Design and analysis of experiments in the presence of network interference

Edo Airoldi

Department of Statistics
Harvard University

Joint with Simon Lunagomez and Don Rubin

Computational Methods in Social Sciences Opening Workshop, 19 August 2013, SAMSI

Overview

- Structured data vs. latent dependence structure

  Leveraging observed (noisy) structure for estimation

- This talk

  Inference from non-ignorable sampling designs

  Estimation of causal peer-influence effects (interference)

- Applications

  Analytics and marketing on social media platforms

  Online mechanisms that affect behavior online/offline

Agenda

- Inference with non-ignorable sampling designs
  1. Theory
  2. Inferential framework
- Estimation of the causal effects of interference, including peer-influence and peer-pressure
- Concluding remarks

Motivating problems

- Surveys on social media platforms

  Potential market size estimation

- Surveys of hard-to-reach populations

  Cell phone users only (young, third-world countries)

  Epidemiology (drug-injection users, MSM)

  Healthcare (rare diseases, diseases with social stigma)

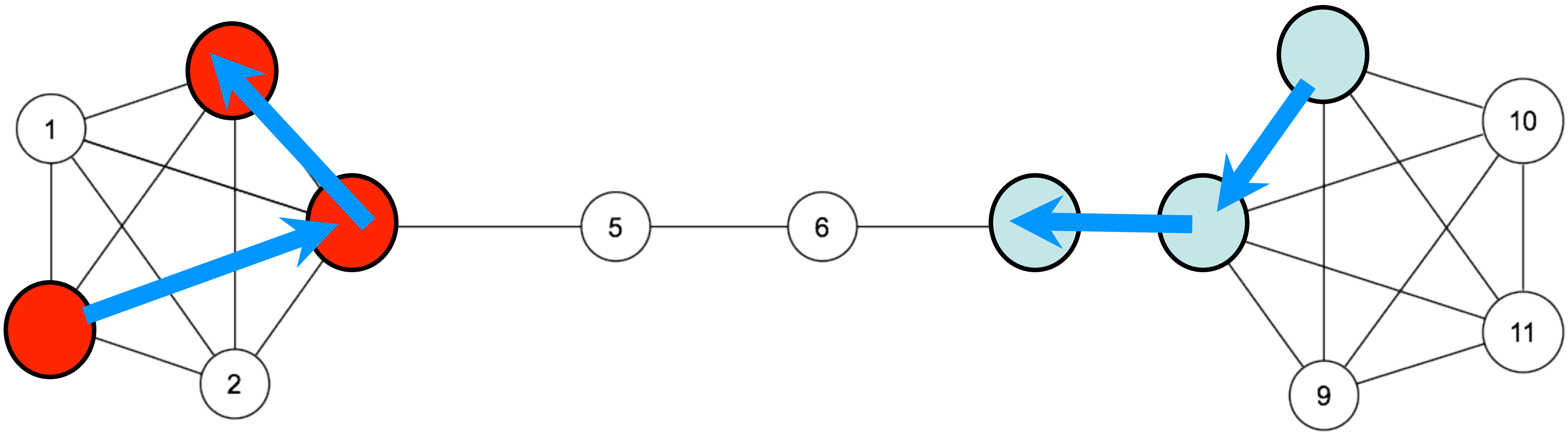
Network sampling designs

- Consider the problem of sampling from hard-to-reach populations or on social media platforms

<u>Idea: leverage social structure to sample population</u>

- Respondent-driven sampling (RDS) is a popular new sampling design that leverages individuals' social network to obtain samples in this setting

An illustration of RDS



- This is not snowball sampling (Goodman, 1961)
- Is RDS ignorable? What role does the graph play in the classical inferential framework?

Agenda

- Inference with non-ignorable sampling designs
  1. Theory
  2. Inferential framework
- Estimation of the causal effects of interference, including peer-influence and peer-pressure
- Concluding remarks

Classical inferential framework

- Y is response
- I is sampling design, implies $Y=(Y_{INC},Y_{EXC})$
- R is missing data mechanism, $Y_{INC}=(Y_{OBS},Y_{MIS})$
- Define $Y_{NOB}=(Y_{EXC},Y_{MIS})$
- X are pre-sampling covariates (e.g., phone book, voter registration lists, …)
- A quantity Q(Y,X) is the estimand of interest

Ignorable sampling designs

A crucial notion is the one of ignorability of the sample mechanism. A sample mechanism I is called ignorable if:

$$\Pr(Y_{NOB} \mid X, Y_{OBS}, R_{INC}, I) = \Pr(Y_{NOB} \mid X, Y_{OBS}, R_{INC}).$$

An equivalent formulation of ignorability for I is the following:

$$\Pr(I \mid X, Y, R_{INC}) = \Pr(I \mid X, Y_{OBS}, R_{INC}).$$

If I does not have this property, it is called non-ignorable design.

The technical challenge

- What role does the graph G play in the classical inferential framework? It is not there.
- G can be thought of providing node-specific covariates. These covariates are only observed for individuals in the sample – like the response
- Introduce X(G), post-sampling covariates. They are used to drive the sample, induce dependence on (and should be kept distinct from) the response

A richer notion of ignorability

Because of this need we introduce the notion of graph ignorability;
We say that a sampling mechanism I is graph ignorable if

$$\Pr(Y_{NOB}, X_{NOB} \mid Y_{OBS}, X_{OBS}, R_{INC}, I)$$

is equal to

$$\Pr(Y_{NOB}, X_{NOB} \mid Y_{OBS}, X_{OBS}, R_{INC}).$$

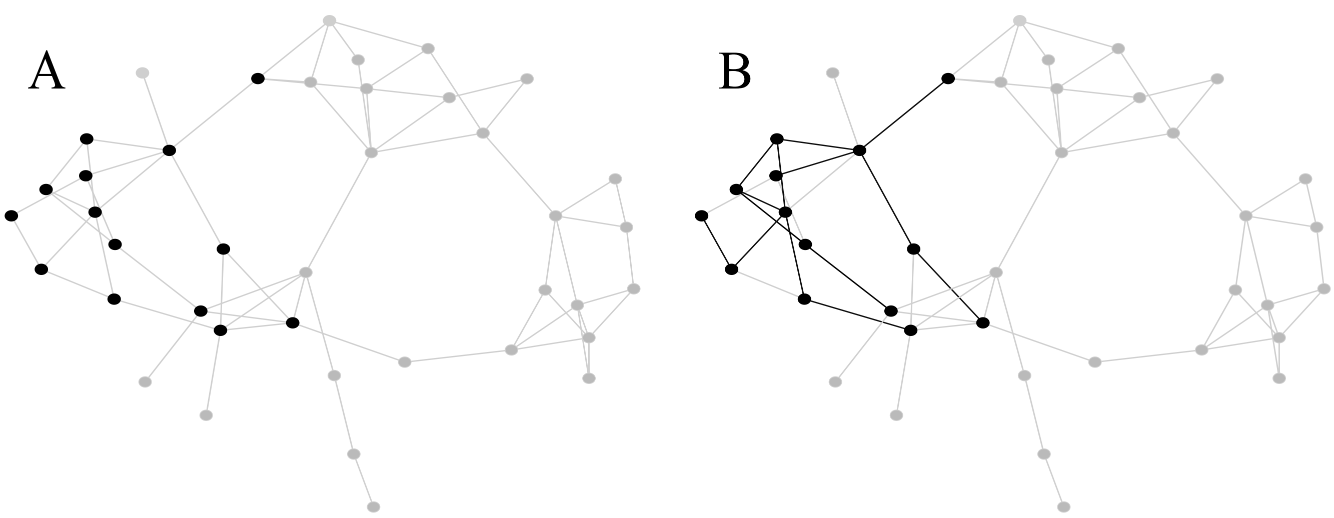
An equivalent expression that may be easier to compute (or to manipulate) for the models we have in mind is:

$$\Pr\{I \mid Y, X, R_{INC}\} = \Pr\{I \mid Y_{OBS}, X_{OBS}, R_{INC}\}.$$

The mathematics are quite simple (just apply Bayes rule).

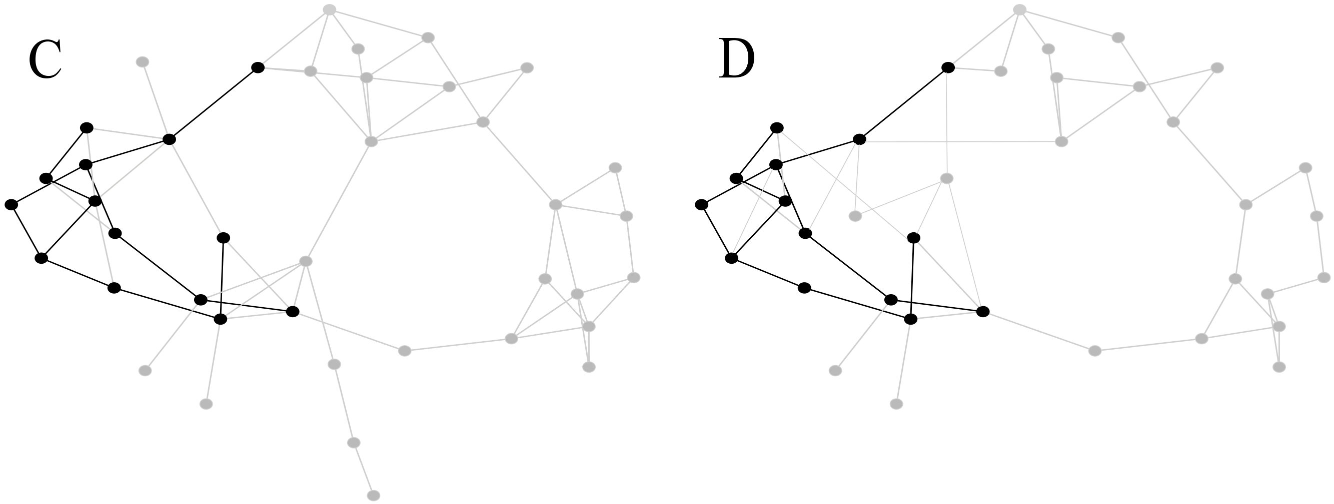
The design *I* as a random variable

- In the classical framework *I* is a vector of 1s and 0s that indicate inclusion and exclusion
- In our setting *I* has a more complicated support



The design *I* as a random variable

- In ***non-ignorable*** network sampling designs, the probability of the observed responses and graph depends on missing nodes and edges



Key remarks

- The graph plays a dual role, on *Y* and *I*
- The standard definition of ignorability and our extension apply to two different settings – post/missing vs pre/obsv
- Only if Y_{NOB} and X_{NOB} are independent a-posteriori, we can distinguish between *Y* and *G* ignorability, but not generally
- If no homophily, $P(Y|G)=P(Y)$, splitting *Y, X(G)* is notation; but homophily is the motivation for non-ignorable designs
- Ignorability of the sampling design is a condition that must be checked, given a joint model – it cannot be assumed

Theorems for popular designs

1. Egocentric sampling (also simple random sampling)
2. Snowball sampling
   – Are ignorable

3. Incomplete egocentric (subset of neighbors)
4. Respondent-driven sampling
   – Are not-ignorable

5. Fixed vs. random population size N
   – No effect on results 1-4

Agenda

- Inference with non-ignorable sampling designs
    1. Theory
    2. Inferential framework
- Estimation of the causal effects of interference, including peer-influence and peer-pressure
- Concluding remarks

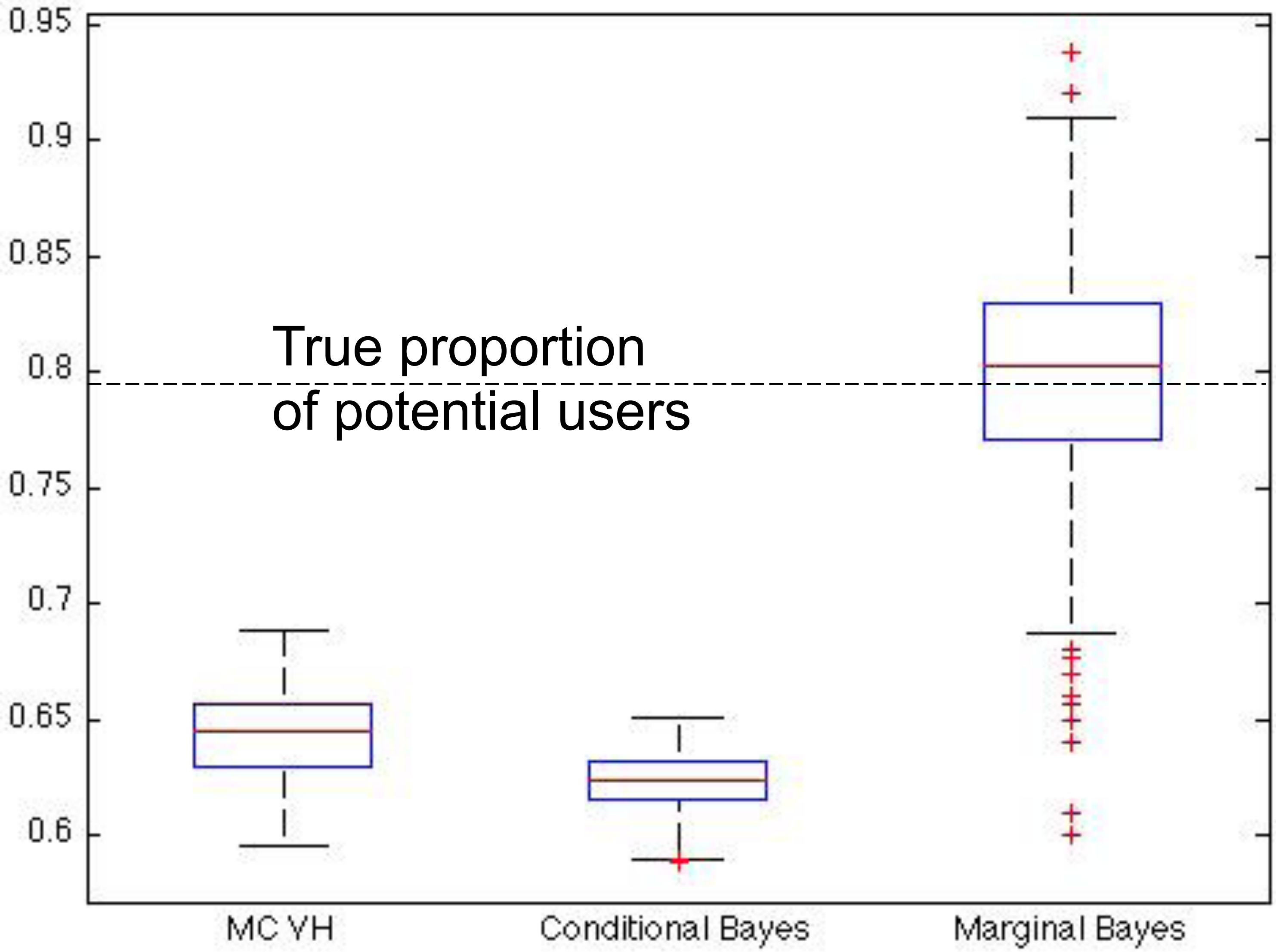
Remarks

- Currently popular Horvitz-Thompson estimators for RDS data are based on inclusion probabilities
- Inclusion probabilities are estimated using various strategies to correct degrees
- Our results suggest that valid inference requires augmenting the sample with both edges and nodes
  1. Devise a reversible-jump MCMC scheme
  2. Propose new Bayes estimators (given choices of Loss)

Toward valid inference

$$
\begin{array}{ccccc}
p(\alpha) & & p(\gamma) & & \\
\downarrow & & \downarrow & & \\
p(\mathcal{G} \mid \alpha) & \rightarrow & p(Y \mid \mathcal{G}, \gamma) & & \\
\downarrow & & \downarrow & & \\
p(I \mid \mathcal{G}) & \rightarrow & p(Y_{INC}, Y_{EXC} \mid \mathcal{G}, \gamma) & & \\
 & & \downarrow & & \\
 & & p(Y_{OBS}, Y_{MISS} \mid \mathcal{G}, \gamma, \eta) & \leftarrow & p(R \mid \eta)
\end{array}
$$

- R is fully observed
- G and Y are partially observed
- This defines a joint distribution P(α,γ,G,Y,I,R)


0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
True proportion
of potential users
MC VH
Conditional Bayes
Marginal Bayes

Agenda

- Inference with non-ignorable sampling designs
- Estimation of the causal effects of interference, including peer-influence and peer-pressure
- Concluding remarks

Motivating problems

Randomized experiments on networks

- Obama for America 2012 campaign
- Leveraging peer-influence for
  - Migrating consumer base from offline to online
  - Increasing ROI by encouraging new product exploration

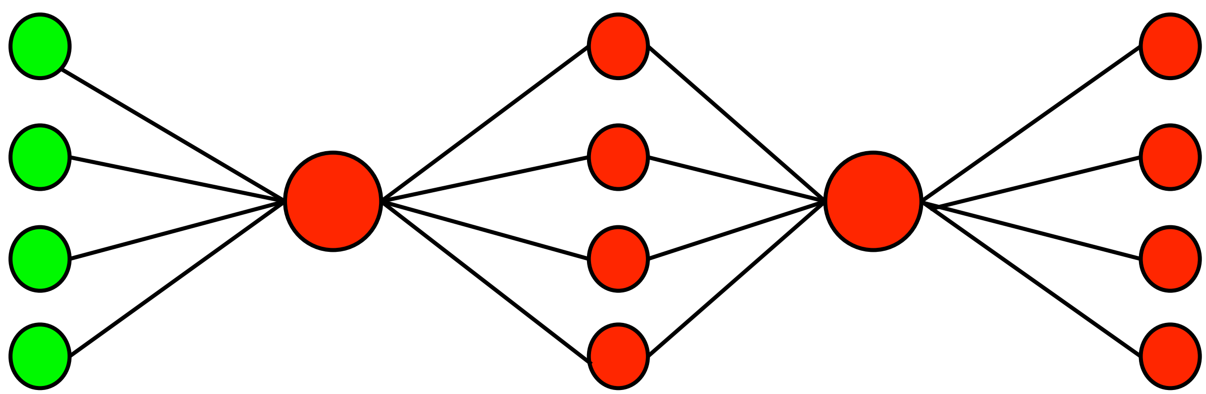
New families of causal estimands

- Prior work (Rosenbaum, Hodgens & Halloran) does not consider social structure explicitly
- Potential outcomes for individual *i* depend on the treatment assignment of its neighbors z_{-i}

$$\delta_k \equiv \frac{1}{|V_k|} \sum_{i \in V_k} \binom{n_i}{k}^{-1} \sum_{\mathbf{z} \in \mathbf{Z}(\mathcal{N}_i;k)} \left(Y_i(0, \mathbf{z}_{-i}) - Y_i(0, \mathbf{z}_{-i} = \mathbf{0})\right)$$

Constrained randomizations

- For δ_k to be estimable, we must observe potential outcomes with both $z_{-i}=0$ and $z_{-i}\neq 0$. This constrains randomizations that lead to valid estimates of δ_k
- We define insulated neighborhood randomizations (INR)



Theory

- We define Sharing Index (SI) as % of nodes that are shared neighbors of at least two other nodes

Thm 1. Number of available INRs $\propto$ 1/SI

Thm 2. INR introduces $bias = SI \times (a - b)$

If we assume additive treatment effects or uniform peer-influence INR leads to unbiased estimates of δ_k

Agenda

- Inference with non-ignorable sampling designs
- Estimation of the causal effects of interference, including peer-influence and peer-pressure
- Concluding remarks

Take home points

- Paired measurements raise statistical problems where the familiar notions of variability, sampling designs, and causal inference are challenged

- Inference from network sampling designs

  Notion of non-ignorability with post-sampling covariates, inferential framework that leads to valid inference

- Causal inference with interference

  New estimands, constrained randomization, theory

Acknowledgements and pointers

CDC, Facebook, Nanigans, Obama for America 2012. J Blitzstein, XL Meng, M Katzoff, J Chang, C Marlow, D Eckles, R Ghani.

1. Valid inference with non-ignorable sampling designs, 2013. Lunagomez & Airoldi.
2. Estimating the causal effect of peer-influence, 2013. Airoldi & Rubin.
3. A survey of statistical network models, *Foundations & Trends in Machine Learning*, 2009. Goldenberg, Zheng, Fienberg & Airoldi.