A/B Testing in Networks with Adversarial Members

Kaleigh Clary
University of Massachusetts Amherst
kclary@cs.umass.edu

David Jensen
University of Massachusetts Amherst
jensen@cs.umass.edu

ABSTRACT

Many researchers attempt to study the effects of interventions in network systems. To simplify experimental design and analysis in these environments, simple assumptions are made about the behavior of its members. However, nodes may not respond to treatment, or may respond maliciously. These *adversarial nodes* influence treatment topology by preventing or altering the expected network effect, but may not be known or detectable. We characterize the influence of adversarial nodes and the bias these nodes introduce in average treatment effect estimates.

In particular, we derive expressions for the bias induced in average treatment effect using the linear estimator from Gui et al (2015). In addition to theoretical bounds, we empirically demonstrate estimation bias through experiments on synthetically generated networks. We consider both the case in which adversarial nodes are dispersed randomly through the network and the case where adversarial node placement is targeted to the highest degree nodes. Our work demonstrates that peer influence makes causal estimates on networks susceptible to the actions of adversaries, and specific network structures are particularly vulnerable to to adversarial responses.

CCS CONCEPTS

•Computing methodologies →Causal reasoning and diagnostics; •Networks →*Social media networks;*

KEYWORDS

casual effect estimation, relational data, social networks, adversarial analysis

1 INTRODUCTION

The effect we have on our peers has been a long-studied question in the social sciences. Social media systems provide a lens through which to study social influence mechanisms and have allowed for deeper analysis and experimentation in the study of peer effects.

The unique role of social media in emotional and social contagion is well documented [9] [16]. Some groups have attempted to harness the power of social systems to influence individual opinions. Political campaigns by special interests groups against particular policy views and fake product reviews from advertisers or competing products all act as adversaries with respect to network effect, attempting to persuade or otherwise manipulate other members of the population. So-called "astroturf" campaigns are one example of this phenomenon [19]. In these campaigns, a single actor determines the behavior of several bot or otherwise artificial accounts in a network (e.g., social media site). These players acts in a specific, coordinated way to convince other members of the network to adopt opinions by simulating grassroots support through the artificial members. These campaigns in special interests, marketing, and political movements have been the subject of recent media and scholarly attention [4] [14] [19] [25]. However, the existence of such individuals is generally ignored in estimation of treatment effects in large-scale online experiments. Given the growing concern over the influence of these campaigns and other adversaries in large social media platforms, we would like to understand the effect of adversaries on effect estimation.

Adversaries can take many forms: bots, competitor-owned accounts, paid individuals, and noncompliers might all function as adversaries in the estimation of treatment effect relative to some network A/B test. The behavior of adversaries in the population can influence the behaviors and outcomes of those exposed to the deviant behavior, which might mask or manipulate the true estimand of interest. These individuals may also distort the treatment exposure topology of the network, further diverting measures of treatment effect.

In this paper, we characterize the effect of adversarial agents on treatment effect estimates in the propositional and network settings. To our knowledge, this is the first exploration of the effect of adversaries on causal estimation in networks. We show that network effect is the primary source of bias from adversaries and identify specific graph structures especially vulnerable to adversary influence. We additionally derive expressions for the bias induced from adversaries and examine the difference between random and targeted placement in the network.

The rest of the paper is structured as follows. In Section 2, we review background on causal effect estimation in both the propositional and relational (network) cases. In Section 3, we define adversaries with respect to experimental design. In Section 4, we derive expressions for the bias in average treatment effect (ATE) estimation due to adversaries in the population. In Section 5, we present simulation results over the increase in ATE bias as the number of adversarial agents increases for three types of random graphs. In Section 6, we review related work in the literature. In Section 7, we conclude and discuss directions for future work.

2 BACKGROUND

A/B testing is the standard method for estimating the effect of some treatment on a particular outcome of interest. The procedure uses random assignments of treatment in a population to determine the difference in outcome after receiving that treatment.

Suppose an administrator for a large-scale online encyclopedia would like to introduce some new feature to increase the time visitors spent on her site. She is considering adding a link to the top of the page that will send the visitor to a random article on the site, which she hopes will increase total browsing time for site visitors. Before deploying this change to all visitors, she would like to quantify the effect of this website change.

For each individual, we are concerned with estimating the effect of some treatment administered. Let z_i be the treatment assignment

to individual i. Here we consider only binary treatments, where each individual either receives treatment ($z_i = 1$), or not ($z_i = 0$). Let Y_{zi} be the outcome of individual i under treatment assignment z.

In the online encyclopedia scenario, treatment is serving a page with the random article link, and outcome is some measure of the visitor's total browsing time over a given period (e.g., minutes browsing per month). This treatment is assigned randomly to site visitors, and the treatment assignment of any user is fixed once assigned.

We let Y_{1i} denote visitor i's minutes per month spent browsing where i was selected to receive pages with the random article link, and Y_{0i} is i's minutes per month spent browsing where i received the control pages.

2.1 Causal Effect Estimation

There are many methods described in the literature to measure the effect of treatment on some population. In this work, we will focus on the estimation of the *average treatment effect (ATE)*, τ, the average difference in outcome under treatment and control:

$$\tau = \frac{1}{N}\sum_{i}^{N}(Y_{1i} - Y_{0i}) \tag{1}$$

We cannot observe both Y_{1i} and Y_{0i} since each individual can only receive a single treatment assignment. Instead, we will estimate τ under the potential outcomes framework of Rubin [22]. This framework relies on the use of *counterfactuals*. A counterfactual value is the outcome of an individual under the alternative treatment assignment.

We would like to quantify the difference between the mean outcomes of the population under global treatment, where every individual receives treatment, and global control, where no treatment is assigned. There are several methods for estimating τ using counterfactuals. The simplest procedure takes the difference between mean outcomes in each treatment group:

$$\hat{\tau} = \frac{1}{N_1}\sum_{i,z_i=1}^{N} Y_{1i} - \frac{1}{N_0}\sum_{i,z_i=0}^{N} Y_{0i} \tag{2}$$

A more sophisticated method learns a model of outcome depending on the unit's treatment assignment and other unit-specific attributes, then estimates each unit's counterfactual outcome [13] [18]. Another alternative matches units in treatment to units in control using e.g., k-means, and uses the outcomes of the matched units to estimate the counterfactual outcomes [20].

The potential outcomes framework assumes that our population samples are individually and identically distributed (*iid*) and the outcome of an individual i is dependent only on i and her treatment assignment. That is, the treatment assignment of other individuals *does not* interfere with i's outcome. This is referred to as the Stable Unit Treatment Value Assumption (SUTVA).

2.2 Causal Effect Estimation in Networks

The causal estimation framework discussed so far has been in the *propositional* setting, where the data is *iid*. Now we consider the network or *relational* case. Returning to our example, suppose the encyclopedia website includes a social sharing component that facilitates article sharing between friends. Now visitors assigned to treatment can easily send interesting random articles to their potentially untreated friends, polluting the time on site estimates of users in the control group. With the addition of this social component, we must revisit the experimental design.

If a treated individual is sharing her randomly served articles with her friends, then her treatment assignment spills over to her friends' treatment assignments. The outcome of an individual, $Y_i(\mathbb{Z} = \{0,1\}^N)$, now depends on the *vector* $\mathbb{Z}$ of treatment assignments across the network rather than just her individual treatment assignment. This is a violation of SUTVA. To estimate treatment effect, we must determine the difference between *global treatment* and *global control*:

$$\tau = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(\mathbb{Z} = \mathbb{1}^N) - Y_i(\mathbb{Z} = \mathbb{0}^N)\right) \tag{3}$$

2.2.1 Graph Cluster Randomization. In the relational paradigm, the properties of one unit are not independent of other units (the data is non-*iid*). Let $G = \langle V, E\rangle$ be an undirected graph representing the relationships among the population, where two nodes v_i, v_j have an edge $e_{i,j}$ if and only if there is a relationship between v_i and v_j. We use A to denote the adjacency matrix of G and D to represent the degree matrix of G.

The edges between nodes are carriers of *treatment exposure*. When units in control are exposed to treatment, the outcome of those exposed units is potentially influenced by exposure. This treatment interference introduces bias in the average treatment effect by attributing outcomes due to treatment exposure to control behavior.

If treatment assignment to nodes across the network is random and uniform, the probability that a node experiences global treatment or global control in their neighborhood is $\frac{1}{2^{N_i}}$ where N_i is the number of nodes adjacent to i, so the probability of a single node being exposed to both treatment assignments is high.

We would like a procedure to minimize this exposure between treatment and control assignments. The graph cluster randomization approach of Ugander et al. [23] accomplishes this by assigning treatment to *clusters* of the graph, which reduces node exposure to the alternative treatment assignment. The general procedure for graph cluster randomization is as follows: (1) cluster the graph, (2) randomly assign treatment to clusters, and (3) estimate causal effect.

The treatment assignment vector over the graph, $\mathbb{Z}$, results in varying levels of exposure for each node in the network. To estimate outcomes for global treatment and global control, the authors assume multiple treatment assignment vectors for a given unit can map to the same potential outcome. $\mathbb{1}[\mathbb{Z} \in \Omega_i^1]$ denotes the indicator function for $\mathbb{Z}$ belonging to the set of treatment assignment vectors under which $Y_{\mathbb{Z}i} = Y_{1i}$. When the function is true, the unit is *network exposed* to treatment t. The analogous definitions hold for units network exposed to control.

Under this assumption, the ATE can be estimated using a Horwitz-Thompson estimator, which uses inverse probability weighting over

outcomes for units network exposed to treatment and network exposed to control:

$$\hat{\tau} = \frac{1}{N}\sum_{i}^{N}\Big(\frac{Y_{i\mathbb{Z}}\mathbb{1}[\mathbb{Z}\in\Omega_i^1]}{Pr(\mathbb{Z}\in\Omega_i^1)} - \frac{Y_{i\mathbb{Z}}\mathbb{1}[\mathbb{Z}\in\Omega_i^0]}{Pr(\mathbb{Z}\in\Omega_i^0)}\Big) \tag{4}$$

The authors identify a number of exposure model definitions for approximating Ω_i^1, Ω_i^0. One definition uses a neighborhood portion threshold q such that $\mathbb{Z}\in\Omega_i^1$ when qN_i of i's neighbors receive treatment, and $\mathbb{Z}\in\Omega_i^0$ when qN_i of i's neighbors are assigned to control. Nodes with treatment exposure outside of either of these ranges are omitted from the estimation, which can lead to high variance over the estimate [23].

2.2.2 Additive Models. An alternative estimation procedure by Gui, et al. [12] uses a linear estimator of outcome g from individual treatment effect and the portion of treated neighbors. Returning to the online encyclopedia example, if we believe the browsing time for one visitor influences browsing time for her friends, then her treatment assignment affects her friends' outcomes *through* its effect on her browsing time. This, too, is a violation of SUTVA.

The linear additive model assumes that ATE is additive in individual and network effects. Instead of binning units according to their network exposure to treatment and control, the portion of treated neighbors σ is used directly in the estimation:

$$g(z_i, \sigma_i) = \alpha + \beta z_i + \gamma\sigma_i \tag{5}$$

ATE is then estimated as the sum of the individual treatment and treatment exposure parameters, $\hat{\beta} + \hat{\gamma}$. This estimation method allows a spectrum of treatment exposure across the network, and is robust to SUTVA violations from outcome interference.

3 ADVERSARIAL NODES

Given this framework for causal inference, we define an *adversary*, or *adversarial node* as an individual in the population who is aware of her treatment assignment in the experiment and acts under a specific behavioral model in order to skew the experimental quantity of interest.

It is unlikely for a single node to have a large effect on estimation in an appropriately dense network. More likely, we are concerned with the effect of a set of adversaries following the same behavioral model. In the propositional setting, the data is *iid*, so bias in estimated treatment effect is determined only by the distorted response of adversaries. In the network setting, however, treatment of a single individual may expose the neighbors of that unit to treatment, so bias is induced both through the adversary's outcome and the peer effect that adversary applies to its neighbors. We consider the case with interference from both treatment and outcome. This means the behavior of the adversary additionally influences the outcome of its neighbors, so adversary bias in the network setting can diffuse through the network via peer influence into nonadversary outcomes.

3.1 Behavioral Models for Adversaries

We will assume that each adversary in the network follows the same behavioral model. Here we assume that the outcome of individuals is bounded. We recognize three possible behavioral models an adversary might follow:

(1) The adversary responds randomly from a uniform distribution over the outcome space, regardless of treatment assignment.
(2) The treated adversary responds with the maximum outcome, and the control adversary responds with the minimum outcome, in order to inflate the estimated treatment effect.
(3) The treated adversary responds with the minimum outcome, and the control adversary responds with the maximum outcome in order to minimize the estimated treatment effect.

The first of these behavioral models results in an increase in variance over the ATE. The adversaries inject noise into the estimate through random behavior. In the network setting, this increases the amount of random noise observed by neighbors, but does not bias neighbor outcomes.

The behavioral models (2) and (3) aim to bias the effect estimation in some direction. In the network setting, these behavioral models have a stronger effect than the random response model. By using extremes in their outcome response function, adversaries push the outcomes of their neighbors by exercising maximum network effect.

Consider behavioral model (2) in the encyclopedia example, and recall that we assume adversaries know their treatment status. An adversary given treatment would spend the maximum time on site and share a large number of pages with their friends, and an adversary in control would stop using the site and social component.

The models described here are not meant to be a complete definition for any possible adversary. The adversary behavioral model can be arbitrarily complex. Indeed, sophisticated adversaries are likely interested in masking some behavior to avoid detection. In this work, however, we are interested in a general exploration of the worst case performance of causal effect estimation in the presence of adversaries and concentrate on extreme models of adversary behavior.

4 BIAS FROM ADVERSARIAL NODES

We are interested in ATE bias due to adversarial behavior in the network. Previous work has shown that variance over ATE estimation using graph cluster randomization is large and sensitive to several choices in the experimental setup: estimation parameters, clustering method, and treatment assignment each influence the variance over the ATE estimate for a particular graph [8]. Given the plurality of factors that influence the variance over ATE estimation, this work focuses on the bias in the ATE estimate due to adversarial behavior. In addition, the bias-inducing behavioral models are of greater interest in the network experimental setting because those behaviors result in stronger peer influence, especially in estimation methods that include the neighborhood outcome mean as a parameter in the effect estimation.

We define $\delta_R(\hat{x}, k)$ as the bias in the estimate of x due to k adversaries in the population. For our framework and analysis, we will assume adversary identities and behavioral functional forms are known.

4.1 Adversaries in the Propositional Setting

In the propositional setting, the only influence an adversary can exact on the estimated ATE is through its own behavior. We can therefore separate the outcomes of adversaries from the outcomes of non-adversary units:

$$\hat{\tau} = \frac{1}{N_1} \sum_{\substack{i=1 \\ z_i=1 \\ i \notin R}}^{N} Y_{1i} - \frac{1}{N_0} \sum_{\substack{i=1 \\ z_i=0 \\ i \notin R}}^{N} Y_{0i} + \frac{1}{N_1} \sum_{\substack{r \in R \\ z_r=1}} Y_{1r} - \frac{1}{N_0} \sum_{\substack{r \in R \\ z_r=0}} Y_{0r} \quad (6)$$

where R is the set of adversaries in the population. Note that the adversary outcomes Y_{1r} and Y_{0r} are defined by the adversary behavioral function. When the adversary outcome function is constant, we can derive the bias due to adversary behavior:

$$\hat{\tau} = \frac{1}{N_1} \sum_{\substack{i=1 \\ z_i=1 \\ i \notin R}}^{N} Y_{1i} - \frac{1}{N_0} \sum_{\substack{i=1 \\ z_i=0 \\ i \notin R}}^{N} Y_{0i} + \frac{|R_1|}{N_1} Y_{R1} - \frac{|R_0|}{N_0} Y_{R0} \quad (7)$$

where R_1, R_0 are sets of adversaries receiving treatment or in control, respectively, and Y_{R1}, Y_{R0} are adversary outcomes under treatment and control. So the bias introduced from adversarial behavior in the propositional case is given by:

$$\delta_R(\hat{\tau}, |R|) = \frac{|R_1|}{N_1} Y_{R1} - \frac{|R_0|}{N_0} Y_{R0} \quad (8)$$

4.2 Adversaries in the Relational Setting

In systems with outcome interference, the treatment effect diffuses through the network. This generally requires a temporal component in the model. In the first time-step, the treatment influences only the the nodes for which it was assigned. For subsequent time-steps t, outcome is a function of both z_i, the unit's treatment assignment, and $Y_{j,t-1}$, the outcome of neighbors, j, of i.

Adversaries in the network setting are particularly interesting because these nodes block the flow of treatment effect through the network. This distorts the treatment topology of the network, and the effect of those blocks deepens as the number of time-steps for treatment propagation increases. In addition, diffusion of treatment effect through the network also allows bias to propagate through the network. This is why astroturf campaigns are effective: artificial accounts target individuals susceptible to peer influence and push them toward a particular outcome, and those individuals in turn influence their neighbors.

In this context, the most influential nodes are not necessarily nodes with highest degree. A node with fewer neighbors experiences a larger effect from any single neighbor's outcome. A high degree node only has a high effect on its neighbors outcomes if its neighbors have low degree. We define *adversarial influence*, ω_i, of a node i in the network G as:

$$\omega_i = D^{-1} A \mathbb{I}_i^N \quad (9)$$

where D is the diagonal degree matrix of G, A is the adjacency matrix of G, and $\mathbb{I}_i$ is a 1xN vector with the ith row equal to 1. Note ω_i is equal to the ith column sum of the transition matrix of G and bounded [0,N]. For comparing total influence of a set of adversaries among graphs, we take the sum of ω_a for each adversary a in the set and divide by N to normalize.

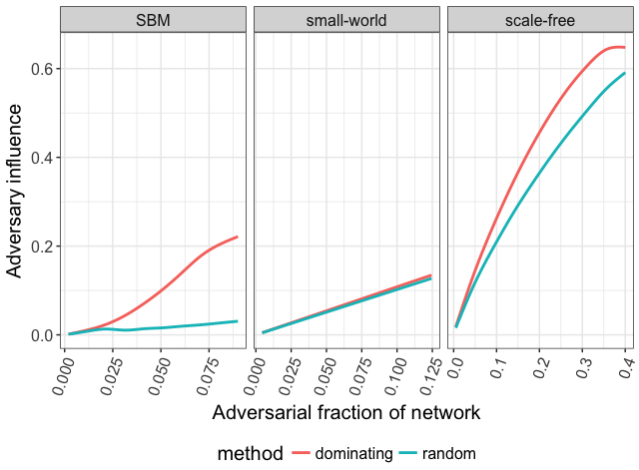


Figure 1: Total normalized influence of adversaries as the number of adversaries increases in stochastic block models, small-world networks, and scale-free networks. We considered cases where adversaries are selected either randomly or greedily based on maximum degree.

If the goal of the set of adversaries is to maximize bias in the experimental estimate, it is unlikely the entire network will consist of adversaries even in the worst case. We will instead consider bias for a set of adversaries up to a *dominating set* of adversaries. A dominating set S of a graph is a set of vertices such that every node in the network shares at least one edge with a member of S. When adversaries form a dominating set over the graph, every node will be adversary-exposed to some degree.

There are a number of ways to construct this set. The simplest procedure is to greedily select nodes from the graph according to some heuristic until the set of chosen adversaries dominates the network. We compare the following selection methods:

(1) Randomly select nodes from the graph.
(2) Greedily select nodes according to number of neighboring uncovered nodes, breaking ties with vertex degree.

An *uncovered* node is one that is not adjacent to an adversary. These two procedures represent random placement of adversaries and targeted placement of adversaries, respectively. The greedy procedure using number of uncovered nodes and degree is the standard method for greedily constructing a dominating set over the graph [7].

Figure 1 shows the increase in total normalized adversary influence as the number of adversaries increases using the random and greedy selection procedures for three different random graph generation procedures: small-world networks [24], scale-free networks [5], and stochastic block models [11]. Our analysis does not include Erdös-Rényi graphs, as these graphs rarely exhibit community structure or other properties consistent with those observed in real-world networks.

Small-world networks are generated by constructing a lattice with a given degree and then rewiring edges to new nodes with

rewiring probability p. A rewiring probability of 0 produces a regular lattice, and a rewiring probability of 1 produces a random (Erdős-Rényi) network. When the rewiring probability falls in the range [0.01, 0.1], the network is considered a small-world network. These networks have large clustering coefficients and short diameters, which are properties found to be consistent with many real-world networks [24].

In a scale-free network, new nodes are connected to existing nodes in proportion γ to the current in-degree of the existing node. Networks generated in this way have degree distribution following a power law $\mathcal{P}(k) \sim k^{-\gamma}$ [5], which is an additional property noted in real-world networks.

Stochastic block models (SBMs) are a widely used benchmark for graph generation in the community detection literature. These models are generated by constructing individual communities of bounded size, each generated using some intracommunity connection probability, and adding edges between communities according to a community mixing probability. SBMs have ground truth communities by construction, and the networks generated follow a power law distribution in node degree and community size.

Our analysis considers three graph generation procedures. For scale-free networks, we generate 200-node graphs with power parameter $\in \{0.1, 0.3, 0.5\}$. We generate 225-node small-world graphs with rewiring parameter $p \in \{0.03, 0.05, 0.1\}$. We generate stochastic block models with 500 nodes, intracommunity attachment probability 0.8, and intercommunity attachment probability $\in \{0.1, 0.2, 0.3\}$, and community size $\in [10, 20]$. For each graph setting, we generate 100 graphs of that type.

Clearly, influence of adversaries is intimately related to the connectivity and degree distribution of the graph. Scale-free networks are the easiest to exploit due to their generating algorithm. A greedily-selected dominating set in scale-free networks total an average of 0.65 influence over the graph, and the curve of total influence increases steeply with the number of adversaries. The construction mechanism of scale-free networks lends these graphs to easy capture. New edges are added in proportion to in-degree of each nodes. This leads to the existence of high-degree hub nodes in the network, which are likely to have high node influence.

Small-world graphs show low total adversary influence, even with a dominating set of adversaries. This, too, is due to the construction procedure for the graph. The degree distribution in small-world graphs is tight, so all nodes have close to the same number of neighbors. As a result, no individual node is likely to have a significantly different influence than any other in the network and there is little difference between the randomly and greedily selected adversaries.

SBMs also show a large difference in total adversary influence between adversary selection methods. Adversaries sets selected randomly total an average of 0.04 adversary influence, while dominating sets selected greedily total an average of 0.25 adversary influence. When SBMs are generated with a moderate level of intercommunity connection, as is typical of real-world graphs, only a small number of adversaries is needed to form a dominating set over the graph. In our experiments, the size of the dominating set was on average 6% of the total network.

The low total adversary influence of SBMs and small-world networks is also partially due to the small size of the dominating set relative to the size of the network.

4.3 Bias in Average Treatment Effect

We will now examine bias in estimated ATE due to adversaries. First, we can derive the bias due to adversaries in the linear estimator from Gui et al. [12], shown in Equation 5. Recall that $\hat{\tau} = \hat{\beta} + \hat{\gamma}$. By definition, the parameters β, γ are estimated as:

$$\hat{\beta} = \frac{(\sum \boldsymbol{\sigma}^2)(\sum \mathbb{Z}Y) - (\sum \mathbb{Z}\boldsymbol{\sigma})(\sum \boldsymbol{\sigma}Y)}{(\sum \mathbb{Z}^2)(\sum \boldsymbol{\sigma}^2) - \sum(\mathbb{Z}\boldsymbol{\sigma})^2} \tag{10}$$

$$\hat{\gamma} = \frac{(\sum \mathbb{Z}^2)(\sum \boldsymbol{\sigma}Y) - (\sum \mathbb{Z}\boldsymbol{\sigma})(\sum \mathbb{Z}Y)}{(\sum \mathbb{Z}^2)(\sum \boldsymbol{\sigma}^2) - \sum(\mathbb{Z}\boldsymbol{\sigma})^2} \tag{11}$$

We can simplify these expressions using the definition of $\mathbb{Z}$. Since treatment is binary, $\sum \mathbb{Z}^2 = \sum \mathbb{Z} = N_1$, the number of units receiving treatment, and $\sum \mathbb{Z}Y$ is the sum of outcomes from units receiving treatment, $\sum Y_1$.

$$\hat{\beta} = \frac{(\sum \boldsymbol{\sigma}^2)(\sum Y_1) - (\sum \mathbb{Z}\boldsymbol{\sigma})(\sum \boldsymbol{\sigma}Y)}{N_1(\sum \boldsymbol{\sigma}^2) - \sum \mathbb{Z}\boldsymbol{\sigma}^2} \tag{12}$$

$$\hat{\gamma} = \frac{N_1(\sum \boldsymbol{\sigma}Y) - (\sum \mathbb{Z}\boldsymbol{\sigma})(\sum Y_1)}{N_1(\sum \boldsymbol{\sigma}^2) - \sum \mathbb{Z}\boldsymbol{\sigma}^2} \tag{13}$$

There are two ways an adversary r influences the estimated parameters: (1) through its own outcome, Y_r, and (2) through neighborhood exposure to its outcome. In this additive framework, we assume that $\gamma \sum \frac{Y_{A_j}}{d_j}$ is the portion of an individual j's outcome, Y_j, due to network effect. Because of the additive functional form, we can separate these two sources of bias. We let $\hat{\beta}_R, \hat{\gamma}_R$ be the bias in estimated parameters due to the outcome from a set of adversaries R, and $\hat{\beta}_Y, \hat{\gamma}_Y$ be the bias in estimated parameters due to neighborhood exposure to outcomes from R, so that

$$\delta_R(\tau, k) = \hat{\beta}_R + \hat{\gamma}_R + \hat{\beta}_Y + \hat{\gamma}_Y \tag{14}$$

Since these parameters are estimated as sums over each unit individually, we can divide $\hat{\beta}$ and $\hat{\gamma}$ into terms accounting for nonadversary outcome and adversary outcomes in the network. Then the estimate of the parameters due to adversary outcomes is:

$$\hat{\beta}_R = \frac{|R_1|Y_{1R}\sum(\boldsymbol{\sigma_r}^2) - (\sum \mathbb{Z}\boldsymbol{\sigma})\left(\sum_{r\in R} \sigma_r Y_r\right)}{N_1(\sum \boldsymbol{\sigma}^2) - \sum \mathbb{Z}\boldsymbol{\sigma}^2} \tag{15}$$

$$\hat{\gamma}_R = \frac{N_1 \sum_{r\in R} \sigma_r Y_r - |R_1|Y_{1R}\sum \mathbb{Z}\boldsymbol{\sigma}}{N_1(\sum \boldsymbol{\sigma}^2) - \sum \mathbb{Z}\boldsymbol{\sigma}^2} \tag{16}$$

So the bias in estimates $\hat{\beta}, \hat{\gamma}$ due to adversary outcome is:

$$\hat{\beta}_R + \hat{\gamma}_R = \frac{|R_1|Y_{1R}(\sum \boldsymbol{\sigma}^2 - \sum \mathbb{Z}\boldsymbol{\sigma}) + (N_1 - \sum \mathbb{Z}\boldsymbol{\sigma})(\sum_{r\in R} \sigma_r Y_r)}{N_1(\sum \boldsymbol{\sigma}^2) - \sum \mathbb{Z}\boldsymbol{\sigma}^2} \tag{17}$$

Note that the only terms of $\hat{\beta}_R + \hat{\gamma}_R$ related to the placement of adversaries r in the network is in exposure to treatment, σ_r. Now we consider the bias due to the adversary network effects.

Reasoning about bias due to adversary network influence through parameter estimation is difficult, since the bias due to the adversary is dependent on the strength of the true network effect, γ, which is only calculated in the linear estimator through the portion of the neighborhood receiving treatment. Further, even if nonadversary outcome is separated into (1) the portion of its outcome independent of its neighbors, (2) the the portion of outcome due to nonadversary peer effects, and (3) the portion of outcome due to peer effects from adversaries, we still must account for the adversarial peer effects in i's adversary-exposed neighbors in (2), which may be exposed to a different adversary than r. Instead of reasoning about the strength of adversarial diffusion, we can approximate the bias induced by a single adversary r's outcome on nonadversary neighbor j's outcome using the fact that Y_r skews Y_j relative to the distance between Y_r and the mean outcome of j's neighbors excluding r, $\overline{Y}_{A_j \setminus r}$. So the bias in Y_j due to Y_r is $\propto \frac{1}{d_j}(Y_r - \overline{Y}_{A_j \setminus r})$, where d_j is the degree of node j, and Y_r is the outcome of adversary r under treatment assignment z_r. So the total bias induced by r on its neighbors outcome is:

$$\begin{aligned}\hat{\beta}_Y + \hat{\gamma}_Y &= \sum_{j \in A_r} \frac{1}{d_j}(Y_r - \overline{Y}_{A_j \setminus r}) \\ &= \omega_r \sum_{j \in A_r} (Y_r - \overline{Y}_{A_j \setminus r}) \approx \omega_r (Y_r - \overline{Y}_{A_{2_r}})\end{aligned} \tag{18}$$

where $\overline{Y}_{A_{2_r}}$ is the mean outcome in r's two-hop neighborhood. Then the total ATE bias due to adversaries in the network is

$$\begin{aligned}\delta_R(\tau, k) = &\frac{|R_1| Y_{1R}(\sum \boldsymbol{\sigma}^2 - \sum \mathbb{Z}\boldsymbol{\sigma}) + (N_1 - \sum \mathbb{Z}\boldsymbol{\sigma})(\sum_{r \in R} \sigma_r Y_r)}{N_1(\sum \boldsymbol{\sigma}^2) - \sum \mathbb{Z}\boldsymbol{\sigma}^2} \\ &+ \sum_{r \in R} \omega_r (Y_r - \overline{Y}_{A_{2_r}})\end{aligned} \tag{19}$$

5 SIMULATION RESULTS

To empirically demonstrate the effect of adversaries in ATE estimation over a network, we simulated outcomes in a network and used the ATE estimation of Gui et al. [12] to estimate treatment effect in networks both with and without adversary interference.

5.1 Graph generation

Graphs are generated using the same procedure reported in section 4.2. We consider scale-free networks with power = 0.3, small-world networks with rewiring parameter $p = 0.5$, and SBMs with community mixing parameter $\mu = 0.2$.

5.2 Simulation model

Observed outcome values were generated using the following linear model adapted from Gui et al. [12], Eckles et al. [10]:

$$Y_{i,t} = \lambda_0 + \lambda_1 z_i + \lambda_2 \frac{A_i Y_{t-1}}{D_{i,i}} + U_{i,t} \tag{20}$$

where z_i is the treatment assignment of unit i, $\frac{A_i Y_{t-1}}{D_{i,i}}$ is the mean outcome units neighboring unit i, and $U_{i,t} \sim \mathcal{N}(0, 0.1)$ is an individual-level noise parameter. Following the parameter assignments of [12], we set $t = 3$, $\lambda_0 = -1.5$, and $Y_{0,i} = 0$ for all i. We consider individual treatment parameters $\lambda_1 \in \{0.25, 0.5, 0.75, 1\}$ and neighborhood treatment parameters $\lambda_2 \in \{0, 0.1, 0.5, 1.0\}$.

The adversary behavior follows (3) from section 3.1 where adversaries respond to minimize the estimated ATE. Adversary outcome is determined by λ_0 and λ_1:

$$Y_r = \begin{cases} \lambda_0 & \text{if } z_r = 1, \\ \lambda_0 + \lambda_1 & \text{if } z_r = 0. \end{cases}$$

We assign clusters to treatment or control according to a binomial distribution $\mathcal{B}(c, 0.5)$, where c is the number of clusters produced by the clustering procedure.

5.3 Methodology

For a given graph, we generate a clustering of the graph using the Infomap algorithm for community detection [21] and assign clusters to treatment or control. We then generate a dominating set of adversaries using greedy selection, sorted by decreasing degree. A second adversary set of the same size is selected uniformly at random from the graph. Given this selection procedure, the randomly selected adversary set may not form a dominating set over the graph.

For specific settings of λ_1 and λ_2, we determine the outcome function for nodes and adversaries. We then iterate over the set of adversaries, starting with an empty set and adding an adversary to the previous set, ending with the entire dominating set. For each adversary subset, we generate outcome $Y_i, 3$ for each node. We use the linear estimator over generated outcomes to estimate ATE for each subset of adversaries. The estimated ATE for the experiment with no adversary interference is used as a baseline of comparison for ATE bias induced by adversaries.

5.4 Synthetic network results

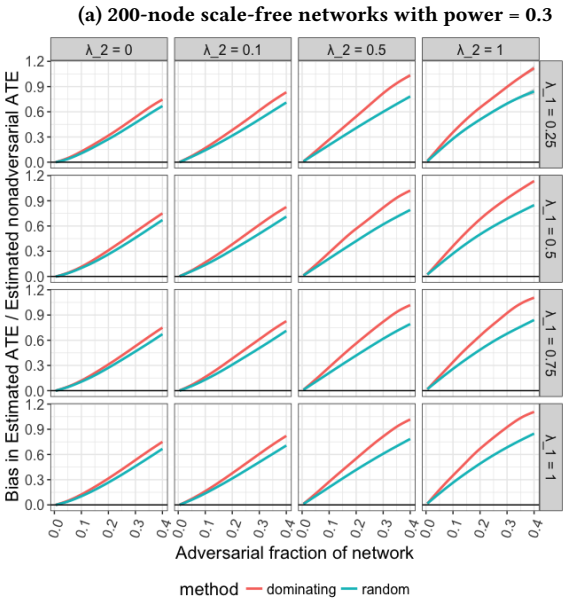
Figure 2 shows the results of our simulations. These results demonstrate the effect of adversaries on the estimated treatment effect. The bias induced by adversaries increases as the number of adversarial nodes increases.

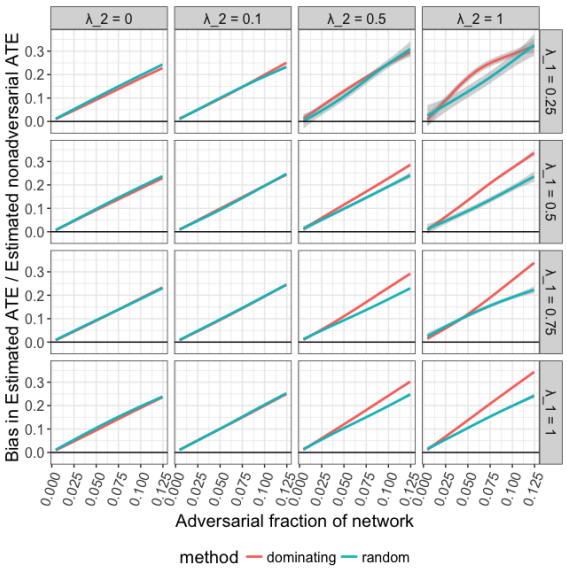
As individual treatment effect λ_1 increases, the slope of the bias increase remains constant. For randomly selected adversaries, the increase in peer effect λ_2 has little effect. The dramatic changes in bias occur with increases to peer effect for greedily selected adversaries.

The importance of peer effect strength also depends on network topology. SBMs and scale-free networks are particularly susceptible to bias in ATE for networks with large peer effects, especially for greedily selected adversaries. SBMs with greedily selected adversaries double the ATE bias of randomly selected adversary sets. These two graph generation procedures are often cited as producing graphs most closely resembling real-world networks, and SBMs in particular are recommended as random graphs most closely replicating real world community structure [5] [17] [26]. It is interesting that these structures are also the most vulnerable to malicious behavior, even for relatively small sets of adversaries.

6 RELATED WORK

Literature in influence maximization is closely connected to this work. Influence maximization is concerned with identifying the set of nodes in a (social) network to target in order to maximize the spread of some quantity of interest [9] [15]. A similar problem is





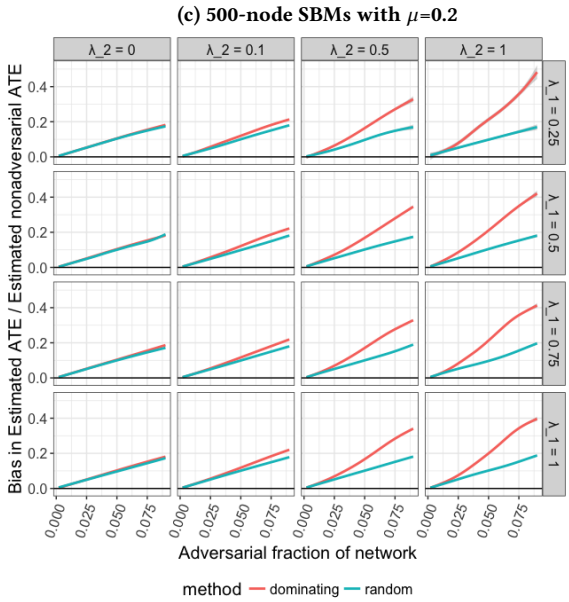


Figure 2: Bias in estimated ATE for (a) scale-free, (b) small-world networks, and (c) SBMs under different assignments of individual treatment and network treatment effects. Rows share individual treatment effect parameter settings, λ_1, and columns share network treatment effect settings, λ_2. Note the difference in scales on the y-axis for adversary bias. Scale-free networks and SBMs exhibit significant bias with increases in the strength of peer effects, even for a small set of adversaries.

the study of diffusion over a network and, in particular, resource-constrained diffusion maximization. In this setting, node selection is associated with some cost, and the total cost of the set of selected nodes cannot exceed some budget [2]. This is similar to reasoning about adversary selection, though our work is interested in the effects of adversary behavior under various selection procedures rather than maximizing the effect.

Some models of adversary behavior can be cast as non-compliance behavior in an experimental study. Recent work by Kang and Imbens introduces peer encouragement designs, a novel approach to estimating causal estimands which allow for analysis of non-compliance behavior in a network setting [14]. This design accounts for dropout non-compliance, though non-compliance through purposefully extreme adversarial behavior is still a concern.

Our work is related to other studies of the effects of adversaries in the network setting. These tend to focus on detection of adversaries [4] [6] or measuring and analyzing the success of adversary integration in the network [1] [3]. This paper is the first to explore the effect of bots and other adversaries in network A/B testing.

7 CONCLUSION AND FUTURE WORK

This work presents an introductory analysis of the effect of adversary behavior on the average treatment effect estimate. Adversarial behavior adds a layer of complexity over peer effects in network experiments that has so far been ignored, though it is a recognized issue in the literature. Our work demonstrates a vulnerability in network A/B testing to manipulation by adversary behavior, especially in networks with easily exploited topology. We have shown that networks with strong peer effects are susceptible to ATE bias from adversary behavior and identified scale-free networks and SBMs as network structures vulnerable to adversary bias.

As future work, we are interested in exploring the space of adversary behavioral models. We have focused on extreme models of adversary behavior, and it is not difficult to imagine applications

in which an adversary will choose their outcome according to not only their treatment assignment, but the treatment assignment and behavior of its neighbors. Alternatively, we might consider less extreme adversary behavior models or mixtures of behavioral models. An interesting approach to exploring this space might be to characterize the trade off between injecting ATE bias and avoiding detection from extreme responses.

Another direction is in identifying experimental designs over networks that are robust to adversary influence. Our work has demonstrated that the ATE estimate using state of the art methods is easily skewed. We might instead consider alternative methods for estimating treatment effect (e.g., average treatment effect on the treated, local average treatment effect), peer encouragement designs, which may be less vulnerable to manipulation or reveal other weaknesses in causal effect estimation in networks.

REFERENCES

[1] N. Abokhodair, D. Yoo, and D. W. McDonald. Dissecting a social botnet: Growth, content and influence in twitter. *CoRR*, abs/1604.03627, 2016.

[2] K. Ahmadizadeh, B. Dilkina, C. P. Gomes, and A. Sabharwal. An empirical study of optimization for maximizing diffusion in networks. In *Proceedings of the 16th International Conference on Principles and Practice of Constraint Programming*, CP'10, pages 514–521, Berlin, Heidelberg, 2010. Springer-Verlag.

[3] L. M. Aiello, M. Deplano, R. Schifanella, and G. Ruffo. People are strange when you're a stranger: Impact and influence of bots on social networks. *CoRR*, abs/1407.8134, 2014.

[4] L. Akoglu, R. Chandy, and C. Faloutsos. Opinion fraud detection in online reviews by network effects. *ICWSM*, 13:2–11, 2013.

[5] A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286:509–512, Oct. 1999.

[6] Q. Cao, X. Yang, J. Yu, and C. Palow. Uncovering large groups of active malicious accounts in online social networks. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 477–488, New York, NY, USA, 2014. ACM.

[7] V. Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of operations research*, 4(3):233–235, 1979.

[8] K. Clary, D. Arbour, and D. Jensen. Does topology matter? The impact of network structure on network A/B testing. In preparation, 2017.

[9] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 57–66, New York, NY, USA, 2001. ACM.

[10] D. Eckles, B. Karrer, and J. Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *arXiv preprint arXiv:1404.7530*, 2014.

[11] K. Faust and S. Wasserman. Blockmodels: Interpretation and evaluation. *Social networks*, 14(1-2):5–61, 1992.

[12] H. Gui, Y. Xu, A. Bhasin, and J. Han. Network A/B testing: From sampling to estimation. In *Proceedings of the 24th International Conference on World Wide Web*, pages 399–409. International World Wide Web Conferences Steering Committee, 2015.

[13] J. J. Heckman. The scientific model of causality. *Sociological Methodology*, 35:1–97, 2005.

[14] H. Kang and G. Imbens. Peer encouragement designs in causal inference with partial interference and identification of local average network effects. *arXiv:1609.04464 [stat]*, Sept. 2016. arXiv: 1609.04464.

[15] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 137–146, New York, NY, USA, 2003. ACM.

[16] A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[17] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E*, 78:046110, Oct 2008.

[18] J. Pearl. *Causality*. Cambridge university press, 2009.

[19] J. Ratkiewicz, M. Conover, M. R. Meiss, B. Gonalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. *ICWSM*, 11:297–304, 2011.

[20] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41, 1983.

[21] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[22] D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 2005.

[23] J. Ugander, B. Karrer, L. Backstrom, and J. Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337. ACM, 2013.

[24] D. J. Watts and S. H. Strogatz. Collective dynamics of'small-world'networks. *Nature*, 393(6684):409–10, 1998.

[25] D. G. a. S. Woolley. How twitter bots are shaping the election. *The Atlantic*, Nov. 2016.

[26] X. Zhang, R. R. Nadakuditi, and M. E. J. Newman. Spectra of random graphs with community structure and arbitrary degrees. *CoRR*, abs/1310.0046, 2013.