Midterm Report
Causal Inference II Spring 2022

March 25, 2022

Team

Reetahan Mukhopadhyay, rm3873

Working Title

Empirical Application and Comparison of Estimating Peer Influence on Social Networks

Problem Background

As people, we are generally social creatures that live in societal structure, whose reach we may find difficult to escape. Our decisions and behaviors are typically not made in isolation, but can stem from additional implicit and external causes. One key question we may wish to understand is the magnitude of influence our friends may have on our decisions compared to other latent factors. Applying a causal lens to the realm of social networks, network effects such as contagion and influence and general network theory, has become a recent phenomenon as it was difficult to study due to being a system with dependence among observations without necessarily having a latent underlying Euclidean geometry [7], and there have been several studies that explore the notion of using this framework to explore this query. One such recent work [3] aims to use a network embedding as a way to address confounding from the latent characteristics that impact individuals as well as homophily - the notion those same characteristics impact the existence of dyadic relationship itself. This work defines and demonstrates the validity of an estimator for peer influence, and presents steps to find a valid embedding, but does not complete providing some empirical demonstration of their theory.

In this project, we aim to first formally discuss several comments and concerns from a theoretical perspective in regards to comparing the causal and statistical assumptions and methodology in [3], to other literature in the area. Next, we will provide an implementation in code of the process described in [3] to develop an embedding that can be used to generate the desired estimate of the causal effect of the treatment assigned to an individual upon the outcome of that individual's neighbor. Finally, we will compare these results to other primarily causal as well as some less causal or non-causal methods exploring the problem, that we can provide empirical validation for by using existing code or additional implementation ourselves of these other methodologies.

Summarization of Paper Theory

The paper defines a scenario in which we have a network G_n of n individuals, where connections between individuals are encoded by the presence of an edge between them - consider the corresponding adjacency matrix A - but we let $A_{ii} = 1$. For each node i we consider the vector $O_i = (Y_i, C_i, T_i)$, where Y_i is the observed outcome, T_i is the treatment C_i are unobserved attributes of the individuals that drive homophily. Let $\{C_i\}_i$ represent the unobserved attributes for all individuals in G_n, let the variables $\epsilon_{\{\}}$ be i.i.d exogenous noise, and s_Y represent some sort of summarizing function (such as an average, median, etc). We can generate the structural equation model as

$$C_i \leftarrow f_C[\epsilon_{C_i}]$$

$$A_{ij} \leftarrow f_A[\{C_i\}_i, \epsilon_{ij}]$$
$$T_i \leftarrow f_T[C_i, \epsilon_{T_i}]$$
$$Y_i \leftarrow f_Y[s_Y(T_j : A_{ij} = 1), C_i, \epsilon_{Y_i}]$$

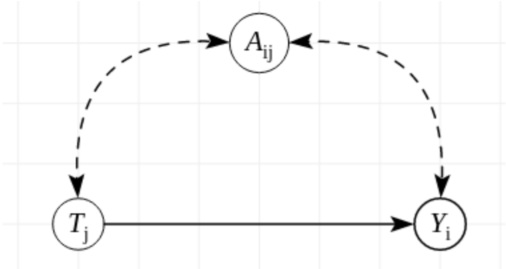
. We define the causal estimand for a treatment $T = t^*$ as

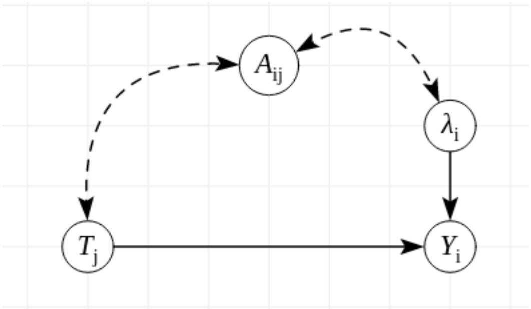
$$\psi_n := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i|do(T = t^*), \{C_i\}_i, G_n]$$

. They provide a theorem and proof that says the estimand ψ_n converges to ψ in probability if $\mathbb{E}[Y_i]^4$ is finite for all nodes i and $D_n = O(n^{1/4})$ where D_n is the maximum degree of any node i in G_n. If we define $\lambda_i \in \mathbf{R}^k$ as the embedding of node i, $V_i := s_Y(T_j : A_{ij} = 1)$, v^* is the value of V under treatment $T = t^*$, and $m_{G_n} := \mathbf{E}[Y_i|V_i, \lambda_i, G_n]$, then the second theorem and proof states that under the assumptions

$$i.A_{ij} \perp Y_i|\lambda_i, T_j$$
$$ii.\forall v_i \mid P(V_i = v^*|\lambda_i(C_i)) > 0$$
$$iii.\lambda_i \ is \ C_i - measurable$$

then we get $\psi_n = \frac{1}{n}\sum_{i=1}^{n} m_{G_n}(t^*, \lambda_i)$. They also defined the causal diagram on the left to represent the current regime, and the right causal diagram to represent the regime after introducing the node embedding





Finally they provide a procedure to generate the estimator $\hat{\psi_n} = \frac{1}{n}\sum_{i=1}^{n} m_{G_n}(t^*, \hat{\lambda_i})$. First, define the sampling algorithm $Sample(G_n, k)$ that returns G_k, a random subgraph of size k from G_n, and randomly divide those nodes into sets I_0 and $I\backslash I_0$. Then define loss function $L(G_l, \lambda, \gamma) = \sum_{i\in I\backslash I_0}(y_i - m(v_i, \lambda_i, \gamma))^2 + \sum_{i,j\in IxI} CrossEntropy(A_{ij}, \sigma(\lambda_i^T\lambda_j))$, and train the model with $\hat{\lambda}, \hat{\gamma} = argmin_{\lambda,\gamma}\mathbf{E}_{G_k}[L(G_k, \lambda, \gamma)]$. Then compute $\hat{m}_{G_n}(t^*, \hat{\lambda_i})$ for each $i \in I_0$. Finally, compute the estimate as $\hat{\psi_n}(I_0) = \frac{1}{|I_0|}\sum_{i\in I_0} \hat{m}_{G_n}(t^*, \hat{\lambda_i})$

Primarily Causal Methods Literature Review

The causal approaches to the matter of understanding take several different approaches to the matter. In [7], which was the source for [3] for their structural equation model setup - although they did not let the network A itself be a random variable and they did allow for the C_j of a neighbor to have impact on the treatment of the individual T_i. There, they aim for semiparamteric (as G_n remains parameterized) estimation of peer influence by developing and proving the validity of consistent and asymptotically normal estimators, that can be used for two sets of situations, one just meant to measure a direct effect and the other to include the presence of homophily. They formally flesh out these out using six assumptions, that also seem to be held true in [3] but they neglected to formally state. They also develop a pathwise differentiable mapping from the space of probability distributions to arise $P(Y)$ to $\mathbb{R}$. They also formalize a confidence interval for their estimator, which is not provided in [3], as well as provide evidence their model makes sense for real social networks, where degree distribution may follow a power law. In [6], they take a simpler approach. They focus on just the graphical criterion - which was covered in [3] and [7] in the causal diagram, for identification through the backdoor. They provide several different causal diagrams to consider, however, to represent the situation at hand. They also suggestions of finding a way to encode direction and magnitude into Y_i.

[21] took an embedding route to find the average treatment effect as well. Like [6] they elected to represent the scenario with a temporal model, unlike [7]. They chose to use network representation learning methods, including matrix factorization-based, random walk-based, and deep learning-based methods, to determine an appropriate network embedding. They also elected to make the graph be non-parametric like [3], although their model for graph generating is not exactly alike. They chose to develop simple linear equations for their structural equations, and then applied Ordinary Least Squares.

[13] chose a more novel paradigm to consider in the search area. They apply a formal framework which leverages a pair of negative control outcome and treatment variables (double negative controls) using a network outcome confounding bridge function to nonparametrically identify causal peer effects, while of course considering homophily. From there, they give a method for a generalized method of moments estimator for the effects, establishing its consistency and asymptotic normality like [7], and give a network heteroskedasticity and autocorrelation consistent variance estimator to generate confidence intervals.

[19] chooses to minimize the number of assumptions it makes in turn for generating just bounds for the causal effect instead of a single point estimand. It highlights the importance of assumptions, pragmatism and realism. They end up formulating an optimization problem that can be resolved using linear programming, similar to our causal bounds process discussed in class.

[14] and [17] take a more traditionally causal and less statistical approach to the matter, and are some of the more well cited literature in the field. They act as general introductions to the notions of homophily, social contagion, and interference, establishing causal diagrams to assess several different scenarios and discuss identifiability in each of them. [14] introduces the notion of several different types of interference, such as direct, contagion, and allocational.

[2] aimed to use the underlying network grid to try forms of clustering in the process of using randomization to the groups to try and eliminate forms of homophily. Simple randomization and assumptions about what the latent homophilic traits are seem to be a common tactic in the non-causal community when addressing this question. [15] also utilizes randomization as well as developing an unbiased estimator, and the notion of a trade-off between network manipulablity and estimator bias. They also elect to form a treatment vector that maximizes Fisher information to optimize asymptotic expected performance.

[5] goes for the route of formalizing peer influence as a causal effect similar to several other aformentioned sources, but then tackles the estimation process using a developed method known as Poisson Influence Factorization (PIF). PIF determines probabilistic factor models to networks and behavior data to infer variables that serve as substitutes for the confounding latent traits. They also state assumptions under which PIF recovers estimates of social influence.

Other Methods Literature Review

Of course, the desire to identify and quantify peer influence has existed outside the realm of a modern causal framework, and they may be just as varied as our causal framework collection. Many related and oft-cited studies discuss the notion of peer influence and showing its effects and displays of the notion of homophiliy and contagion, but do not try and quantify the effect of peer influence itself, like [10]. However, we were able to collect a few that do elect to foray into the space.

Several of the studies simply go for a route of randomization in the experiment among the relevant groups. They assume this will resolve and eliminate discrepancies introduced by latent traits, as shown at the core of [18] or [16], where they do actually conduct large-scale randomized trials of peer-to-peer communications with the aim of getting an unbiased estimate. They make interesting claims like, when "marketing is controlled for, contagion effects disappear”, and that logically, homophily explains clustering in product adoptions in social graphs. They claim that, if a randomized trial is correctly designed and implemented, it can be shown that simple OLS estimation provides unbiased estimates of the treatment that are internally valid.

The notion of looking at peer influence as a process across a network and not just individual effects is also an area of focus, and is well discussed in [12], where he considers an iterated adjustment process on a matrix representing opinions in the space, and crafting simple structural equations, with the key assumption that every trait here is observable, that show the progression of opinions diffusing in the space. [8] considers utilizing the Voter Model for their system, and then applying propensity score ratios on the outcomes to determine a definition for the degree of influence.

This somewhat is a simplified version, or ties into the stochastic-actor based (SABA) models, that are run often with simulation investigation for empirical network analysis (SIENA) software. They aid in statistical framework for co-evolving nodal and relational variables and their interactions over time, as discussed in [4].

[11] was a very interesting read, it aimed to study the proliferation of the iPhone 3G across a number of communities. They attempt to control social clustering, gender, previous adoption of mobile Internet data plans, ownership of technologically advanced handsets, and heterogeneity in the regions where subscribers move during the day and spend most of their evenings to get the impact of peer influence. Unlike most other studies, they do consider the notion of higher hop influences, the idea of a peer impact, causally, beyond just your own neighbors. They key notion the utilize that does somewhat tie into causality is their use of instrumental variables. They needed to specifically find where in the population where they witnessing this notion of "one-side non-compliance" - subscribers adopting first, but then their friends did not. They were able to draw measures on average treatment effect on the treated and the average partial effect.

[1] similarly made considerations about what possiblity there may be of multi-hop influencing. More so, they built a model that extended the multi-network auto-probit (mNAP) model, which allows for multiple network autocorrelation terms with no restriction on the signs of the estimated parameters, that additionally relies spatial autoregressive model, which allows them to identify the direct and indirect peer influences on each of the extracted subpopulations. They use meta-analysis to summarize the estimated parameters from all subpopulations.

Datasets

Now, we must consider datasets that can be used in our empirical review. We do not have the means to run a study to generate our own large scale real-world dataset or access to proprietary internal datasets with most social networks that could provide datasets large enough to be of interest. However, through Stanford Network Analysis Project (SNAP), we do have some access to anonymized social network data that can be of use. In particular, I believe I will go with the Pokec dataset - representing the most popular social network in Slovakia (even after the introduction of Facebook!). It contains over 1.6 million profiles with data on gender, age, hobbies, interest, education, etc, and of course, friendship connections between users. This dataset is often cited in social network analysis, and also happen to be analyzed in [22], cited by [3].

Potential Experimental Design

Given the data, we can prepare a simulation to run the process described in [3]. We can select variables in the data to focus on for the treatment and outcome, as well as the covariates, and leave the remainder as confounders. We can simulate each individual accepting the treatment of their neighbors as input as well as the confounder, and decide an appropriate set of functions to utilize. We can directly implement the embedding procedure as stated in [3] as it seems fairly straightforward. We allow the outcome to be a function of the treatment, as well as the computed embedding. We can generate an average treatment effect and run the setup as many times as needed under different functions and different initial conditions. We can experiment with different embedding procedures, as well as of course, different causal and non-causal methodologies.

Further Steps

At this point, we have conducted a thorough literature review to better understand the field of causal effects in social networks, in particular, quantification of peer influence. We will first finish formalizing causal discussion, analysis and critique of the original paper using this literature review, as well as notes from more broad causal inference and network science literature. In this formal discussion, we will denote that there are several key causal assumptions that [3] in the papers that it cites that it seems to utilize but not explicity state. We will gave a clear formation of [3] as a straightforward SCM. We will also provide additional details about options for the embedding process - [3] provides a framework for doing so, which is entirely inspired by [20]. Options like

GraRep, SDNE and Node2Vec may be worth investigating. But beyond that, network embedding is a wide field with several options to consider, with [9] having some quality suggestions. Then, we will implement in code the potential experimental design, and test to make sure it is correct. We will generate estimates with the code, and then find available code from a select few of the other cited literature, and compare the estimates from this process versus the alternative processes. If needed, we may elect to try and implement a couple of the other alternatives if code is unavailable for them.

References

[1] R. Krishnan B. Zhang, P. Pavlou. On Direct Versus Indirect Peer Influence in Large Social Networks. 2017.

[2] Niloy Biswas and Edoardo M. Airoldi. Estimating Peer-Influence Effects Under Homophily: Randomized Treatments and Insights. 2020.

[3] Veitch Cristali. Using Embeddings to Estimate Peer Influence on Social Networks. In *WHY-21 Workshop at NeurIPS 2021*, 2021.

[4] Nayan G. Ramirez James Moody Scott D. Gest Daniel T. Ragan, D. Wayne Osgood. A Comparison of Peer Influence Estimates from SIENA Stochastic Actor–based Models and from Conventional Regression Approaches. May 2019.

[5] David Blei Dhanya Sridhar, Caterina De Bacco. Estimating Social Influence from Observational Data. 2022.

[6] Susan Halford Dimitra Liotsiou, Luc Moreau. Social Influence: from Contagion to a Richer Causal Understanding. 2016.

[7] Ivan Diaz Mark J. van der Laan Elizabeth L. Ogburn, Oleg Sofrygin. Causal Inference for Social Network Data. 2018.

[8] Francisco C. Santos Flávio L. Pinheiro, Marta D. Santos and Jorge M. Pacheco. Origin of Peer Influence in Social Networks. 2014.

[9] Rami Al-Rfou Haochen Chen, Bryan Perozzi and Steven Skiena. A Tutorial on Network Embeddings. 2018.

[10] Marco Gonzaleza Kevin Lewisa and Jason Kaufman. Social selection and peer influence in an online social network. 2011.

[11] Pedro Ferreira Miguel Godinho de Matos and David Krackhardt. Peer Influence in the Diffusion of iPhone 3G over a Large Social Network. 2014.

[12] James Moody. Peer influence groups: identifying dense clusters in large networks. May 2001.

[13] Eric J. Tchetgen Tchetgen Naoki Egami. Identification and Estimation of Causal Peer Effects Using Double Negative Controls for Unmeasured Network Confounding. 2021.

[14] Elizabeth L. Ogburn and Tyler J. VanderWeele. Causal Diagrams for Interference. 2015.

[15] Edward Kao Panos Toulis. Estimation of Causal Peer Influence Effects. 2019.

[16] Akhmed Umyarov Ravi Bapna. Do Your Online Friends Make You Pay? A Randomized Field Experiment in an Online Music Social Network. 2012.

[17] Cosma Rohilla Shalizi and Andrew C. Thomas. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. May 2011.

[18] Dylan Walker Sinan Aral. Creating Social Contagion Through Viral Product Design: A Randomized Trial of Peer Influence in Networks. 2011.

[19] Greg Ver Steeg and Aram Galstyan. Statistical Tests for Contagion in Observational Social Network Studies. 2013.

[20] Blei Veitch, Wang. Using Embeddings to Correct for Unobserved Confounding in Networks. December 2019.

[21] Cheng Zhang Xi Chen, Yan Liu. Distinguishing Homophily from Peer Influence Through Network Representation Learning. 2022.

[22] Volker Tresp Yunpu Ma. Causal Inference under Networked Interference and Intervention Policy Enhancement. January 2019.