Some challenges and results for causal and statistical inference with social network data

Elizabeth L. Ogburn

Department of Biostatistics,
Johns Hopkins University

May 10, 2013
• Network data evince complex forms of dependence among observations, making statistical inference difficult.

• Two challenges for causal inference using data sampled from a single social network:
  • nonparametric identification of causal effects with interference,
  • valid statistical inference under network dependence.
• Computer scientists, physicists and mathematicians have been researching networks for decades:
  • topology, diffusion properties, generation,...
  • very little statistics for outcomes on network nodes.

• Statisticians have been researching dependent data for decades:
  • time series
  • spatial data
  • interference
  • surprisingly little of this work is directly applicable to networks.
1. Introduction

2. HopeNet

3. Inference

Outline

- Introduction and background
- Motivating example: HopeNet
  - Design and aims.
  - Briefly describe three distinct types of interference.
- Inference in the presence of network dependence.
  - Explain why results for spatial dependence are not immediately applicable.
  - Point towards some possible solutions to the problem of statistical inference in networks.
Interference requires a new framework and new tools for causal inference.

Existing methods for causal inference in the presence of interference have limitations in the social network context.

- Generally assume that treatment is randomized.
- Generally assume multiple independent groups are observed.
Recent availability of social network data has inspired methods and analyses, but with caveats:

- GEEs and GLMs have been used extensively but are usually inappropriate for network dependence. (Cf. Lyons, 2011; Shalizi & Thomas, 2011; VanderWeele, Ogburn & Tchetgen Tchetgen, 2012; Christakis & Fowler, 2012; Shalizi, 2012)

- Spatial autoregressive (SAR) parameterize an equilibrium state and generally lack causal interpretations.

- Recent work by van der Laan (2012) harnesses independence assumptions that are reasonable if dependence is due to contagion and network evolution is observed over time; other recent work relies on randomization-based inference (Cf. Bowers et al., 2013; Lunagomez & Airoldi, 2014; Rosenblum, 2007)
HopeNet Study

- **Health outcomes, progressive entrepreneurship, and Networks**
- **Nyahabare Parish, Mbarara Province, SW Uganda**
  - 8 villages
  - 2000 adults
HopeNet Study

- Complete social network census for the adult population of the parish.
- Clean water intervention.

- Microenterprise intervention.
effects of interest

What types of effects are we interested in?

- Under what conditions are these effects nonparametrically identifiable?
  - Ogburn & VanderWeele (2012)

- Under what conditions can we perform inference about them?
definitions

- Stationarity: features of the distribution of observations do not depend on location in the network.
- Distance $\|i, j\|$ between nodes $i$ and $j$: length of shortest path connecting $i$ and $j$.
  - Other definitions are possible, e.g. taking into account number of paths or average path length between $i$ and $j$.
- M-dependence: $W_i \perp W_j$ if $\|i, j\| > m$.
- Mixing conditions: $\text{Cov}(W_i, W_j) \to 0$ as $\|i, j\| \to \infty$. 

inference with network dependence

- Sources of network dependence.
- Limitations of methods for spatial dependence.
  - local dependence
  - weakly dependent clusters
  - subsampling
  - k-dependence
sources of network dependence

**Latent variables** cause outcomes among close social contacts to be more correlated than among distant contacts. (E.g. homophily, geography, shared culture.)
sources of network dependence

Contagion implies information barrier structures

- e.g. $Y_1^t \perp Y_2^t \mid Y_1^{t-2}, Y_2^{t-2}, Y_1^{t-1}, \text{ and } Y_2^{t-1}$ and $Y_1^{t-2} \perp Y_3^{t-1}$.

- When a network is observed at a single time point, this will resemble latent variable dependence.
- If network is observed before information has had the opportunity to diffuse, m-dependence would be reasonable.
Why can’t we use spatial dependence results?

Network topology doesn’t naturally correspond to Euclidean space.

- In order to embed a network in $\mathbb{R}^d$, we would have to let $d$ grow with sample size $n$.
  - Spatial results require $d$ to be fixed or to grow slowly with $n$.
- Population growth is usually assumed to occur at the boundaries of the $d$-dimensional space.
  - It’s not clear how to define boundaries in networks (nor how to define population growth).
- Mixing assumptions and $m$-dependence don’t imply bounded correlation structure.
  - In spatial data most observations are distant from one another.
  - The maximum network-based distance between two observations may be very small.
  - The distance distribution may not be right-skewed enough.
inference with network dependence

- Focusing for now on a traditional frequentist inferential framework, point estimates are generally unbiased; the challenge is for s.e. and interval estimation. Estimates of s.e. may be unbiased but not consistent.
inference with network dependence

- For simplicity, suppose our target parameter is a population mean, $\mu$.
- The sample mean, $\bar{W} = \sum_{i=1}^{n} W_i$, is unbiased for $\mu$.
- The problem is consistent estimation of the asymptotic variance $Avar(\bar{W})$.
- Agnostic about sources of dependence (contagion vs. latent variables).
- Population growth must preserve key features of the network and data.
• Local dependence: for each observation $W_i$, there is a set of indices $I$ such that $W_i \perp W_j$ for $j \in I^C$.

• M-dependence is a special case.

• Using Stein’s method, it is easy to show that CLTs hold for locally dependent data, under restrictions on the size of $I$. (Cf. Chen, 1978; Barbour, Karonski & Rucinski, 1989; Rinott & Rotar, 1996; Raic, 2002; Chen & Shao, 2005)

• What types of networks are consistent with the restrictions on $I$?
weakly dependent clusters

If $K$ clusters are asymptotically mean independent from one another, there are two approaches we might consider:

1. $T$-distribution based confidence intervals (Ibragimov & Muller, 2010; Bester, Conley & Hansen, 2011).
   - Requires asymptotic normality and mean stationarity at the cluster level.

2. Bootstrap the weakly dependent communities.
   - Stationarity is required only at the cluster level.

In the spatial dependence literature mean independence is justified with conditions on the relative size of the boundaries and interiors of the clusters; growth in $d$ dimensions uniformly.
   - These conditions don’t translate into the network setting...
Subsampling has been used in many spatial dependence contexts (cf. Lahiri, 2003; Politis, Romano & Wolf, 1999), but neither the implementation nor the conditions under which it is appropriate are immediately applicable to networks.

Under mild stationarity and dependence conditions, we can subsample to estimate $Avar(\bar{W})$:

1. Select $B$ subsamples of “consecutive” observations.
2. In each subsample, calculate the subsample variance estimator $\hat{\sigma}_b^2$.
3. Estimate $Avar(\bar{W})$ with the average of the subsample estimators:
   $$\hat{\sigma}_W^2 = \frac{1}{B} \sum_{b=1}^{B} \hat{\sigma}_b^2.$$
This is reasonable if, as $n_b \to \infty$ and $n \to \infty$,

1. $\hat{\sigma}_b^2$ is asymptotically unbiased for $\text{Avar}(\bar{W})$.
   - This will hold if key features of the network and the mean and variance of $W$ are stationary over groups smaller than the subsample sizes.

2. $\text{Var}(\hat{\sigma}_W^2) \to 0$, where

   $$\text{Var}(\hat{\sigma}_W^2) = \frac{1}{B^2} \sum_{b=1}^{B} \text{Var}(\hat{\sigma}_b^2)$$

   $$+ \frac{1}{B^2} \sum_{\|l_b, l_d\| \leq m} 2\text{Cov}(\hat{\sigma}_b^2, \hat{\sigma}_d^2)$$

   $$+ \frac{1}{B^2} \sum_{\|l_b, l_d\| > m} 2\text{Cov}(\hat{\sigma}_b^2, \hat{\sigma}_d^2)$$
k-dependence

In some settings it may be expedient to estimate $\text{Cov}\left(\mathcal{W}\right)$ directly.

- **K-dependence:** $\text{Cov}(W_i, W_j) = \sigma_k$, where $k = \|i, j\|$.
- Under k-dependence, m-dependence, and mean stationarity, we can get an unbiased and consistent estimate of $\text{Cov}\left(\mathcal{W}\right)$ by this procedure:
  1. For each $k < m$, select pairs of nodes that are $k$ units apart, such that the pairs themselves are at least $m$ units apart from one another.
  2. Estimate $\hat{\sigma}_k$ with the average covariance across the selected pairs.
  3. Estimate $\text{Cov}\left(\mathcal{W}\right)$ with the plug-in estimator.
- This doesn’t demand as much from m-dependence as other procedures do...
other directions

- Different types of asymptotics:
  - combine infill and increasing domain asymptotics,
  - “fractal” asymptotics.

- Identify low-level conditions for CLTs and LLNs in network settings.

- Learn a new, latent distance metric. (E.g. work by Adrian Raftery & colleagues)

- Finite sample results using bounded influence:
  \[ \text{Var}(W_i) \gg \sum_j \text{Cov}(W_i, W_j). \]
Thank you


sources of network dependence

1. **Latent variables** cause outcomes among close social contacts to be more correlated than among distant contacts. (E.g. homophily, geography, shared culture.)
   - This is the type of dependence considered in the context of spatial data.

   - Outcomes are independent conditional on exposure, unless other forms of dependence are also present.
3. **Contagion** implies an information barrier structure:

\[
Y_t^1 \perp Y_t^2 \mid Y_{t-2}^1, Y_{t-2}^2, Y_{t-1}^1, \text{ and } Y_{t-1}^2\] \quad \text{and} \quad \left[ Y_{t-2}^1 \perp Y_{t-1}^3 \right].
\]

- When a network is observed at a single time point, this will resemble latent variable dependence.
- If network is observed before information has had the opportunity to diffuse, m-dependence would be reasonable.