KU LEUVEN

CESO
Centre for Sociological Research

Probing causal inference in the face of interference

Simulations of social networks with ERGMs

Martin Lukac

martin.lukac@kuleuven.be

@mblukac

BACKGROUND

The **potential outcomes** (PO) framework (Rubin 1974) is largely considered the golden standard in the social sciences; formal theory for inferring **causation in social networks** is, however, still in its infancy.

PO commonly employs a **"no interference" assumption** (SUTVA; Cox 1958), stating that outcomes of one individual are not influenced by the exposures of others. Such assumption is implausible to hold in many social scientific settings (Sobel 2006; VanderWeele & An 2013).

Various adjustments have been suggested in the literature, but so far very few can be realistically **used in observational survey research** due to lack of survey data that explicitly collect information on networks.

In this project, we propose a **computational methodology** for the social sciences by **using simulation for estimating lower and upper bounds of interference in causal inference** studies based on large survey observational datasets.

EMPIRICAL EXAMPLE

The method is illustrated on the case of **attitudes towards welfare** and **redistribution** to the **unemployed**.

Economic self-interest theory postulates that an individual is more likely to support redistribution towards a needy group if they are themselves part of it (see Alt & Iversen 2016). The sociological literature, however, maintains that **self-interest covers also people from respondents' social circles**—friends and family (see van Oorschot 2013). Unemployment of a close person is as well likely to change our attitudes in favour of redistribution. This question can be framed as a **problem of interference**, as treatment (unemployment) of close relatives can influence our own attitudes towards redistribution. Moreover unemployment is not randomly assigned and is more likely to occur for some social groups—e.g. with low education, that are in return more likely to be friends/family due to **homophily** and **assortative mating**.

To date, this question remains unsolved mainly due to **lack of data on redistribution attitudes and social networks**. Hence, estimation of the self-interest effect continues to be confounded by interference:

$Y(\boldsymbol{z})$	Potential outcome
$= Y(0,\boldsymbol{0})$	Baseline
$+ \big(Y(1,\boldsymbol{0}) - Y(0,\boldsymbol{0})\big)$	Direct effect
$+ \Big(Y\big(0,\boldsymbol{z}_{\mathcal{N}_i}\big) - Y(0,\boldsymbol{0})\Big)$	Interference

(see Sussman & Airoldi 2017)

Proposed methodology aims to solve this short-coming by **combining information from two distinct datasets via a set of overlapping auxiliary variables**.

Boils down to network prediction problem — how can we predict the unobserved network between individuals?

Figure 1. *Illustration of interference in a small network.*

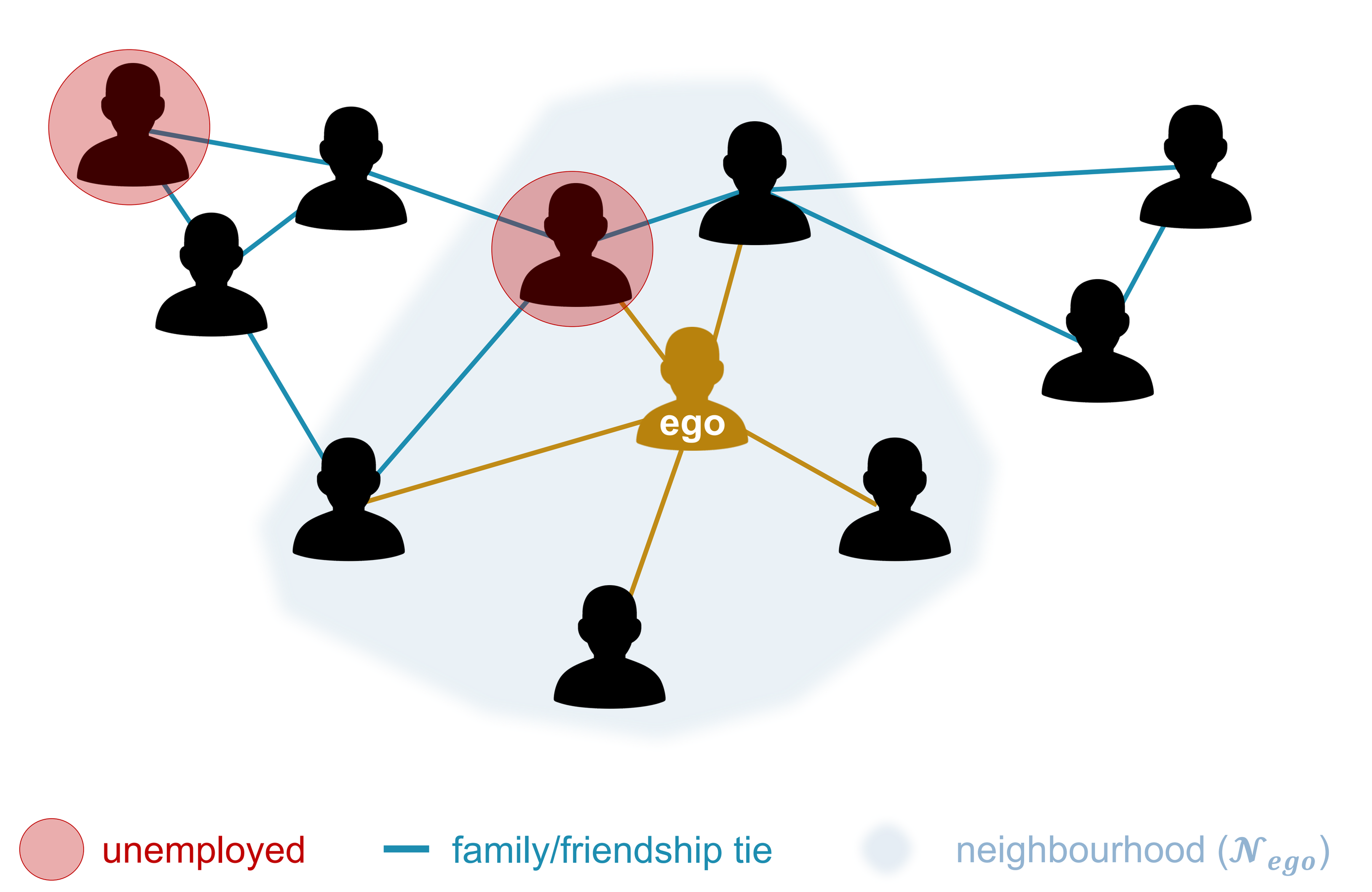
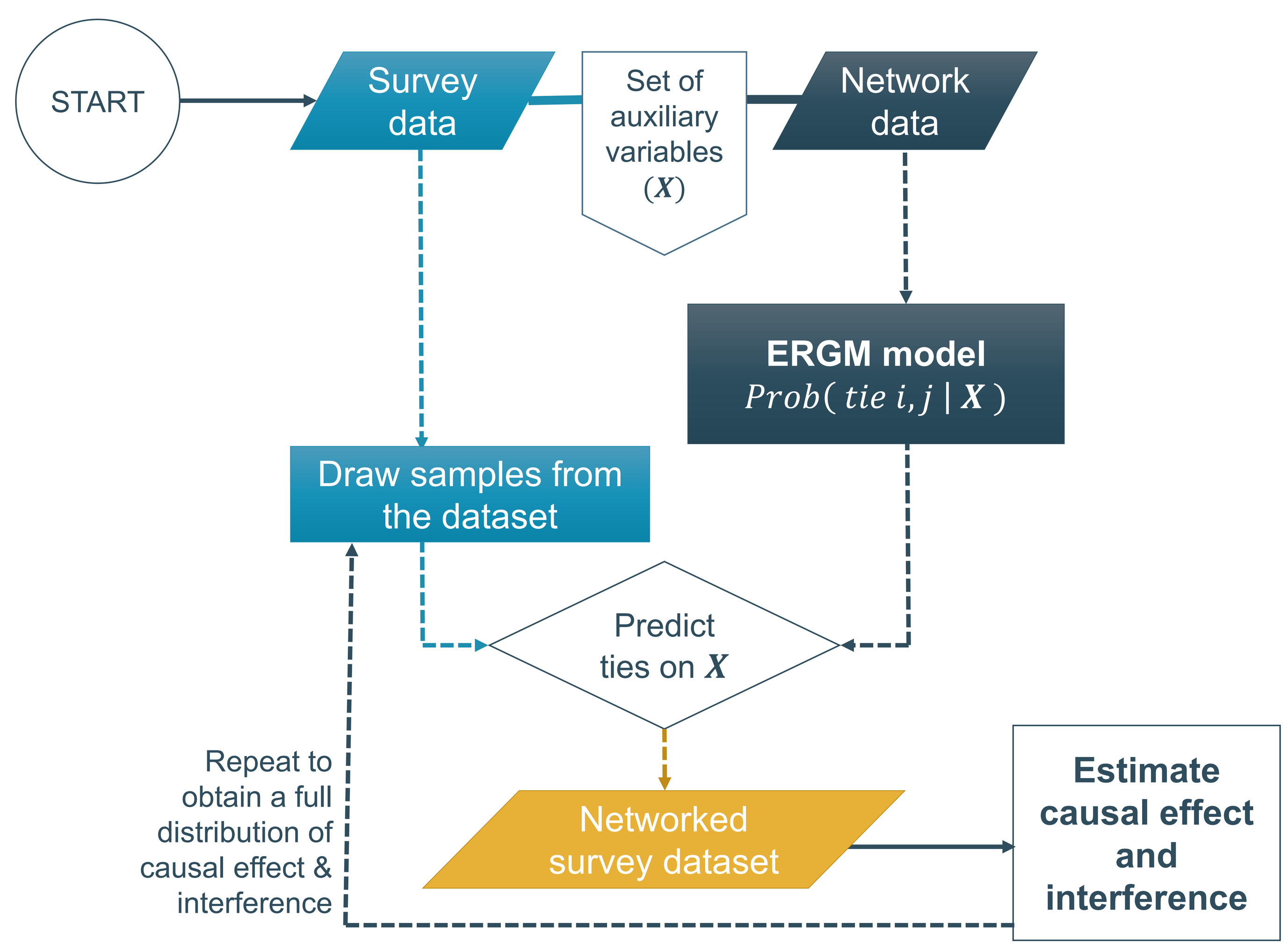


Figure 2. *Simulation of networked survey data from a social survey and network data that overlap on a set of auxiliary variables ($\boldsymbol{X}$).*



METHOD & ESTIMATION

To combine survey data with network data (see Fig. 2), we first estimate an **Exponential Random Graph Model** (ERGM) on a set of auxiliary variables that occur in both datasets.

Second, we **draw a sample** from the survey data to create an artificial population with empirical distributions of the treatment and outcome.

Third, we use the ERGM to **calculate probabilities of a tie occurring** between all observed pairs in the sampled data. Based on these probabilities we simulate datasets with different realizations of ties given the probabilities.

Finally, in each of these simulation runs, we estimate the **causal effect and the effect of interference**. This results in a distribution that allows us to estimate the lower and upper bounds of interference on the causal effect of interest.

Given the survey is **representative** of the population, observations of the dataset can be used as **artificial stand-ins for the unobserved part of the population**. Although it is unlikely that these individuals would be actually connected, the **repeated simulation and representativeness of the sample in aggregate provide similar properties to the population**. This allows us to effectively combine information about the distributions of the treatment and outcome variables in the population with information from data on social networks.

Estimation of the causal effect and interference effect can draw on the available literature (Sussman and Airoldi 2017; Hudgens and Halloran 2008; VanderWeele 2015).

KEY QUESTIONS & CHALLENGES

In the presence of interference, two key questions arise:

(1.) Which units' treatment can affect ego's outcome?

(2.) How can treatments affect ego's outcomes?

Challenges to the method:

- **Computational time** — estimating ERGMs on large graphs is computationally expensive
- **Small-world property** — estimation problem for larger neighbourhoods
- **Network model misspecification** — unknown biases potentially introduced to the model