures (e.g., treatment, isolation and culling) focus on those members of the population that are known to have the in-

fection. Because there are often many infected people who are unobserved, control can be ineffective (e.g., HIV (Potterat et al., 1989). The alternative of applying control measures to the entire population can be economically infeasible or ineffective e.g., instances of safe sex education (Potterat et al., 1989, Klinkenberg et al., 2006). Contact tracing is the hybrid approach of treating both the known infected individuals and those who may have been infected by them (Potterat et al., 1989, Klinkenberg et al., 2006). In U.S. public health, health clinics are required by state law to notify those at risk from infection due to their sexual relations with individuals tested, and found to be infected by the clinic. The process of locating, notifying and then testing partners that may have been exposed to an infectious agent allows additional information about the partners to be collected. While the primary purpose of contact tracing is disease control via partner notification and partner services, it is also a form of data collection that is rarely utilized. Such approaches are used most commonly for syphilis and HIV/AIDS, but also for other STIs such as gonorrhea and chlamydia (Golden et al., 2004), as well as routinely for tuberculosis and infectious disease outbreaks. Contact tracing has also been applied in many recent epidemics (Fenner et al., 1988, Ferguson et al., 2001, Donnelly et al., 2003). In positive contact tracing, we follow all edges from infected nodes, but edges from uninfected nodes are not followed.

While the process varies from state to state and also by disease, we consider the following biased seed link tracing process:

1. Select s_{-i} seed subjects at random from among the non-infected population, observe them.

2. Select s_i seeds subjects at random from among the infected population, observe them.

3. Choose the next infected seed at random.

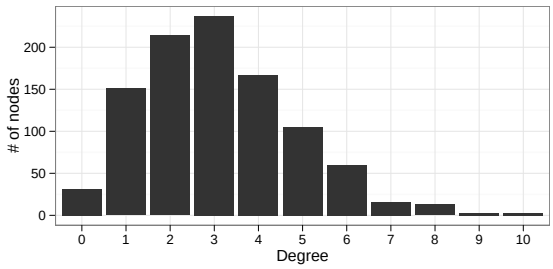


Figure 2.4: Degree distribution of the networked population.

4. Observe all edges from the selected subject, and the infection status of these subjects.

5. For all infected neighbors of the selected subject, go to step 4.

6. If all the seeds have not been chain sampled, go to step 3

We simulated a networked population of $n = 1000$ people from the simple homophily model of Section 2.1.1 with natural parameters of $\eta = (-5.8, .7, -1.95)$. The number of infected nodes was fixed at 150. The generated network had a mean degree of 3.1, and its degree distribution is displayed in Figure 2.4. There were 296 infected to non-infected ties, with the mixing distribution displayed in Figure 2.5 indicating moderate homophily.

Starting with $s_i = 40$ infected seeds, we simulated 100 positive link tracing samples for each of $s_{-i} = (0, 45, 90, 135, 180, 225)$. Figure 2.6 displays a histogram of the sizes of the samples when there are no non-infected seeds (i.e., $s_{-i} = 0$).

To provide a comparison for our method we considered two estimators that could be utilized. Neither of them uses a model for the networked population but is motivated by approximations to the sampling design. The first treats the sample as a simple random sample

$$Naive = n\frac{n_i}{n_i + n_u},$$

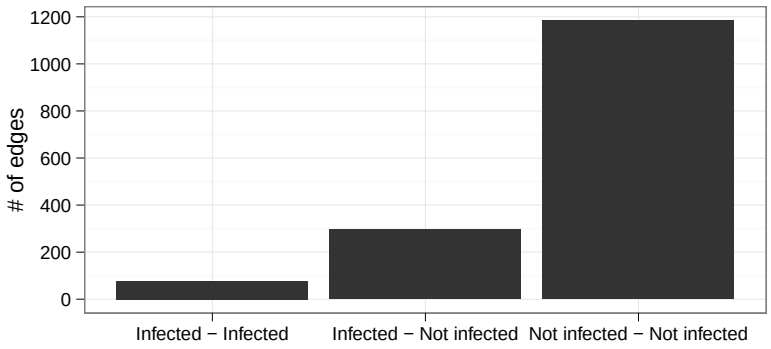


Figure 2.5: Mixing statistics: Counts of the numbers of edges by the infection status of the incident nodes for the networked population.

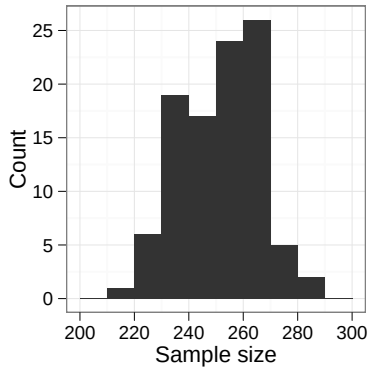


Figure 2.6: Sizes of the contact-traced samples based on 40 seed subjects ($s_i = 40, s_{-i} = 0$).

where n_i and n_u are the number of infected and uninfected in the sample respectively. The second adjusts for the sampling of the seeds

$$Naive\ (seed\ adj.) = (n - s_i - s_{-i})\frac{n_i - s_i}{n_i - s_i + n_u - s_{-i}} + s_i.$$

Our approach is to fit an ERNM to the contact tracing data. In this situation the contact tracing sampling design is clearly informative. For comparison, we compute two estimates of the model. The first takes into account the informativeness of the contact tracing design (MNAR) and the other assume it is ignorable (MAR). These are based on the likelihoods 2.2 and 4.4, respectively, and the algorithm in Section 2.3.

Figure 2.7 shows the results for each of the estimators over the samples. The median of the MNAR estimator is centered around the true value of 150 in all sampling scenarios, while the MAR estimator performs poorly with all infected seeds ($s_{-i} = 0$) and increasingly well as the number of non-infected seeds increases to $s_{-i} = 225$. This is somewhat expected as the proportion of infected in the seeds approximately matches that of the population when $s_{-i} = 225$. The two naive estimators are significantly biased across all samples. This is especially true for the sample mean which is biased both by the seed selection and by the link-tracing design. The adjusted sample mean corrects somewhat for the seed bias but does not represent the link-tracing.

This application illustrates the advantage of the model-based approach over the ad hoc estimators. By representing the structure of the networked population, the model-based approach can leverage the information in the data more efficiently.

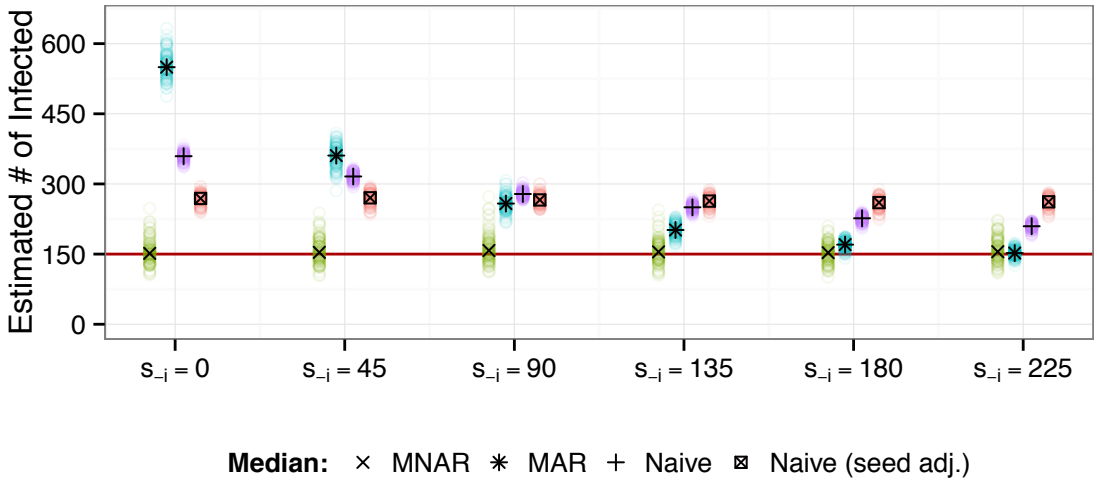


Figure 2.7: Estimates via contact tracing with $s_i = 40$ infected seeds and varying numbers of non-infected seeds.

2.5 Discussion

In this chapter we have given a concise and systematic statistical framework for dealing with partially observed network data when some knowledge is available on the sampling design. The framework includes, but is not restricted to, ignorable sampling designs. We have also shown that likelihood-based inference is practical under partial observation for ERN models, and that the likelihood framework naturally accommodates standard sampling designs.

We developed and implemented algorithms to compute Monte Carlo approximations to the likelihood, and showed how these can be used in practice. Three important special cases of these designs were demonstrated in Section 2.4. In Subsection 2.4.1 we consider a missingness process which randomly selected dyads and nodal attributes to be missing. Sub-section 2.4.1 considers the case where *a*ll nodal attributes are missing, thus introducing a novel form of the latent cluster model.

In Sub-section 2.4.3 we consider non-ignorable sampling in the context of con-

tact tracing data, a case of vital importance to public health. At present, this is the first statistically defensible approach to inference in this form of data. The example presented here shows that the MLE estimation task is robust, but is limited by the fact that inference was performed on a simulated network. Whether the model presented here would provide a good fit for real public health data remains an important research question that we hope to address in the future.

2.6 Appendix: Algorithmic and Computational Details

2.6.1 A.1: Alternate MLE Formulation

While the algorithm outlined in Section 2.3 works well, there are some situations where an alternate formulation using equation (2.5) may be useful. First let us consider the case where $\theta = \theta_0$, then the likelihood is

$$\ell(\eta) - \ell(\eta_0) = \log(E_{\eta_0}(e^{(\eta-\eta_0)\cdot g(T)}|t_{obs})) - \log(E_{\eta_0}(e^{(\eta-\eta_0)\cdot g(T)})) + \log(\tfrac{E_\eta(P(W=w|T,\theta)|T_{obs}=t_{obs})}{E_{\eta_0}(P(W=w|T,\theta)|T_{obs}=t_{obs})} \quad (2.7)$$

The first expectation, and the expectation in the denominator of the third term, can be calculated using an MCMC sample from $p(t|t_{obs}, \eta_0)$. The second can be approximated with an MCMC sample from $p(t|\eta_0)$. The numerator of the third term can be approximated by importance sampling.

$$E_\eta(P(W = w|T, \theta)|T_{obs} = t_{obs}) \approx \frac{1}{k}\sum_i^k p(w|t^{(i)}, \theta)\omega^{(i)}$$

where $t^{(i)} \sim p(t|t_{obs}, \eta_0)$ and

$$\omega^{(i)} = \frac{e^{(\eta-\eta_0)\cdot g(t^{(i)})}}{\sum_j^k e^{(\eta-\eta_0)\cdot g(t^{(j)})}}$$

If the sampling process is ignorable, then the third term drops out of the likelihood ratio. The first and second derivatives of the likelihood are useful in the

maximization process. For notational convenience let $\Delta_i(t) = g_i(t) - E(g_i(T))$.

$$\begin{aligned}
\frac{\delta \ell}{\delta \eta} &= \frac{\delta}{\delta \eta_i} \log(\sum_{t_{miss}} p(W = w|T = t) P(T_{miss} = t_{miss}|\eta, T_{obs} t_{obs}) P(T_{obs} = t_{obs}|\eta)) \\
&= \frac{\sum_{t_{miss}} p(W = w|T = t) \Delta_i(t) P(T_{miss} = t_{miss}|\eta, T_{obs} = t_{obs}) P(T_{obs} = t_{obs}|\eta)}{\sum_{t_{miss}} p(W = w|T = t) P(T_{miss} = t_{miss}|\eta, T_{obs} = t_{obs}) P(T_{obs} = t_{obs}|\eta)} \\
&= \frac{E(p(W = w|T) \Delta_i(T)|T_{obs} = t_{obs})}{E(p(W = w|T)|T_{obs} = t_{obs})}
\end{aligned}$$

$$\begin{aligned}
\frac{\delta^2 \ell}{\delta \eta_i \delta \eta_j} &= \frac{\delta}{\delta \eta_j} \frac{\sum_{t_{miss}} P(W = w|T = t) \Delta_i(t) P(T_{miss} = t_{miss}|\eta, T_{obs} = t_{obs}) P(T_{obs} = t_{obs}|\eta)}{\sum_{t_{miss}} P(W = w|T = t) P(T_{miss} = t_{miss}|\eta, T_{obs} = t_{obs}) P(T_{obs} = t_{obs}|\eta)} \\
&= -cov(g_i(T), h_j(T)) + \frac{E(p(W = w|T) \Delta_i(T) \Delta_j(T)|T_{obs} = t_{obs})}{E(p(W = w|T)|T_{obs} = t_{obs})} \\
&\quad - \frac{E(p(W = w|T) \Delta_i(T)|T_{obs} = t_{obs}) E(p(W = w|T) \Delta_j(T)|T_{obs} = t_{obs})}{E(p(W = w|T)|T_{obs} = t_{obs})^2}
\end{aligned}$$

And if the missingness process is ignorable, these equations simplify to

$$\frac{\delta \ell}{\delta \eta} = E(\Delta_i(T)|T_{obs} = t_{obs})$$

$$\frac{\delta^2 \ell}{\delta \eta_i \delta \eta_j} = -cov(g_i(T), g_j(T)) + cov(g_i(T), g_j(T)|T_{obs} = t_{obs})$$

If we fix η, then the observed likelihood of θ

$$\begin{aligned}
L(\theta|t_{obs}, w, \eta) &\propto P(t_{obs}|\eta) E(P(W = w|T, \theta)|T_{obs} = t_{obs}) \\
&= E(P(W = w|T, \theta)|T_{obs} = t_{obs}, \eta)
\end{aligned}$$

can be maximized to find the MLE of θ.

This motivates the following algorithm for maximizing the observed data likelihood.

1. Let $k = 0$ and choose initial parameter values $\eta^{(0)}$, θ_0.

2. Use MCMC to generate k samples, $t_{miss}^{(i)}$ from $P(t_{miss}|\eta^k, t_{obs})$.

3. Use MCMC to generate m samples $t^{(i)}$ from $P(t|\eta^k)$.

4. Set $\theta^{k+1} = argmax(E(P(w|T,\theta)|T_{obs} = t_{obs}, \eta))$, with samples from step 2 used to approximate the expectation.

5. Using the samples from steps 2 and 3 to approximate the relevant expectations, find η^{k+1} maximizing equation (2.7) subject to $||\eta^{k+1} - \eta^k|| < \epsilon$.

6. Set $k = k + 1$, and go to step 2.

The disadvantage of this method is that if the networks generated by the MNAR process are very different from those generated assuming MAR, the estimates of the last expectation in equation (2.7) can become unstable. The benefit of using this method is that the sampling probability ($P(W = w|T = t, \theta)$) only needs to be calculated for networks included in the sample, and not at every MCMC step as is required by the algorithm in Section 2.3, so if the sampling probability is computationally expensive to calculate, this method can be significantly faster than the one outlined in Section 2.3

2.6.2 A.2: Estimating Network Statistics

We can use MCMC samples from $p(t_{miss}|t_{obs}, \eta)$ to estimate the network statistics of the sampled network. Suppose that we have used MCMC to draw k samples $t^{(i)}_{miss}$ from the distribution $p(t_{miss}|t_{obs}, \eta)$, and $t^{(i)} = (t_{obs}, t^{(i)}_{miss})$. Then we can estimate the expectation of a set of network statistics g as

$$E(g(T)|t_{obs}, \eta) \approx \frac{1}{k} \sum_{i=0}^{k} g(t^{(i)}).$$

However, this equation ignores the possible bias introduced by our sampling process w. The distribution that we should be sampling from is the full conditional distribution of t_{miss},

$$p(T_{miss} = t_{miss}|T_{ob} = t_{obs}, W = w, \eta) \propto p(T_{miss} = t_{miss}|T_{obs} = t_{obs}, \eta)p(W = w|T = t, \theta).$$

We then use importance sampling to estimate the relevant quantity

$$E(g(T)|t_{obs}, w, \eta, \theta) \approx \frac{\sum_{i=0}^{k} g(t^{(i)})p(W = w|T = t^{(i)}, \theta)}{\sum_{i=0}^{k} p(W = w|T = t^{(i)}, \theta)}.$$

CHAPTER 3

Implementing MCMC-MLE in Exponential-Family Models

3.1 Introduction

In the previous chapters we have heavily relied on the use of Markov Chain Monte Carlo Maximum Likelihood (MCMC-MLE; Geyer and Thompson 1992) to fit our ERNM models, both when the full network is observed, and in the partially observed case. This algorithm is central to both ERGM and ERNM models, and is non-trivial to implement in a robust and reliable manner.

In this chapter we show how the exponential family likelihood permits a useful approximation to the normalizing constant that increases the stability and accuracy of the MCMC-MLE algorithm. We also show how MCMC standard errors can be used as a measure of when to trust this approximation. Though we are particularly interested in ERNM, the methods outlined in this chapter pertain to any exponential family distribution.

3.2 The Geyer-Thompson Likelihood Ratio Formulation

Let T be random variate with realization t, then the general exponential family model for T is expressed as

$$P(T = t|\eta) = \frac{1}{c(\eta)} e^{\eta \cdot h(t) + g(t)}, \tag{3.1}$$

where h is a vector valued function generating sufficient statistics for T, g is an offset statistic, and c is the normalizing constant. If T is high dimensional, calculating c becomes increasingly intractable, and so the likelihood becomes difficult or impossible to evaluate.

If we consider the log likelihood ratio of η versus η_0, we can transform the likelihood into a more tractable form

$$\ell(\eta) - \ell(\eta_0) = (\eta - \eta_0) \cdot h(t) - \log[E_{\eta_0}(e^{(\eta - \eta_0) \cdot h(T)})]. \tag{3.2}$$

The first term is a simple function of the sufficient statistics and is thus easy to evaluate. The second term requires the calculate of the expectation of a quantity under η_0. This is difficult to calculate in general, but if we have an MCMC sample from the model at η_0, we can use the sample to approximate the expectation.

If only part of T is observed and the missingness process is ignorable in the sense of Rubin (1976) then the log likelihood ratio may be written as

$$\ell(\eta) - \ell(\eta_0) = \log(E_{\eta_0}(e^{(\eta - \eta_0) h(T)}) | T_{obs} = t_{obs}) - \log(E_{\eta_0}(e^{(\eta - \eta_0) h(T)})), \tag{3.3}$$

where T_{obs} is the observed part of T (Handcock and Gile, 2010, Geyer, 1994, Gelfand and Carlin, 1993). The second term is identical to the second term in the fully observed case, and the first term can be approximated by an MCMC sample conditional upon the observed data. This formulation is exactly analogous to the development in Section 2.2.

3.3 Approximating the Expectation Via MCMC

Given a sample t_i for $i \in (1, ..., n)$ from the relevant distribution, Geyer and Thompson (1992) suggested approximating the log expectations with

$$\log(E_{\eta_0}(e^{\eta^* h(T)})) \approx \log(\frac{1}{n}\sum_{i}^{n} e^{\eta^* h(t_i)})$$

where $\eta^* = (\eta - \eta_0)$

Hummel et al. (2012) found that (in the context of exponential family random graph models) this approximation degrades quickly as η moves away from η_0, and suggested an alternate approximation. If $\eta^* h(T)$ is normally distributed, then the log expectation is

$$\log(E_{\eta_0}(e^{\eta^* h(T)})) = E(\eta^* h(T)) - \text{var}(\eta^* h(T))/2 \approx \hat{E}(\eta^* h(T)) - \hat{\text{var}}(\eta^* h(T))/2.$$

In the case of the Erdos-Renyi ERGM model this approximation was shown to outperform the Geyer-Thompson approximation, but there is no reason to think think that X will be distributed normally in general. In fact if any of the statistics h are distributed non-normally, then there exist a set of η^* in the maximization process that yield a non-normal $\eta^* h(T)$.

3.3.1 The Cumulant Generating Function Approximation

The log expectation that we are attempting to evaluate is known as the cumulant generating function. Performing a Taylor expansion around 0 of e^x and then of

$\log(x)$ we obtain

$$\begin{aligned} log(E(e^{\eta^* h(T)})) &= log(E(1 + \eta^* h(T) + \eta^* h(T)^2/2 + ...)) \\ &= log(1 + E(\eta^* h(T)) + E((\eta^* h(T))^2)/2 + ...) \\ &= (E(\eta^* h(T)) + E((\eta^* h(t))^2)/2 + ...) - \\ &\quad (E(\eta^* h(T)) + E((\eta^* h(T))^2)/2 + ...)^2/2 + ... \\ &= \sum_i^\infty \frac{\kappa_i}{i!} \\ &\approx \sum_i^m \frac{\hat{\kappa}_i}{i!} \end{aligned} \tag{3.4}$$

where κ_i is the ith cumulant of $\eta^* h(T)$, and $\hat{\kappa}_i$ is the sample cumulant based on the MCMC sample. If we let $m = 2$ we obtain the log normal approximation, so this derivation provides a justification for its use even when the distribution is significantly non-normal. As η^* moves away from 0 the approximation degrades. The more terms that are used, the less error there is in the Taylor expansions, but higher order cumulants have more sampling error, so a balance must be struck between these two competing error sources. We have had success using $2 \leq m \leq 4$.

3.4 When to Trust the Sample

No matter what approximation is used to estimate the cumulant generating function, the approximation degrades as η^* deviates from 0. The solution to this is to take an iterative approach, maximizing the log likelihood subject to the constraint that η^* is not to large. Then setting $\eta_0 \leftarrow \eta$, and repeating the maximization until convergence. The question of course is how large should η^* be allowed to get before new MCMC samples are drawn.

Geyer and Thompson (1992) Suggested that the maximization should be performed subject to $\eta^* < \delta$ but provided no guidelines as to what δ should be.

The default settings of the R package ergm (Handcock et al., 2012) restrict the maximization such that $\ell(\eta) - \ell(\eta_0) < 20$.

More recently Hummel et al. (2012) developed a new approach where the maximization is done unconditionally, but the likelihood is altered. Specifically, the likelihood that is maximized is

$$\ell(\eta) - \ell(\eta_0) = (\eta - \eta_0) \cdot (pE_{\eta_0}(h(T)) + (1-p)h(t)) - \log[E_{\eta_0}(e^{(\eta-\eta_0)\cdot h(T)})], \quad (3.5)$$

where $p \in (0,1)$ is chosen such that the point $pE_{\eta_0}(h(T)) + (1-p)h(t)$ is inside the convex hull of the MCMC sample. This restricts the target mean value parameters (μ) to be within the convex hull of the sample, and thus the natural parameters η are also restricted to be near η_0.

Hummel et al. (2012) show that using this restriction leads to more robust estimation, especially in cases where the solution is near a degenerate region. One issue with this approach is that it only is applicable in the complete data case. If our data has missing values, the method can not be applied.

3.4.1 Effective Sample Size Restriction

In exponential family models, in addition to the natural parameters (η), there is an alternate expression of the model using the so called mean value parameters $\mu(\eta) = E_\eta(h(T))$. Given an MCMC sample from η_0, we can estimate the mean value parameters associated with η using importance sampling. Specifically, given an MCMC sample $t^{(i)}$ for $i \in 1, , k$ from the model at η_0 we can estimate the mean value parameters as

$$\hat{\mu}(\eta) = \frac{\sum_i^k \omega_i h(t_i)}{\sum_i^k \omega_i}$$

where $\omega_i = \exp((\eta - \eta_0) \cdot h(t_i))$. As η deviates from η_0 the estimation becomes increasingly poor, and we can measure the degradation of the approximation using

MCMC standard errors. If the mean value parameters can not be estimated accurately, we should have little confidence in our ability to estimate the likelihood ratio. If we divide our k samples into a batches of size b, then for each batch we can calculate the estimated mean values as

$$\hat{\mu}_j(\eta) = \frac{\sum_{i=(j-1)b+1}^{jb} \omega_i h(t_i)}{\sum_{i=(j-1)b+1}^{jb} \omega_i} \qquad \text{for} \quad j = 1, \ldots, a.$$

The MCMC batch mean standard error is then defined as

$$\hat{\sigma}_\mu(\eta) = \sqrt{\frac{b}{a-1} \sum_{j=1}^{a} (\hat{\mu}_j(\eta) - \hat{\mu}(\eta))^2}.$$

For any fixed batch size b, $\hat{\sigma}_\mu$ is not a consistent estimator of the true standard error (Glynn and Iglehart, 1990), but if we let $b = \sqrt{k}$, then consistency is achieved (Jones et al., 2006). We can then express the effective sample size, relative to a simple random sample at η_0 as

$$\hat{ess}(\eta) = k \frac{\hat{\sigma}_\mu(\eta)^2}{\frac{\hat{var}(h(T))}{k}}.$$

where $\hat{var}(h(T)) = \frac{1}{k} \sum_i^k (h(t_i) - \hat{\mu}(\eta_0))^2$. This motivates maximizing the likelihood subject to the constraint that

$$\hat{ess}(\eta) > s,$$

where s is the minimal acceptable sample size. Our empirical investigations with various ERNM models suggest that using a small number (such as 5) allows for fairly large jumps in the likelihood, while maintaining stability.

CHAPTER 4

A Model Based Analysis of Respondent Driven Sampling Data

The respondent driven sampling (RDS) design is a widely successful method of obtaining samples from difficult to reach populations (Johnston and Sabin, 2010). It is used extensively in the surveillance of infectious diseases such as HIV, both in industrialized nations and in the developing world (Malekinejad et al., 2008).

While it is a popular method in practice, the statistical validity of estimators based on the design have been met with some criticism based both on their theoretical underpinnings (Gile and Handcock, 2010) and on simulations in real networks (Goel and Salganik, 2010). Also, the types of inferences that are available to the researcher are limited to very simple quantities such as means and proportions. More complicated inferences such as group comparisons and regression modeling are unaddressed by current methodology.

This lack of a methodology has not stopped researchers on the ground from using a variety of methods ranging from simply ignoring the sampling process in the regression to mixed effect models (see Johnston et al. (2008) and references therein). This indicates that there is significant demand with the epidemiological community for a way to address regression problems within a network sampling framework.

In this chapter we frame the RDS design as a missing data process over an underlying social network. By modeling the network using Exponential Family

Random Network Models (ERNM), and the RDS design process we develop a novel framework for the analysis of RDS data. This framework allows us to estimate not only proportions, but also multivariable hypotheses such as logistic regression.

4.1 Respondent Driven Sampling

Respondent driven sampling is a variant of the link tracing design (Handcock and Gile, 2007), and proceeds as follows:

1. Recruit a group of subjects from the population (know as seeds). These subjects are chosen by convenience, and may represent a very biased group.

2. Each subject is asked how many people within the population of interest they have that they might be willing to recruit into the study. These people are referred to as the alters or neighbors of the subject. They are then asked to approach them and recruit them into the study. The number of recruits per subject is usually limited to two or three alters.

3. As subjects are recruited, they are in turn asked to recruit their alters until the desired sample size is reached.

The advantage of using RDS is that it exploits the fact that even though the researcher can't obtain a list of members of the population (which would be required to obtain a simple random sample), the members of the population do know each other. Thus by traversing the social network we can reach most or all people, and because members of the population will tend to trust their referrer, they are more likely to participate in the study.

The disadvantage of RDS is that specialized estimators adjusting for the sampling process must be used, and even then estimates can be highly biased (Gile and

Handcock, 2010) due to assumption violations. These estimators must contend with adjusting for both the network structure and initial seeds that are chosen by convenience and therefore may not be representative. In this article we will address two network structures which can be sources of bias. First, for many traits, two subjects with similar values of the trait may be more likely to be alters. This is known as homophily in the network literature. Secondly, members with a trait may be more socially active (i.e. have higher degrees) than average, leading to these subjects being more likely to show up in the sample than average. This is known as differential activity, and it can severely bias naive estimators (Gile and Handcock, 2010). Several estimators have been developed to account for the biases resulting from this non-random sampling process.

4.1.1 Unadjusted Mean (mean)

If the recruitment graph structure is independent of the outcome variable, then the sample mean is a correct and unbiased estimator of the population mean (Heckathorn, 1997). Formally, let $i \in \{1, ..., n\}$ be an RDS sample of size n from a population of size N, and $z \in \{0, 1\}$ be a binary trait, then the simple mean estimator is

$$\hat{\mu}_{mean} = \frac{1}{n} \sum_{i=1}^{n} z_i.$$

4.1.2 The Salganik-Heckathorn Estimator (rds-i)

Let y be an N by N matrix representing the recruitment graph, where $y_{j,k} = 1$ if subject k is a possible recruit of subject j, and $y_{j,k} = 0$ otherwise. This matrix is assumed to be symmetric ($y_{j,k} = y_{k,j}$). Further, let $t_{01} = \sum_{j=1}^{N} \sum_{k=1}^{N} I(z_j = 0) y_{j,k} I(z_k = 1)$ be the total number of edges in the recruitment graph connecting nodes with differing traits where I is the indicator function. We also define $N_a = \sum_{j=1}^{N} I(z_j = a)$ to be the number of nodes with trait a, and $D_a =$

$\frac{1}{N_a}\sum_{j:z_j=a}\sum_{k>j} y_{j,k}$ be the average degree (network size) of group a. We can then write the proportion of ties from nodes with $z_j = 0$ to nodes with $z_k = 1$ as

$$c_{01} = \frac{t_{01}}{N_0 D_0}.$$

Similarly we may write the reverse as

$$c_{10} = \frac{t_{01}}{N_1 D_1},$$

and by rearranging the terms we obtain

$$\mu = \frac{N_1}{N} = \frac{D_0 c_{01}}{D_1 c_{10} + D_0 c_{01}}.$$

This equation can be used as the basis of an estimator by inserting estimates for the quantities on the right hand side. We can estimate the average degree in each group with a generalized Horvitz-Thompson estimate with probability proportional to nodal degree

$$\hat{D}_a = \frac{\sum_{i=1}^{n} I(z_i = a)}{\sum_{i=1}^{n} \frac{1}{d_i}}.$$

The use of probability proportional to degree can be justified along the same lines as the Voltz-Heckathorn estimator in section 4.1.3.

If one assumes that the recruitment process is at the equilibrium of a Markov random walk through the graph, then we may estimate c as

$$\hat{c}_{01} = \frac{r_{01}}{r_{00} + r_{01}} \qquad \hat{c}_{10} = \frac{r_{10}}{r_{11} + r_{10}},$$

where r_{ab} is the observed number of recruiter-recruit in groups a and b respectively. Plugging these estimators in we obtain the Salganik-Heckathorn estimator

(Salganik and Heckathorn, 2008)

$$\hat{\mu}_{SH} = \frac{\hat{D}_0 \hat{c}_{01}}{\hat{D}_1 \hat{c}_{10} + \hat{D}_0 \hat{c}_{01}}.$$

4.1.3 The Voltz-Heckathorn Estimator (rds-ii)

The Voltz and Heckathorn estimator (Volz and Heckathorn, 2008) is defined as:

$$\hat{\mu}_{VH} = \frac{\sum_i \frac{z_i}{d_i}}{\sum_i \frac{1}{d_i}}$$

where z is the outcome and d is the individual's network degree (i.e. the number of recruitable acquaintances).

This estimate is a generalized Horvitz-Thompson estimator assuming that the probability of inclusion into the sample is proportional to nodal degree. Gile (2011b) showed that this is in fact the case if:

1. The probability model for the graph follows the configuration model.
2. Seeds are chosen proportional to degree.
3. $N >> n$

The configuration network model is a simple parsimonious probability model for networks, particularly popular in the physics literature (Molloy and Reed, 1995). It is defined as follows

1. Let $R = \{R_0, ..., R_N\}$ be an arbitrary but fixed degree distribution such that R_i is the number of nodes with degree i.
2. Randomly assign degrees to each node from R
3. Select edge ends completely at random and assign an edge between the two nodes selected.

This estimator is appealing because of its simplicity, ease of computation and because there is an actual underlying network model. It has also faced considerable criticism because of its unrealistic assumptions and high variance (Gile, 2011b, Goel and Salganik, 2010).

4.1.4 Gile's Sequential Sampling Estimator (gile)

Gile (2011b) developed a finite-population correction to the Voltz-Heckathorn estimator, relaxing the $N >> n$ assumption. Like rds-ii the gile estimator models the sampling process with inclusion probabilities proportional to degree, but does so without replacement. This leads to an estimator with no closed form solution, though it is relatively easy to compute. Gile (2011b) showed that this estimator outperforms the above estimators when the sample fraction is large, and converges to rds-ii when the sample fraction is small.

4.2 The RDS Design

Section 4.1 gives an informal definition of the RDS sampling process, which we will now formalize into a probability model. Consider the following idealized RDS process:

1. Choose seeds from the population. The method of choosing may be biased toward any of the nodal covariates. Seeds are all surveyed at the same time.

2. Each subject then randomly recruits $C_i \sim min(d_i', Multinomial(c_{max}, \theta))$ children into the study, where c_{max} is the number of coupons received, d_i' is the number of alters not yet recruited, and θ are the multinomial probabilities.

3. Each child waits an independent and identically distributed time before coming in to be surveyed.

4. Once the required sample size has been met the recruitment process stops.

Consider a network Y to be an n by n matrix whose entries $Y_{i,j}$ indicate whether subject i and j are connected, where n is the size of the population and X be a matrix of nodal covariates. We define a network as the union of the nodal covariates and the graph structure (i.e. $T = \{X, Y\}$), and W be a random variable representing the RDS sampling process. More specifically, W consists of the set of nodes sampled, their children, and the time at which they are observed, with realization w. Further, let W_i be the ith observed (node, children, time) triplet, W_{-i} be the set of nodes observed before the ith node, and $r(i)$ be the set of nodes surveyed after the ith node, but whose recruiter was recruited before the $(i-1)$th node, and $n(i)$, c_i, and s_i be the node, recruiter and time respectively, then we can recursively define the RDS sample probability as

$$p(W_i = (n(i), c_i, s_i)|w_{-i}, p) = p(C_i = c_i|d_i', c_{max}, \theta)b(s_i, s_{i-1}) \prod_{j \in r(i)} 1 - \int_{s_{i-1}}^{s_i} b(z, s_{i-1})dz,$$

where $b(s_i, s_{i-1})$ is the probability, given the node is chosen by its parent node, and that the subject was not observed before s_{i-1}, that the subject is observed at a time s_i. So if we let w_s represent the seeds, k be the number of nodes sampled, and n_s be the number of seeds, then conditional upon the seeds the sampling probability is

$$p(W = w|x, y, w_s, \theta) = \prod_{i=n_s+1}^{k} p(W_i = w_i|w_{-i}, \theta).$$

Considerable simplification is achieved when looking at the likelihood ratio of θ versus θ_0, as the terms involving b drop out

$$\frac{p(W|\theta, w_s)}{p(W|\theta_0, w_s)} = \prod_{i=n_s+1}^{k} \frac{p(C_i = c_i|d_i', c_{max}, \theta)}{p(C_i = c_i|d_i', c_{max}, \theta_0))}. \tag{4.1}$$

4.3 A Joint Model for the Network and RDS Recruitment Process

Given a fixed network, equation 4.1 allows us to calculate the recruitment probability, but the network is only partially observed. In RDS we observe the degree of each subject recruited (d_i), their nodal covariates (x_i) and the ties connecting recruiters to recruits. The approach that we take to solve this missing data problem is to model the underlying network using ERNM, which are a flexible class of network models, and represent the sampling design by the probability model in the previous section. This leads to a missing data process that is missing not at random (MNAR; Rubin 1976).

In Chapter 1 we defined the exponential family model of the network T as

$$P(T = t|\eta) = \frac{1}{c(\eta)} e^{\eta g(t)},$$

where η is a vector of parameters, g is a vector valued function defining a set of sufficient statistics, and c is a normalizing constant. For the analysis of RDS data, we make a slight modification by conditioning upon the nodal covariates of the seeds x_s

$$P(T = t|\eta, X_s = x_s) \quad = \quad \frac{1}{c(\eta, x_s)} e^{\eta g(t)}. \tag{4.2}$$

We can then express the observed data distribution as

$$P(T_{obs} = t_{obs}, W = w|\eta, \theta, w_s, x_s) = \sum_{t \in \Psi_{miss}} p(W = w|x, y, w_s, \theta) \frac{1}{c(\eta, x_s)} e^{\eta g(t)}, \tag{4.3}$$

where T_{obs} is the observed part of T and Ψ_{miss} is the set of all networks with observed values identical to T_{obs}. Following Chapter 2, we can write the likelihood

ratio of (η, θ) versus (η_0, θ_0) as

$$\begin{aligned}\ell(\eta,\theta) - \ell(\eta_0,\theta_0) &= \log(E_{\eta_0,\theta_0}(\frac{p(W=w|T,\theta,w_s)}{p(W=w|T,\theta_0,w_s)}e^{(\eta-\eta_0)\cdot g(T)})|W=w, T_{obs}=t_{obs}) \\ &\quad - \log(E_{\eta_0}(e^{(\eta-\eta_0)\cdot g(T)}|X_s=x_s)). \end{aligned} \tag{4.4}$$

4.3.1 The Maximum Likelihood Algorithm

Using the MCMC algorithms developed in Chapters 1 and 2, we can generate samples from $P(T=t|\eta,\theta,X_s=x_s)$ and $P(T=t|\eta,\theta T_{obs}=t_{obs},W=w)$, motivating the following algorithm to find approximate maximum likelihood estimates for η and θ.

1. Let $k=1$ and choose initial parameter values $\eta^{(1)}$, θ_0.
2. Use MCMC to generate k samples, $t_{miss}^{(i)}$ from $P(T=t|\eta^{(k)},\theta^{(k-1)}T_{obs}=t_{obs},W=w)$.
3. Find $\theta_k = argmax_\theta(\hat{E}(\frac{p(W=w|T,\theta,w_s)}{p(W=w|T,\theta^{(k-1)},w_s)}))$.
4. Use MCMC to generate k samples, $t_{miss}^{(i)}$ from $P(T=t|\eta^{(k)},\theta^{(k)},T_{obs}=t_{obs},W=w)$.
5. Use MCMC to generate m samples, $t^{(i)}$ from $P(T=t|\eta^{(k)},\theta^{(k)},X_s=x_s)$.
6. Using the samples from step 4 and 5 to estimate the expectations in equation 4.4, find η^{k+1} maximizing the likelihood ratio, subject to $||\eta^{(k+1)}-\eta^{(k)}|| < \epsilon$.
7. Set $k=k+1$ and go to step 2.

4.3.2 A Basic Model for Estimating a Proportion

Up to now we have developed the theory in a very general form, but the choice of g gives us great flexibility in our modeling of the network. The estimators

presented in section 4.1 all focus on the estimation of a proportion, and so it is desirable to create a parsimonious network model that can estimate proportions, correcting for likely sources of bias. The simulations in section 4.4 include the following network statistics for estimation

$$\begin{aligned}
\#\text{ edges} = g_1(t) &= \sum_{i<j} y_{i,j} \\
\#\text{ with trait} = g_3(t) &= \sum_i I(x_i = 1) \\
\text{diff. activity} = g_4(t) &= \sum_i d_i I(x_i = 1) - (\#\text{ with trait}) \cdot (\text{mean degree}) \\
\text{homophily} = g_5(t) &= \sum_{k \in \{0,1\}} \sum_{i:x_i=k} \sqrt{d_{i,k}} - E_{\perp\!\!\!\perp}(\sqrt{d_{i,k}})
\end{aligned}$$

where $d_{i,j} = \sum_k y_{i,k} I(x_k = j)$ be the degree of node i to category j, and $E_{\perp\!\!\!\perp}(\sqrt{d_{i,j}})$ be its expected value assuming no homophily.

One particularly important thing to note about this model is that as the sample fraction becomes small, the probability of any observed subject being tied to an already recruited subject goes to 0. Thus d_i' in equation 4.1 converges to the nodal degree, which is an observed quantity. The result is that the sampling process becomes ignorable as the sampling fraction becomes small, providing us comfort that even if the recruitment process does not play out exactly as outlined in section 4.2, our inferences about the network will remain valid.

4.3.2.1 Estimating the Proportion from the Model

Just as we can estimate the expectations in equation 4.4 using MCMC, give k samples (t_i^{sim}) from $P(T = t|\hat{\eta}, \hat{\theta} T_{obs} = t_{obs}, W = w)$ we can calculate the estimated proportion as

$$\hat{p} = \frac{\frac{1}{k}\sum_i^k g_3(t_i^{sim})}{n}.$$

4.3.3 Logistic Regression

Multivariable logistic regression fits into this framework quite naturally with the appropriate choice of g. Suppose that Z is a nodal variate of particular interest as an outcome variable. Then if we choose our statistics g such that we can write equation 4.2 as

$$P(Z = z, X = x, Y = y|\eta, \beta, \lambda, x_s, z_s) = \frac{1}{c(\beta, \eta, \lambda, x_s, z_s)} e^{z \cdot x\beta + \eta \cdot g_1(y,x) + \lambda \cdot g_2(y,z)}. \tag{4.5}$$

Then the distribution of z_i conditional upon all other variables is

$$P(z_i = 1|z_{-i}, x_i, Y = y, \beta, \lambda) = \frac{e^{x_i\beta}}{e^{\lambda \cdot [h(y,z^-) - h(y,z^+)]} + e^{x_i\beta}}, \tag{4.6}$$

where z_{-i} represents the set of z not including z_i, z^+ represents the variant of z where $z_i = 1$, z^- is the variant of z where $z_i = 0$, and x_i represents the ith row of X. Suppose all variables remain fixed at their value except for x_i, which changes to x_i^*, then using equation 4.5, we can write the log odds ratio as

$$\text{logodds}(z_i = 1|z_{-i}, x_i, Y = y, \beta, \lambda) - \text{logodds}(z_i = 1|z_{-i}, x_i^*, Y = y, \beta, \lambda) = \beta(x_i - x_i^*).$$

Thus, the coefficients β may be interpreted as a conditional logistic regression model (i.e. conditional upon the rest of the network, a unit change in x_i leads to a β change in the log odds). Though the interpretation of the coefficients is familiar, the usual algorithms for estimating a logistic regression can not be used because the distribution of z_i depends on z_{-i}, and thus the independence assumption does not hold.

4.4 Simulation Study

We evaluated the effectiveness of our estimators using both simulated and real data. We generated simulated RDS samples over these networks using the process in section 4.2. In each case, we set our sample size to 350, the distribution of the times between recruitment and survey administration were drawn from a unit lognormal distribution, and the multinomial probabilities were $\theta = (\frac{1}{16}, \frac{9}{16}, \frac{3}{16}, \frac{1}{16})$. These θ values were chosen so that the recruitment chains would be relatively long in all networks, and that there would only rarely be the need to draw additional seeds because the sampling process terminated prior to the sample size being achieved. 20 seeds were used in the simulations, and two methods of seed choice were performed. The first was a simple random sample from the population, and the second was a simple random sample from only those subjects with the trait under investigation. For each network, simulated and real, 500 RDS samples were drawn, and the model in section 4.3.2 was fit. The model based proportion estimate was calculated and compared to the previously developed estimators in section 4.1.

4.4.1 Simulated Networks

For our simulated networks, we wished to evaluate the effect of differing levels of differential activity and homophily on our proportion estimates. To do so, we generated 1000 node networks from ERNM models with terms for the number of edges, differential activity, and homophily. The number of nodes with the trait was fixed at 20%. Table 4.1 shows the η parameters.

These parameter settings allow us to assess the robustness of our estimators over a variety of possibly biasing population characteristics. Figure 4.1 shows the distribution of our 500 simulations at various levels of differential activity. With random seeds, we see that the unadjusted mean is a very poor estimate of the

	Edges	Diff. activity	Homophily
Equal activity	-5	0	1.5
Higher activity	-5	.5	1.5
Much higher activity	-5	1	1.5
No homophily	-5	0	0
Moderate homophily	-5	0	1.5
High homophily	-5	0	3

Table 4.1: ERNM η values for simulated networks.

true proportion, and that Gile's SS estimator and our new model based estimator have less bias than rds-i and rds-ii at the higher differential activity levels. One explanation for this is that the sample fraction (350 out of 1000) is significant and both the Gile estimator and our model based estimator adjust for the finite population effect.

When the seeds are biased, the model based estimator remains nearly unbiased. The other estimators do not fare so well, the rds-i estimator underestimates the proportion, while the mean and gile estimators over-estimate. The rds-ii estimator underestimates in the "Much higher activity" case, but overestimates in the other cases. Table 4.2 shows the design effects of the estimators, defined as the ratio of the mean squared error of the estimator, over the mean squared error of a simple random sample of equivalent size. We see that the model based estimator maintains a design effect between 1.0 and 2.5.

Figure 4.2 represents the performance of the estimators at different levels of homophily. When the seeds are random, the mean performed well when no homophily was present, but displayed bias when homophily was present. All of the other estimators remained approximately unbiased. With non-random seeds we see the mean, rds-ii and gile over-estimating the proportion, while rds-i underestimates it. The model based estimator is approximately unbiased at the no and moderate homophily levels. It displays a bit of upward bias at the high homophily level.

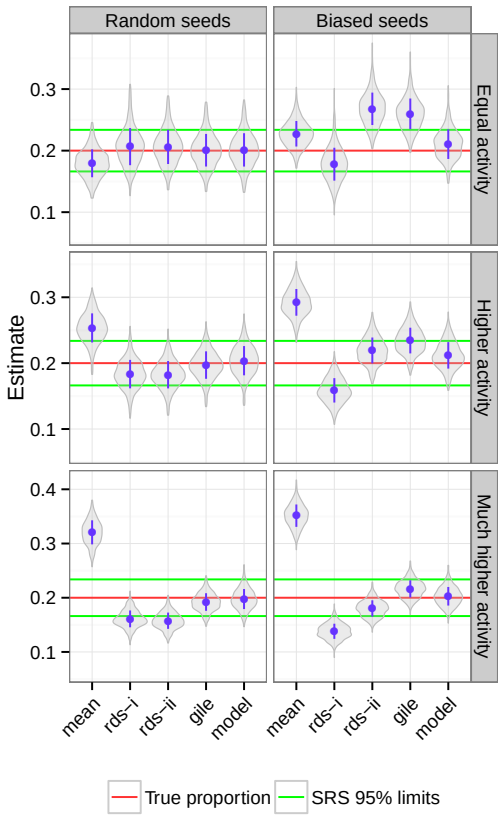


Figure 4.1: Effect of activity on estimator accuracy.

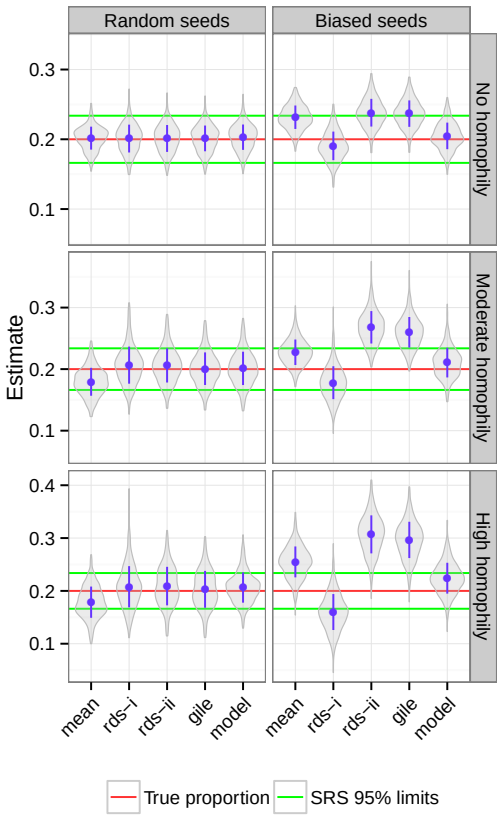


Figure 4.2: Effect of homophily on estimator accuracy.

Seeds	Estimator	Diff. activity Much higher	Higher	Equal
Random	mean	50.6	11.3	3.1
	rds-i	5.9	2.5	3.2
	rds-ii	6.8	2.5	2.7
	gile	1.1	1.5	2.4
	model	1.2	1.7	2.5
Biased	mean	78.5	30.0	4.0
	rds-i	13.5	6.9	4.0
	rds-ii	2.1	2.5	17.9
	gile	1.7	5.2	14.1
	model	1.0	1.8	2.3

Table 4.2: Differential activity networks: Design effects of the various estimators compared to a simple random sample of the same size.

Table 4.3 displays the design effects in the homophily networks. With random seeds, higher levels of homophily lead to larger design effects for all estimators. We also see that the model based estimator has much lower design effect in the high homophily network than the other even though they are all nearly unbiased. With biased seeds, the model based estimator is the only one that maintains even remotely reasonable design effects. This is likely due to the homophily magnifying the effect of the biased seeds, because with homophily, not only the first wave is biased, but also subsequent waves due to their correlation with their recruiter.

Seeds	Estimator	Homophily High	Moderate	None
Random	mean	4.5	3.1	0.9
	rds-i	5.3	3.2	1.3
	rds-ii	4.7	2.7	1.2
	gile	4.1	2.4	1.1
	model	2.9	2.5	1.1
Biased	mean	13.0	4.0	4.3
	rds-i	9.3	4.0	1.7
	rds-ii	42.8	17.9	6.2
	gile	35.3	14.1	5.8
	model	4.8	2.3	1.3

Table 4.3: Homophily networks: Design effects of the various estimators compared to a simple random sample of the same size.

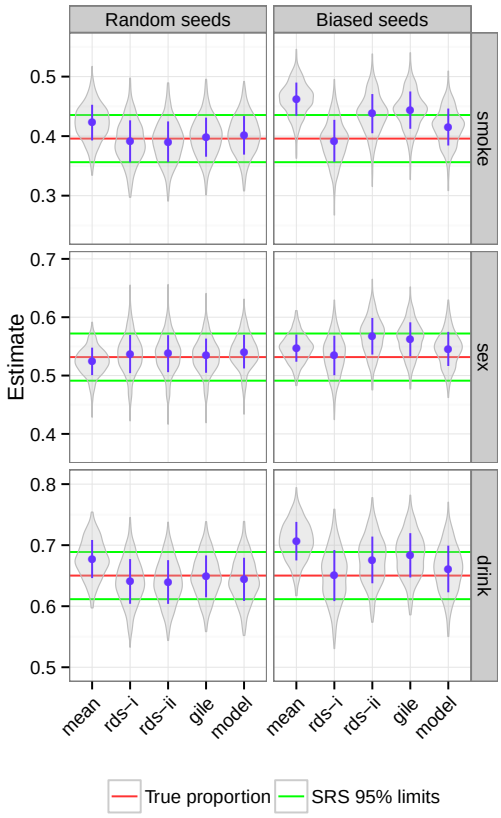


Figure 4.3: Simulated PNS samples from the Add Health network.

4.4.2 Adolescent Health Simulation Study

For the real data, we used the social relationships in a school enrolled in the National Longitudinal Study of Adolescent Health (Add Health) (Harris et al., 2003a). In this network of students in grades 9 to 12, subjects were asked to nominate up to five boys and five girls as friends, forming a relational graph. In our network students were considered connected if either of them nominated the other. In addition to the social graph, students were asked a number of questions about themselves, including their sex, and whether they had used alcohol and/or tobacco. Finally, we restricted our attention to the giant component of the graph, yielding a network of 869 students.

Figure 4.3 displays the simulations in the ADD Health dataset. When the seeds were random, we see that the unadjusted mean does display some bias in the "smoke" and "drink" variables. Rather reassuringly, all of the other estimators are approximately unbiased. When the seeds are biased, the mean, rds-ii and gile estimators all display upward bias, while rds-i and the model based estimator perform roughly equivalently.

		Variable		
Seeds	Estimator	drink	sex	smoke
Random	mean	4.4	1.4	4.0
	rds-i	3.7	2.6	3.1
	rds-ii	3.6	2.4	2.9
	gile	3.0	2.0	2.6
	model	3.3	2.1	2.6
Biased	mean	10.7	1.8	12.6
	rds-i	4.5	2.7	3.1
	rds-ii	5.5	5.3	6.9
	gile	6.2	4.2	8.0
	model	4.1	2.5	3.3

Table 4.4: Add Health Network: Design effects of the various estimators compared to a simple random sample of the same size

Table 4.4 shows design effects between 2.0 and 3.7 for the various RDS aware estimators in the random seed case. These effects are larger than the effects considered by Salganik (2006) who suggests to design RDS studies with a design effect of 2 as a rule of thumb. However, they are significantly smaller than those reported by Goel and Salganik (2010) whose simulations found design effects ranging from 5.7 to 58.3. With biased seeds, rds-i and the model based estimator both display similar design effects that are lower than any of the other estimators. It is unclear why the rds-i performed so well with biased seeds in this network, as the results of 4.4.1 suggest that in many networks rds-i can be highly biased if the seeds are biased.

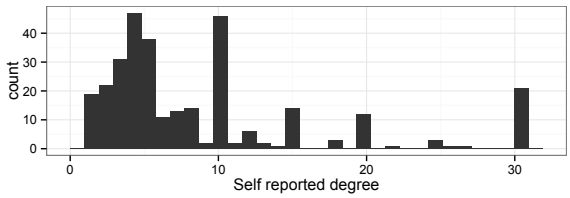


Figure 4.4: Simulated PNS samples from the Add Health network.

4.5 Example: The Dominican Republic

The national strategic HIV surveillance plan for the Dominican Republic surveyed drug users in four of its major cities (Santo Domingo, Santiago, Barahona and Higuey). In this section we focus on the drug user survey in Santo Domingo, where 310 users were surveyed and tested. Of these, 286 tested negative for HIV, 23 tested positive, and 1 subject's HIV status was unavailable.

Individuals degrees were measured with a series of survey questions designed to elicit as accurate a measure of degree as possible (Gile et al., 2012). The median degree reported was 5 with a maximum of 500. Given the questionable reliability of the very large self-reported degrees, we decided to top code degree at 30, which affected 15 subjects. Figure 4.4 shows the full self reported degree distribution.

Despite the care that was taken in eliciting degrees, 41 subjects reported degrees inconsistent with their recruiting activities in that they recruited more subjects into the study than would be possible with their stated degree. The questionable accuracy of subject reported network data has been well established (Bernard et al., 1984), and Gile et al. (2012) investigated the effect of self-reported degree accuracy on the the various RDS estimators in the Dominican Republic data, finding that differences were small in an absolute sense but could introduce large relative changes in low prevalence variables.

4.5.1 Results

The primary variable of interest in this dataset is HIV status. Which displayed a homophily of 1.26 by the definition of Handcock et al. (2012). The population size was unknown, so it was estimated from the data using a hierarchical bayesian model yielding an estimated population size of 2468 (Handcock et al., 2012). We fit the basic model described in section 4.3.2 to estimate both the rate of HIV infection and the proportion of drug users with a history of imprisonment.

The model estimated proportions were then compared to the estimators described in section 4.1. The model based estimator's standard error was obtained using 200 parametric bootstraps. The standard error for the mean was calculated assuming a simple random sample, while rds-i/rds-ii used the methodology of Salganik (2006). The standard error for Gile's sequential sampling estimator was calculated using the methods outlined in Gile (2011b).

Method	Estimate	Standard error	Nominal design effect
mean	0.074	0.015	1.0
rds-i	0.059	0.016	1.4
rds-ii	0.058	0.016	1.5
gile	0.059	0.015	1.3
model	0.079	0.025	2.7

Table 4.5: HIV rate estimates

Table 4.5 shows the estimates and their standard errors for the prevalence of HIV infection. We see that rds-i, rds-ii and gile all have similar estimates which are lower than the mean estimate, while the model based estimator has an estimate slightly higher than the mean. A higher estimate for the model based estimator is expected, as it is the only estimator that corrects for seed bias, and all of the seeds are HIV negative. We can also see that the standard errors of all but the model based estimator are very similar to the standard error of the mean, indicating little or no design effect. Given that we do see homophily in this data, and all of the seeds come from HIV negative persons, this seems unlikely. The

model based estimator standard error by contrast shows a large design effect. The model based estimated design effect go 2.7 is more in line with the effects that we saw in the Add Health simulations, perhaps indicating that the other standard error are understating the amount of extra variation caused by the RDS design.

Method	Estimate	Standard error	Nominal design effect
mean	0.52	0.028	1.0
rds-i	0.46	0.040	2.0
rds-ii	0.46	0.042	2.2
gile	0.46	0.037	1.7
model	0.52	0.043	2.3

Table 4.6: Imprisonment rate estimates

The estimates for imprisonment are displayed in table 4.6. Again we see a model based standard error and design effect (2.3) which is in line with the effects we saw in the Add health simulations.

Finally we fit a combined model with terms for both HIV status, and imprisonment in the model, as well as a logistic regression term modeling the relationship between imprisonment and HIV status (see: section **??**). Table 4.7 displays the maximum likelihood estimates for the combined model parameters. The logistic regression term may be interpreted as a standard conditional log-odds ratio of HIV status versus no imprisonment. Thus according to the model, imprisoned drug users have a $exp(0.97) = 2.63$ times higher odds of being HIV infected.

Term	$\hat{\eta}$	s.e.
Edges	-5.69	0.02
HIV Homophily	0.53	0.35
HIV- Activity	-0.03	0.08
HIV-	2.25	0.36
Imprisonment Homophily	0.23	0.16
Imprisonment Activity	0.08	0.04
Imprisonment	-0.02	0.15
Logistic	0.97	0.53

Table 4.7: ERNM model for the relationship between HIV and Imprisonment

	Estimate	Standard error	p-value
Naive GLM	1.02	0.49	0.037
Full Model	0.97	0.53	0.070

Table 4.8: Logistic Regression

We can compare this estimate to a simple logistic regression ignoring the sampling design and network structure, leading to an estimate of 1.02, which is nearly identical to the ERNM model estimate. The ERNM model does have a higher standard deviation, indicating uncertainty introduced by the design and network structure (see: table 4.8), which is enough to change a significant result to and insignificant one (based on a significance cut off of 0.05).

4.6 Discussion

Respondent driven sampling is an important part of the public health disease monitoring infrastructure. The current estimators such as RDS-II and Gile's SS, have an unrealistic implicit network model underlying them. In this paper we have shown that by increasing the sophistication of the underlying network model, we can account for factors such as seed bias, which can seriously bias all other estimators.

The model based estimator accounts for the structure of both the design and the underlying recruitment network. We found that the resulting maximum likelihood problem was well formed, and can reliably be solved using MCMC methods. Our estimator shows decreased bias, especially when the recruitment seeds are biased. Furthermore, in our simulation studies we found that the variance of the estimator is never much larger than the variance of the other estimators.

CHAPTER 5

A New Link Tracing Design for Hard-to-Reach Populations

In the last chapter we saw that ERNM models can successfully be used to analyze RDS data, but is RDS the best (and most rigorous) design to use? Given the strict assumptions that RDS requires regarding the progression of coupon passing and accuracy of self-reported degrees, it is useful to consider whether our ERNM formulation motivates any more rigorous survey design.

In this chapter we will outline a new sampling design, which we call privatized network sampling (PNS). PNS addresses two of the major concerns with regard to RDS data, namely the assumption that coupons are passed at random among alters, and that subjects can accurately report the number of alters that they have. We also will note that, as PNS is closely related to RDS, the standard RDS estimators may be used on data collected with the PNS design.

5.1 The Privatized Network Sampling Design (PNS)

RDS data contains very minimal information about the recruitment graph. In it we observe the edge between recruiter and recruit, and the degree of each observed subject, but nothing else. The sampling process for PNS proceeds almost identically, but collects more data about the graph. Each subject that is asked not just the number of alters that they have, but for identifiers for each of their alters. Thus the sampling process is an example of biased seed link tracing introduced in

Chapter 2. Specifically the sampling process follows these steps:

1. Recruit a group of subjects from the population (know as seeds). These subjects are chosen by convenience, and may represent a very biased group.

2. Each subject is asked for an identifier for each alter that they might be willing to recruit into the study. The researcher then selects a fixed number of these identifiers at random, and the subject is asked to recruit them.

3. As subjects are recruited, they are in turn asked to recruit their alters until the desired sample size is reached.

PNS differs from RDS in two fundamental ways. First, we observe all edges connected to the subject, giving us a true measure of the subject's degree with much less opportunity for recall bias. The edge information also provides much more information about the structure of the network, which we can then model using ERNM. Secondly, by having the researcher randomly pick which alters will be approached by the subject for recruitment, the opportunity for the research subject to bias the results with their recruitment choices is greatly reduced.

Any attempt to add rigor to an experimental design will undoubtably come with additional implementation difficulties, and this is especially true when the populations under study are stigmatized or otherwise difficult to sample. Some populations may not respond well to the researcher randomly selecting their alters, even with appropriate motivating compensation for their recruitment efforts. In these populations, in the interest of practicality, this part of the design may be dropped by allowing subjects to recruit at will from among their laters. This modification would, of course, admit the possibility of recruitment bias. Heckathorn (1997) noted that in stigmatized populations, collecting the identities of alters is not a practical strategy in some populations (such as heroin users in the United States) as it would violate their culture of "not snitching." Thus it is vital to collect identifiers that respect the population's expectation of privacy.

5.1.1 Choosing an Identifier that Preserves Privacy

The PNS design requires that subjects provide an identifier for each of their possible recruits. One possible identifier would be the name, address and phone number of the alter. This would constitute a unique and verifiable quantity, but would not be particularly good to use because it could violate that individual's privacy. To alleviate this concern we borrow an idea from cryptography called a hash function (Paar and Pelzi, 2009) to create subject identifiers which are unique and verifiable, but would be difficult to use to single out a member of the population.

A hash function is a process that transforms an identifier that we wish to keep private (i.e. a name or phone number) into a unique public identifier called a hash. The interesting property of these hashed identifiers is that it is very easy to test whether an identifier matches a hash (i.e. that the identifier was used to create the hash), but it is nearly impossible to reconstruct the private information from the hash. Hashes are used throughout the computer world to store important information like passwords and ensure that files have not been tampered with.

In out study design, the researcher would only be required to keep the hash of the private identifiable information, not the actual identifiable information, thus preserving the privacy of the alters of the recruited subjects. Of course there are many possible hash functions that could be used ranging from high tech to very low tech. The following examples will illustrate several possible methods, using the name and phone number as the private identifier.

The SHA-1 hash function: The SHA-1 hash function (Gallagher et al., 2008) is a secure cryptographic hash function representing a gold standard method to protect alter privacy.

Private identifier: ian fellows 6195556565

Hashed identifier: d143e809684af303550fcf47bbac103e1274586f

Unordered phone number: A lower tech solution is to reorder the phone number from smallest to largest digit.

Private identifier: 6195556565

Hashed identifier: 1555556669

Unordered phone number: Another lower tech solution is to just keep the first letter of the first name and the last four digits of the phone number.

Private identifier: Ian Fellows 6195556565

Hashed identifier: I6565

In each of these cases, given the private identifier, it is easy to calculate the hashed identifier, but it is (nearly) impossible to recover the sensitive private identifier given the hashed identifier.

5.2 Analysis of PNS Data

One of the benefits of using the PNS design is that existing estimators developed for RDS data (including the ERNM model in Chapter 4) may be applied to PNS data as well, but with the added benefit of ensured random recruitment, and accurate degree values. Additionally, we may leverage the edge information collected in the PNS design to create a more accurate model of the underlying network.

Like the positive contact tracing example in Chapter 2, PNS is a biased seed link tracing design, so the model in Section 2.4.4 is directly applicable. A more elegant approach, however, is to follow the lead of Chapter 4 and notice that, conditional upon the seed covariates, PNS is a missing at random (MAR) process, and therefore ignorable. More specifically, we consider the conditional ERNM

model

$$P(T = t|\eta, X_s = x_s) \quad = \quad \frac{1}{c(\eta, x_s)} e^{\eta \cdot g(t)}, \tag{5.1}$$

where x_s are the observed nodal covariate values for the seeds. The methods of Chapter 2 may then be applied to perform inference on the model.

5.3 Discussion

In this chapter we have introduced a novel new design for sampling hard to reach populations. The practicality of the design will come down to in-the-field performance. The fact that it is a modest modification of the RDS design, which has been wildly successful as a method to recruit individuals in difficult to reach populations, provides a basis to believe that PNS will be a workable design.

CHAPTER 6

Conclusion

Nodal variates are a fundamental part of the structure of social networks. Prior to this work, they were either treated as fixed quantities (as in ERGM), or as random variables over a fixed graph structure (Gibbs fields). We have found that by jointly modeling the nodal covariates along with the graph structure, we have created a framework that can fit a rich set of useful models that meet the real world needs of researchers.

In Chapter 1, we saw how ERNM can be exploited to model the relationship between the graph structure and a categorical covariate using a new definition of homophily, allowing us to generate reasonable simulated networks and perform inference. Through the appropriate choice of statistics included in the model, the natural parameters of an ERNM were shown to be interpretable as conditional logistic regression parameters.

Chapter 2 dealt with missing data, and provided two major contributions to the literature. First, a new type of latent class model was presented, and was shown to have good agreement with previous classifications in real data. The new model has the benefit of being able to assign probabilities of class membership to each node, allowing for an accounting of the level of certainty and clustering in the data, and can automatically select the number of classes in a natural way. Secondly, ERNMs with missing data may be formulated to allow for the analysis of a whole class of sampling designs where the social relations are not of particular interest, but rather are exploited as a method of obtaining a sample.

In Chapter 3, we explored some of the practical computational issues relevant to the implementation MCMC-MLE in exponential family distributions. We showed that the "log-normal" approximation to the likelihood ratio used in the ergm R package (Handcock et al., 2012) is in general valid even when the distribution of the model statistics is non-normal. Secondly, a new procedure for detecting when a new MCMC sample must be drawn was presented.

Chapter 4 tackles a difficult (though popular) link tracing design (RDS). Utilizing the methods in Chapter 2, we explore the performance of ERNM models compared to previous approaches in both simulated and real data. ERNM was shown to outperform previous methods, especially when the initial recruits came from a biased sample. It was noted that the logistic regression modeling term in Chapter 1 can be applied to perform multivariable inference in RDS data, though this model was not fit in the simulation studies.

Finally in Chapter 5 we described a new sampling design, based on RDS. This new design, which is dubbed privatized network sampling (PNS), is similar enough to RDS that existing estimators may be used, but improves the rigor of the design by having the researcher perform a randomization determining the recruitment procedure, and recording the identifiers of all subject alters (in a way that preserves subject privacy).

The ERNM framework presented here represents not an inference problem that is now solved, but rather the foundation for a rich set of theoretical and practical advances. Many of the examples here are of vital importance to researchers, and some represent types of designs previously impossible to analyze (for example the positive contact tracing example in Chapter 2). More work will need to be done to deal with the data quality and structure issues that are sure to arise if our ERNM are widely applied to these designs. On the theoretical front, there are a number of avenues of research that look promising. In particular, fitting the model with maximum likelihood is not the only approach possible. Pseudo-likelihood

solutions have shown themselves useful in ERGM models (at the very least as starting values for maximum likelihood), and this, along with quasi-likelihood or Bayesian fits could be explored in relation to ERNM.

Bibliography

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

Baxter, R. (1982). *Exactly Solved Models In Statistical Mechanics*. Academic Press Inc., San Diago, CA, USA.

Bearman, P. S., Moody, J., and Stovel, K. (2004). Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, 110:44–91.

Bernard, H. R., Killworth, P., Kronenfeld, D., and Sailer, L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology*, 13:pp. 495–517.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society. Series B*, 36:192–236.

Breiger, R. L., Boorman, S. A., and Arabie, P. (1975). An algorithm for clustering relational data, with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology*, 12:328–383.

Donnelly, C. A., Ghani, A. C., Leung, G. M., et al, and Anderson, R. M. (2003). Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. *Lancet*, 361(9371):1761–1766.

Dyson, F. J. (1969). Existence of a phase-transition in a one-dimensional ising ferromagnet. *Communications in Mathematical Physics*, 12:91–107. 10.1007/BF01645907.

Erdos, P. and Renyi, A. (1959). On random graphs. *Publicationes Mathematicae*, 6:290–297.

Fellows, I. and Handcock, M. S. (2012). Exponential-family Random Network Models. *ArXiv e-prints*.

Fenner, F., Henderson, D. A., Arita, I., Jezek, Z., and Ladnyi, I. (1988). Smallpox and its eradication. Technical report, Geneva: World Health Organization.

Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (2001). Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature*, 413(6855):542–548.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842.

Gallagher, P., Foreword, D. D., and Director, C. F. (2008). Fips pub 180-3 federal information processing standards publication digital signature standard (dss).

Gelfand, A. E. and Carlin, B. P. (1993). Maximum-likelihood estimation for constrained- or missing-data models. *Canadian Journal of Statistics*, 21(3):303–311.

Georgii, H.-O. (1988). *Gibbs measures and phase transitions*. Berlin: De Gruyter.

Geyer, C. and Thompson, E. (1992). Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B*, 54(3):657–699.

Geyer, C. J. (1994). On the convergence of monte carlo maximum likelihood calculations. *Journal of the Royal Statistical Society, Series B*, 56:261–274.

Gile, K., Johnston, L., and Salganik, M. (2012). Diagnostics for respondent-driven sampling. *Working paper*.

Gile, K. J. (2011a). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association*, 106(493):135–146.

Gile, K. J. (2011b). Improved inference for respondent-driven sampling data with application to hiv prevalence estimation. *Journal of the American Statistical Association*, 106(493):135–146.

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

Gile, K. J. and Handcock, M. S. (2011). Network model-assisted inference from respondent-driven sampling data. *ArXiv Preprint*.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors (1996). *Markov Chain Monte Carlo in Practice*. New York: Chapman and Hall.

Glynn, P. and Iglehart, D. (1990). Simulation output analysis using standardized time series. *Mathematics of Operations Research*, 15:1–16.

Goel, S. and Salganik, M. J. (2010). Assessing respondent-driven sampling. *Journal of Computational and Graphical Statistics*, 107(15):6743–6747.

Golden, M. R., Hogben, M., Potterat, J. J., and Handsfield, H. H. (2004). HIV partner notification in the United States: a national survey of program coverage and outcomes. *Sex Transm Dis*, 31(12):709–712.

Goodreau, S. M., Kitts, J., and Morris, M. (2009). Birds of a feather, or friend of a friend? using statistical network analysis to investigate adolescent social networks. *Demography*, 46:103–125.

Handcock, M. S. (2003). Statistical models for social networks: Inference and degeneracy. In Breiger, R., Carley, K., and Pattison, P., editors, *Dynamic Social Network Modeling and Analysis*, volume 126, pages 302–322. Committee on Human Factors, Board on Behavioral, Cognitive, and Sensory Sciences, National Academy Press, Washington, DC.

Handcock, M. S. and Gile, K. J. (2007). Modeling social networks with sampled or missing data. Working paper #75, Center for Statistics and the Social Sciences, University of Washington.

Handcock, M. S. and Gile, K. J. (2010). Modeling networks from sampled data. *Annals of Applied Statistics*, 272(2):383–426.

Handcock, M. S., Gile, K. J., and Mar, C. M. (2012). Estimating Hidden Population Size using Respondent-Driven Sampling Data. *ArXiv e-prints*.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., and Morris, M. (2012). *ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks*. Seattle, WA. Version 3.0-1. Project home page at urlstatnet.org.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2006). Model-based clustering for social networks. *Journal of the Royal Statistical Society Series A*, 170:1–22.

Harris, K. M., Florey, F., Bearman, P. S., Jones, J., and Udry, R. J. (2003a). The national longitudinal of adolescent health: Research design.

Harris, K. M., Florey, F., Tabor, J., Bearman, P. S., Jones, J., and Udry, J. R. (2003b). The national longitudinal of adolescent health: Research design [WWW document]. Technical report, Carolina Population Center, University of North Carolina at Chapel Hill, Available at: http://www.cpc.unc.edu/projects/addhealth/design.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2):pp. 174–199.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches

to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Hummel, R., Hunter, D., and Handcock, M. (2012). Improving simulation-based algorithms for fitting ergms. *Journal of Computational and Graphical Statistics*.

Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008). Goodness of fit for social network models. *Journal of the American Statistical Association*, 103:248–258.

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

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift fur Physik*, 31:253–258.

Johnston, L., Malekinejad, M., Kendall, C., Iuppa, I., and Rutherford, G. (2008). Implementation challenges to using respondent-driven sampling methodology for hiv biological and behavioral surveillance: Field experiences in international settings. *AIDS and Behavior*, 12:131–141. 10.1007/s10461-008-9413-1.

Johnston, L. and Sabin, K. (2010). Sampling hard-to-reach populations with respondent driven sampling. *Methodological Innovations Online*, 5:38–48.

Jones, G., Haran, M., Caffo, B., and Neath, R. (2006). Fixed-width output analysis for markov chain monte carlo. *ournal of the American Statistical Association*, 101:1537–1547.

Klinkenberg, D., Fraser, C., and Heesterbeek, H. (2006). The effectiveness of contact tracing in emerging epidemics. *PLoS ONE*, 1(1):e12.

Krivitsky, P. N. (2011). Exponential-Family Random Graph Models for Valued Networks. *ArXiv e-prints*.

Leenders, R. (1997). Longitudinal behavior of network structure and actor attributes: modeling interdependence of contagion and selection. In Doreian, P. and Stokman, F., editors, *Evolution of Social Networks*, pages 165–184. Gordon and Breach, Amsterdam.

Malekinejad, M., Johnston, L., Kendall, C., Kerr, L., Rifkin, M., and Rutherford, G. (2008). Using respondent-driven sampling methodology for hiv biological and behavioral surveillance in international settings: A systematic review. *AIDS and Behavior*, 12:105–130. 10.1007/s10461-008-9421-1.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC.

Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180.

Morris, M., Handcock, M. S., and Hunter, D. R. (2008). Specification of exponential-family random graph models: Terms and computational aspects. *Journal of Statistical Software*, 24(4).

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087.

Paar, C. and Pelzi, J. (2009). *Understanding Cryptography, A Textbook for Students and Practitioners*. Springer, Berlin, Heidelberg.

Potterat, J. J., Spencer, N. E., Woodhouse, D. E., and Muth, J. B. (1989). Partner notification in the control of human immunodeficiency virus infection. *American Journal of Public Health*, 79(7):874–876.

R Development Core Team (2012). *R: A Language and Environment for Statistical*

Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Resnick, M. D., Bearman, P. S., Blum, R. W., Bauman, K. E., Harris, K. M., Jones, J., Tabor, J., Beuhring, T., Sieving, R. E., Shew, M., Ireland, M., Bearinger, L. H., and Udry, J. R. (1997). Protecting adolescents from harm. Findings from the National Longitudinal Study on Adolescent Health. *Journal of the American Medical Association*, 278(10):823–832.

Robins, G., Elliott, P., and Pattison, P. (2001a). Network models for social selection processes. *Social Networks*, 23(1):1–30.

Robins, G., Pattison, P., and Elliott, P. (2001b). Network models for social influence processes. *Psychometrica*, 66(2):161–190.

Rubin, D. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Salganik, M. (2006). Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Journal of Urban Health*, 83:98–112. 10.1007/s11524-006-9106-x.

Salganik, M. J. and Heckathorn, D. D. (2008). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34:193–240.

Sampson, S. F. (1969). *Crisis in a Cloister*. PhD in Sociology, Cornell University.

Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496):1361–1370.

Steglich, C., Snijders, T. A. B., and Pearson, M. (2010). Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology*, 40(1):329–393.

Strauss, D. (1986). On a general class of models for interaction. *SIAM Review*, 28:513–527.

Thompson, S. K. (2002). *Sampling*. Wiley, Second edition.

Udry, J. R. (2003). The national longitudinal of adolescent health: (add health), waves I and II, 1994-1996; wave III, 2001-2002 [machine-readable data file and documentation]. Technical report, Carolina Population Center, University of North Carolina at Chapel Hill.

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

Volz, E. and Heckathorn, D. D. (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 24(1):79–97.

Zhu, S.-C. and Liu, X. (2002). Learning in gibbsian fields: How accurate and how fast can it be? *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(7):1001–1006.