Modeling of Networked Populations with Exponential-Family Random Network Models when data is Sampled or Missing

Mark S. Handcock Ian E. Fellows

Department of Statistics
University of California - Los Angeles
`https://faculty.stat.ucla.edu/handcock`

UCLA

Some joint work with

Andrea Wang Krista J. Gile

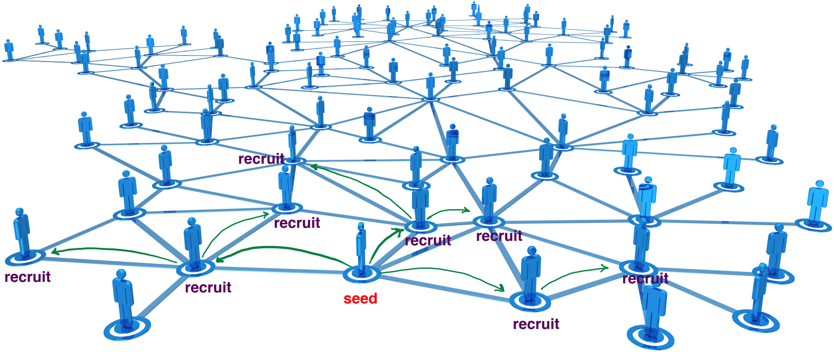
Supported by NIH Grant HD041877, NSF awards CCF-2200197, MMS-0851555, SES-1357619, IIS-1546259.

Sunbelt 2023, June 29

Networked Population

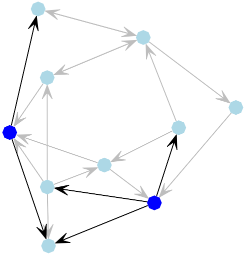
UCLA

Contact other people via the seed’s social network



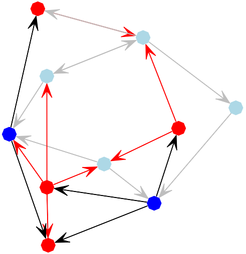
Mechanisms for Partial Observation of Social Networks

- *Sampling Design:* Known mechanism
  - *Egocentric*



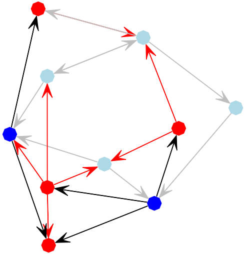
Partial Observation of Social Networks

- *Sampling Design:* Known mechanism
  - Egocentric
  - *Adaptive*



Partial Observation of Social Networks

- *Sampling Design:* Known mechanism
  - Egocentric
  - *Adaptive*
- Out-of-design Missing Data: Unknown mechanism



Adaptive Network Sampling

Strengths:

- Exploits information in the network of relationships
- Network structure used to improve the design
- Increases the range of possible designs
- Adjusts for discovered features in the population
- Leads to increased efficiency of sampling

Issues:

- Seed Dependence: final sample depends on sampling mechanism of seeds
- Privacy: some populations prefer to stay "hidden"
- Link-tracing can be challenging: confidentiality, logistics
- Estimation: The sample and sampling probabilities depend on the unknown network

Statistical Models for Social Networks

Consider a networked population with a set of n social “actors", social relationship between each pair of actors, and a set of variables on those actors/pairs.

- a set of n social “actors"
- a social relation Y_{ij} between pairs of actors.
- call $Y \equiv [Y_{ij}]_{n \times n}$ a *graph*
- X be $n \times q$ matrix of actor and dyadic covariates
- call (Y, X) a *network*
- The basic problem of stochastic modeling is to specify a distribution for X, Y i.e.,

$$P(Y = y, X = x)$$

The ERGM Framework for Network Modeling

Let $\mathcal{Y}$ be the sample space of Y e.g. $\{0,1\}^N$
and $\mathcal{X}$ be the sample space of X.
Model the multivariate distribution of Y given X via:

$$P_\eta(Y = y|X = x) = \frac{\exp\{\eta{\cdot}g(y|x)\}}{c(\eta, x, \mathcal{Y})} \qquad y \in \mathcal{Y}, \quad x \in \mathcal{X}$$

Frank and Strauss (1986)

- $\eta \in \Lambda \subset R^d$ d-vector of parameters
- $g(y|x)$ d-vector of *graph statistics*.
  $\Rightarrow$ $g(Y|x)$ are jointly sufficient for the model
- $c(\eta, x, \mathcal{Y})$ distribution normalizing constant

$$c(\eta, x, \mathcal{Y}) = \sum_{y \in \mathcal{Y}} \exp\{\eta{\cdot}g(y|x)\}$$

Extensive development of conditional models

- Classes of $g(y|x)$ (Generative Theory, Structural signatures)
- Inference on the log-likelihood function,

$$\ell(\eta|y_{\rm obs}; x_{\rm obs}) = \eta{\cdot}g(y_{\rm obs}|x_{\rm obs}) - \log c(\eta|x_{\rm obs})$$

$$c(\eta|x_{\rm obs}) = \sum_{z \text{ in } \mathcal{Y}} \exp\{\eta{\cdot}g(z|x_{\rm obs})\}$$

- For computational reasons, approximate the likelihood via Markov Chain Monte Carlo (MCMC)

Exponential-family Random Network Models

Joint modeling of Y and X Fellows and Handcock (2012)
Let $\mathcal{N}$ be the sample space of Y, X

Model the multivariate distribution of Y, X
via the form:

$$P_\eta(Y = y, X = x) = \frac{\exp\{\eta \cdot g(y, x)\}}{c(\eta, \mathcal{N})} \qquad y, \;\; x \in \mathcal{N}$$

- $\eta \in \Lambda \subset R^q$ q-vector of parameters
- $g(y, x)$ q-vector of network statistics.
  $\Rightarrow$ $g(Y, X)$ are jointly sufficient for the model
- $c(\eta, \mathcal{N})$ distribution normalizing constant

$$c(\eta, \mathcal{N}) = \int_{y,\; x \in \mathcal{N}} \exp\{\eta \cdot g(y, x)\} \cdot dP_0(y, x)$$

Interesting model-classes of ERNM

Relationship to ERGM and Gibbs Random Fields

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

$$\begin{array}{rrcll} \text{ERGM} & P(Y = y|X = x; \eta) & = & \frac{1}{c(\eta; x)} e^{\eta \cdot h(x,y)} & y \in \mathcal{N}(x) \\ \text{Gibbs measure} & P(X = x|Y = y; \eta) & = & \frac{1}{c(\eta; y)} e^{\eta \cdot h(x,y)} & x \in \mathcal{N}(y) \\ \text{ALAAM} & & & & \end{array}$$

- The first model is the ERGM for the network conditional on the nodal attributes.
- The second model is an exponential-family for the field of nodal attributes conditional on the network.

Example: Joint Ising Models

Suppose X is univariate and binary $x_i \in \{-1, 1\}$. One measure of homophily on x is

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

A simple model for the network is

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

where density(y) = $\frac{1}{n} \sum_i \sum_j y_{i,j}$

$$\begin{array}{llcll} \text{GLM} & P(Y_{i,j} = y_{i,j} | X = x, \eta_1, \eta_2) & \propto & e^{\eta_1 \frac{1}{n} y_{i,j} + \eta_2 x_i y_{i,j} x_j} & y \in \{0, 1\},\ x \in \mathcal{X} \\ \text{Ising model} & P(X = x | Y = y, \eta_2) & \propto & e^{\eta_2 \sum_i \sum_j x_i y_{i,j} x_j} & (y, x) \in \mathcal{N} \end{array}$$

- So we have a simple joint Ising model

Comparing ERGM to ERNM Conceptually

- Social selection process: conventional network analysis with nodal attributes fixed

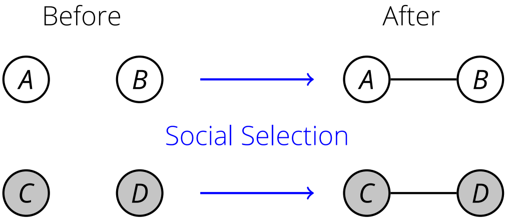


Figure: Illustration of Social Selection: Color of nodes: nodal attributes

Comparing ERGM to ERNM Conceptually

- Social influence process: the network is fixed and the nodal attributes can vary

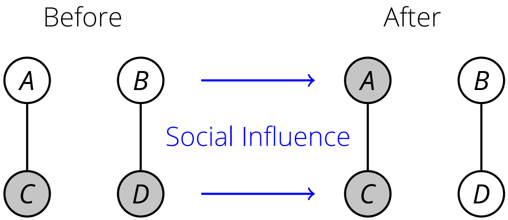


Figure: Illustration of Social Influence: Color of nodes: nodal attributes

Comparing ERGM to ERNM Conceptually

- Social selection process: ERGM, SBM, etc
- Social influence process: Gibbs fields, Ising, ALAAM, etc
- Social selection and influence jointly: ERNM

Fitting Models to Partially Observed Social Network Data

- Focus on the joint distribution of $Z = (Y, X)$.
- Types of data: Observed relations, nodal and dyadic variables ($z_{obs} = (y_{obs}, x_{obs})$), and D, indicators of relations and covariates being observed
- $Z = (Z_{obs}, Z_{unobs})$

$$\begin{aligned}
L(\boldsymbol{\eta}, \psi) &\equiv P(Z_{obs} = z_{obs}, D | \boldsymbol{\eta}, \psi) \\
&= \sum_{z_{unobs}} P(Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs}, D | \boldsymbol{\eta}, \psi) \\
&= \sum_{z_{unobs}} P(D | Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs}, \psi) P_{\boldsymbol{\eta}}(Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs}) \\
&= \sum_{z_{unobs}} P(D | Z, \psi) \times P_{\boldsymbol{\eta}}(Z = z)
\end{aligned}$$

sampling design × network model

- $\boldsymbol{\eta}$ is the network model parameter (“super population")
- ψ is the sampling parameter

Adaptive Sampling Designs

- A sampling design adaptive if:

$$P(D = d|Z_{obs}, Z_{mis}, \psi) = P(D = d|Z_{obs}, \psi) \qquad \forall z \in \mathcal{Z}.$$

that is, it uses information collected during the survey to direct subsequent sampling, but the sampling design depends only on the observed data.

- adaptive sampling designs satisfy a "*missing at random*" condition from Rubin (1976) in the context of missing data.
- **Result:** standard network sampling designs such as conventional, adaptive web, and multi-wave link-tracing sampling designs are adaptive $\Rightarrow$ Thompson and Frank (2000), Handcock and Gile (2006, 2010, 2016).

Adaptive Sampling Designs and their Amenable Models

Definition: Consider a sampling design governed by parameter $\psi \in \boldsymbol{\Psi}$ and a stochastic network model $P_{\boldsymbol{\eta}}(Z = z)$ governed by parameter $\boldsymbol{\eta} \in \Xi$. We call the sampling design amenable to the model if the sampling design is adaptive and the parameters ψ and $\boldsymbol{\eta}$ are distinct.

Result: If the sampling design is amenable to the model the likelihood for $\boldsymbol{\eta}$ and ψ is

$$L[\boldsymbol{\eta}, \psi | Z_{obs} = z_{obs}, D = d] \propto L[\psi | D = d, Z_{obs} = z_{obs}] L[\boldsymbol{\eta} | Z_{obs} = z_{obs}]$$

sampling design likelihood×face-value likelihood

$$L[\psi | D = d, Z_{obs} = z_{obs}] = P(D | Z_{obs} = z_{obs}, \psi)$$

$$L[\boldsymbol{\eta} | Z_{obs} = z_{obs}] = \sum_{z_{unobs}} P_{\boldsymbol{\eta}}(Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs})$$

Adaptive Sampling Designs and their Amenable Models

Result: If the sampling design is ***not*** amenable to the model the likelihood for $\boldsymbol{\eta}$ and ψ is

$$L(\boldsymbol{\eta}, \psi) = \sum_{z_{unobs}} P(D|Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs}, \psi) P_{\boldsymbol{\eta}}(Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs})$$

and the design will need to be represented.

Clearly $P(D|Z, \psi)$ can be modeled when it is unknown.

Approximating the loglikelihood

- Let $\color{blue}{Z = (Y, X)}$, $Z_{obs} = (Y_{\cancel{un}obs}, X_{\cancel{un}obs})$, $Z_{unobs} = (Y_{unobs}, X_{unobs})$
- missing data log-likelihood

$$
\begin{aligned}
\ell(\boldsymbol{\eta}, \psi | Z_{unobs} = z_{unobs}) &= \\
\log\Big[P(D|z_{obs}, z_{unobs}, \psi) P_{\boldsymbol{\eta}}(z_{obs}, z_{unobs})\Big] &- \log\Big[c(z_{unobs}, \boldsymbol{\eta}, \psi)\Big]
\end{aligned}
$$

- $Z_1, Z_2, \ldots, Z_m$ i.i.d. $P_{\eta_0}(Z = z)$ for some η_0 via MCMC.
  $Z_1^c, Z_2^c, \ldots, Z_m^c$ i.i.d. $P_{\eta_0}(Z = z | Z_{obs} = z_{obs})$ via MCMC.
- Using the LOLN, the difference in observed data log-likelihoods is

$$
\begin{aligned}
\ell(\eta, \psi) - \ell(\eta_0, \psi_0) &= \log \frac{c(z_{unobs}, \eta, \psi)}{c(z_{unobs}, \eta_0, \psi_0)} - \log \frac{c(\eta)}{c(\eta_0)} \\
&\approx \text{weighted sample means over } \{Z_k\}_{k=1}^m \text{ and } \{Z_k^c\}_{k=1}^m
\end{aligned}
$$

Ex 1: Ignorably missing nodal attributes and ties

Sampling Sampson's monk's ties and Cloisterville attendance

Y

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	X
1	🟥	⬜	⬜	⬜	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥 NA	🟥	🟥	🟥	🟥	🟥	🟥
2	⬜	🟥	⬜	🟥	🟥	🟥	🟥	🟥	🟥	🟥 NA	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
3	⬜	⬜	🟥	🟥	🟥	🟥	🟥	🟥	🟥	⬜	🟥	🟥 NA	🟥 NA	🟥	🟥 NA	🟥	🟥	🟥	🟥
4	⬜	⬜	⬜	🟥	🟥	🟥	🟥 NA	🟥	🟥 NA	🟥	🟥	🟥	🟥 NA	🟥	🟥 NA	🟥	⬜	⬜	🟥
5	🟥	🟥	🟥	🟥 NA	🟥	⬜	🟥	🟥 NA	🟥	🟥	🟥	🟥 NA	⬜	🟥	🟥	⬜	🟥	🟥	🟥
6	⬜	🟥	🟥	⬜	⬜	🟥	⬜	🟥	⬜	🟥	⬜ NA	🟥 NA	🟥	🟥	🟥 NA	🟥	🟥	🟥	🟥
7	🟥	🟥	🟥	🟥	⬜	🟥	🟥	⬜	⬜	🟥 NA	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
8	🟥	🟥	🟥 NA	🟥 NA	🟥 NA	🟥	⬜	🟥	⬜	⬜	⬜	🟥	⬜	🟥	🟥	🟥	🟥	⬜	🟥
9	⬜	⬜	⬜	🟥	⬜	⬜	⬜	⬜	🟥	⬜	⬜	⬜	🟥 NA	🟥	🟥	🟥	🟥	🟥	🟥
10	⬜	🟥	⬜	🟥	⬜	⬜	⬜ NA	⬜	⬜	🟥	⬜	⬜	🟥	🟥	⬜	🟥	⬜	🟥	🟥
11	🟥	🟥	🟥	🟥 NA	⬜	⬜	⬜	⬜	⬜	⬜ NA	🟥	🟥	🟥	🟥	🟥	⬜ NA	🟥	🟥	🟥 NA
12	🟥	🟥	🟥	🟥	🟥	⬜	🟥 NA	🟥 NA	🟥	⬜ NA	🟥	🟥	⬜	🟥	🟥	⬜ NA	🟥	⬜	⬜
13	🟥	🟥	🟥 NA	🟥	🟥	🟥 NA	🟥	⬜	🟥	🟥	🟥	🟥 NA	🟥	⬜	🟥 NA	🟥	⬜ NA	🟥	🟥
14	🟥	🟥	🟥 NA	🟥	🟥 NA	🟥	🟥 NA	🟥	🟥 NA	🟥	🟥	⬜	⬜	🟥	⬜	🟥	⬜	⬜	⬜
15	🟥	🟥	🟥	🟥	🟥	🟥 NA	🟥	🟥	🟥	🟥	🟥	⬜	🟥	⬜ NA	🟥	🟥	🟥 NA	🟥 NA	⬜ NA
16	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	⬜	🟥	🟥	⬜	🟥	⬜	⬜	⬜
17	🟥	🟥	🟥	⬜	🟥	🟥	⬜	🟥	🟥	⬜	🟥	🟥	⬜	⬜	⬜	⬜ NA	🟥	⬜	⬜ NA
18	🟥	🟥	🟥	🟥	🟥	🟥 NA	🟥	🟥	🟥	🟥	🟥 NA	⬜	🟥 NA	⬜	🟥 NA	🟥	🟥	🟥	⬜

Ex 2: Biased seed link-tracing

In disease modeling, take the disease status as the nodal covariate.

If seeds are chosen as a convenience sample, followed by link-tracing then likelihood inference is amenable.

e.g., seeds picked at random from among the infected individuals, convenience sample of uninfected seeds

Ex 3: Positive contact tracing

Contact tracing that follows all ties from infected nodes only

Clearly sampling is informative and the design is non-amenable.

Still the design can be modeled and likelihood inference based on ERNM is very effective.

Summary

- We present a concise and systematic statistical framework for dealing with partially observed network data mechanisms
  - missing relational ties and nodal covariates
  - adaptive sampling: link-tracing
  - non-amenable sampling designs (e.g., positive contact tracing)
- likelihood-based inference is practical under partial observation
- We develop MCMC-MLE algorithms and show they are computationally feasible
- We give three important special cases:
  - ignorably missing nodal attributes and relational ties
  - link-tracing with a convenience sample of seeds: e.g., seeds picked at random from among the infected individuals, convenience sample of uninfected seeds
  - positive contact tracing: follow all ties from infected nodes only
- Made available open-source, powerful, general, *user-friendly software* to do all of the above.