# Causal Inference from Network Data

Elena Zheleva
ezheleva@uic.edu
University of Illinois at Chicago
Chicago, IL

David Arbour
darbour26@gmail.com
Adobe Research
San Jose, CA

## ABSTRACT

This tutorial presents state-of-the-art research on causal inference from network data in the presence of interference. We start by motivating research in this area with real-world applications, such as measuring influence in social networks and market experimentation. We discuss the challenges of applying existing causal inference techniques designed for independent and identically distributed (i.i.d.) data to relational data, some of the solutions that currently exist and the gaps and opportunities for future research. We present existing network experiment designs for measuring different possible effects of interest. Then we focus on causal inference from observational data, its representation, identification, and estimation. We conclude with research on causal discovery in networks.

## CCS CONCEPTS

• **Computing methodologies** → **Causal reasoning and diagnostics**; • **Mathematics of computing** → **Causal networks**.

## KEYWORDS

causal inference; interference; graphs; social networks

**ACM Reference Format:**
Elena Zheleva and David Arbour. 2021. Causal Inference from Network Data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3447548.3470795

## TUTORIAL OVERVIEW

Causal inference is central to a vast number of scientific and industrial applications. The goal of causal inference is to estimate the effect of an unseen intervention on one or more variables of interest (commonly referred to as causes or treatments) on another set of variables of interest, commonly referred to as outcomes [18, 27, 33]. The fundamental problem of causal inference is that that while we can observe the factual outcome for a given unit under one treatment assignment, the counterfactual, the outcome under any other treatment assignment, is unobserved by definition [17, 36]. For example, we can measure whether a person got sick after getting vaccinated but cannot measure whether they would have gotten sick without receiving the vaccine. Multiple disciplines have developed rich literatures in causal inference, including statistics [18], economics [2], epidemiology [16], philosophy [43], computer science [33, 34], and the social sciences [27]. This tutorial largely focuses on methods of interest to the computer science community.

To capture the noise, heterogeneity, and complex relationships in real-world data, a common practice is to model data sources as relational systems and to reason about them probabilistically. Relations in data can be represented through heterogeneous networks in which nodes represent interdependent entities, such as people, companies, websites, and diseases, while edges denote different relationships between these entities, such as friendship, hyperlink, contribution, and spread of disease. Some example applications of causal inference in networks include measuring influence in social networks [5, 8, 28], information diffusion [46], and marketplace experimentation [19].

*Interference* (also known as spillover or network effects), where the outcome of a treated node depends not only on its treatment but also on the treatment and outcome of neighboring nodes, is commonly observed in relational systems. For example, a social media campaign about the benefits of vaccines may lead to a user deciding not to vaccinate themselves if that user is also seeing anti-vaccination posts by their friends on social media. Interference breaks the Stable Unit Treatment Value Assumption (SUTVA) of causal inference, which requires that the outcome of a given unit depends only on the treatment to which they were assigned, and can lead to biased causal effect estimation. There are three main types of interference: direct interference, interference by contagion, and allocation interference [32]. Direct interference refers to the treatment of one or more nodes in the neighborhood of an ego node affecting the outcome of that ego node (e.g., being shown an anti-vaccination post and in turn sharing it with friends and affecting their decision whether to vaccinate). Contagion refers to the outcome of one or more nodes affecting the outcome of another node (e.g., deciding not to vaccinate and affecting a friend's decision whether to vaccinate). Allocational interference is the most complex of the three, and it refers to group composition influences on individual outcomes. The presence of interference also introduces a new set of causal estimands which range from the individual treatment effect to peer effects and total treatment effect.

Randomized controlled experiments, also known as A/B tests, are considered the gold standard for inferring causality. However, accounting for interference is challenging even in the context of randomized controlled trials. A number of network experiment designs have been developed to address interference in the general network setting [3, 7, 10–14, 29, 30, 37, 47] and the case of bipartite networks which commonly occurs in marketplace experimentation [9, 35].

It is not always possible to design and run network experiments due to ethical concerns, cost, or mere impossibility (e.g., due to immutable characteristics). To avoid these limitations, researchers resort to *quasi-experimental designs (QED)* which aim to estimate


*KDD '21, August 14–18, 2021, Virtual Event, Singapore*

ACM ISBN 978-1-4503-8332-5/21/08.
https://doi.org/10.1145/3447548.3470795

causal effects from observational data which can suffer from model dependence [38]. Central to estimation of causal effects is the causal model representation. There are three main relational representations included in this tutorial: blocks, abstract ground graphs, and segregated graphs. A block is a set of variables which define the causal graph and are repeatable (e.g., a pair of connected nodes). Causal effects are estimated in expectation across blocks. Abstract ground graphs [23, 24] provide a lifted representation of directed acyclic multi-relational systems such that conditional independence semantics on the lifted representation faithfully represent conditional independence facts on the individual level, which enables scalable causal reasoning for multi-relational data. Finally, researchers have also studied modeling relational data through chain graphs and segregated graphs [6, 31, 40–42], where explicitly noncausal undirected edges represent feedback between nodes. Dependent on the causal model, causal effects may be fully identifiable, partially identifiable, or not identifiable at all [4, 32, 39–41, 45].

When the causal model is unknown a priori, it can be learned from data under certain assumptions. The goal of causal discovery is to learn a causal graph in which the causal relations are asymptotically correct and describe the causal process that generated the data [44]. Multiple algorithms have been developed to learn causal model structure from data but these algorithms typically assume i.i.d. data [15, 44]. When the data breaks the SUTVA assumption and instances can influence each other's treatments or outcomes, these algorithms no longer apply. Causal structure learning algorithms for relational data, also known as relational causal discovery (RCD) algorithms, aim to learn the abstract ground graph from the relational skeleton. Existing RCD algorithms [20, 21, 24–26] focus on adapting the constraint-based PC algorithm [43] for i.i.d. data to relational domains. These algorithms assume that all relevant variables have been measured and either rely on a relational conditional independence oracle or on tests for i.i.d. data. Relational dependence tests have been developed recently to make network causal discovery more practical [1, 22].

## REFERENCES


[1] Ragib Ahsan, Zahra Fatemi, David Arbour, and Elena Zheleva. 2021. Nonparametric inference of relational dependence. *arXiv preprint* (2021).
[2] Joshua D Angrist and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.
[3] David Arbour, Drew Dimmery, and Anup Rao. 2021. Efficient Balanced Treatment Assignments for Experimentation. *AISTATS* (2021).
[4] David Arbour, Dan Garant, and David Jensen. 2016. Inferring network effects in observational data. In *KDD*.
[5] Eytan Bakshy, Dean Eckles, Rong Yan, and Itamar Rosenn. 2012. Social influence in social advertising: evidence from field experiments. In *EC*.
[6] Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. 2019. Causal Inference Under Interference And Network Uncertainty. In *UAI*.
[7] Alex Chin. 2019. Regression adjustments for estimating the global treatment effect in experiments with interference. *Journal of Causal Inference* 7, 2 (2019).
[8] Nicholas A Christakis and James H Fowler. 2007. The spread of obesity in a large social network over 32 years. *NEJM* 357, 4 (2007), 370–379.
[9] Nick Doudchenko, Minzhengxiong Zhang, Evgeni Drynkin, Edoardo Airoldi, Vahab Mirrokni, and Jean Pouget-Abadie. 2020. Causal inference with bipartite designs. *arXiv preprint arXiv:2010.02108* (2020).
[10] Dean Eckles, Brian Karrer, and Johan Ugander. 2017. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference* 5, 1 (2017).
[11] Dean Eckles, René F Kizilcec, and Eytan Bakshy. 2016. Estimating peer effects in networks with peer encouragement designs. *PNAS* (2016).
[12] Zahra Fatemi and Elena Zheleva. 2020. Minimizing Interference and Selection Bias in Network Experiment Design. In *ICWSM*.
[13] Zahra Fatemi and Elena Zheleva. 2020. Network Experiment Design for Estimating Direct Treatment Effects. In *MLG*.
[14] Huan Gui, Ya Xu, Anmol Bhasin, and Jiawei Han. 2015. Network a/b testing: From sampling to estimation. In *WWW*.
[15] Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. 2018. Causal structure learning. *Annual Review of Statistics and Its Application* 5 (2018), 371–391.
[16] Miguel A Hernán and James M Robins. 2020. Causal inference: what if. *Boca Raton: Chapman & Hill/CRC* 2020 (2020).
[17] Paul W Holland. 1986. Statistics and Causal Inference. *J. Amer. Statist. Assoc.* 81, 396 (1986), 945–960.
[18] G. Imbens and D. Rubin. 2015. *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction.* Cambridge Univ Press.
[19] Ramesh Johari, Hannah Li, Inessa Liskovich, and Gabriel Weintraub. 2020. Experimental design in two-sided platforms: An analysis of bias. *arXiv preprint arXiv:2002.05670* (2020).
[20] Sanghack Lee and Vasant Honavar. 2016. On Learning Causal Models from Relational Data. In *AAAI*. 3263–3270.
[21] Sanghack Lee and Vasant Honavar. 2019. Towards Robust Relational Causal Discovery. In *UAI*.
[22] Sanghack Lee and Vasant G Honavar. 2017. A kernel Conditional Independence Test for Relational Data. In *UAI*.
[23] Marc Maier. 2014. *Causal discovery for relational domains: Representation, reasoning, and learning.* Ph.D. Dissertation. University of Massachusetts, Amherst.
[24] Marc Maier, Katerina Marazopoulou, David Arbour, and David Jensen. 2013. A Sound and Complete Algorithm for Learning Causal Models from Relational Data. In *UAI*.
[25] Mark Maier, Brian Taylor, Hüseyin Oktay, and David Jensen. 2010. Learning causal models of relational domains. In *AAAI*.
[26] Katerina Marazopoulou, Marc Maier, and David Jensen. 2015. Learning the structure of causal models with relational and temporal dependence. In *UAI*.
[27] Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference.* Cambridge University Press.
[28] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.
[29] Preetam Nandy, Kinjal Basu, Shaunak Chatterjee, and Ye Tu. 2019. A/B testing in dense large-scale networks: design and inference. *arXiv preprint arXiv:1901.10505* (2019).
[30] Molly Offer-Westort and Drew Dimmery. 2021. Experimentation for Homogenous Policy Change. *arXiv preprint arXiv:2101.12318* (2021).
[31] Elizabeth L Ogburn, Ilya Shpitser, and Youjin Lee. 2018. Causal inference, social networks, and chain graphs. *arXiv preprint arXiv:1812.04990* (2018).
[32] Elizabeth L Ogburn, Tyler J VanderWeele, et al. 2014. Causal diagrams for interference. *Statistical science* 29, 4 (2014), 559–578.
[33] Judea Pearl. 2009. *Causality.* Cambridge Univ Press.
[34] Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms.* MIT Press.
[35] Jean Pouget-Abadie, Kevin Aydin, Warren Schudy, Kay Brodersen, and Vahab Mirrokni. 2019. Variance Reduction in Bipartite Experiments through Correlation Clustering. In *NeurIPS*.
[36] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 5 (1974), 688.
[37] Martin Saveski, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M Airoldi. 2017. Detecting network effects: Randomizing over randomized experiments. In *KDD*.
[38] Usman Shahid and Elena Zheleva. 2019. Empirical Study of Model Dependence in Counterfactual Learning from Networks. In *AAAI Spring Symposium on Beyond Curve Fitting: Causation, Counterfactuals, and Imagination-based AI (AAAI-WHY)*.
[39] Cosma Shalizi and Andrew Thomas. 2011. Homophily and contagion are generically confounded in observational social network studies. *SMR* 40 (2011).
[40] Eli Sherman, David Arbour, and Ilya Shpitser. 2020. General Identification of Dynamic Treatment Regimes Under Interference. In *AISTATS*.
[41] Eli Sherman and Ilya Shpitser. 2018. Identification and estimation of causal effects from dependent data. In *NeurIPS*. 9424–9435.
[42] Eli Sherman and Ilya Shpitser. 2019. Intervening on Network Ties. In *UAI*.
[43] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, prediction, and search.* MIT press.
[44] Peter Spirtes and Kun Zhang. 2018. Search for causal models. *Handbook of Graphical Models* (2018), 457–488.
[45] Eric J Tchetgen Tchetgen, Isabel Fulcher, and Ilya Shpitser. 2017. Auto-g-computation of causal effects on a network. *arXiv preprint arXiv:1709.01577* (2017).
[46] Christopher Tran and Elena Zheleva. 2020. Heterogeneous Threshold Estimation for Linear Threshold Modeling. In *MLG*.
[47] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. 2013. Graph cluster randomization: Network exposure to multiple universes. In *KDD*.