Optimal design of experiments in the presence of network interference

Edo Airoldi
Department of Statistics
Network data

Units of analysis in the population, $i=1, 2, \ldots$

- Networks as tools: measurements on single units, $Y_i$
  Example: real-valued medical test scores for person $i$
  $Y_i \sim N_p(0, \Omega^{-1})$ and $\{pq : \Omega \neq 0\}$ defines a GMRF [R+11]

- Network data: measurements on pairs of units, $Y_{ij}$
  Example: number of emails person $i$ sends to $j$
  $\{ij : Y_{ij} > 0\}$ is often displayed as a network

Notions of sample size and variability differ [KK11]
This talk at a glance

• Two approaches to optimal design: model-based (Kiefer) vs. randomization-based (Fisher/Kempthorne)

• Define and estimate network interference effects

• Optimal design with network interference
  1. Summarize observed interference with a network model
  2. Find a set of randomizations that maximizes identification and minimizes variance of network interference effects
Agenda

• Interference in applications
• Good design principles and formalisms
• Optimal design with network interference
What is network interference?

Conceptually. Distinct statistical problem settings

\[ Y_i = \text{final grade of student } i \text{ (outcome)} \]
\[ Z_i = \text{student } i \text{ gets private tutoring (intervention)} \]

Interference proper: \[ Y_{\text{Joe}} = f(Z_{\text{Joe}}, Z_{\text{Joe's buddies}}) \]
Spillover/carryover: \[ Y_{\text{Joe}} = f(Y_{\text{Joe's buddies}}) \]
Combinations: \[ Y_{\text{Joe}} = f(Z_{\text{Joe}}, Z_{\text{Joe's buddies}}, Y_{\text{Joe's buddies}}) \]

More generally, we have network-correlated outcomes; connections/covariates play a role in defining scenarios.
Causal interference/network effects

- **Education (EdX)**
  - Designing weekly study groups to maximize learning

- **Labor economics (LinkedIn)**
  - Effects of social structure on labor market dynamics

- **Healthcare interventions (Yale and Harvard)**
  - Disease prevention campaigns that leverage social structure

- **Advertising on social media (Facebook, Google, …)**
  - Increase volunteer time, donations, voter turnout
  - Design repeated auctions
Why is interference so interesting?

From a methodological perspective
   It is the intellectual frontier; 50+ years of work on causal inference assume it away! No longer tenable

From an applied perspective
   Not only we want to control for interference when estimating the effects of an intervention; we want to estimate the effects of interference and leverage them for affecting future events and outcomes
Agenda

• Interference in applications

• **Good design principles and approaches**
  Leveraging an observed network
  Multiple networks
  Imputing aspects of the network

• Optimal design with network interference
Potential outcomes and causality

- Table of potential outcomes $Y_i(Z)$ for $M$ units
  i runs over units and $Z$ runs over all treatment assignments

No interference (SUTVA, ITR) assumes $Y_i(Z_i)$

- Relevant outcomes for $i$ are $Y_i(Z_i=0)$ and $Y_i(Z_i=1)$
  Table of potential outcomes simplifies to $M \times 2$

- Each inferential target is a function of this table
  Effect of treatment on unit $i$ is: $Y_i(Z_i=1) - Y_i(Z_i=0)$
Inferential targets

• Estimable inferential targets are often defined as averages over units and treatment assignments

• Popular example is average treatment effect
  \[ \text{ATE} \equiv \text{Average } Y_i(Z_i=1) - Y_i(Z_i=0) \]

• Simple restrictions on the inferential target may be dictated by the question or statistical considerations
  \( \text{ATE for male freshmen at Harvard} \) (restricts units)
  \( \text{ATE for } Z \) with 50% treated units (restricts assignments)
Agenda

- Interference in applications
- Good design principles and approaches
  - Leveraging an observed network [ATKR12, AR15]
  - Multiple networks
  - Imputing aspects of the network
- Optimal design with network interference
Potential outcomes revisited

Interference (SUTNVA) assumes $Y_i(Z_i, Z_{Ni})$

• Different number of relevant outcomes for unit $i$ as a function of the size of its neighborhood $N_i$

• Example: $i$ is exposed if it has 3 treated neighbors
Inferential targets revisited

- Estimable inferential targets are still often defined as averages over units and randomizations

Average interference effect (AIE) such as the average effect of being exposed to 3 treated neighbors

$$\text{AIE}(3\text{-exposure}) \equiv \text{Average } Y_i(3 \text{ exposed}) - Y_i(\text{control})$$
Inferential targets revisited

- Estimable inferential targets are still often defined as averages over units and randomizations
  Average interference effect (AIE) such as the average effect of being exposed to 3 treated neighbors
  \[ \text{AIE}(3\text{-exposure}) \equiv \text{Average } Y_i(3 \text{ exposed}) - Y_i(\text{control}) \]

- What is the role of observed interference \( G \) ?

- A new set of combinatorial restrictions, due to \( G \), needs to be imposed for statistical considerations
Restrictions on egos and alters

- Example: Inferential target is AIE of 3-exposure
- Inferential target identifies units as eligible to be subjected to effect of interest (egos) or not (alters)
- For any given $Z$ only some egos are relevant
  AIE is average restricted to red and black egos, for each $Z$
Good design pipeline for AIE

1. Assume SUTNVA, …
2. Table of science $Y$ for relevant assignments $Z$
3. Declare causal inferential target(s) of interest
4. Identify egos and alters
5. Restrict randomizations/egos to increase balance and probabilities of inclusion
6. Randomize/re-randomize to insulate neighborhoods
7. Allocate treatment, collect outcomes and estimate
Remarks

Restricting inferential targets (units/randomizations) is a good design principle when interference is present

- In computational social science we often target many small causal effects asserting big data will help
- But finite-population inference based on observed network interference leads to combinatorial issues
- Big data cannot help reduce standard errors if we neglect combinatorial issues during design
Agenda

• Interference in applications
• Good design principles and approaches
  Leveraging an observed network
  Multiple networks [TAR15]
  Imputing aspects of the network
• Optimal design with network interference
Multiple aspects of social structure

• A few years back, we looked at the overlap between declared social relations, on an online social media platform, and social relations revealed by multiple users being co-tagged in uploaded pictures.

• Focused on undergraduates at 130+ US colleges.

• We found surprisingly little overlap.

• Issues of differential coverage due to early stages of the system, many users not being tagged, …
Available network data may not help

• Designed a disease prevention campaign in a number of small villages in South America

• Elicited a social network pre-intervention

• After the intervention, most social coupons were passed outside of the social network …

In a small village everyone knows everyone else; fully connected social structure. Lack of alignment between elicited network and target interference
Experimentally revealed interference

- The observed/available network may not be well suited for the interference effect of interest
- Alternate formalism for estimating interference effects from experimentally revealed network
- Only minimally restricted inferential targets are estimable in a finite-population setting
Agenda

- Interference in applications
- Good design principles and approaches
  - Leveraging an observed network
  - Multiple networks
  - Imputing aspects of the network [SSA15]
- Optimal design with network interference
Imputing interference

Key aspects of interference may not be available

- Oyster card data from Transport for London
- Targets are the effects of line and station closures on traffic from/to other lines and stations
- O/U/DLR map and passenger survey data
- Impute origin-destination traffic volumes on the O/U/DLR network to get at interference effects
Remarks

• Similar design principles apply to observational studies, together with novel matching strategies

• Combinatorial issues in finite-population inference, lack of alignment between network data and the target notion of interference, uncertainty and heterogeneity in the observed network …

Let’s use a model to describe network interference
Agenda

• Interference in applications

• Good design principles and approaches

• Optimal design with network interference [A15]
  Summarizing observed interference with a network model
  Finding a class of optimal treatment allocation strategies
Two approaches

• Finds 1 optimal set of design points for estimating a complex response surface, e.g., nanotechnology, 3D printing, or 1 treatment allocation (Kiefer)
  – Model for the surface, several criteria of optimality

• Finds an optimal collection of treatment allocations to be used for randomization (Fisher)
  – Assumptions on complexity of potential outcomes, minimize variance of an estimator for given inferential target
Agenda

• Interference in applications
• Good design principles and approaches
• Optimal design with network interference
  - Summarizing observed interference with a network model
  - Finding a class of optimal treatment allocation strategies
Formalisms for modeling networks

- Exchangeable graph models [H79, A81]
  \[ P(Y|\theta) = \int \ldots \int \prod_{ij} \text{Bern}(Y_{ij}|u_i, v_j, \theta) \, du \, dv \]
e.g., block models, latent space models [H07, BC09]

- Exponential random graph models [FS86, SI90]
  \[ P(Y|\theta) = \text{Exp}\{ \sum_k \theta_k S_k(Y) + A(\theta) \} \]
where \( S_k(y) \) count motifs, e.g., edges, triangles, …

- Models for dense graphs; what if sparse networks?
Exchangeable graph models & graphons

\[ u_i \sim \text{Uniform (0,1)} \]
\[ w_{ij} \equiv P(G_{ij}=1 \mid u_i, u_j) \]
\[ G_{ij} \sim \text{Bernoulli (} w_{ij} \text{)} \]

where, \( w_{ij} \) is graphon

[H79, A81, S78, LS06, DJ08]
Modeling heterogeneity

- Modeling observed heterogeneity as a stochastic blockmodel; a coarser summary of the network
- Structurally equivalent nodes are collapsed into the same block, statistically exchangeable; not clustering!

\[
\text{Network data } = \text{ Blockmodel} + \text{ Nodes-to-blocks map}
\]

- **From**
  - A | B | C  
  - 1.0 | 0 | 0.3  
  - 0.3 | 1.0 | 0  
  - 0 | 0.3 | 0

- **To**
  - A | B | C  
  - 1.0 | 0 | 0  
  - 0 | 0.3 | 0  
  - 0 | 0 | 0  

- **Node**
  - A | B | C  
  - 1 | 0 | 0  
  - 2 | 0 | 0  
  - 3 | 0 | 0  
  - 4 | 0 | 0  
  - 5 | 0 | 0  
  - 6 | 0 | 0  
  - 7 | 0 | 0  
  - 8 | 0 | 0  
  - 9 | 0 | 0
Stochastic blockmodel approximation

- Posit network $G$ is generated by a fairly general (exchangeable) graph process $\mathcal{G}$
- Stochastic blockmodel approximation $\mathcal{G}_{\text{SBA}}$ is a good estimate of process $\mathcal{G}$ from network data $G$
Sparsity, choice of norm and more

• 1 large sample vs. 2+ small replicated samples [A14]
• Sparse graphons \( \tilde{w}_{ij} = \rho_n \cdot w_{ij} \) (where \( \rho_n \to 0 \) as \( n \uparrow \infty \)) work in asymptopia, but marginals are not amenable to Kolmogorov extension theorem.
• Cut-norm (natural, but abysmal rates) vs. \( L_2 \)
• Theory is for infinitely exchangeable graphs, but finite nodes in practice. [Sharp TV bounds in VA14]
• Alternative multi-stage strategies available. [+++A]
Comparative performance analysis

Process $G$

$G_{SBA}$

MSE = $9.0 \times 10^{-4}$

Process $G$

$G_{SBA}$

MSE = $7.4 \times 10^{-4}$

USVT

MSE = $3.8 \times 10^{-4}$

MSE = $1.2 \times 10^{-3}$

TV min.

MSE = $5.8 \times 10^{-5}$

MSE = $1.3 \times 10^{-4}$
$G_{SBA}$ reduces dimensionality …
... and leads to analytical insights

• Difficult tasks for arbitrary networks [A14]
  1. Counting sub-graphs
  2. Percolation threshold

• Typically, compute approximate answers on the exact observed network $G$

• Instead, perform exact calculations on the stochastic blockmodel approximation ($G_{SBA}$) obtained from $G$
Agenda

• Interference in applications
• Good design principles and approaches
• Optimal design with network interference
  Summarizing observed interference with a network model
  Finding a class of optimal treatment allocation strategies
Set-up

• Assumptions
  Interference as $Y_i(Z_i, Z_{\text{Neighborhood of } i})$
  Observed interference $G$ from exchangeable graph model $G$

Additive treatment and interference effects lead to
$$Y_i = \alpha + \beta \cdot Z_i + \gamma \cdot f_i(Z, G) + \varepsilon_i$$

where $\varepsilon \sim \text{mv-Normal (0, } \Sigma)$, the matrix $\Sigma$ can depend on $G$, and $f_i(Z, G)$ is percentage (or number) of neighbors of $i$ treated
Optimal design procedure

1. Map restricted inferential targets to $\beta$ and $\gamma$ in the linear model: $\text{ATE}(\beta, G)$, $\text{AIE}(\gamma, G)$

2. Estimate $\mathcal{G}$ from $G$ using the stochastic blockmodel approximation, $\mathcal{G}_{\text{SBA}} \equiv (\text{map of units to blocks and block-to-block network matrix})$

3. Find analytical expressions for $f_i(\mathbf{Z}, \mathcal{G}_{\text{SBA}})$

4. Optimal treatment allocation strategies, $\mathbf{p}_{\text{opt}}$ (or $\mathbf{r}_{\text{opt}}$), minimize the variance inferential target estimates
Comparative performance analysis

Given a budget on percentage (or number) of treated:

1. Completely at random, with $p = \frac{1}{2}$

2. Block randomization using $G_{SBA}$. Pick a block at random then treat all units in it (s.t. budget)

3. Sub-optimal. Asymptotic expression for $f_i(Z, G_{SBA})$ assumes units in a block have same no. of neighbors

4. Optimal. Expression for $f_i(Z, G_{SBA})$ accounts for variability in no. neighbors among units in a block
Illustrative simulation results

ATE = $10, AIE = $3, and SE is over randomizations

<table>
<thead>
<tr>
<th>Allocation</th>
<th>ATE = $10</th>
<th>Rndmztn SE</th>
<th>AIE = $3</th>
<th>Rndmztn SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>10.005</td>
<td>0.099</td>
<td>2.860</td>
<td>5.249</td>
</tr>
<tr>
<td>Block $G_{SBA}$</td>
<td>10.000</td>
<td>0.115</td>
<td>2.999</td>
<td>0.748</td>
</tr>
<tr>
<td>ATE sub-opt.</td>
<td>9.993</td>
<td>0.103</td>
<td>2.843</td>
<td>2.129</td>
</tr>
<tr>
<td>ATE optimal</td>
<td>9.988</td>
<td>0.091</td>
<td>3.105</td>
<td>1.910</td>
</tr>
<tr>
<td>AIE sub-opt</td>
<td>10.044</td>
<td>0.133</td>
<td>3.135</td>
<td>0.401</td>
</tr>
<tr>
<td>AIE optimal</td>
<td>10.004</td>
<td>0.115</td>
<td>2.965</td>
<td>0.356</td>
</tr>
</tbody>
</table>
Insight for estimating the AIE

• The optimal strategy $p_{opt}$ treats a distinct percent of units in each block; units are indistinguishable (i.e., exchangeable, or structurally equivalent)

• Globally, together with the blockmodel, the optimal strategy $p_{opt}$ induces a multimodal distribution for the percent of neighbors treated, $f_{i}(Z,G_{SBA})$

• In turn, this increases identifiability of AIE.
Percent of neighbors treated
Analytical insights for AIE and ATE

Terms that appear in bias and variance formulas

- $\beta \cdot \left( \frac{1}{n_t} \sum_i Z_i \cdot \text{degree}_i - \frac{1}{n_c} \sum_i (1-Z_i) \cdot \text{degree}_i \right)$

- $-\gamma \cdot \frac{m}{n(n-1)}$

- $\gamma \cdot \left( \frac{1}{n_t + 1/n_c} \right)$

- $\frac{\sigma^2}{n_t} \cdot \sum_{ij} Z_i Z_j \cdot |\text{neighbors i and j share}|$

- $\frac{\sigma^2}{n_c} \cdot \sum_{ij} (1-Z_i)(1-Z_j) \cdot |\text{neighbors i and j share}|$

- $-2\frac{\sigma^2}{n_t n_c} \cdot \sum_{ij} Z_i(1-Z_j) \cdot |\text{neighbors i and j share}|$

[BA15, KA15, …]
Concluding remarks

• Big data cannot overcome poor experimental design, especially when targeting (small) interference effects

• Good design principles
  Define restrictions on inferential targets when relying on observed interference, or use alternative formalisms

• Optimal designs
  1. Summarize observed interference with a blockmodel
  2. Find optimal randomization strategies given blockmodel
Acknowledgements

Toulis, Kao, Costa, Basse, Karwa, Rubin, Parkes, King, Tarokh (Harvard), Imbens (Stanford), Christakis (Yale), Manski (Northwestern), Silva, Wolfe (UCL), Fienberg, Shalizi (Carnegie Mellon), Scott, Lambert, Azari, Pregibon (Google), Agarwal (LinkedIn), Marlow*, Bakshy, Eckles (Facebook), Calvillo, Denton, Ezzat, optimization team (Nanigans), Daneshmand (AT&T) and Katzoff* (CDC). *formerly at

Contact: airoldi@fas.harvard.edu