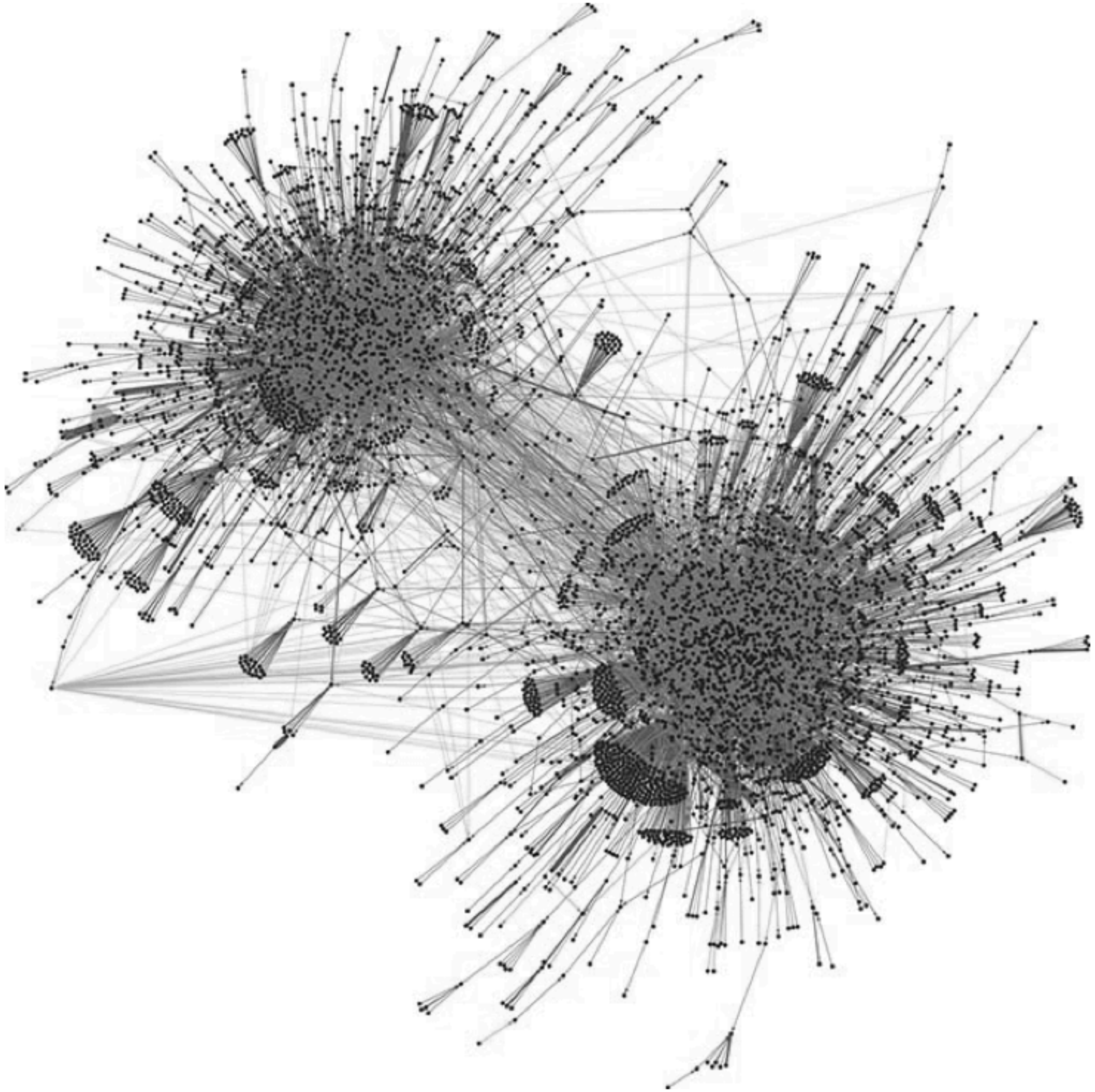
On Causal Analysis for Heterogeneous Networks

Katerina Marazopoulou, David Arbour, David Jensen

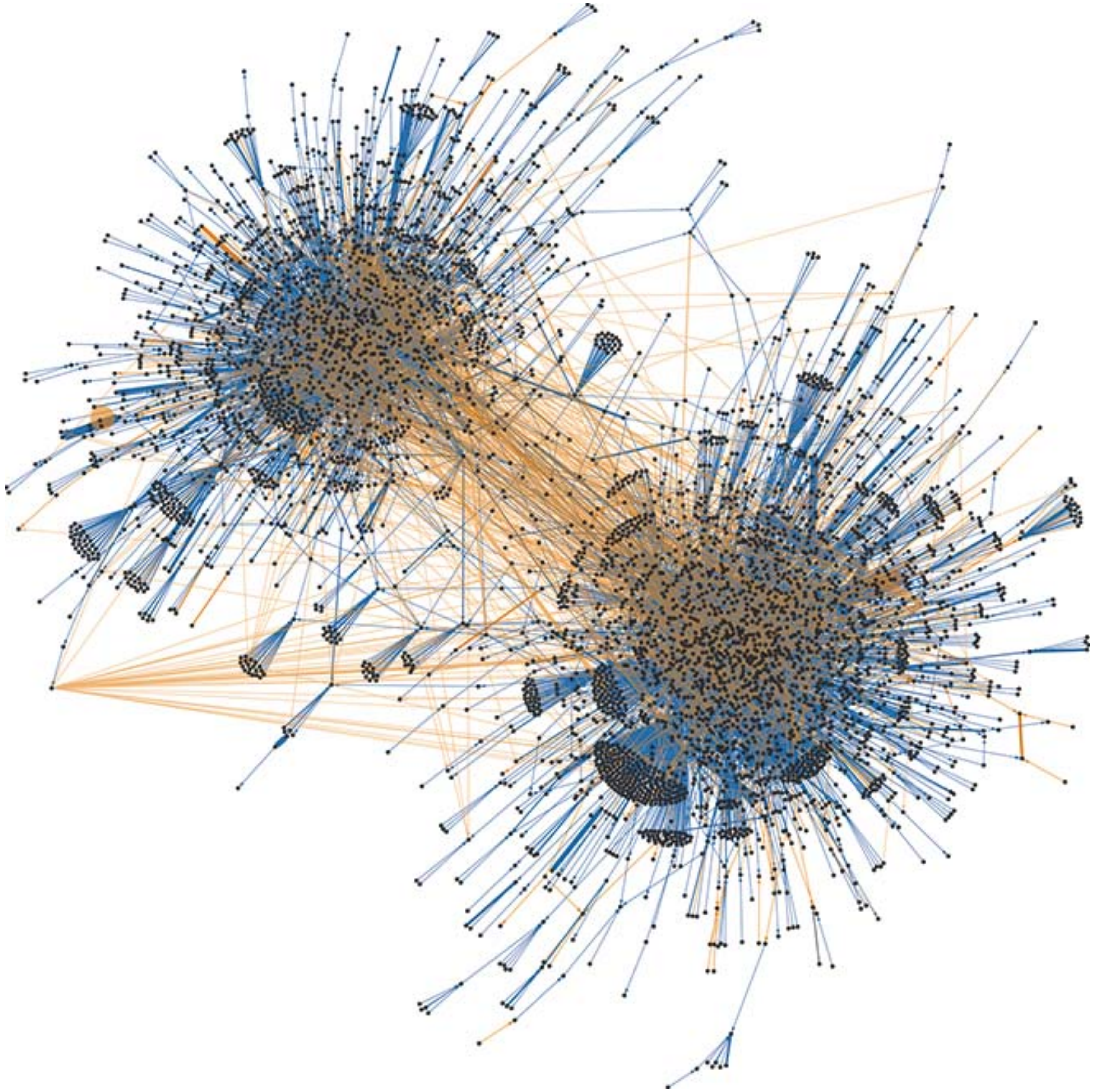
KDD Workshop on Causal Discovery

August 2017



Causal inference in networks: How is the behavior of an individual affected by his/her peers?

source: Visual Complexity



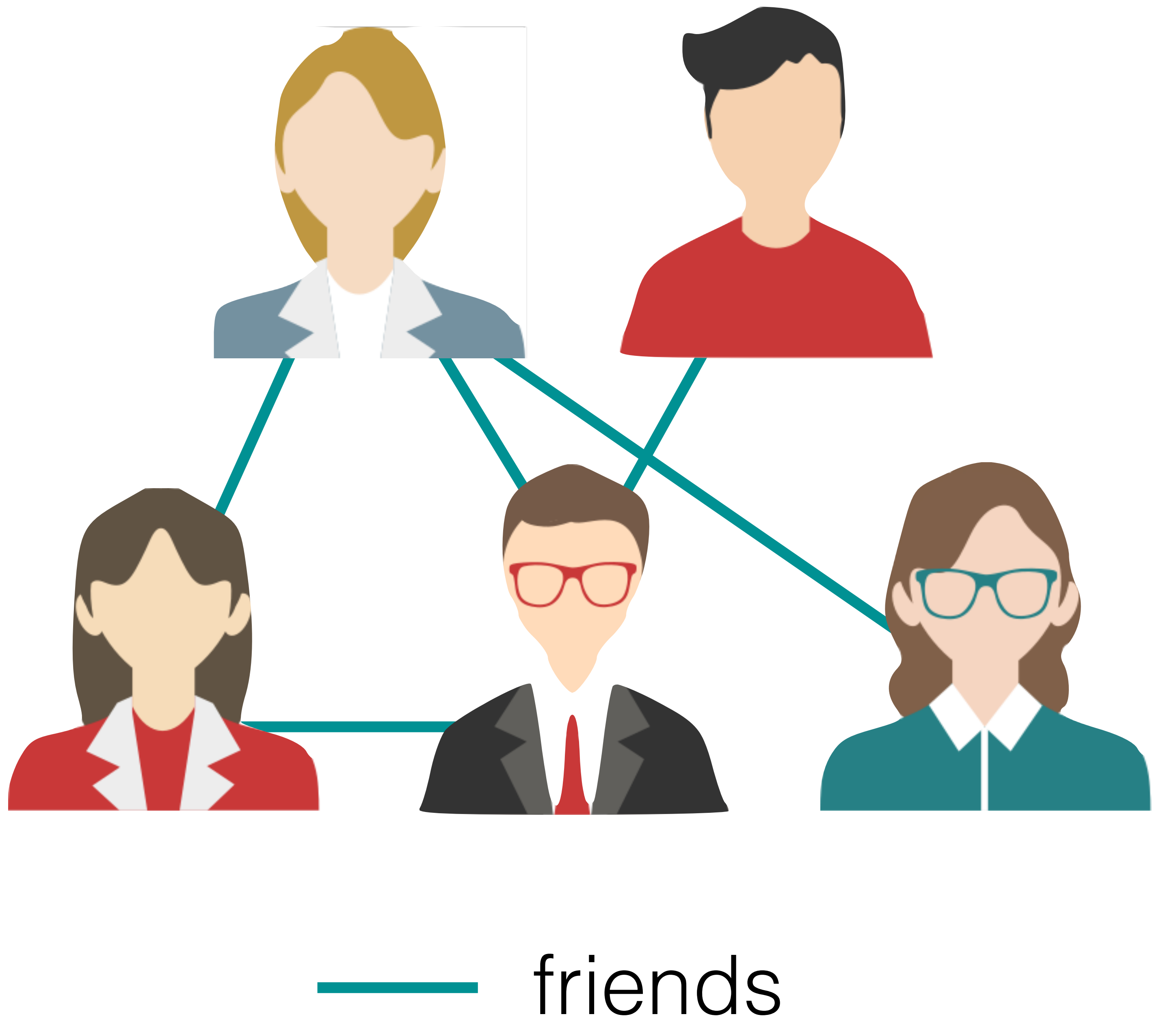
How does the presence of multiple relationship types affect causal analysis?

source: Visual Complexity

Outline

- Background: Causal effect estimation on networks
- Causal effect estimation in heterogeneous networks
- Experiments on synthetic data
- Application on real-world dataset

Causal Effect Estimation in Networks



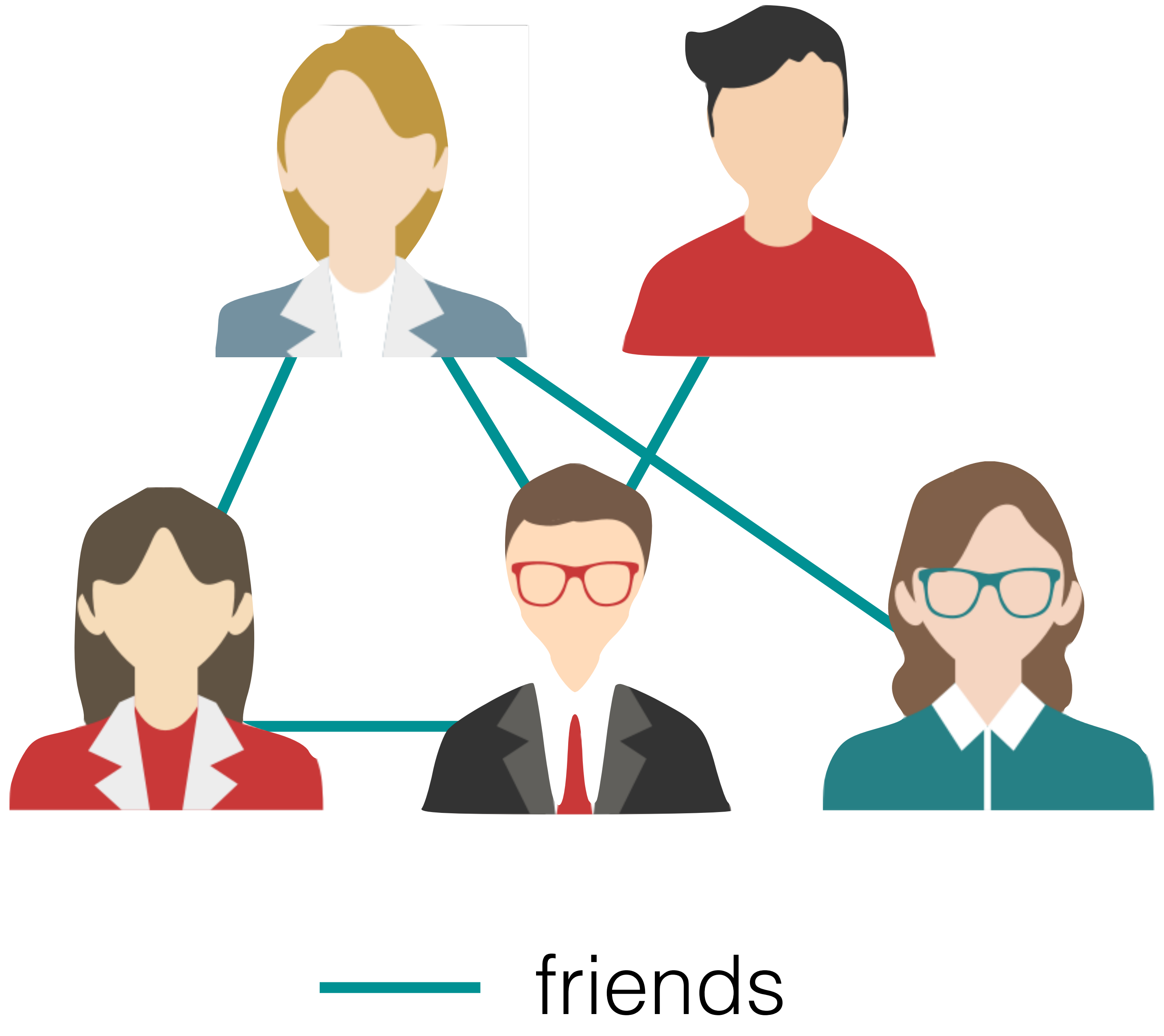
Causal Effect Estimation in Networks

- Population of n individuals that form an undirected graph
- Binary treatment T and outcome O
- The outcome of a node depends on the global treatment assignment:

$$O_i(\mathbf{T} = \mathbf{t}) \quad \text{where} \quad \mathbf{t} \in \{0, 1\}^n$$

- ATE between global treatment and global control

$$\tau(\mathbf{1}, \mathbf{0}) = \frac{1}{n} \sum_{i=1}^{n} E[O_i(\mathbf{T} = \mathbf{1}) - O_i(\mathbf{T} = \mathbf{0})]$$



Causal effect estimation

Estimation procedure for causal inference:

1. Treatment assignment
2. Exposure model: When an individual is considered to be treated
3. Analysis: How to estimate the causal quantity of interest

Causal effect estimation

Estimation procedure for causal inference:

1. **Treatment assignment**
2. **Exposure model**: Fraction neighborhood exposure [Gui et al. 2015]
3. **Analysis**: Linear regression adjustment [Gui et al. 2015]

Gui, Basin, Han. WWW 2015

The Gui et al. framework

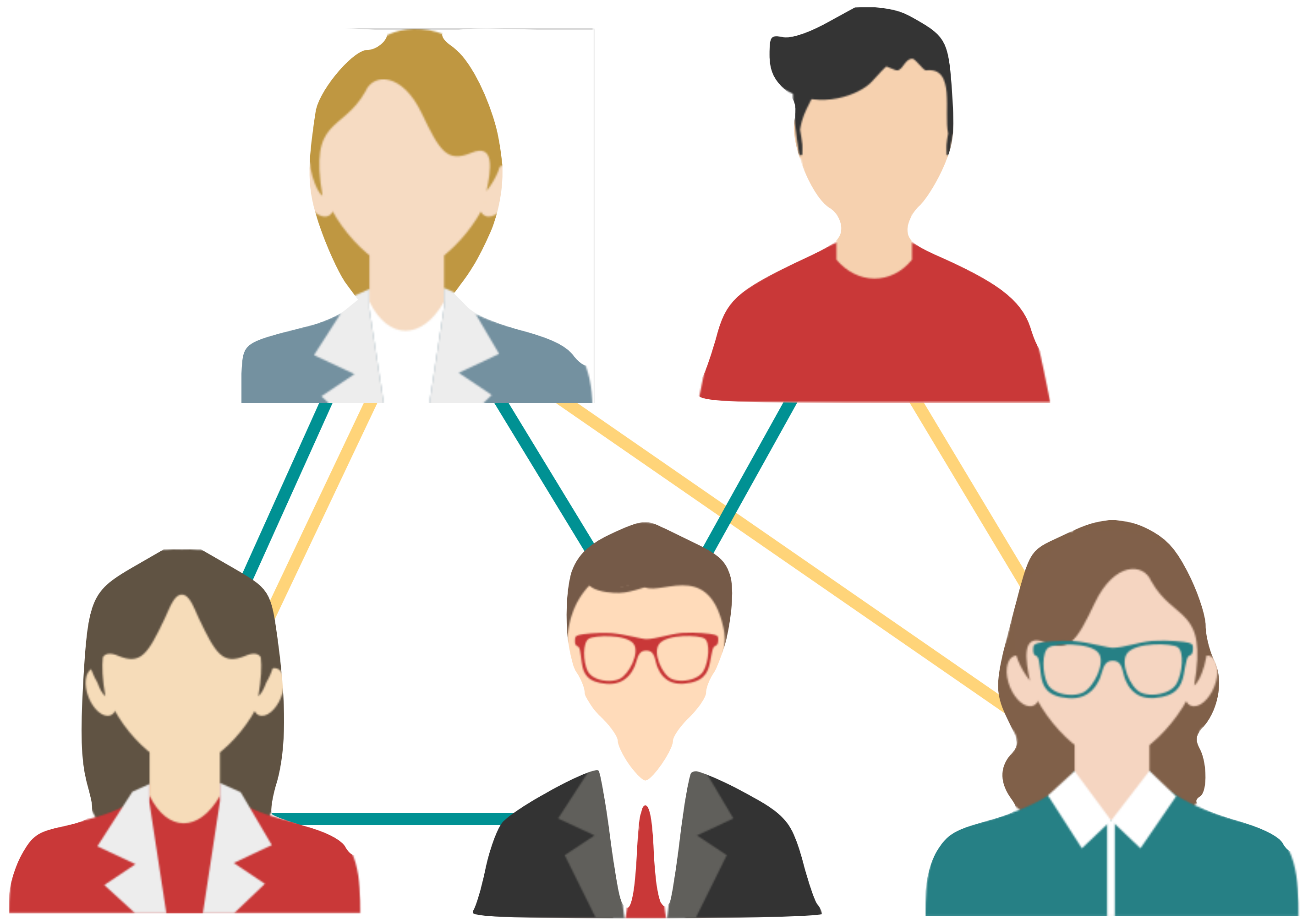
- Fraction neighborhood exposure model:

  The response function depends on a node's own treatment assignment and the **proportion of its treated peers**

$$g(T_i, \lambda_i) = \alpha + \beta T_i + \gamma \lambda_i$$

- ATE: $\tau(\mathbf{1}, \mathbf{0}) = g(T_i = 1, \lambda_i = 1) - g(T_i = 0, \lambda_i = 0) = \beta + \gamma$

Heterogenous Network



friends

coworkers

Response function

Homogeneous networks:

$$g(T_i, \lambda_i) = \alpha + \beta T_i + \gamma \lambda_i$$

Heterogeneous networks:

$$g_{f,c}(T_i, \lambda_i) = \alpha + \beta T_i + \gamma^f \lambda_i^f + \gamma^c \lambda_i^c$$

Sets of peers

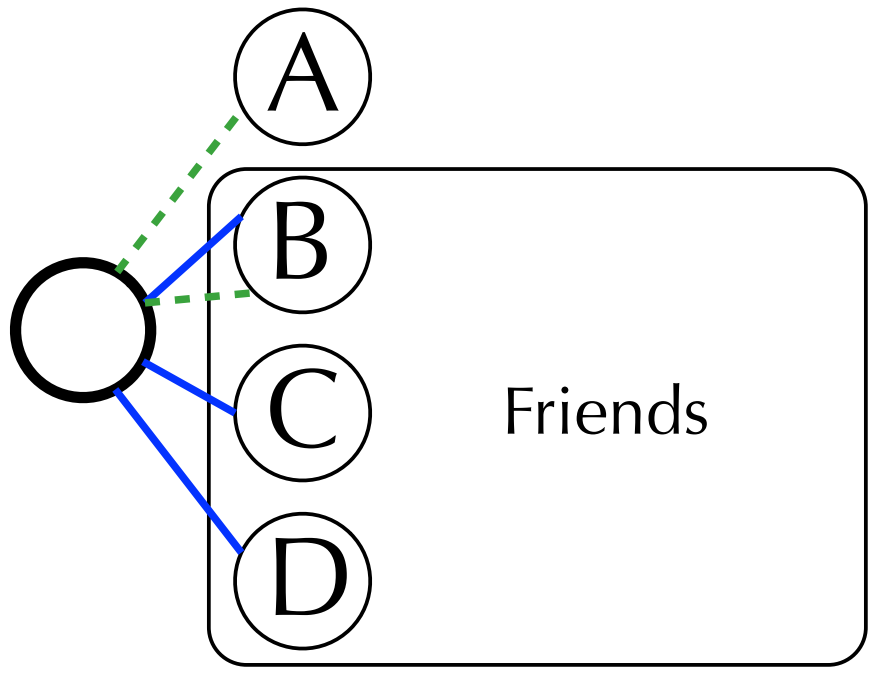
$$g_{f,c}(T_i, \lambda_i) = \alpha + \beta T_i + \gamma^f \lambda_i^f + \gamma^c \lambda_i^c$$

- There are more options than friends and coworkers.
- We can consider any combination of non-overlapping sets of peers
  - friends and coworkers
  - friends only
  - friends or coworkers but not both

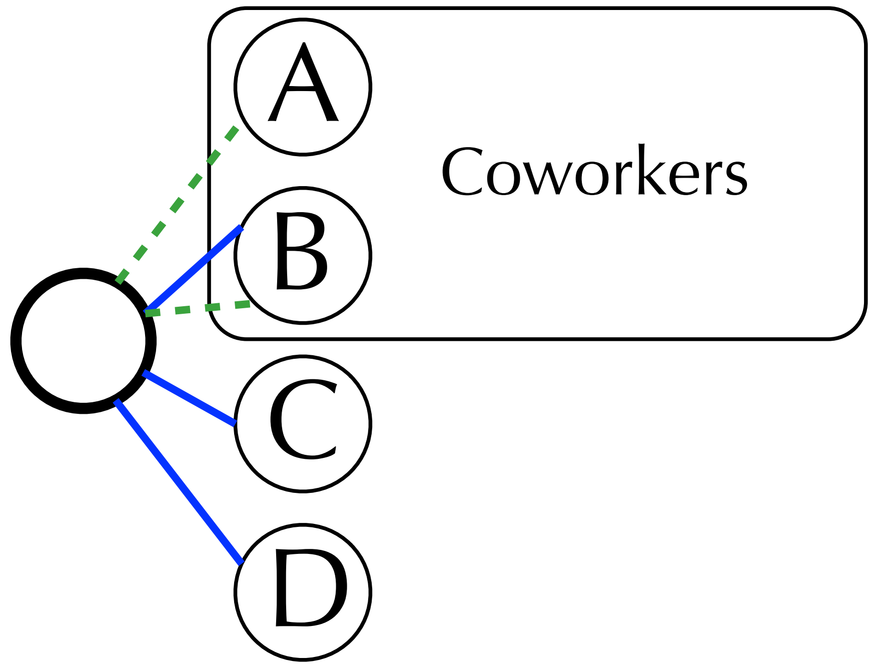
Peer-sets of interest

- Friends (homogeneous network)
- Coworkers (homogeneous network)
- Friends or coworkers (union as a homogeneous network)
- Disjoint
- Friends-coworkers

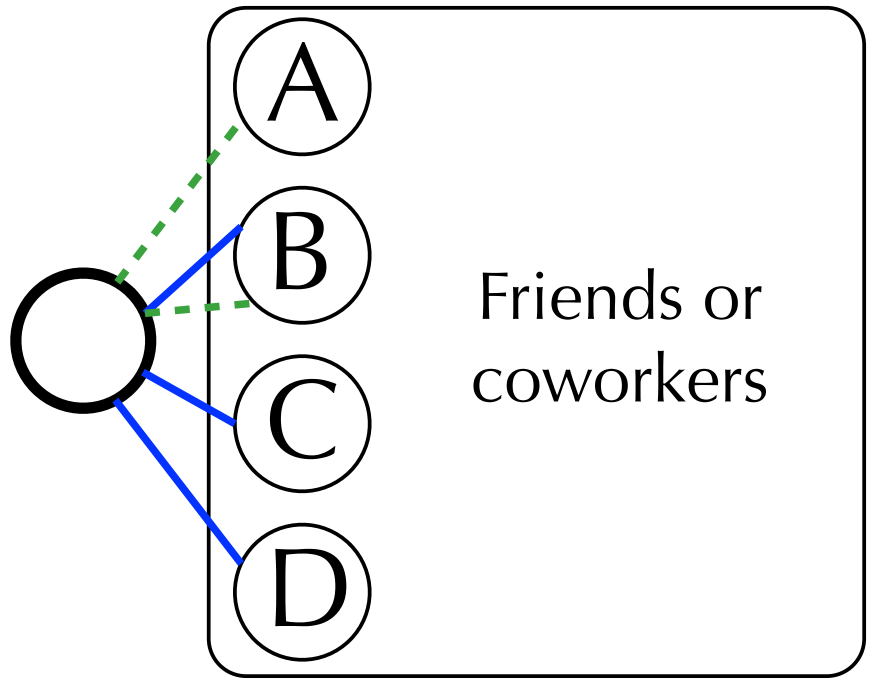
Sets of peers we consider



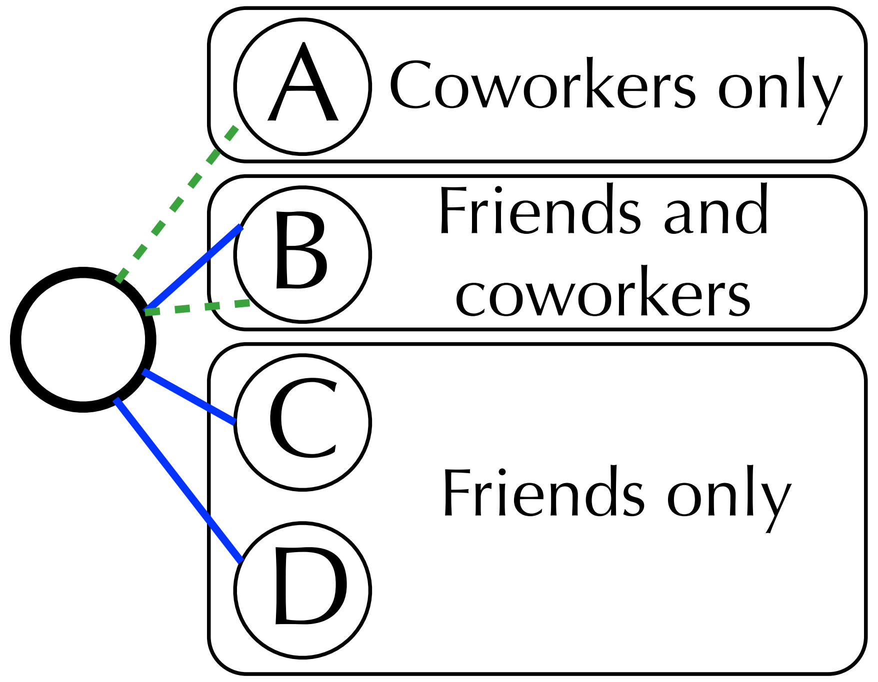
Friends



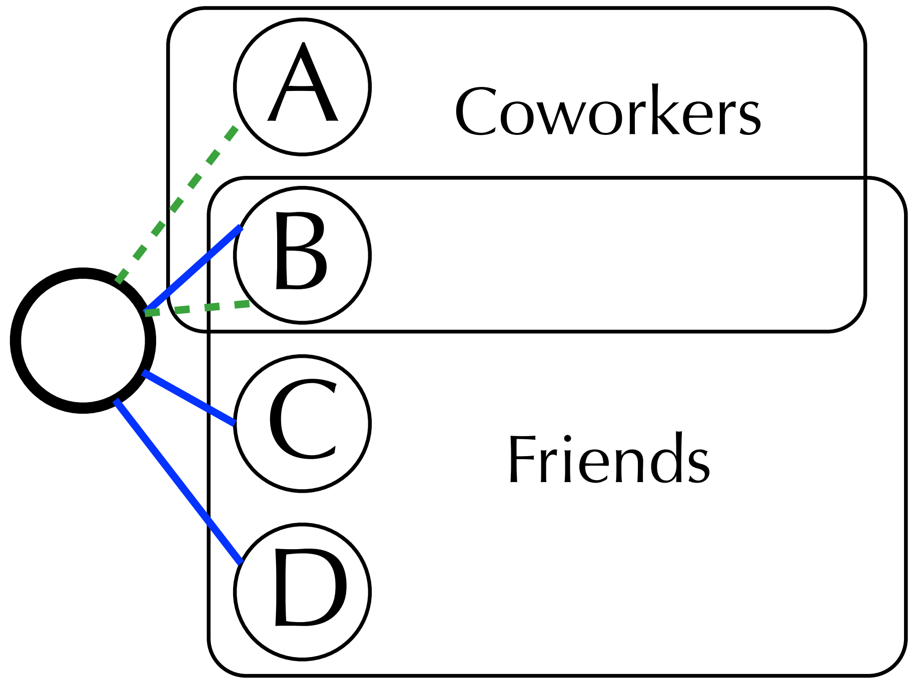
Coworkers



Friends or coworkers



Disjoint



Friends-coworkers

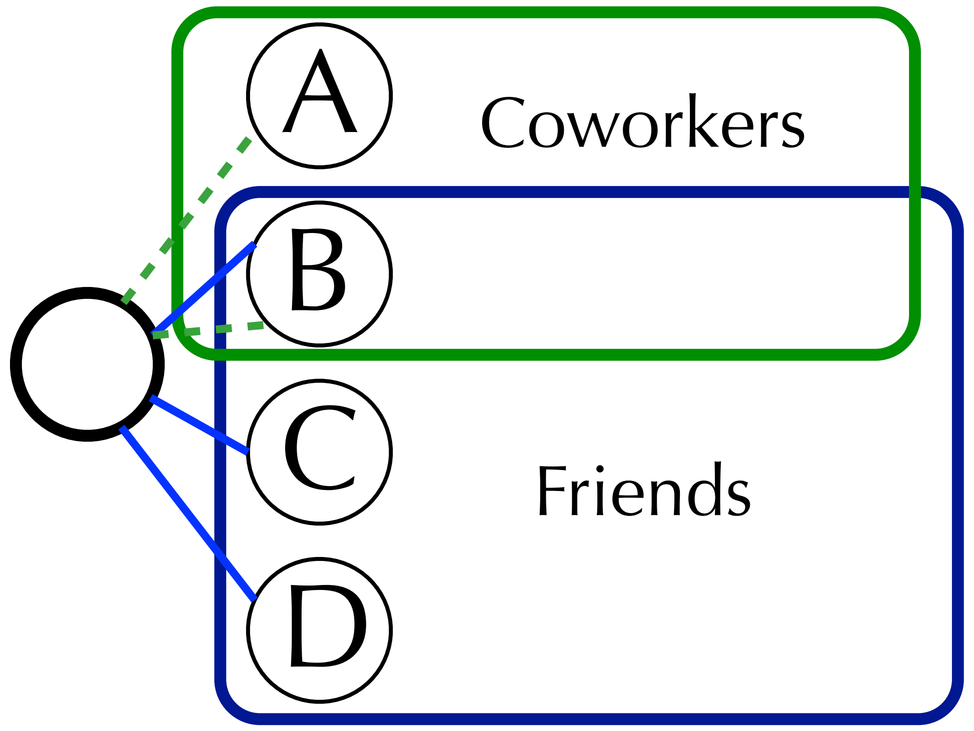
friends
coworkers

Peer sets of interest: Where are they used?

Used for:

- Response functions
- ATE estimators
- Outcome generation

Peer sets of interest: Where are they used?



Friends-coworkers

Used for:

- Response functions

$$g_{f,c}(T_i, \lambda_i) = \alpha + \beta T_i + \gamma^f \lambda_i^f + \gamma^c \lambda_i^c$$

- ATE estimators

$$\tau_{f,c} = \beta + \gamma^f + \gamma^c$$

- Outcome generation

$$O_i = w_0 + w_1 T_i + w_2^f \frac{F[\cdot, i]^\top \mathbf{O_t}}{D_i^F} + w_2^c \frac{C[\cdot, i]^\top \mathbf{O_t}}{D_i^C} + \epsilon$$

How does ignoring/mis-specifying the type of relationships affect estimation of causal effects?

Experiments (synthetic data)

Goal: impact on estimation of causal effects

- Generation of graphs
  - Erdos-Renyi
  - Watts-Strogatz
  - Stochastic block model
- Generation of treatment values
  1. Independent assignment for every node
  2. Graph cluster randomization [Ugander et al. 2013]

Ugander, Karrer, Backstrom, Kleinberg. KDD 2013

Experiments (synthetic data)

- Generation of outcome values

1. Outcome Interference

$$O_{i,t+1} \sim w_0 + w_1 T_i + f(O_{peers\_of\_i,t}) + \epsilon$$

2. Treatment Interference

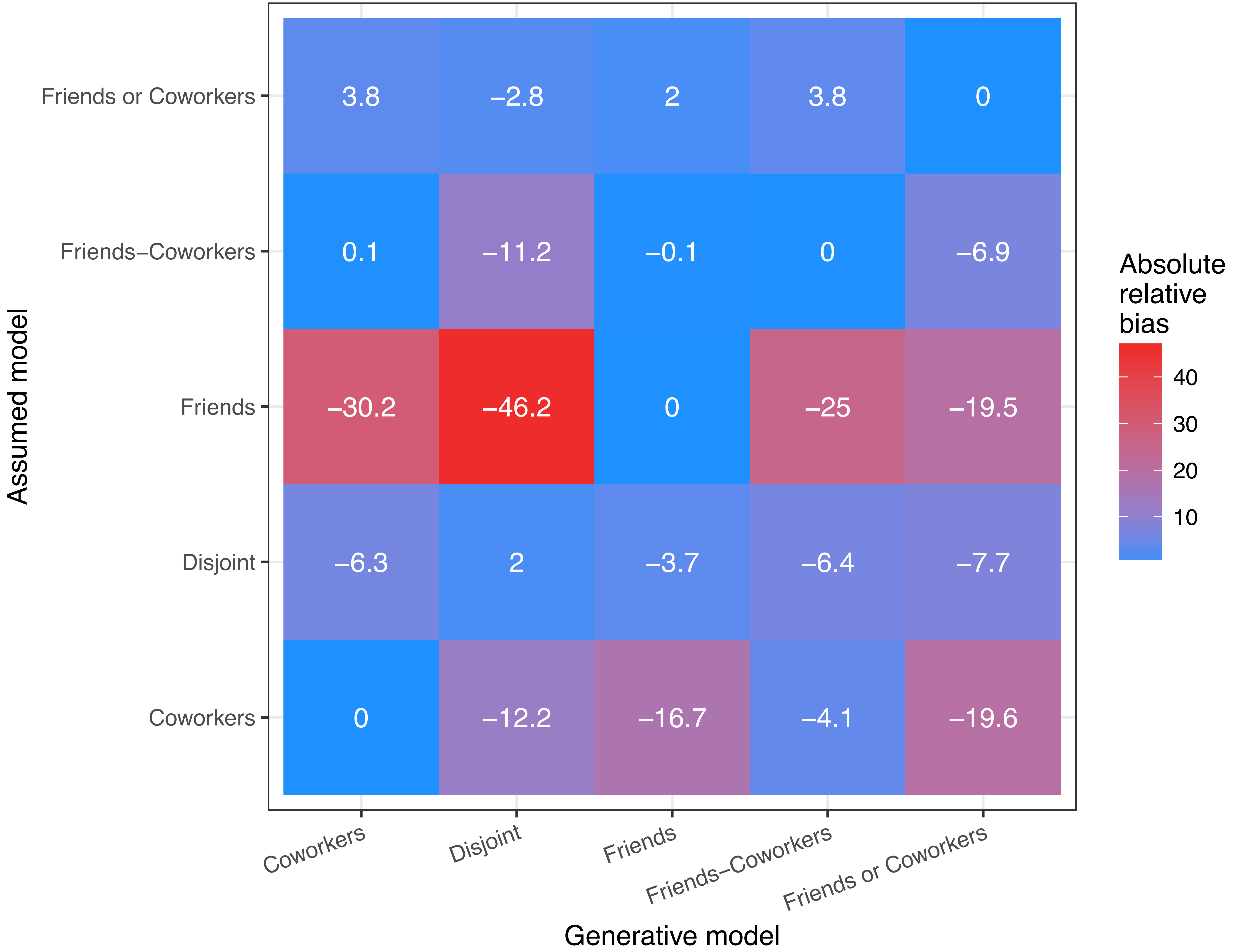
$$O_i \sim w_0 + w_1 T_i + f(T_{peers\_of\_i}) + \epsilon$$

where: $\epsilon = \beta_\epsilon \mathcal{N}(0,1)$

Results

Experiment configuration:

- Graph model: Watts-Strogatz
- Treatment assignment: Graph cluster randomization
- Treatment probability: 0.5
- Outcome generation: Treatment interference


Assumed model
Friends or Coworkers
Friends–Coworkers
Friends
Disjoint
Coworkers
3.8
–2.8
2
3.8
0
0.1
–11.2
–0.1
0
–6.9
–30.2
–46.2
0
–25
–19.5
–6.3
2
–3.7
–6.4
–7.7
0
–12.2
–16.7
–4.1
–19.6
Coworkers
Disjoint
Friends
Friends–Coworkers
Friends or Coworkers
Generative model
Absolute
relative
bias
40
30
20
10

Results

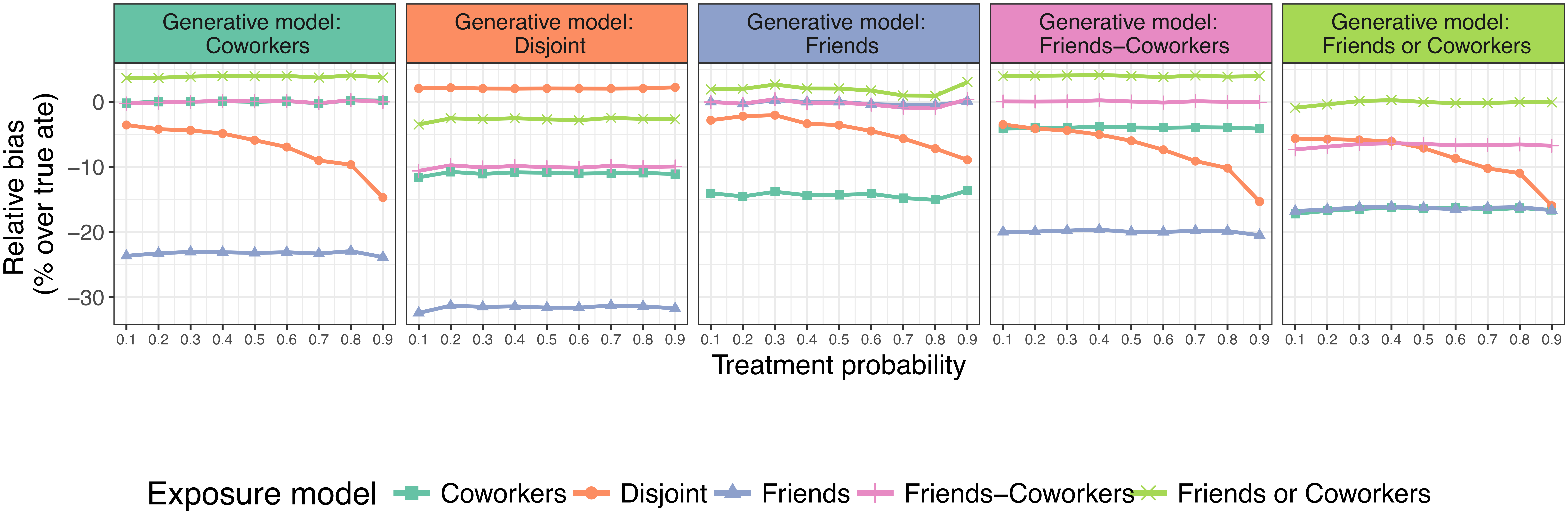
Experiment configuration:

- Graph model: Watts-Strogatz
- Treatment assignment: Graph cluster randomization
- Treatment probability: 0.5
- Outcome generation: Treatment interference

Results

Experiment configuration:

- Graph model: Watts-Strogatz
- Treatment assignment: Graph cluster randomization
- Treatment probability: **varying**
- Outcome generation: Treatment interference


Generative model: Coworkers
Generative model: Disjoint
Generative model: Friends
Generative model: Friends–Coworkers
Generative model: Friends or Coworkers
Relative bias (% over true ate)
0
−10
−20
−30
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
Treatment probability
Exposure model
Coworkers
Disjoint
Friends
Friends–Coworkers
Friends or Coworkers

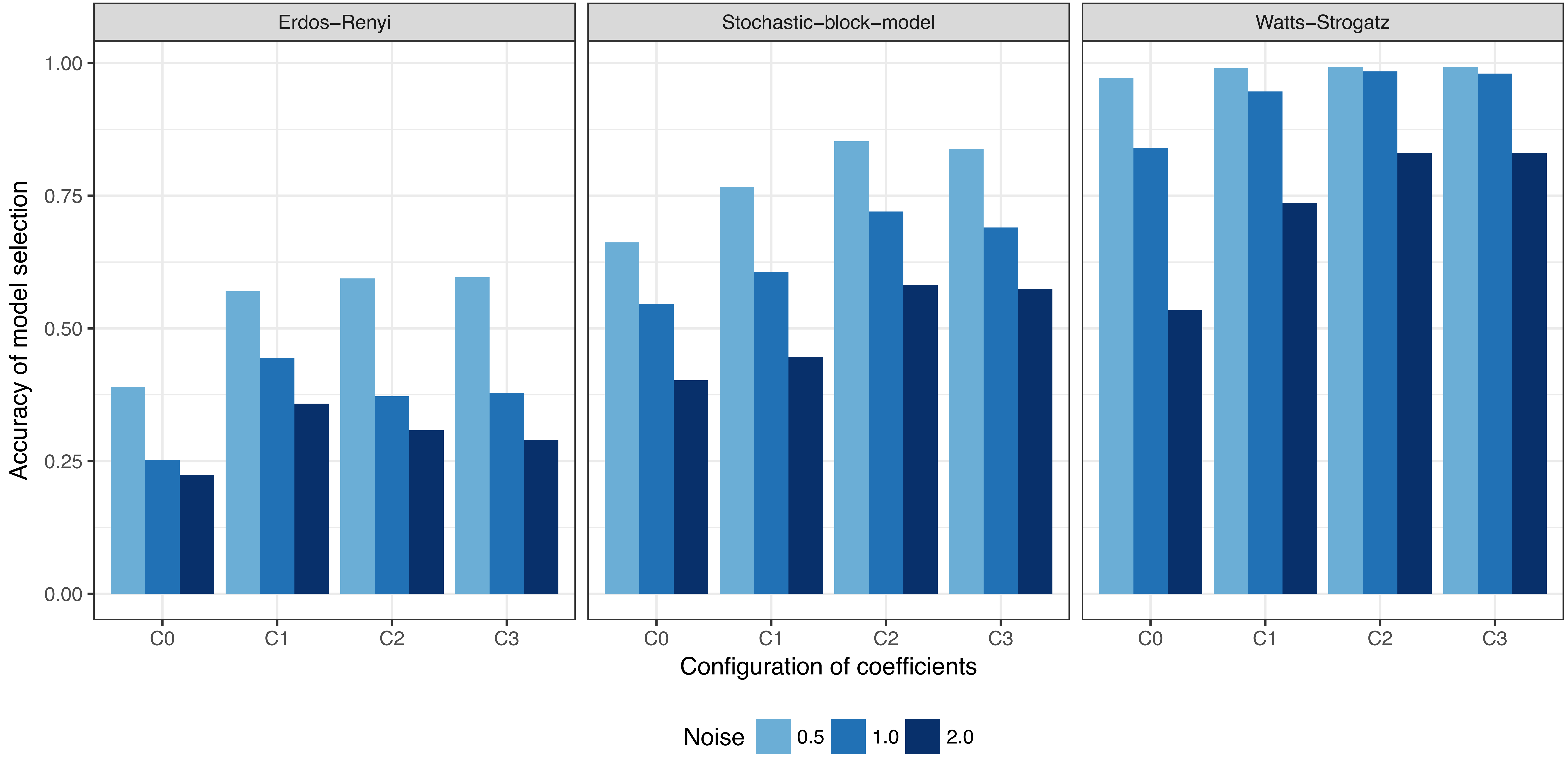
Model selection

Given a set of alternative models,
is it possible to identify the true generating model?

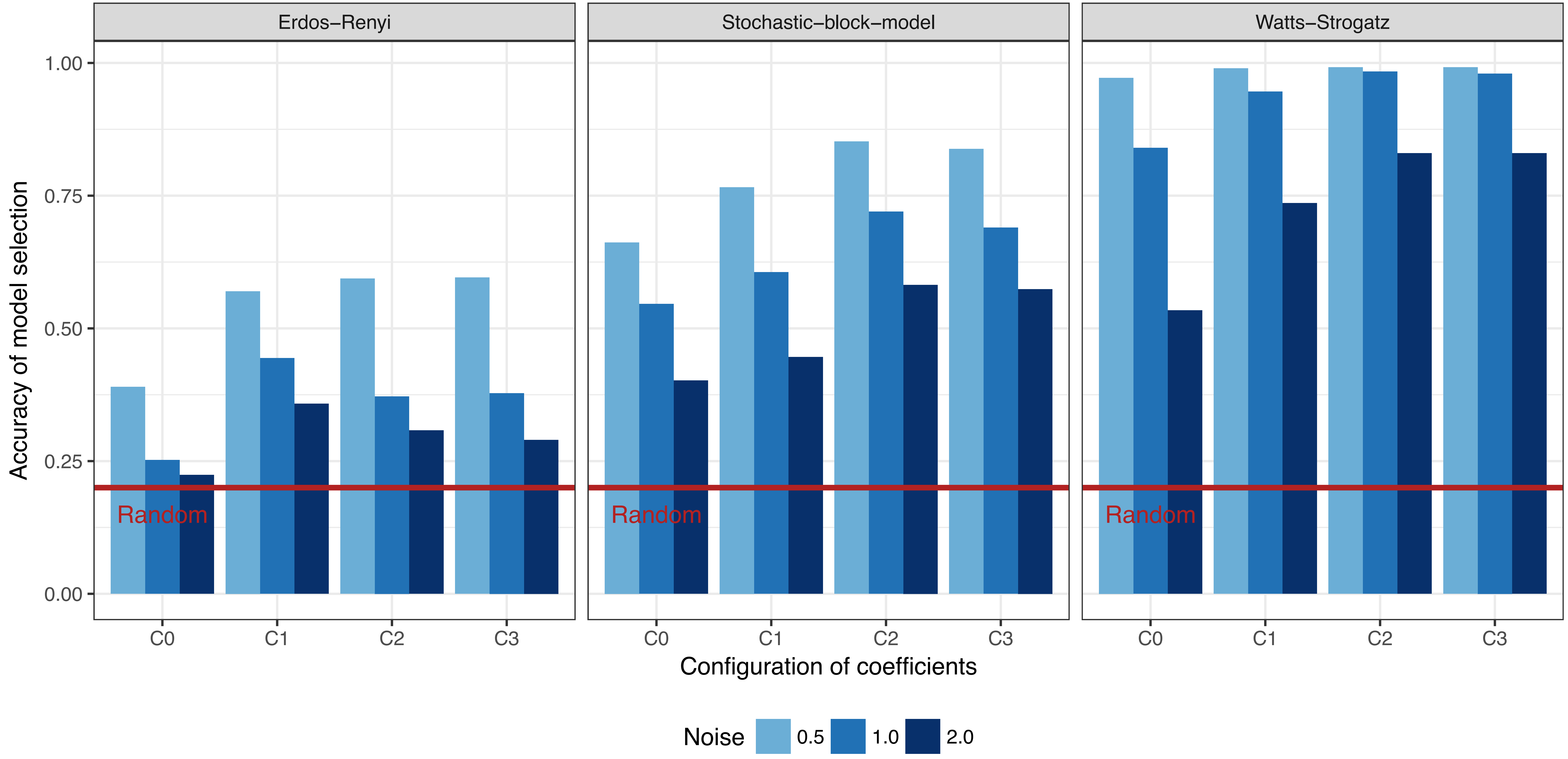
Procedure:

- Generate synthetic networks and synthetic data (as before).
- Compute BIC for each of the five alternative models.
- Select model with the lowest BIC.

Model selection



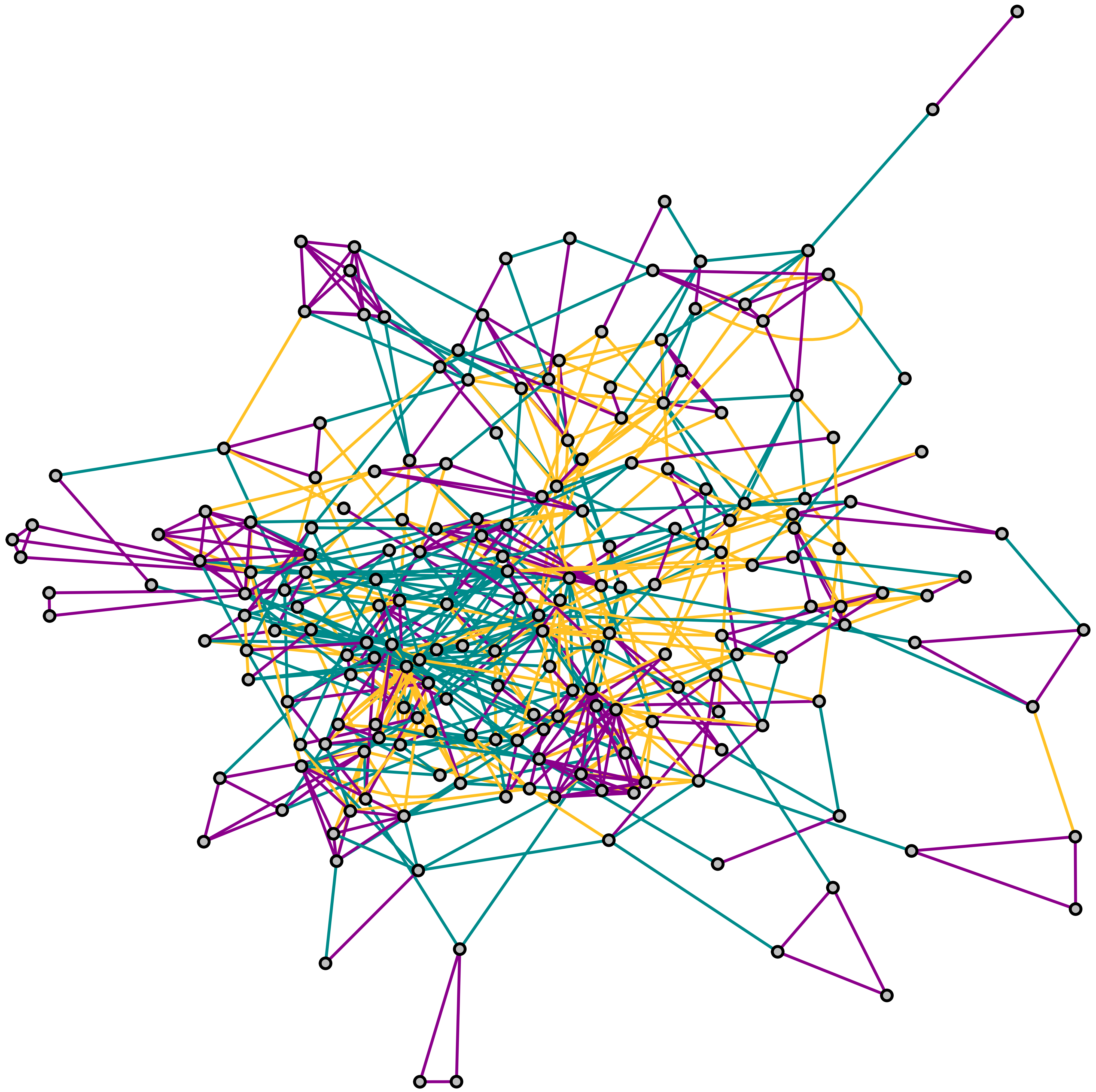
Model selection



Real data

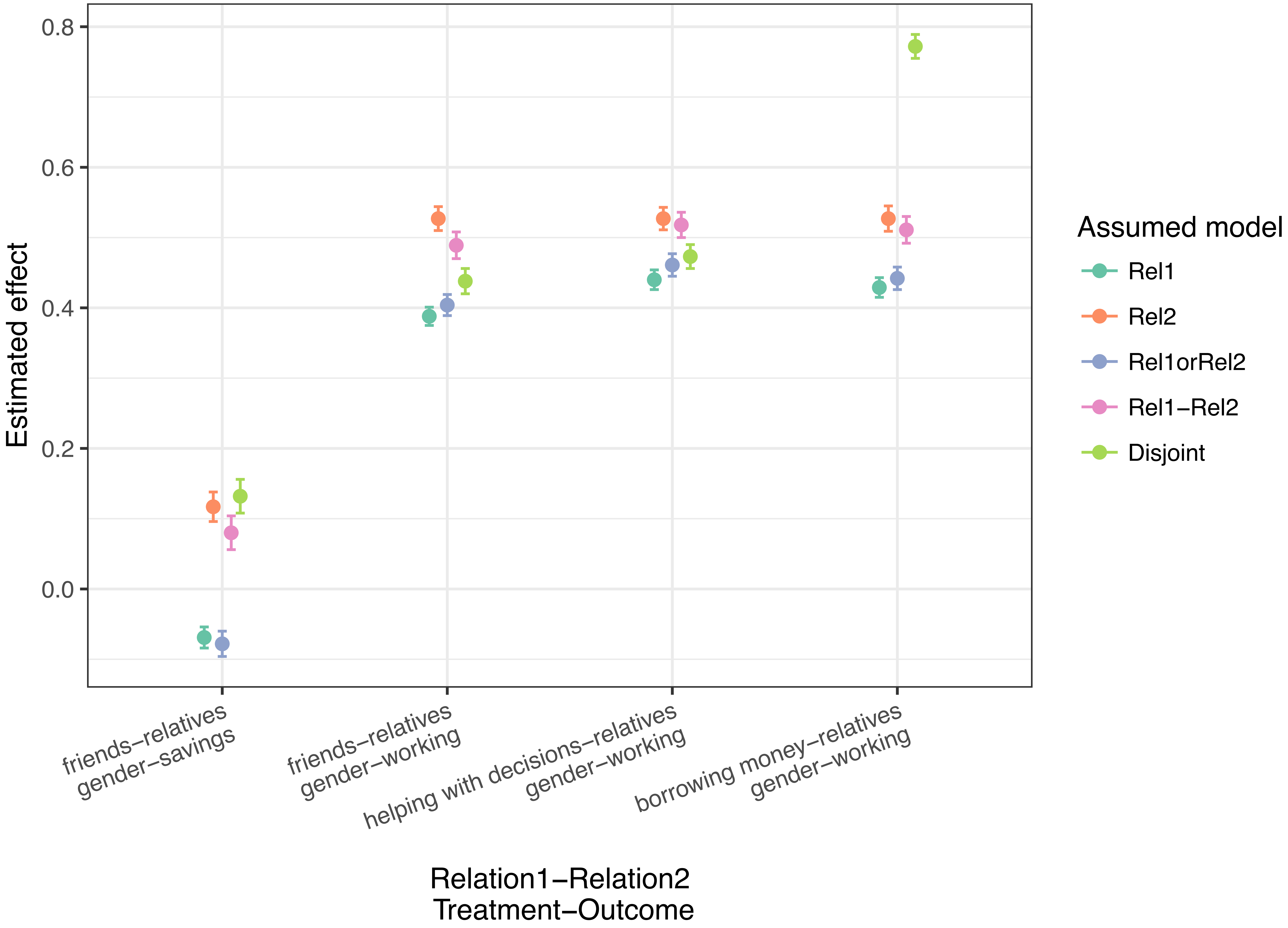
- Study on the diffusion of micro financing loans through various social networks
- Survey conducted in 75 villages in southern India
- Village-level survey and follow-up survey on a subsample of individuals for each village
- Individual surveys identify 13 types of social relationships (e.g., friends, relatives, borrowing money from, going to temple with)
- Individual's attributes (age, gender, etc)

Real heterogeneous network



Experimental setup for real data

- Several pairs of social relationships
- Combinations of treatment-outcome variables
- Estimate effect using different response functions


0.8
0.6
0.4
0.2
0.0
Estimated effect
friends–relatives
gender–savings
friends–relatives
gender–working
helping with decisions–relatives
gender–working
borrowing money–relatives
gender–working
Relation1–Relation2
Treatment–Outcome
Assumed model
Rel1
Rel2
Rel1orRel2
Rel1–Rel2
Disjoint

Summary

- Recent work has extended causal inference frameworks for network data.
- We address the case of heterogeneous networks and causal effect estimation in this framework.
- Mis-specifying the relational structure of causal dependence can lead to significant bias.
- Model selection for distinguishing among candidate response functions.

Directions for future work

- Formal characterization of bias and variance of ATE estimators for heterogeneous networks
- Interactions of relational semantics (effect present from multiple relational phenomena)
- Measure of model selection for relational data
- Fully automated methods for choosing appropriate response functions
- Extending A/B testing framework for heterogeneous networks

Questions?

Thank you!