Misspecification, Sparsity, and Superpopulation Inference with Large-Scale Social Network Data

Alexander D'Amour

UPenn Wharton Department of Statistics

February 8, 2016

Overview: Foundations of Applied Statistics

Research agenda: A theory of Applied Statistics.

Overview: Foundations of Applied Statistics

Research agenda: A theory of Applied Statistics.

"Is this the right method to use to answer my question or make my decision?"

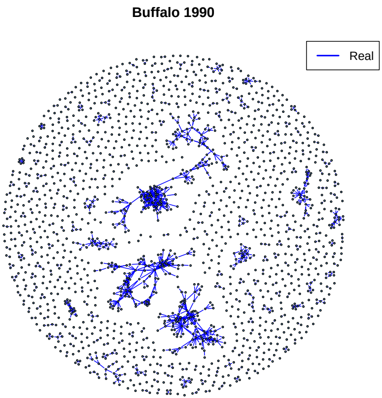
Overview: Foundations of Applied Statistics

Research agenda: A theory of Applied Statistics.

"Is this the right method to use to answer my question or make my decision?"

New desiderata that we can work into modeling decisions.

Generative Network Models



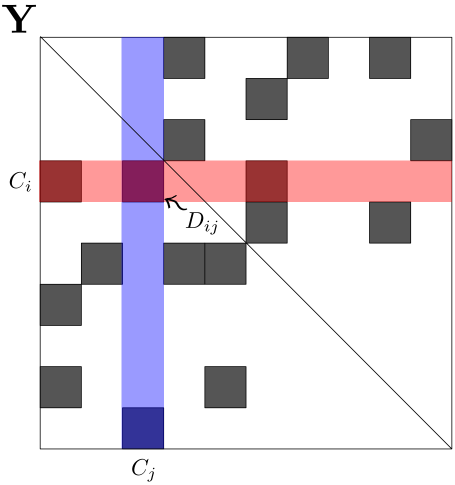
Obtain a set V of n actors.

Explain or predict pairwise outcomes Y_V, potentially using pairwise covariates X_V.

X_V may be observed or latent.

Running example: inventor collaboration network.

Data Representation



Y_V entries in **arbitrary sample space** $\mathcal{Y}$.

Covariates X_V combine observed, latent attributes,

$$X_V^{ij} = f(C_V^i, C_V^j, D_V^{ij}).$$

Social Network Questions

Given the history of a set of actors, how will they behave in the future? (e.g., Mikes karate club)

Social Network Questions

Given the history of a set of actors, how will they behave in the future? (e.g., Mikes karate club)

Given a partially observed set of interactions among a set of actors, can we impute unobserved interactions? (e.g., network security)

Social Network Questions

Given the history of a set of actors, how will they behave in the future? (e.g., Mikes karate club)

Given a partially observed set of interactions among a set of actors, can we impute unobserved interactions? (e.g., network security)

Do two sets of individuals interact in noticeably different ways? (e.g., US vs Canada, Boston vs Silicon Valley)

Social Network Questions

Given the history of a set of actors, how will they behave in the future? (e.g., Mikes karate club)

Given a partially observed set of interactions among a set of actors, can we impute unobserved interactions? (e.g., network security)

Do two sets of individuals interact in noticeably different ways? (e.g., US vs Canada, Boston vs Silicon Valley)

Given our understanding of social interactions among one set of actors, what can we say about behavior among another set of actors? (e.g., sociological theory-building)

Social Network Questions

Given the history of a set of actors, how will they behave in the future? (e.g., Mikes karate club)

Given a partially observed set of interactions among a set of actors, can we impute unobserved interactions? (e.g., network security)

Do two sets of individuals interact in noticeably different ways? (e.g., US vs Canada, Boston vs Silicon Valley)

Given our understanding of social interactions among one set of actors, what can we say about behavior among another set of actors? (e.g., sociological theory-building)

Can we borrow strength between different actor-sets to obtain better resolution on the behavior of both? (e.g., regional random effects)

Attempt 1: Network Regression

Cox PH regression. (Perry and Wolfe, 2013)
Inventor coauthorships in Michigan's motor industry 1982-1988.

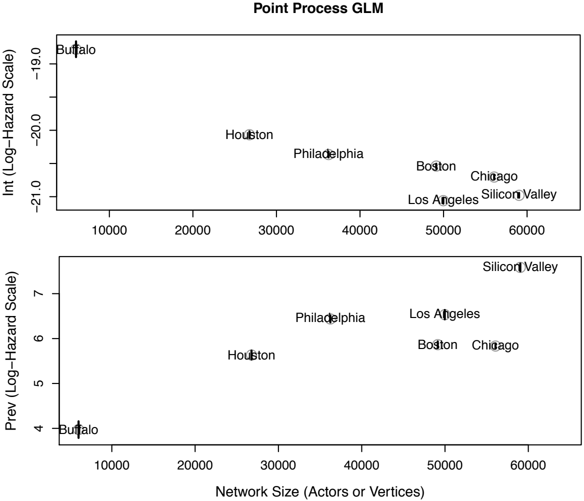
Covariates (Coefs are **log-ratios**):

- `post85`: After 1985.
- `asgnum`: Work for same firm.
- `prev`: Have worked together before.

	lower	est	upper
post85	15.49	15.84	16.20
asgnum	4.65	4.83	5.02
pre	11.36	11.73	12.10
post85:asgnum	-4.77	-4.40	-4.03
post85:prev	-14.57	-14.00	-13.44
asgnum:prev	-5.56	-5.16	-4.76
post85:asgnum:prev	3.91	4.52	5.13

Attempt 2: Regional Comparison Regression

Point process regression. Same time window, different regions.



Social Network Questions

Given the history of a set of actors, how will they behave in the future? (e.g., Mikes karate club)

Given a partially observed set of interactions among a set of actors, can we impute unobserved interactions? (e.g., network security)

Do two sets of individuals interact in noticeably different ways? (e.g., US vs Canada, Boston vs Silicon Valley)

Given our understanding of social interactions among one set of actors, what can we say about behavior among another set of actors? (e.g., sociological theory-building)

Can we borrow strength between different actor-sets to obtain better resolution on the behavior of both? (e.g., regional random effects)

Social Network Questions

Given the history of a set of actors, how will they behave in the future? (e.g., Mikes karate club)

Given a partially observed set of interactions among a set of actors, can we impute unobserved interactions? (e.g., network security)

Do two sets of individuals interact in noticeably different ways? (e.g., US vs Canada, Boston vs Silicon Valley)

Given our understanding of social interactions among one set of actors, what can we say about behavior among another set of actors? (e.g., sociological theory-building)

Can we borrow strength between different actor-sets to obtain better resolution on the behavior of both? (e.g., regional random effects)

Social Network Questions

Given the history of a set of actors, how will they behave in the future? (e.g., Mikes karate club)

Given a partially observed set of interactions among a set of actors, can we impute unobserved interactions? (e.g., network security)

Do two sets of individuals interact in noticeably different ways? (e.g., US vs Canada, Boston vs Silicon Valley)

Given our understanding of social interactions among one set of actors, what can we say about behavior among another set of actors? (e.g., sociological theory-building)

Can we borrow strength between different actor-sets to obtain better resolution on the behavior of both? (e.g., regional random effects)

Inference Paradigms

What sort of question are you asking?

"Single-Sample"

For a **fixed set** of actors

- Project forward in time.
- Impute unmeasured links.

Inference Paradigms

What sort of question are you asking?

"Single-Sample"

For a **fixed set** of actors

- Project forward in time.
- Impute unmeasured links.

"Superpopulation"

For **differing sets** of actors

- Compare network samples.
- Predict or pool information across networks.
- Scale local intuition to global network.

Inference Paradigms

What sort of question are you asking?

"Single-Sample"

For a **fixed set** of actors

- Project forward in time.
- Impute unmeasured links.

Replications restricted to V.

"Superpopulation"

For **differing sets** of actors

- Compare network samples.
- Predict or pool information across networks.
- Scale local intuition to global network.

Replications for any $V \subset \mathbb{V}$.

Inference Paradigms

What can theory tell us?

Except under restrictive assumptions, criteria for superpopulation and single-sample inference are non-equivalent.

Inference Paradigms

What can theory tell us?

Except under restrictive assumptions, criteria for superpopulation and single-sample inference are non-equivalent.

- Each outcome Y_V^{ij} is conditionally independent.
- If model is correctly specified.

Inference Paradigms

What can theory tell us?

Except under restrictive assumptions, criteria for superpopulation and single-sample inference are non-equivalent.

- Each outcome Y_V^{ij} is conditionally independent.
- If model is correctly specified.

Theory so far covers single-sample inference, giving little guidance for superpopulation questions .

The Question: Superpopulation Inference

How do we represent a network superpopulation?

The Question: Superpopulation Inference

How do we represent a network superpopulation?

Infinite random graph object defined **top-down** as stochastic process (Shalizi and Rinaldo 2013).

The Question: Superpopulation Inference

How do we represent a network superpopulation?

Infinite random graph object defined **top-down** as stochastic process (Shalizi and Rinaldo 2013).

Definition 1 (Random Graph Process).

A random interaction process $Y_{\mathbb{V}}$ is a stochastic process indexed by a countably infinite vertex set $\mathbb{V}$ whose finite-dimensional distribution for any finite subset $V \subset \mathbb{V}$ defines an interaction graph Y_V with vertex set V. Denote the law of $Y_{\mathbb{V}}$ as $\mathbb{P}_{\mathbb{V}}$ and the law of a finite-dimensional projection Y_V as $\mathbb{P}_V$.

The Question: Superpopulation Inference

How do we represent a network superpopulation?

Infinite random graph object defined **top-down** as stochastic process (Shalizi and Rinaldo 2013).

Definition 1 (Random Graph Process).

A random interaction process $Y_{\mathbb{V}}$ is a stochastic process indexed by a countably infinite vertex set $\mathbb{V}$ whose finite-dimensional distribution for any finite subset $V \subset \mathbb{V}$ defines an interaction graph Y_V with vertex set V. Denote the law of $Y_{\mathbb{V}}$ as $\mathbb{P}_{\mathbb{V}}$ and the law of a finite-dimensional projection Y_V as $\mathbb{P}_V$.

Statistical interpretation: Observed samples are finite subgraphs of population graph. Population graph is of scientific interest.

The Question: Superpopulation Inference

Differences vs single-sample objects

The Question: Superpopulation Inference

Differences vs single-sample objects

Role of infinite objects.

The Question: Superpopulation Inference

Differences vs single-sample objects

Role of infinite objects.

Single-sample: Random graph sequence

- Constructed to induce particular limit.
- Limit is deterministic analogy for a single, large sample.
- Internal consistency unimportant.

The Question: Superpopulation Inference

Differences vs single-sample objects

Role of infinite objects.

Single-sample: Random graph sequence

- Constructed to induce particular limit.
- Limit is deterministic analogy for a single, large sample.
- Internal consistency unimportant.

Superpopulation: Random graph process

- Induces Kolmogorov consistency on constructed sequences.
- Focus on relationships between finite-dimensional distributions.
- Infinity is useful, but limit is unimportant.

The Method: Parametric MLE

Operational procedure

Let $\mathbb{P}_{0,\mathbb{V}}$ be the law of the true population process; $\mathbb{P}_{0,V}$ be the distribution of the sample Y_V.

The Method: Parametric MLE

Operational procedure

Let $\mathbb{P}_{0,\mathbb{V}}$ be the law of the true population process; $\mathbb{P}_{0,V}$ be the distribution of the sample Y_V.

Likelihood inference procedure

1. Propose a model family $\mathcal{P}_{\Theta,\mathbb{V}}$ of models $\mathbb{P}_{\theta,\mathbb{V}}$.
2. $\mathcal{P}_{\Theta,\mathbb{V}}$ implies a likelihood $\mathbb{P}_{\theta,V}$ on the sampled index set V for each $\theta \in \Theta$. Compute

$$\hat{\theta}_V = \arg\max_{\Theta} \log \mathbb{P}_{\theta,V}(Y_V). \tag{1}$$

3. Interpret $\hat{\theta}_V$ as a superpopulation parameter estimate.

The Method: Parametric MLE

Operational procedure

Let $\mathbb{P}_{0,\mathbb{V}}$ be the law of the true population process; $\mathbb{P}_{0,V}$ be the distribution of the sample Y_V.

Likelihood inference procedure

1. Propose a model family $\mathcal{P}_{\Theta,\mathbb{V}}$ of models $\mathbb{P}_{\theta,\mathbb{V}}$.
2. $\mathcal{P}_{\Theta,\mathbb{V}}$ implies a likelihood $\mathbb{P}_{\theta,V}$ on the sampled index set V for each $\theta \in \Theta$. Compute

$$\hat{\theta}_V = \arg\max_{\Theta} \log \mathbb{P}_{\theta,V}(Y_V). \tag{1}$$

3. Interpret $\hat{\theta}_V$ as a superpopulation parameter estimate.

Note: Step 3 is the only difference between single-sample and superpopulation.

Models and Misspecification

Model-Building Tradeoffs

No parsimonious model can fully represent complex network structure.

Models and Misspecification

Model-Building Tradeoffs

No parsimonious model can fully represent complex network structure.

Choose one:

- **Local structure.** Homophily, Heterophily, Transitivity, etc.
- **Global structure.** Sparsity, Percolation, etc.

Models and Misspecification

Common local approaches

Conditionally independent dyads (regression):

$$P(Y \mid X) = \prod_{i<j<n} P(Y_V^{ij} \mid X_V^{ij}).$$

Infinitely exchangeable dyads (Aldous-Hoover):

$$P(Y \mid X) = \int_{\mathcal{C}} \prod_{i<j<n} P(Y_V^{ij} \mid W_V^{ij}(C_V^i, C_V^j))dF(C).$$

Misspecified Models and Science

All models are wrong but...

How do we make sense of a misspecified model?

Misspecified Models and Science

All models are wrong but...

How do we make sense of a misspecified model?

Parameter estimates as **measurements** of $\mathbb{P}_{0,V}$.

Misspecified Models and Science

All models are wrong but...

How do we make sense of a misspecified model?

Parameter estimates as **measurements** of $\mathbb{P}_{0,V}$.

Minimal criterion for "usefulness": **Stability**.

Misspecified Models and Science

All models are wrong but...

How do we make sense of a misspecified model?

Parameter estimates as **measurements** of $\mathbb{P}_{0,V}$.

Minimal criterion for "usefulness": **Stability**.

"Similar" inputs Y_V yield "similar" estimates $\hat{\theta}_V$.

"Similar input" means replications of Y_V from the same finite distribution $\mathbb{P}_{0,V}$.

Misspecified Models and Science

Stability: Single sample case

"Similar input" means replications of Y_V from the same finite distribution $\mathbb{P}_{0,V}$.

Huber 1967 showed MLE is large-sample consistent for a pseudo-true parameter (naming due to Sawa 1978), satisfying

$$\bar{\theta}_V = \arg\max_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_0}[\log \mathbb{P}_{\theta,V}(Y_V)] \tag{2}$$

"Similar output" defined by concentration of $\hat{\theta}_V$ in large samples.

Misspecified Models and Science

Stability: Superpopulation case

"Similar input" means any sample Y_V drawn from the same superpopulation $\mathbb{P}_{0,\mathbb{V}}$.

Intuitively, outputs $\hat{\theta}_V$ are similar if they **effectively estimate** the same thing.

What does the MLE $\hat{\theta}_V$ effectively estimate when the model is misspecified?

The Effective Estimand of the MLE

Define the effective estimand of the MLE as the finite-sample pseudo-true parameter.

$$\bar{\theta}_V = \arg\max_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_0}[\log \mathbb{P}_{\theta,V}(Y_V)] \tag{2}$$

The Effective Estimand of the MLE

Define the effective estimand of the MLE as the finite-sample pseudo-true parameter.

$$\bar{\theta}_V = \arg\max_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_0}[\log \mathbb{P}_{\theta, V}(Y_V)] \tag{2}$$

Justifications:

- Finite-sample concentration (e.g., Spokoiny 2012).
- Fisher-consistency inversion.
- Estimating equation unbiased.
- KL projection plug-in.

Superpopulation Stability Criterion

Criterion 1.

A procedure is superpopuatlion stable for making inferences about a superopulation process $\mathbb{P}_{0,\mathbb{V}}$ only if, for any finite sample Y_V generated according to $\mathbb{P}_{0,V}$, the effective estimand $\bar{\theta}_V$ of the estimator $\hat{\theta}_V$ is invariant to the indexing set V.

Superpopulation Stability Criterion

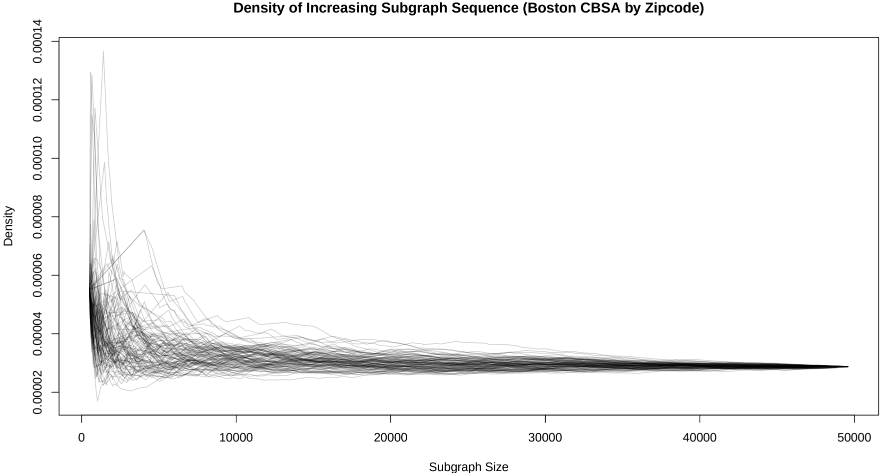
Criterion 1.

A procedure is superpopuatlion stable for making inferences about a superopulation process $\mathbb{P}_{0,\mathbb{V}}$ only if, for any finite sample Y_V generated according to $\mathbb{P}_{0,V}$, the effective estimand $\bar{\theta}_V$ of the estimator $\hat{\theta}_V$ is invariant to the indexing set V.

Test the criterion with top-down specification of superpopulation properties, e.g., sparsity.

The Wrinkle: Sparsity

Illustration



The Wrinkle: Sparsity

Formally

The Wrinkle: Sparsity

Formally

Define the **density operator**

$$D(Y_V) = \frac{\sum_{ij} \mathbb{I}\{Y_V^{ij} \neq 0\}}{\binom{|V|}{2}}.$$

The Wrinkle: Sparsity

Formally

Define the **density operator**

$$D(Y_V) = \frac{\sum_{ij} \mathbb{I}\{Y_V^{ij} \neq 0\}}{\binom{|V|}{2}}.$$

Definition 1 (Sparse Graph Process).

Let $Y_{\mathbb{V}}$ be a random graph process on $\mathbb{V}$. $Y_{\mathbb{V}}$ is *sparse* if and only if for any $\epsilon > 0$ there exists an n such that for any subset of vertices $V \in \mathbb{V}$ with $|V| > n$ the corresponding finite dimensional random graph Y_V has the property $\mathbb{E}(D(Y_V)) < \epsilon$.

The Wrinkle: Sparsity

The Wrinkle: Sparsity

Differences vs single-sample sparsity.

Single-sample

- Defined in terms of random graph sequences.
- Often defined in explicitly non-Kolmogorov-consistent terms.
- Analogy for single sample with very few observed interactions.

Superpopulation

- Property of a random graph process, not a random graph.
- Defines an assumption about the system, not a theoretical object.

Sparsity Misspecification

Definition

Sparsity Misspecification

Definition

A model family $\mathcal{P}_{\theta,\mathbb{V}}$ is **sparsity misspecfied** iff for every $\theta \in \Theta$ and every increasing sequence of vertex sets (V_n),

$$\frac{\mathbb{E}_\theta(D(Y_{V_n}))}{\mathbb{E}_0(D(Y_{V_n}))} \to 0 \text{ or } \infty.$$

Sparsity Misspecification

Definition

A model family $\mathcal{P}_{\theta,\mathbb{V}}$ is **sparsity misspecfied** iff for every $\theta \in \Theta$ and every increasing sequence of vertex sets (V_n),

$$\frac{\mathbb{E}_\theta(D(Y_{V_n}))}{\mathbb{E}_0(D(Y_{V_n}))} \to 0 \text{ or } \infty.$$

Sparsity Misspecification
Definition

A model family $\mathcal{P}_{\theta,\mathbb{V}}$ is **sparsity misspecfied** iff for every $\theta \in \Theta$ and every increasing sequence of vertex sets (V_n),

$$\frac{\mathbb{E}_\theta(D(Y_{V_n}))}{\mathbb{E}_0(D(Y_{V_n}))} \to 0 \text{ or } \infty.$$

For example,

- For CID (under regularity) and exchangeable models, population extension is **dense** or **empty** (e.g., Orbanz and Roy, 2013).
- For process models, most lock in a given form for $\epsilon(n)$ (e.g., power law for preferential attachment).

Main Result: Moving Target

Theorem 1 (Moving target theorem).

Let (V_n) be an increasing sequence of vertex sets from $\mathbb{V}$. Suppose the following hold:

1. *For some finite n, $\mathbb{E}_0(D(Y_{V_n})) > 0$.*
2. *The inferential family $\mathcal{P}_{\Theta,\mathbb{V}}$ is sparsity misspecified for the true population process $\mathbb{P}_{0,\mathbb{V}}$.*
3. *The inferential model is responsive to the sample density $D(Y_{V_n})$ under the true population process and*

$$|\mathbb{E}_{\bar{\theta}}(D(Y_{V_n})) - \mathbb{E}_0(D(Y_{V_n}))| = O(\epsilon_(n)). \tag{3}$$

Then, $\bar{\theta}_{V_n}$ varies with n in the sense that for any n, there exists an $n' > n$ such that $\bar{\theta}_{V_n} \neq \bar{\theta}_{V_{n'}}$, and the MLE of the model violates Criterion 1.

Example: Poisson Regression

Setup

Question: How do firms influence collaboration dynamics?

Y_V collaboration counts; X_V^{ij} indicates shared firm.

Assumptions:

- The true collaboration-generating process $Y_{0,\mathbb{V}}$ is sparse .
- All firms have finite size.
- A non-vanishing fraction of firms have a positive number of expected within-firm interactions.

Example: Poisson Regression

Model and effective estimand

Model:

$$Y^{ij}_{V_n} \overset{\perp\!\!\!\perp}{\sim} \mathrm{Pois}(\exp(\theta_1 + X^{ij}_{V_n}\theta_2)), \tag{4}$$

Effective Estimands:

$$\bar{\theta}_{1V} = \log\left(\frac{\sum_{ij}\mathbb{E}_0(Y^{ij}_V \mid X^{ij}_V = 0)(1 - X^{ij}_V)}{\sum_{ij}(1 - X^{ij}_V)}\right) \tag{5}$$

$$\bar{\theta}_{2V} = \log\left(\frac{\sum_{ij}\mathbb{E}_0(Y^{ij}_V \mid X^{ij}_V = 1)X^{ij}_V}{\sum_{ij}X^{ij}_V} \Bigg/ \frac{\sum_{ij}\mathbb{E}_0(Y^{ij}_V \mid X^{ij}_V = 0)(1 - X^{ij}_V)}{\sum_{ij}(1 - X^{ij}_V)}\right). \tag{6}$$

What's wrong with this picture?

Partial Resolution

Salvaging conditional independence

What *can* we estimate with local models?

Two-stage process. Complex **relationship** structure R. Y simple conditional on relationships.

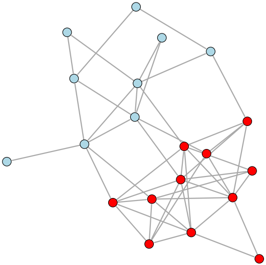
$$\mathbb{P}_{0,V} = \mathbb{P}_0(R) \prod_{i<j<n} \mathbb{P}_0(Y_V^{ij} \mid R_V, X_V).$$

Changes question:

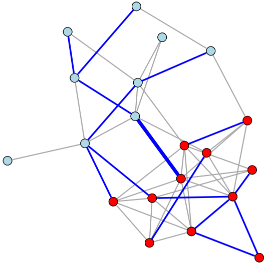
$$\mathbb{P}_{\theta,V}(Y_V \mid X_V) \to \mathbb{P}_{\beta,V}(Y_V \mid R_V, X_V).$$

Partial Resolution

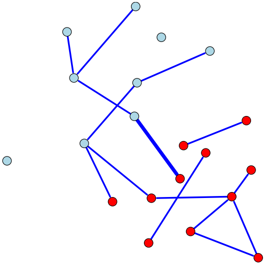
Conditionally Independent Relationship model



Unobservable
Relationship/Risk



Both



Observable
Interactions

Partial Resolution

Partial likelihood inference

Let $A_V = \mathbb{I}\{Y_V \neq 0\}$.

Exploit conditional distribution

$$\mathbb{P}_\beta(Y_V \mid A_V, X_V) = \frac{\mathbb{P}_\beta(Y_V \mid R_V, X_V)}{1 - \mathbb{P}_\beta(Y_V = 0 \mid R_V X_V)},$$

or the **zero-truncated likelihood**.

Partial Resolution

Partial likelihood inference

Let $A_V = \mathbb{I}\{Y_V \neq 0\}$.

Exploit conditional distribution

$$\mathbb{P}_\beta(Y_V \mid A_V, X_V) = \frac{\mathbb{P}_\beta(Y_V \mid R_V, X_V)}{1 - \mathbb{P}_\beta(Y_V = 0 \mid R_V X_V)},$$

or the **zero-truncated likelihood**.

Partial Resolution

Partial likelihood inference

Let $A_V = \mathbb{I}\{Y_V \neq 0\}$.

Exploit conditional distribution

$$\mathbb{P}_\beta(Y_V \mid A_V, X_V) = \frac{\mathbb{P}_\beta(Y_V \mid R_V, X_V)}{1 - \mathbb{P}_\beta(Y_V = 0 \mid R_V X_V)},$$

or the **zero-truncated likelihood**.

By **modeling less**, obtain an effective estimand $\bar{\beta}_V$ invariant to the marginal distribution of R_V.

Partial Resolution

Partial likelihood inference

Let $A_V = \mathbb{I}\{Y_V \neq 0\}$.

Exploit conditional distribution

$$\mathbb{P}_\beta(Y_V \mid A_V, X_V) = \frac{\mathbb{P}_\beta(Y_V \mid R_V, X_V)}{1 - \mathbb{P}_\beta(Y_V = 0 \mid R_V X_V)},$$

or the **zero-truncated likelihood**.

By **modeling less**, obtain an effective estimand $\bar{\beta}_V$ invariant to the marginal distribution of R_V.

Partial Resolution

Partial likelihood inference

Let $A_V = \mathbb{I}\{Y_V \neq 0\}$.

Exploit conditional distribution

$$\mathbb{P}_\beta(Y_V \mid A_V, X_V) = \frac{\mathbb{P}_\beta(Y_V \mid R_V, X_V)}{1 - \mathbb{P}_\beta(Y_V = 0 \mid R_V X_V)},$$

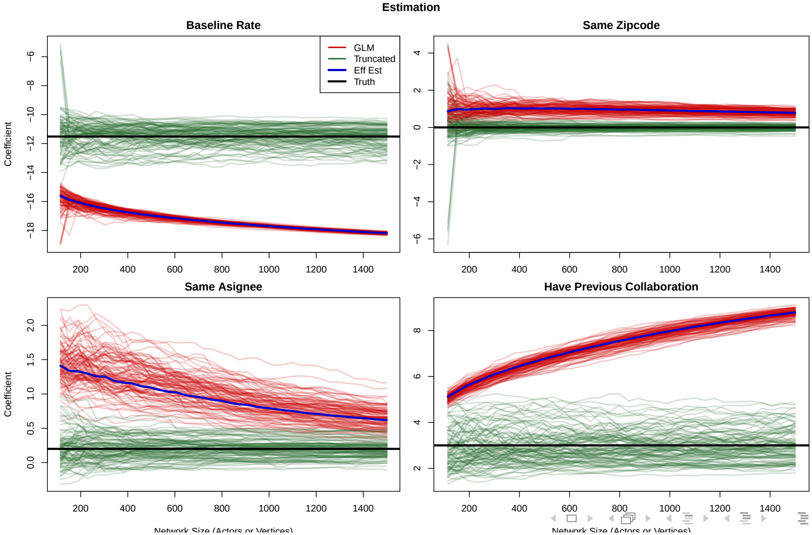
or the **zero-truncated likelihood**.

By **modeling less**, obtain an effective estimand $\bar{\beta}_V$ invariant to the marginal distribution of R_V.

Bonus: Computation is $O(\sum_{ij} A_V)$.

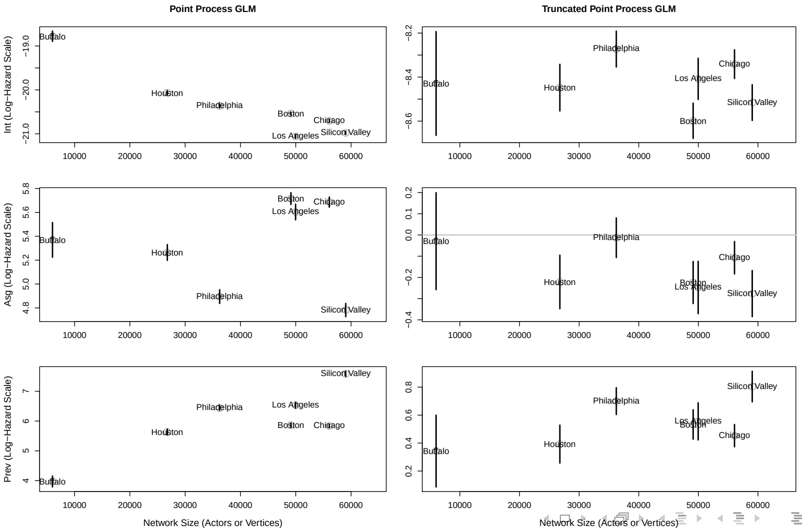
Partial Resolution

Simulated success



Partial Resolution

Real success



Discussion

GEP Box: “Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.”

Discussion

GEP Box: "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."

We need a formal language of **usefulness**.

Discussion

GEP Box: "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."

We need a formal language of **usefulness**.

Estimated objects are tools for decision-making. Estimators must recover decision-critical properties. **Effective estimand** makes this formal.

Discussion

GEP Box: "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."

We need a formal language of **usefulness**.

Estimated objects are tools for decision-making. Estimators must recover decision-critical properties. **Effective estimand** makes this formal.

Modeling *less* of the system by seeking invariances make a model *more* scientifically relevant. **Invariance** can be better than a bad explanation.