Stick It to The Man: Correcting for Non-Cooperative Behavior of Subjects in Experiments on Social Networks

Kaleigh Clary[1] Emma Tosch[2] Jeremiah Onaolapo[2] David D. Jensen[1]

[1] *College of Information and Computer Sciences, University of Massachusetts Amherst*
[2] *College of Engineering and Mathematical Sciences, University of Vermont*

Abstract

A large body of research in network and social sciences studies the effects of interventions in network systems. Nearly all of this work assumes that network participants will respond to interventions in similar ways. However, in real-world systems, a subset of participants may respond in ways purposefully different than their true outcome. We characterize the influence of *non-cooperative nodes* and the bias these nodes introduce in estimates of average treatment effect (ATE). In addition to theoretical bounds, we empirically demonstrate estimation bias through experiments on synthetically generated graphs and a real-world network. We demonstrate that causal estimates in networks can be sensitive to the actions of non-cooperative members, and we identify network structures that are particularly vulnerable to non-cooperative responses.

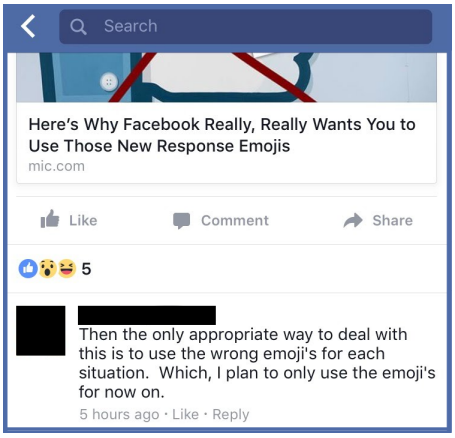


Figure 1: An instance of non-cooperative-but-not-malicious behavior on Facebook, a large online social network.

1 Introduction

Experimentation is an important facet of responsible governance, especially in the digital public sphere [46]. Without experimentation, firms that control the platforms we use to power the public sphere will make changes based on observational data or by fiat [48]. While experimentation on platforms such as online social networks (OSNs) has at times been controversial, experimentation is not harmful *per se*, and can in fact be the only mechanism for mitigating harm when implementing design changes [39, 47].

In OSNs, users often do not know whether they are currently receiving an experimental treatment. In this setting, consent to participate in the experiment is *not* decoupled from participation in the service. Furthermore, users may be aware that the platform makes extensive use of A/B tests, and that may even be explicit in the OSN terms of service. Non-participation in an OSN means not using the service, which for many users is unacceptable.

Others have studied and addressed consent models in these contexts [25, 33, 34]; we will not address consent here. Instead, we recognize that there are experiments that may not require the level of consent or control that users desire, while still being ethical. We focus on a direct consequence of this phenomenon: non-cooperative behavior when the user believes they are being experimented upon.

Figure 1 illustrates such behavior. This seemingly innocuous example belies two major concerns: (1) users may believe they are currently in an experiment due to external information or events (e.g., an upcoming election or a global pandemic), causing them to engage in non-cooperative behavior, and (2) users may interact with others *who are also in the experiment*, causing any estimation of treatment effect to be tainted by *spillover* or *peer effects*. While peer effects have been studied extensively, the models that correct for spillover on average treatment effect (ATE) typically assume that the measured outcome for an individual is some combination of their true outcome and some additional effect resulting from exposure to other experimental subjects. When participants respond in ways purposefully different than their true outcome, their non-cooperative outcomes are observed by neighbors, resulting in a change in behavior of neighboring participants due to spillover effects.

Contributions. We examine how non-cooperative participant (NCP) behavior and its resulting peer effects can influence effect estimates. We compare the effect that subsets of non-cooperative nodes can have on treatment effect estimates in network (relational) and non-network (propositional) settings. To our knowledge, this is the first exploration of the effect of non-cooperative behavior on cluster-randomized designs for causal estimation in networks. We make the following contributions:

1. We introduce a framework to unify the study of non-cooperative behavior and the estimation of treatment effect in network settings. (Section 4)

2. We derive terms for the bias in ATE resulting from non-cooperative behavior using a standard linear estimator of individual and peer treatment effect. (Section 5.3)

3. We demonstrate empirically that non-cooperative behavior can result in biased estimates of ATE for the same linear estimator using simulations on both synthetic and real-world graphs. (Section 6.4, Section 6.5)

4. We derive expected bias and both theoretically and empirically explore the difference between random and targeted placement of non-cooperative nodes in the network. (Section 5.4)

5. We identify specific graph topologies that are particularly vulnerable to non-cooperative influence. (Section 5.2, Section 6.4)

In this work, we study the organization of *non-cooperative participants* (NCPs) and whether the spillover effects of their behavior can bias the effect estimates of A/B tests conducted on the network. Our analysis assumes the set of non-cooperative participants is known. For logical consistency, we will assume NCPs are placed before treatment is assigned. We use uniform edge-weighting designs, i.e. peer effects weight outcomes of neighbors equally. We assume NCPs behave according to the same behavioral model, which may be conditioned on the treatment received.

Non-cooperative behavior has been documented in OSNs, both in the academic literature and news media under a variety of non-cooperative user behaviors: privacy-preserving data obfuscation methods e.g. k-subscription [52], teens using account-sharing rings on Instagram to protect their privacy while applying for college [51], and shopper participation in grocery loyalty card swapping pools [17] as cited in Brunton and Nissenbaum [14].

We define our threat model with respect to experimental design in Section 4 and provide a set of realistic real-world scenarios in Section 7. In Section 8, we consider the ethical implications of the use of bias correction methods to address the effects of non-cooperative behavior.

2 Problem Formulation: Peer Effects of Non-Cooperative Behavior

Peer effects describe the phenomenon in which interaction between subjects (often people) causes the treatment or outcome of an experiment to differ from the treatment or outcome of a subject in isolation. Understandably, network and social scientists care a great deal about quantifying and controlling for peer effects. OSNs allow for unprecedented study of peer effects; they are a unique source of data on social influence, and the role of social media platforms themselves in emotional and social contagion is well documented [20, 39].

Academic social scientists are not the only entities interested in understanding social behavior on OSNs. The popularity of social networking platforms has also encouraged some advertising campaigns to use fake reviews and bot accounts to sway public opinion by simulating grassroots support for products and ideas. Commercial organizations have used fake product reviews to influence consumer purchasing patterns [4, 55, 72], and political organizations have recruited real and automated users in attempts to influence voting behavior and policy positions [15, 54, 64]. So-called "astroturf" campaigns are one example of this phenomenon [54]. The proliferation of disinformation by powerful actors has led to a study of discord-maximization agents in OSNs [27]. Non-cooperative behavior, as explored in this work, is not always high stakes, as Figure 1 demonstrates. However, real-world instances of adversarial non-cooperation could also lead to devastating effects; we discuss two hypothetical instances in Section 7.

Given the growing concern over the influence of non-cooperative and sometimes targeted adversarial behavior in large social media platforms [4, 37, 54, 71], we would like to better understand the potential effects of participant non-cooperation on effect estimation. This work seeks to answer the following research questions:

RQ1 How can we measure the effects of a non-cooperative node?

RQ2 Under what circumstances can these effects be large?

RQ3 What types of networks are structurally vulnerable to the amplification of non-cooperative peer effects?

One of the challenges to developing accurate estimates of peer effects is that non-cooperative and even adversarial behavior can take many forms. Bots, competitor-owned accounts, paid individuals, non-compliers, and discord-maximizers might all function as adversaries in the estimation of treatment effect relative to some network experiment. Non-cooperative behavior can influence the behaviors and outcomes of those exposed to non-cooperative outcomes, which might mask, influence, or otherwise manipulate the true estimand of interest. These individuals may also distort the

treatment exposure topology of the network, further biasing measures of treatment effect.

Detecting specific behaviors of non-cooperative users requires custom code tailored to the behavior of each form of non-cooperation, yet different behavioral models may have similar implications for effect estimation. Furthermore, non-compliance behavior does not need to be adversarial to result in bias effects, and the effect estimate bias may even be the goal of a non-malicious user, e.g., in order to preserve their privacy [52]. We discuss ethical aspects of non-compliance behaviors, their detection, and the implications of their removal in Section 8. We revisit the *detection* and *prevention* tasks in Section 9.

2.1 Network Topology and Diffusion of Peer Effects

Non-cooperative nodes block the flow of treatment effect through a behavioral outcome intervention. This will have different implications depending on the properties of the local graph. We consider edge-weighting effect estimation models out of scope, but note one may study similar effects of edge-weighting by considering graph samples with variations in graph degree distributions. We discuss this connection further in Section 4.2 and conduct an empirical study in Section 5.2.

We specifically study the impact of graph topology in this setting as any estimate of peer effect will be gated by the connectivity of the network. The study of the relationship between graph topology and the diffusion of treatment effect (and the effects of non-cooperative participation) has connections with work studying e.g. information diffusion across networks [8, 9, 56]. 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 [20, 35, 38]. A similar problem is 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 NCP node selection, though our work is interested in studying the effects of non-cooperative behavior under various selection procedures rather than maximizing the effect.

If the goal of the set of non-cooperative participants is to maximize bias in the experimental estimate, it is unlikely the entire network will consist of NCPs even in the worst case. We instead consider bias for a set of NCPs up to a *dominating set* of non-cooperative participants. A dominating set X of a graph is a set of vertices such that every node in the network is covered by, or shares an edge with, a member of X. This is a similar coverage model as a Sybil attack, where attackers seek to weaken redundancy protections and subvert reputation systems by controlling a disproportionate share of user identities in a peer-to-peer system [21, 67].

2.2 Related Work: Estimating Effects with Non-Cooperative Participants

Some models of non-cooperative behavior can be cast as non-participatory or non-compliance behavior in an experimental study. For example, Kang and Imbens [37] introduce peer encouragement designs under one-party compliance, an approach to estimating causal estimands which is robust to some forms of non-cooperative behaviors in a network setting achieved through non-compliance.

We treat adoption of more robust estimators as an orthogonal concern, since ATE and its variants are by far the most commonly used estimators for causal effects. 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 non-cooperation biases but introduce other weaknesses in causal effect estimation in the non-cooperative setting.

3 Estimating Causal Effects

A/B testing is the standard method for estimating the effect of treatment on a particular outcome of interest.[1] The procedure uses random assignments of treatment in a population to determine the difference in outcome after receiving that treatment. Consider two example experiments:

Exp1 An OSN serves video content to a worldwide audience. Developers at the OSN may want to select between software-defined high and low video bitrates to see if bitrates affect the percentage of videos watched to completion (modeled after an experiment in Tosch et al. [66]).

Exp2 The same OSN provides users with a stream of personalized curated content. The administrators of the service might ask how the sentiment of curated content (e.g., positive content vs. neutral content) influences the amount of time a user spends with the service (modeled after an experiment in Kramer et al. [39]).

In both experiments, the administrators would like to conduct an experiment to measure the relationship between two outcomes by exposing some users U_A to one treatment (e.g., content with positive-leaning sentiment or high bitrates), and exposing others U_B to another treatment (e.g., content measured to be neutral or low bitrates).

If the treatment is served to each user in individual silos, we can directly compare the outcome (e.g., average amount of time spent browsing or video completion rates) between the two groups U_A and U_B. For each individual, we are concerned

[1] We use the term *treatment* to refer generally to the assignment status of a unit to some experimental protocol, which may include multiple treatment arms or the control arm.

with estimating the effect of treatment on outcome. Let N be the size of the user population, and z_i be the treatment assignment to user i. Here we consider only binary treatments (i.e., $z_i = 1$ or $z_i = 0$).

Let Y_{zi} be the outcome of user i under treatment assignment z (also denoted $Y_i(Z = z)$). Treatment is assigned randomly to users, and the treatment assignment of each user is fixed once assigned:

Exp1	Y_{0i}	Average daily video completion rate for user i, where i's software bitrate was set to the low value
	Y_{1i}	Average daily video completion rate for user i, where i's software bitrate was set to the high value
Exp2	Y_{0i}	Average number of minutes per day user i spent browsing, where i received neutral sentiment content
	Y_{1i}	Average number of minutes per day user i spent browsing, where i was received positive sentiment content

3.1 Propositional Setting

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 contrasting treatments:

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

Each individual unit (user) can only receive a single treatment assignment, so we cannot observe both Y_{1i} and Y_{0i}. Instead, we will estimate τ under the potential outcomes framework of Rubin [62]. This framework relies on the use of *counterfactuals*. A counterfactual value is the outcome of an individual under the alternative treatment assignment.

Our aim is to quantify the difference between the mean outcomes of the population under global treatments (where every individual receives treatment A *and* in a parallel universe, every individual receives treatment B). There are several methods for estimating τ using counterfactuals. The simplest procedure takes the difference between mean outcomes in each treatment group:

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

Structural estimation methods learn a model of outcome depending on the unit's treatment assignment and other unit-specific attributes, then estimate each unit's counterfactual outcome [32, 53]. Matching designs pair units in treatment to units in control using e.g., nearest neighbor, and use the outcomes of the matched units to estimate the counterfactual outcomes [58].

The causal estimation framework discussed so far has been in the *propositional* setting, where the data is independent and identically distributed (*iid*). The potential outcomes framework assumes that our population samples are *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) [61].

It is easy to see how SUTVA holds for **Exp1**: if each user is randomized into a software bitrate treatment implemented as a software configuration setting, then the bitrate only affects views on that device. Unfortunately, SUTVA may not hold for **Exp2**. We assumed that the accounts in **Exp2** were siloed, but this may not always be true. If users can share content to others' content streams, it becomes possible for a user assigned one treatment to effectively receive both treatments due to the exposure from their peers. Fortunately, there are statistical methods for correcting this kind of spillover or peer effect.

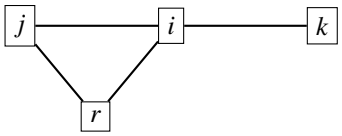
3.2 Relational Setting

When units can interact, and through interaction expose other units to additional treatments, we are in a network or *relational* context. In **Exp2** this happens when the OSN content stream includes the curated content mixed with other user-generated content. Additional positive content served to users in U_A may increase the amount of positive user content created and shared by those users, potentially affecting the time-on-site estimates of users in U_B. With the addition of this social component, we must revisit the experimental design.

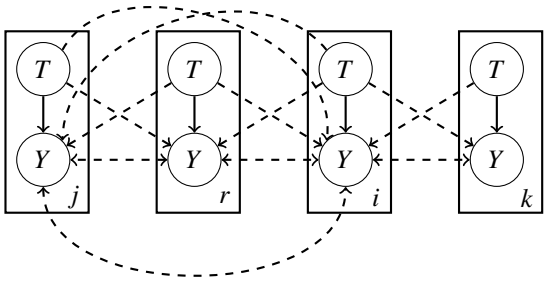
3.2.1 SUTVA Violations

In the relational paradigm, the properties of one unit (e.g., user) are typically 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.

The edges between nodes are avenues of *treatment exposure*. When units in treatment B are exposed to treatment A, the outcome of those exposed units is potentially influenced by that exposure. This treatment *spillover* or *interference* introduces bias in estimates of ATE by mixing outcomes due to differing treatments [57]. Figures 2a and 2b show an example network of units in a relational context and the corresponding causal-relational model for SUTVA violations in that network, respectively.



(a) Example network of units, e.g., users in an OSN, in a relational context. Units i, j, and k represent compliant participants; r is a non-cooperative participant (NCP). This network structure undergirds the peer effects of treatment and outcome on connected nodes depicted in Figures 2b (no NCPs) and 3 (r is an NCP).



(b) Causal-relational model corresponding to treatment assigned over the sample network shown in Figure 2a. In the networked environment, the treatment (T) *and* outcome (Y) of unit j can cause the outcome of another unit i, which in turn can influence the outcome of another unit k. Solid lines correspond to causal relationships in both the propositional and relational environments; dashed lines correspond to causal relationships that occur only in the relational environment and that imply SUTVA violations. Double-headed arrows between outcomes represent temporal unrolling over multiple timesteps.

Figure 2: Contrasting propositional and relational models.

If a user whose treatment was positive content shares that content with their friends, then their treatment assignment *spills over* to their friends' treatment assignments. Now the outcome of each user depends on the *vector* $\mathbf{Z}$ of treatment assignments across the network rather than just individual treatment assignment z_i. This is a violation of SUTVA.

To estimate treatment effect, we must determine the difference between *global treatment assignments* $Y_i(\mathbf{Z} = \mathbf{1}^N)$ and $Y_i(\mathbf{Z} = \mathbf{0}^N)$ for every unit *in parallel universes*:

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

That is, we wish to control for spillover by ensuring that all units that could affect this unit's treatment receive the same treatment as our unit of interest. However, each unit can only receive one treatment assignment. If the network graph has a single component and $\mathbf{Z}$ assigns both treatments, at least one unit will experience spillover from a unit assigned a different treatment.

3.2.2 Measuring Spillover via Treatment Exposure

The treatment assignment vector $\mathbf{Z}$ over the graph results in varying levels of treatment *exposure* for each node in the network; that is, $\mathbf{Z}$ captures both the assigned treatment and any spillover due to networked nodes [6]. If treatment assignment to nodes across the network is assigned uniformly at random, the probability that a node's *neighborhood* (i.e., set of adjacent nodes) is assigned global treatment A or global treatment B is 2^{-d_i} where d_i is the number of nodes adjacent to i. This probability becomes very small very quickly, so the probability of a single node being exposed to both treatment assignments is high. To minimize this exposure between differing treatment assignments, Ugander et al. [68] introduce a cluster-randomized treatment design which clusters the graph and assigns treatment to entire clusters, a procedure termed *graph cluster randomization*. This treatment design reduces exposure to the alternative treatment assignment.

To estimate unit-level outcomes for global treatment A (Y_{1i}) and global treatment B (Y_{0i}), the authors assume multiple treatment assignment vectors for a given unit can map to the same potential outcome. We use $\mathbb{I}[\mathbf{Z} \in \Omega_i^1]$ to denote the indicator function for $\mathbf{Z}$ belonging to the set of treatment assignment vectors under which $Y_{\mathbf{Z}i} = Y_{1i}$ (i.e., assigned to treatment A). When the function is true, unit i is *network-exposed* to treatment. The analogous definitions hold for units network-exposed to treatment B.

Under this assumption, the ATE can be estimated using a Horvitz-Thompson estimator, which uses inverse probability weighting over outcomes for units network-exposed to treatment A and network-exposed to treatment B [68]:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_{\mathbf{Z}i}\mathbb{I}[\mathbf{Z}\in\Omega_i^1]}{\mathcal{P}(\mathbf{Z}\in\Omega_i^1)} - \frac{Y_{\mathbf{Z}i}\mathbb{I}[\mathbf{Z}\in\Omega_i^0]}{\mathcal{P}(\mathbf{Z}\in\Omega_i^0)}\right) \tag{4}$$

Ugander et al. 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 $\mathbf{Z} \in \Omega_i^1$ when at least qd_i of i's neighbors receive treatment, and $\mathbf{Z} \in \Omega_i^0$ when at least qd_i of i's neighbors are assigned to control.

3.2.3 Additive Exposure Models

Gui et al. [30] introduce an estimator which separates treatment effect into both *individual* treatment effect β, and *peer* or *network* effect γ, i.e., the effect of neighbors' outcomes on that unit's outcome.[2] Individual and peer effect estimates are taken from the coefficients of a linear model of outcome g using individual treatment assignment z_i and the portion of treated neighbors σ_i. Returning to the content curation example, if we believe the browsing time for one user influences

[2]Estimation of peer effects is an active area of research. New estimators have been proposed since Gui et al. (e.g., [65]), some of which are not linear, making decomposition of peer effects into non-cooperative and compliant components quite challenging. As this is the first analysis of non-cooperative units in A/B tests, we sought a more tractable estimator.

browsing time for their friends, then their treatment assignment affects that user's friends' outcomes *through* its effect on that user's browsing time. This, too, is a SUTVA violation.

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 as in Ugander et al. [68], the portion of treated neighbors σ_i is used directly in the effect 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 without throwing out outcomes from partially-exposed units, and is robust to SUTVA violations from outcome interference (e.g., peer effects).

4 Threat Model

Given this framework for causal inference, we define a *non-cooperative participant* (NCP), or *non-cooperative node* as an individual in the population who responds under an intentionally different response model. Non-cooperative outcomes may bias estimates of the experimental quantity of interest. There are a number of models for non-cooperative behavior. For example, bots may not respond to treatment in the same way as human users, or competitor-owned accounts and paid users may engage in user manipulation unrelated to an experiment in progress, where the effects of that manipulation may interfere with the outcome of interest.

Combating non-cooperation is fundamentally an arms race [5, 10]; it is therefore impossible to produce a generic model that covers all possible models of non-cooperative behavior. The behavioral model can be arbitrarily complex. Indeed, sophisticated NCPs are likely interested in masking some behavior to avoid detection. NCPs may be acting in isolation or coordinating with other participants. It is even possible for a group of non-cooperative participants to interfere with an experiment without knowing one is underway.

We leave the characterization of real-world behavior to other researchers who are better positioned to assess end-user behavior in networked environments such as OSNs. Our focus remains on exploring whether current network A/B testing methodology is robust to non-cooperative behavior. Thus, we focus our analysis on the worst-case models of non-cooperative behavior. In the propositional setting, the data is *iid*, so bias in estimated treatment effect is determined only by the distorted response of non-cooperative participants. Correcting for this bias is mathematically straightforward.

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 NCP's outcome and the peer effect that NCP applies to its neighbors. We consider the case with interference from both treatment and outcome (see Figure 2b). This means the behavior of the NCP additionally influences the outcome of its neighbors, so non-cooperative bias in the network setting can diffuse through the network via peer effects into outcomes of neighboring participants.

Assumptions. Recall this work conditions on the fact that there is access to a high-quality classifier for non-cooperative participants. OSNs have established systems to identify non-cooperative behavior. Our system provides the follow-up analysis once NCPs are identified. OSNs can use the theoretical approach in this work to provide a foundation for solutions deployed on live data. OSNs may employ tests which detect the presence of peer effects [7, 63].

We assume that each non-cooperative participant in the network follows the same behavioral model and that the outcome of individuals is bounded.

4.1 Non-Cooperative Behavioral Models

We consider the following set of possible behavioral models a non-cooperative participant might follow:

Uniform-Random The participant responds randomly from a uniform distribution over the outcome space, regardless of treatment assignment.

Maximum The treated participant responds with the maximum outcome, and the control participant responds with the minimum outcome, in order to inflate the estimated treatment effect.

Minimum The treated participant responds with the minimum outcome, and the control participant responds with the maximum outcome in order to minimize the estimated treatment effect.

Pooling Several participants coordinate to exchange or pool their outcome responses, resulting in incorrectly recorded outcomes (swapped with other individuals in the NCP pool).

The Uniform-Random behavioral model results in an increase in variance over the ATE. The non-cooperative participants inject noise into the estimate through randomly sampled outcomes. In the network setting, this increases the amount of random noise observed by neighbors, but does not systematically bias neighbor outcomes.

The Maximum and Minimum behavioral models potentially bias the effect estimation in some direction. Because these behavioral models produce extremes in their outcome response function, participants can bias the outcomes of their neighbors in settings with sufficiently large peer effects.

The Pooling model is a special case of bootstrap sampling: unit outcomes are resampled from a given sample of outcomes (i.e., outcomes of other units in the pool). This can act as a

method for data poisoning (it can result in an incorrect record of treatment assignment), and works to obfuscate personal data (i.e., in self-reported outcomes). The NCP behavioral models describing the loyalty-card-swapping and Instagram-sharing-ring examples fall under this category.

4.2 Non-Cooperative Exposure Coverage

When NCPs form a dominating set over the graph, every node is exposed to at least one non-cooperative outcome. Determining whether a set of vertices is a dominating set reduces from the vertex cover problem and is an NP-complete decision problem [31]. To approximate the smallest dominating set, we use a standard greedy procedure to select nodes from the graph using a simple heuristic from the number of uncovered nodes, breaking ties with node degree [19]. Following the additive treatment exposure framework (Section 3.2.3), we use uniform edge-weighting in the heuristic.

In the uniform edge-weight setting, nodes providing the greatest amount of coverage are not necessarily nodes with highest degree. Uniform edge-weighting in the peer-effects outcome model results in an inverse-degree-weighted influence from neighbors: a node with fewer neighbors has a larger inverse-degree weighting per neighbor. We define *non-cooperative influence*, ω_i, of a node i in the network G as:

$$\omega_i = D^{-1} A \mathbf{1}_i^N \tag{6}$$

where D is the degree matrix of G, A is the adjacency matrix of G, and $\mathbf{1}_i$ is a column 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 NCPs between different graphs, we take the sum of ω_r over all non-cooperative participants r in the set and divide by N to normalize.

To maximize non-cooperative influence over the graph, we construct an heuristic to capture both the number of nodes covered by a set of non-cooperative participants and the relative strength of effect they hold over their neighbors. We define *participant degree influence* ω' as node influence ω calculated over non-NCP neighbors. That is,

$$\omega'_i = \sum_{j \in Nb(i) - R} \frac{1}{d_j} \tag{7}$$

where $Nb(i)$ is the set of nodes adjacent to i, and d_j is the degree of node j.

NCP selection methods. We compare the following selection methods for constructing the dominating set:

RNCP Selects nodes to represent non-cooperative participants uniformly at random.

GNCP Greedily selects nodes to represent non-cooperative participants, using participant degree influence ω' as a heuristic.

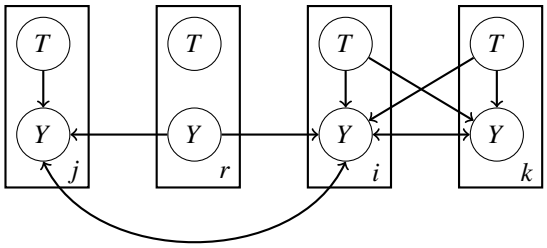


Figure 3: In networked environments with non-cooperative participants (NCPs), the non-cooperative node r can partially block the peer effects acting between unit j and unit i (note the arrows from Y_r are no longer double-headed as in Figure 2b). Because the NCP behaves according to some external protocol (e.g., Uniform-Random, Minimum, etc.), its outcome is independent of treatment (i.e., there is no arrow from T to Y). Furthermore, r's behavior can influence the outcomes of adjacent nodes, and *this influence is independent of treatment*.

RNCP selection procedures up to dominating sets result in larger sets containing nodes with a range of neighborhood sizes, which in this setting control the strength of ties. GNCP selection is equivalent to a heuristic greedily selecting nodes with strong ties in the degree-weighted setting.

5 Bias from Non-Cooperative Participants

We are interested in exploring how non-cooperative behavior in network systems can influence ATE estimation (**RQ1**). Previous work has shown that variance over ATE estimation using graph cluster randomization is large and sensitive to several choices in the experimental setup: e.g., estimation parameters, clustering method, and treatment assignment each influence the variance over the ATE estimate for a particular graph [23, 30, 65, 68]. Given the plurality of factors that influence the variance over ATE estimation, we focus our analysis on the *bias* in the ATE estimate due to non-cooperative behavior. Since we are interested in examining bias in the most extreme case, the Maximum and Minimum behavioral models are of greater interest as those extreme-response behaviors result in stronger peer effect, especially in estimation methods that include the neighborhood outcome mean as a parameter in the effect estimation. We constrain our analysis to a single behavioral model shared between all non-cooperative participants in the network. We assume NCP node identities and behavioral functional forms are known; see Section 9 for a discussion relaxing these assumptions. For an example of how NCPs disrupt peer effects, contrast the causal relational diagram of Figure 2b, which highlights spillover effects, with the causal relational diagram of Figure 3, which highlights *non-cooperative* spillover.

Let $\delta_R(\hat{x}, k)$ denote the bias in the estimate of x due to k non-cooperative participants in the population.

5.1 Non-Cooperative Participation in the Propositional Setting

In the propositional setting, the only influence an NCP can exert on the estimated ATE is through its own behavior. We can therefore separate the outcomes of non-cooperative participants from the outcomes of compliant units:

$$\hat{\tau} = \underbrace{\frac{1}{N_1}\sum_{\substack{i=1\\ z_i=1\\ i\notin R}}^{N} Y_{1i} - \frac{1}{N_0}\sum_{\substack{i=1\\ z_i=0\\ i\notin R}}^{N} Y_{0i}}_{\text{ATE for compliant units}} + \underbrace{\frac{1}{N_1}\sum_{\substack{r\in R\\ z_r=1}} Y_{1r} - \frac{1}{N_0}\sum_{\substack{r\in R\\ z_r=0}} Y_{0r}}_{\text{ATE for NCPs}} \quad (8)$$

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

$$\hat{\tau} = \frac{1}{N_1}\sum_{\substack{i=1\\ z_i=1\\ i\notin R}}^{N} Y_{1i} - \frac{1}{N_0}\sum_{\substack{i=1\\ z_i=0\\ i\notin R}}^{N} Y_{0i} + \underbrace{\frac{|R_1|}{N_1}Y_{R1} - \frac{|R_0|}{N_0}Y_{R0}}_{\text{Bias}} \quad (9)$$

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

$$\delta_R(\hat{\tau}, |R|) = \frac{|R_1|}{N_1}Y_{R1} - \frac{|R_0|}{N_0}Y_{R0} \quad (10)$$

5.2 Non-Cooperative Participation in the Relational Setting

In systems with outcome interference, the treatment effect diffuses through the network. This requires a temporal component in the model. In the first time-step, the treatment influences only 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 $Y_{z_i(j,\,t-1)}$, the outcome of each neighbor j of i at time $t-1$.

5.2.1 Non-Cooperative Influence and Graph Topology

Non-cooperative participants in the network setting are particularly interesting because these nodes can block the flow of treatment effect through the network. Non-cooperative behavior defined irrespective of $Y_{z_i(j,\,t-1)}$ blocks peer effect that would normally spill-over and affect the outcome of j and, through j, j's neighbors. This distorts the diffusion of treatment effect over the network, and the degree of that distortion increases as the number of time-steps increases and treatment further propagates across the network. Naturally, diffusion of treatment effect through the network also allows exposure to adversary behavior to propagate through the network. This is why astroturf campaigns are effective: artificial accounts target individuals susceptible to peer effects and push them toward a particular outcome, and those individuals in turn affect their neighbors.

Given the relationship between non-cooperative participants, neighborhood outcomes, and ATE bias, we would like to examine how placement of NCP nodes in the network graph relates to the strength of non-cooperative bias.

5.2.2 Empirical Study of Influence and Topology

It is likely that differences in network topology can affect the strength of total non-cooperative influence (**RQ3**). To examine the relationship between topology and NCP set selection, we empirically analyzed the increase in total normalized non-cooperative influence as the number of NCPs increases using the random and greedy selection procedures for three different random graph generation procedures: small-world networks [70], forest-fire models [42], and stochastic block models [26]. All parameters were chosen either from examples in the corresponding papers (i.e., [70], [42], and [26]), or tuned to match median edge counts for stochastic block models, which are designed to model real-world community structure. Generating graphs with the same number of nodes and median edge count ensures that comparisons between generated graphs are specific to differences in graph topology alone.

Small-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_{sw}. 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 [70].

Forest-fire models. Forest-fire models are a type of preferential attachment graph. Graphs are constructed by adding nodes one at a time. When a node n is added to the network, an edge is added between n and some other node a chosen uniformly at random from the current set of nodes. Then, some number x, y of additional edges are added between a and other nodes in the graph: x is the number of new outgoing edges from a, and y is the number of new incoming edges to a. These parameters are controlled by parameters p_{ff}, r and assigned by sampling from geometric distributions with means $p_{\text{ff}}/(1-p_{\text{ff}})$ and $rp_{\text{ff}}(1-rp_{\text{ff}})$, respectively. Networks generated in this way have long-tailed in- and out-degree distributions, community structure, shrinking graph diameter,

and densification following a power law, each of which are properties identified in real-world networks [42].

Stochastic block models. 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 [26].

Our analysis considers these three graph generation procedures for graphs of size 500, 1000, and 5000. We selected graph generation parameters such that the median number of edges in generated graphs fell within 10% of the same edge count (1350, 4600, and 125,000 edges for each graph size, respectively). For forest-fire models, we generate graphs with forward-burning probability p_{ff} and backward burning ratio r pairs $(p_{\text{ff}}, r) \in \{(0.32, 1.031), (0.37, 0.892), (0.37, 0.946)\}$. For small-world networks, we generate graphs with rewiring parameter $p_{\text{sw}} \in \{0.03, 0.05, 0.1\}$. For stochastic block models, we generate graphs with intracommunity attachment probability 0.8 and intercommunity attachment probability $\in \{0.1, 0.2, 0.3\}$. For each graph setting, we generate 100 graphs of that type. Results were consistent across size and parameter settings within each graph type, so we report results for 1000-node graphs with a single parameter setting. For forest-fire models, we set $(p_{\text{ff}}, r) = (0.37, 0.892)$. For small-world networks, we set $p_{\text{sw}} = 0.05$. For SBMs, we set $\mu = 0.2$.

Findings. Figure 4 shows the increase in total normalized non-cooperative influence as the number of non-cooperative participants increases for SBMs, small-world graphs, and forest-fire models with 1000 nodes. Both GNCP and RNCP selection procedures are considered. Influence of NCPs is closely related to the connectivity and degree distribution of the graph.

Small-world graphs show low total non-cooperative influence, even under GNCP selection of NCPs. This is due to the construction procedure for the graph. The degree distribution in small-world graphs is tight, so all nodes have nearly the same number of neighbors. As a result, no individual node is likely to have a significantly greater influence than any other in the network and there is little difference between the NCP sets constructed under GNCP or RNCP selection procedures.

Forest-fire models require the largest dominating sets (35% of the nodes), and naturally reach the highest non-cooperative influence. As the size of the NCP set increases, there is a significant and increasing difference between the non-cooperative influence of NCP-sets selected under GNCP compared to NCP-sets selected under RNCP. A dominating set selected under GNCP in the forest-fire simulations total an average of 0.69 influence over the graph, and the curve of total influence increases linearly in the number of NCPs.

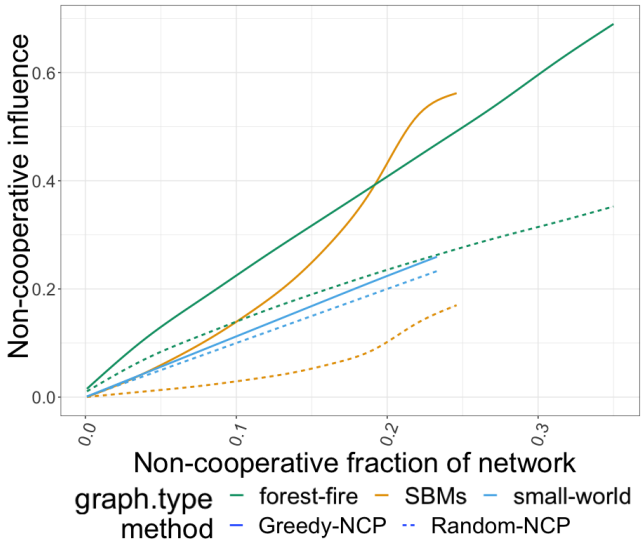


Figure 4: Total normalized influence of non-cooperative participants as the number of NCPs increases in stochastic block models, small-world networks, and forest-fire models. We considered cases where NCPs are selected either uniformly at random (RNCP, dashed) or greedily using the participant degree influence as a coverage heuristic (GNCP, solid).

SBMs also show a large difference in total non-cooperative influence between NCP selection methods. When NCP sets are selected under RNCP, their average non-cooperative influence (ω_r) is 0.17, whereas dominating sets constructed under GNCP selection have an average non-cooperative influence of 0.57.

In forest-fire models, the difference in total non-cooperative influence between NCP-sets selected under GNCP and RNCP is a consequence of the long tails of the degree distributions. Since nodes of high degree must be drawn from the tails of the degree distribution, they are less likely to be selected under RNCP. Nodes with low degree are more likely to be selected under RNCP as the probability density of low degree is much higher in these degree distributions.

5.3 Bias in Average Treatment Effect (ATE)

We will now examine bias in estimated ATE due to non-cooperative participation (**RQ1**). First, we can derive the bias due to NCPs in the linear estimator from Gui et al. [30], shown in Equation 5. Let σ be a vector containing the portion of treated neighbors for each node in the network. Recall that

$\hat{\tau} = \hat{\beta} + \hat{\gamma}$. The parameters β, γ are estimated as (see Eq. 5):

$$\hat{\beta} = \frac{(\sum \sigma^2)(\sum \mathbf{Z}Y) - (\sum \mathbf{Z}\sigma)(\sum \sigma Y)}{(\sum \mathbf{Z}^2)(\sum \sigma^2) - \sum(\mathbf{Z}\sigma)^2} \tag{11}$$

$$\hat{\gamma} = \frac{(\sum \mathbf{Z}^2)(\sum \sigma Y) - (\sum \mathbf{Z}\sigma)(\sum \mathbf{Z}Y)}{(\sum \mathbf{Z}^2)(\sum \sigma^2) - \sum(\mathbf{Z}\sigma)^2} \tag{12}$$

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

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

$$\hat{\gamma} = \frac{N_1(\sum \sigma Y) - (\sum \mathbf{Z}\sigma)(\sum Y_1)}{N_1(\sum \sigma^2) - \sum \mathbf{Z}\sigma^2} \tag{14}$$

There are two ways a non-cooperative participant 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 non-cooperative participants R, and $\hat{\beta}_Y, \hat{\gamma}_Y$ be the bias in estimated parameters due to neighborhood exposure to outcomes from R, so that

$$\delta_R(\tau, k) = \hat{\beta}_R + \hat{\gamma}_R + \hat{\beta}_Y + \hat{\gamma}_Y \tag{15}$$

Since these parameters are estimated as sums over each unit individually, we can divide $\hat{\beta}$ and $\hat{\gamma}$ into terms accounting for outcomes from compliant units separately from outcomes of non-cooperative participants. Then the estimate of the parameters due to non-cooperative outcome is given:

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

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

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

$$\hat{\beta}_R + \hat{\gamma}_R = \frac{|R_1|Y_{1R}\left(\sum \sigma^2 - \sum \mathbf{Z}\sigma\right) + (N_1 - \sum \mathbf{Z}\sigma)\left(\sum_{r\in R} \sigma_r Y_r\right)}{N_1(\sum \sigma^2) - \sum \mathbf{Z}\sigma^2} \tag{18}$$

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

Reasoning about bias due to non-cooperative network influence through parameter estimation is difficult, since the bias due to non-cooperation 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 non-cooperative outcome is separated into (1) the portion of its outcome independent of its neighbors, (2) the portion of outcome due to unbiased peer effects, and (3) the portion of outcome due to peer effects from NCPs, we still must account for the non-cooperative peer effects in i's NCP-exposed neighbors in (2), which may be exposed to a different non-cooperative participant than r. Instead of reasoning about the strength of non-cooperative diffusion, we can approximate the bias induced by a single NCP r's outcome on non-NCP 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: $\bar{Y}_{A_j \setminus r}$. Then the bias in Y_j due to Y_r is $\propto \frac{1}{d_j}(Y_r - \bar{Y}_{A_j \setminus r})$, where d_j is the degree of node j, and Y_r is the outcome of non-cooperative participant r under treatment assignment z_r. So the total bias induced by r on its neighbors' outcome is approximated by:

$$\begin{aligned} \hat{\beta}_Y + \hat{\gamma}_Y &= \sum_{j\in A_r} \frac{1}{d_j}(Y_r - \bar{Y}_{A_j \setminus r}) \\ &= \omega_r \sum_{j\in A_r} (Y_r - \bar{Y}_{A_j \setminus r}) \approx \omega_r (Y_r - \bar{Y}_{A_r^2}) \end{aligned} \tag{19}$$

where $\bar{Y}_{A_r^2}$ is the mean outcome in r's two-hop neighborhood. Then the total ATE bias due to non-cooperative participants in the network is:

$$\begin{aligned} \delta_R(\tau, k) = &\frac{|R_1|Y_{1R}\left(\sum \sigma^2 - \sum \mathbf{Z}\sigma\right) + (N_1 - \sum \mathbf{Z}\sigma)\left(\sum_{r\in R} \sigma_r Y_r\right)}{N_1(\sum \sigma^2) - \sum \mathbf{Z}\sigma^2} \\ &+ \sum_{r\in R} \omega_r \left(Y_r - \bar{Y}_{A_r^2}\right) \end{aligned} \tag{20}$$

5.4 Expected Bias

We now examine the expected bias under random and dominating non-cooperative node placement (**RQ2**).

Recall the assumptions that NCPs behave according to their treatment assignment, and are placed either uniformly at random across the graph (RNCP), or greedily according to graph structure (GNCP). That is, the GNCP procedure does not consider treatment assignment across the graph in its heuristic. For logical consistency, we will assume NCPs are placed before treatment is assigned. Thus, in our expected bias analysis, we will not optimize NCP placement with respect to treatment assignment.

We specifically consider the bias induced by an NCP r on its neighbors' outcome, $\hat{\beta}_Y + \hat{\gamma}_Y$. We use this narrowed view because the bias induced by non-cooperative outcome alone,

$\hat{\beta}_R + \hat{\gamma}_R$, is related to the placement of NCPs only through exposure to treatment. In particular, it is related to the treatment assignments across r's neighborhood. Reasoning over behavior on nodes exposed to both treatments is beyond the scope of this work. For detailed discussion of the relationship between clustering, neighborhood exposure, and ATE estimation, see Ugander et al. [68]. For our analysis, we assume that a node and its neighbors receive the same treatment assignment.

Here we discuss the expected ATE bias induced by k non-cooperative participants in the graph. We will give expressions for both random and dominating NCP placement. Note however that due to our assumption that each node has the same treatment assignment as each of its neighbors, the only term affected by NCP placement in expectation over $\hat{\beta}_Y + \hat{\gamma}_Y$ is ω_r.

5.4.1 Random Placement

Under the RNCP placement protocol, we would expect the average case bias from non-cooperative participants. We use $\bar{\omega}$ to denote expected influence of a randomly placed node, whose expression is given by:

$$\bar{\omega} = \mathbb{E}[\omega] = \sum_{i=1}^{N} \mathcal{P}(\omega_i)\omega_i \tag{21}$$

where $\mathcal{P}(\omega_i)$ is the probability of observing node influence ω_i, determined by the degree distribution in the graph.

So the expected value of bias in network treatment effect from k non-cooperative participants placed uniformly at random is:

$$\mathbb{E}\left[\hat{\beta}_Y + \hat{\gamma}_Y\right] = k\bar{\omega}\left[\frac{N_1}{N}\left(Y_{R1} - \bar{Y}_{\left(1,\, A^2_G\right)}\right) + \frac{N_0}{N}\left(Y_{R0} - \bar{Y}_{\left(0,\, A^2_G\right)}\right)\right] \tag{22}$$

5.4.2 Greedy Coverage Placement

The expected value of ATE bias from NCPs placed under the GNCP protocol follows a similar formulation except that we maximize the total non-cooperative influence for a set of k NCPs. Let ω^*_k be the maximum total influence for a dominating set of size k. Then the expected value of bias in network treatment effect for non-cooperative participants forming a dominating set over the graph is as follows:

$$\mathbb{E}\left[\hat{\beta}_Y + \hat{\gamma}_Y\right] = \omega^*_k\left[\frac{N_1}{N}\left(Y_{R1} - \bar{Y}_{\left(1,\, A^2_G\right)}\right) + \frac{N_0}{N}\left(Y_{R0} - \bar{Y}_{\left(0,\, A^2_G\right)}\right)\right] \tag{23}$$

Finding the dominating set that maximizes the total non-cooperative influence for a set of k NCPs can be formulated as a weighted maximum coverage problem, where the universe U is the set of all nodes, each subset S_i is a set containing the node i and its neighbors, $S_i = \{a \in V(G) \mid a = i \vee \exists (a,i) \in E(G)\}$, and each set's weight $w(i) = \sum_{a \in S_i} \frac{1}{d_a} = \omega'_a$, where d_a is the degree of node a. The weighted maximum coverage problem is NP-hard but can be approximated with a greedy algorithm with an approximation ratio of 1 - $\frac{1}{e}$, with e the base of the natural logarithm. [49].

6 Simulation Study

To empirically demonstrate the effect of non-cooperative participation on OSN effect estimation, we simulated outcomes in a network and used the ATE estimation method of Gui et al. [30] to estimate treatment effect in networks both with and without non-cooperative interference. The experiments consider outcome simulations over both synthetic and real-world graph structures. We use simulations of outcomes over a real-world graph structure as the closest analog to real-world data available for this setting. Simulation studies like the one presented in this work are not uncommon in the network science or statistical relational literatures, where real-world data measuring effect estimation is not available due to e.g., digital privacy concerns, medical record privacy protections, etc. In other cases, real-world data may be available, but the sample size is small or otherwise inappropriate [24].

Where possible, we have substituted synthetic aspects of the simulation study with publicly available empirical data to improve the external validity of the study. This is in line with best practice in evaluating causal effect estimation [28]. The real-world simulation experiments use real-world OSN structure, published individual and peer effect estimates for the same OSN platform, and the outcome simulation model is drawn from prior work developed at a competing OSN.

6.1 Methodology

For a given graph, we generate a clustering of the graph and assign clusters to treatment or control. We then generate a dominating set of non-cooperative participants using a greedy algorithm (GNCP). Recall that this procedure is guaranteed to approximate the maximum total non-cooperative influence within (1-e). A second NCP set of the same size is selected uniformly at random from the graph (RNCP). Given this selection procedure, the RNCP selected non-cooperative nodes are not guaranteed to form a dominating set over the graph.

For specific settings of the individual treatment effect parameter λ_1 and peer effect parameter λ_2, we determine the outcome function for compliant and non-cooperative nodes under the outcome simulation model proposed in Eckles et al. 2014 [23]. We then simulate outcomes iterating over the set of NCP nodes—starting with an empty set and adding an NCP to the previous set—ending with the entire dominating set. For each NCP subset, we simulate outcomes $Y_{i,t}$ for each node i up to $t = 3$. We use the linear estimator over generated outcomes to estimate ATE for each subset of non-cooperative participants. The estimated ATE for the experiment with no non-cooperative interference is used as a baseline of comparison for ATE bias induced by NCPs.

6.2 Network Structure

Synthetic graphs are generated under algorithms designed to match properties of real-world networks, and which result in some form of community structure. Several induce a power-law distribution over node degrees in generated graphs, consistent with standard practice in the network science literature. This power-law relationship has been consistently reported in real-world networks [50].

In addition to the synthetic graph experiments, we also include an experiment simulation with non-cooperative participants using empirical data.

6.2.1 Synthetic Graph Generation

Synthetic network graphs are generated using the same procedure reported in Section 5.2.2. Our results are consistent within graph type, so we report single parameter settings. We consider forest-fire models with forward burning probability $p_{\text{ff}} = 0.37$ and backward burning ratio $r = 0.892$, small-world networks with rewiring parameter $p_{\text{sw}} = 0.05$, and SBMs with community mixing parameter $\mu = 0.2$. Interestingly, the size of the dominating set relative to the graph decreases as the size of the graph increases for graphs with the same parameter values. However, graphs of different size also produce the same pattern of bias increases within graph type, so we report results for graphs with 1000 nodes.

6.2.2 Real-World Graph Structure

For the real-world graph structure, we use a public release of the Facebook subgraph [43] available through the SNAP library [44]. This subgraph was collected by choosing a 10-clique set of users and taking the subgraph induced by this set of users and their neighborhoods. [3] This network sample has 4039 nodes and 88,234 edges.

6.3 Outcome Simulation Model

The simulation model defines a protocol for treatment assignment and uses a linear model of outcome incorporating individual, peer, and neighborhood-level effects on unit response.

Treatment assignment. The experiment simulations use cluster-randomized assignment to minimize treatment exposure between treatment groups. We use the Infomap algorithm for community detection [59] to generate a set of clusters over the network with c clusters. We assign clusters to treatments according to a binomial distribution $\mathcal{B}(c, 0.5)$.

[3] As a result of this subgraph sampling procedure, the 10-clique set of ego nodes forms a dominating set over all 4039 nodes in the network sample.

Participant outcome model. Observed outcome values were generated using the following linear model introduced by Eckles et al. [23] and further adapted from Gui et al. [30] :

$$Y_{i,t} = \lambda_0 + \lambda_1 z_i + \lambda_2 \frac{A_i Y_{t-1}}{D_{i,i}} + U_{i,t} \tag{24}$$

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. At $t = 0$, we set $Y_{0,i} = 0$ for all i.

Outcome parameter values. Following the parameter assignments of [30], we set $t = 3$, $\lambda_0 = -1.5$, and we set individual treatment parameters $\lambda_1 \in \{0.25, 0.5, 0.75, 1\}$ and peer effect parameters $\lambda_2 \in \{0, 0.1, 0.5, 1.0\}$.

In the simulation study using empirical data, we also consider parameter values drawn from published treatment effect estimates from experiments on Facebook [22]: for individual treatment effect we set $\lambda_1 \in \{1.845\text{e}{-4}, 1.929\text{e}{-4}, 1.995\text{e}{-4}\}$, and for peer treatment effect we set $\lambda_2 \in \{1.086\text{e}{-3}, 1.111\text{e}{-3}, 1.136\text{e}{-3}\}$.

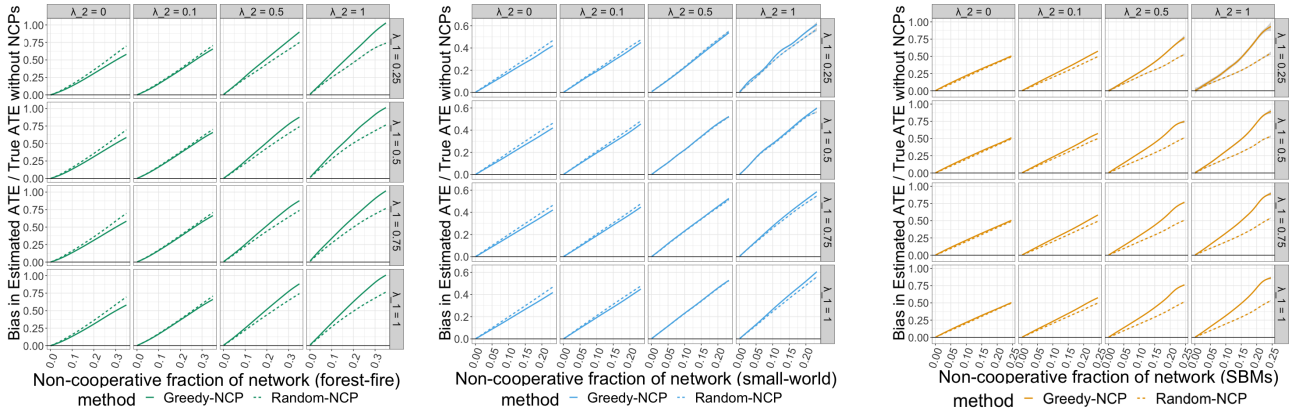
Non-cooperative outcome. Under the non-cooperative behavioral model Minimum, non-cooperative participants respond to minimize the estimated ATE (described in Section 4). Non-cooperative outcome is determined by λ_0 and λ_1:

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

6.4 Synthetic Network Results

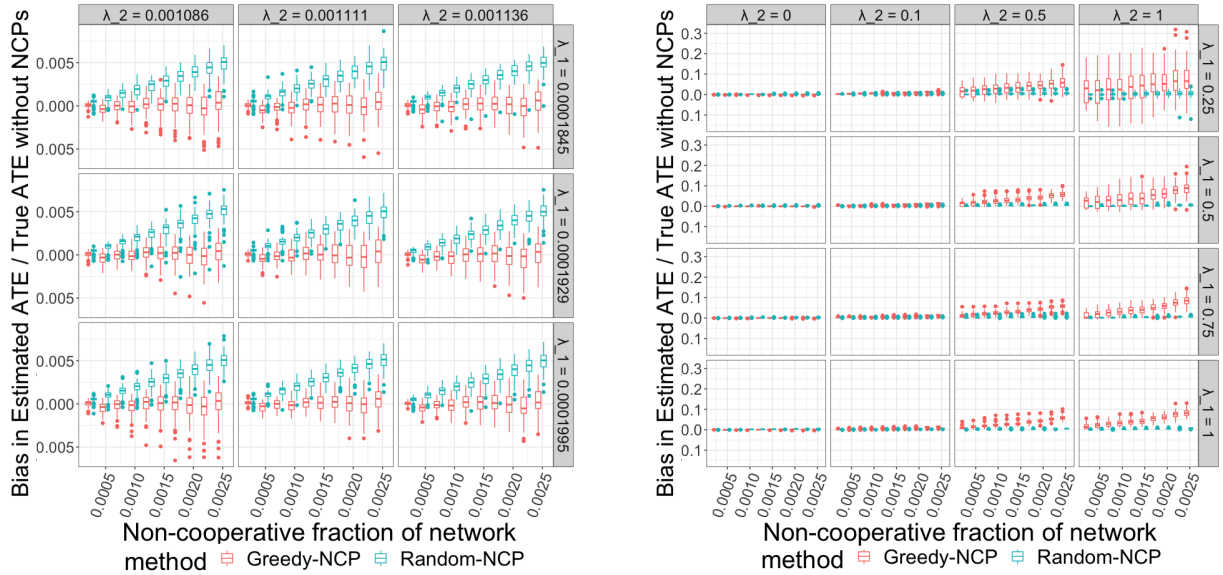
The simulation results in Figure 5 demonstrate the effect of NCPs on the estimated treatment effect. The bias induced by NCPs increases as the number of non-cooperative nodes increases. GNCP placement generally results in greater bias than RNCP placement, especially for large peer effect, λ_2. As individual treatment effect λ_1 increases, the slope of the bias increase remains constant. Under RNCP, the increase in peer effect λ_2 has little effect. The dramatic changes in bias occur with increases to peer effect for NCPs selected under GNCP. For λ_2=1, SBMs with GNCP selection nearly double the ATE bias of NCP-sets selected under RNCP.

The importance of peer effect strength also depends on network topology. SBMs and forest-fire models are particularly susceptible to bias in ATE for networks with large peer effects, especially under GNCP selection procedures. These two graph generation algorithms are often cited as producing graphs closely resembling real-world networks, and SBMs in particular are recommended as random graphs most closely replicating real world community structure [40, 42, 74]. It is interesting that these structures are also the most vulnerable to non-cooperative biasing, even for relatively small sets of non-cooperative participants.



(a) forest-fire networks (N = 1000, p_{ff}=0.37, r=0.892) (b) small-world graphs (N = 1000, p_{sw}=0.05) (c) SBMs (N = 1000, μ=0.2)

Figure 5: Bias in estimated ATE for (a) forest-fire models, (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. Forest-fire networks and SBMs exhibit significant bias with increases in the strength of peer effects, even for a small set of non-cooperative participants.



(a) 4039-node Facebook subgraph with published effect parameters (b) 4039-node Facebook subgraph with simulated effect parameters

Figure 6: Bias in estimated ATE on a subgraph from the Facebook graph using (a) published individual and peer effect estimates [22], and (b) simulated individual and peer effect estimates [30]. Rows share individual treatment effect parameter settings, λ_1, and columns share network treatment effect settings, λ_2. When effect size is small, RNCP placement results in greater bias in ATE than GNCP placement. In general, there is greater variability in ATE bias under GNCP placement.

6.5 Real-World Network Results

Figure 6 shows the results of the Facebook graph simulation under two sets of outcome simulation model parameters. In general, there is greater variability in ATE bias under GNCP placement. When the true individual and peer effect size are small as in Figure 6a, RNCP results in greater bias in ATE than under GNCP. This is consistent with the synthetic network simulations with small peer effect. Results in Figure 6b also show a similar pattern of bias increases as the synthetic network experiments (Section 6.4), which use the same settings for outcome parameters λ_1, λ_2.

7 Implications

In this section, we outline two hypothetical but realistic scenarios to emphasize the implications of our work in both the propositional and relational settings.

Propositional non-cooperative effects. In 2020, Twitter introduced labels on tweets that may contain questionable content [60]. Two such labels were "Get the facts about COVID-19." (*Label A*) and "Some or all of the content shared in this Tweet conflicts with guidance from public health experts regarding COVID-19. Learn more." (*Label B*); if these two labels are randomly assigned, they would constitute an A/B test, or simple experiment. The outcome of this experiment is the rate of spread of misinformation.

End-users would be aware of the new feature and may even suspect that they were in an experiment. Now consider an agent that controls a large number of fake or compromised Twitter accounts (a botherder [13], for instance). Suspecting Twitter's goals, the agent may choose to behave in a manner that propagates misinformation (and disinformation) about COVID-19 vaccines. Such an adversary might instruct the bots to behave in a way that influences the outcome of the A/B test (e.g., by flooding one label with clicks).

Relational non-cooperative effects. Now consider Facebook's election-cycle voter engagement campaigns, in which users received messages with reminders about election events and notices about friends that had checked-in as voters [12,36]. In this scenario, the treatment is a dedicated post in the experimental subject's news feed. Following the design of Jones et al. [36], the posts prominently feature a general-audience message, "It's election day. Tell friends you're voting in the 2012 Election and find out where to vote." (*propositional; Condition P*), or may list specific users if at least one friend had checked-in as having voted, "Kaleigh and 7 other friends are voters. Find out where to vote." (*relational; Condition R*).

Suppose Facebook is interested in evaluating click-through rates of posts over both conditions P and R.[4] Click-through rates of users exposed to R may depend on the accounts featured in the message. Users exposed to P may still influence their neighbor's voting outcomes by sharing their own voting status or other voting messaging as a result of their exposure under P. If some accounts skip the check-in to obfuscate their true voting status, that may influence the probability of adoption of users that may have otherwise adopted if exposed to the factual outcome. This is especially a concern when the *treatment cluster assignments are correlated with existing polarization in the network communities*.

[4]This is known as a 2x2 experimental design and it allows for the calculation of the effects of P only, R only, both P and R together, with no treatment as the control condition.

Summary. These two example scenarios are based on real-world experiments performed by social networking companies; the non-cooperative behavior is hypothetical, but demonstrates *how* A/B testing can be subverted, leading to potentially damaging real-world consequences (i.e., the inadvertent sub-optimal stemming of COVID-19 misinformation in the propositional scenario and the effects non-cooperative peer-engagement on voting in the relational scenario). In view of this, it is important for A/B testing operations to account for ATE bias as discussed in Section 5.

8 Ethics of Correction

It is important to note that non-cooperation does not automatically imply malicious or adversarial action. A user can act as an adversary of the OSN system *without being malicious*. In general, non-cooperative action may just as likely be non-malicious as not, and we should be careful about strictly adversarial framing of non-cooperative behavior [5]. Non-cooperative behavior may be a means of e.g. denying information to an OSN platform, which may protect against an automated system using information about the user that leads to biased application of services downstream. Non-cooperative behavior need not be directly related to the experiment underway (users may not know they are in an experiment), but those behavioral choices may still affect outcome estimates, including and especially when the treatment is cluster-randomized and the non-cooperative behavior is observed by peers.

As further ethical consideration for calculating non-cooperative bias terms for ATE correction, we note that this is, necessarily, removing the outcomes of nodes identified as non-cooperative. If those nodes share some feature (are acting to e.g., preserve their privacy), this induces a data-missingness selection bias on the corrected ATE estimates which may algorithmically perpetuate existing bias and injustice [11, 29].

9 Discussion

Our work is related to other studies of non-cooperative behavior in the network setting. Some of these investigate the spread of infection [45] or misinformation [15] in a network. Other work examines detection of adversaries [4, 16, 69] or measuring and analyzing the success of adversary integration in the network [1, 3]. The most similar work to ours in structure is Yildiz et al. [73], who model opinion networks that have "stubborn agents" placed in various configurations. A graphical representation of this work would overlay the network connectivity of Figure 2a with edges representing the flow of opinion; there is no notion of treatment and outcome in their framework, which fundamentally changes the problem formulation. This paper is the first to explore the effect of bots and other non-cooperative participants in network A/B testing.

When handling non-cooperative units in experiments in networked environments, there are three main problems or tasks: non-cooperative *detection*, statistical bias *correction*, and poisoning *prevention*. This work focuses on correction, relying on assumptions about our ability to identify non-cooperative network members and know the model of their behavior.

Clearly there is a need for future work on the detection and modeling of non-cooperative behavior in relational settings, e.g., OSNs. A major challenge in such detection is differentiating between strong opinion-holders and truly adversarial or trollish behavior [18]. Classifying non-cooperative behavior is a socio-technical issue and raises ethical concerns, especially if it is done behind walled gardens.

That said, there is still a great deal to study about the interaction of non-cooperative behavior and peer effects. As future work, we are interested in the effects of additional non-cooperative behavioral models. We have focused on extreme-response models of non-cooperative behavior, but non-cooperative participants may act under a wide range of other behavioral models. NCPs may condition their outcome behaviors according to not only their treatment assignment, but the treatment assignment and behavior of their neighbors. We might also consider less extreme non-cooperative 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.

This work also does not address the task of prevention. While there are certainly sociological approaches to prevent non-cooperative behavior, an alternative is to simply design experiments that are robust to such behavior. Our work has demonstrated that the ATE estimate using standard network A/B testing protocols can be biased by non-cooperative behavior. While there may be more robust estimators (e.g., median treatment effect [41]), these estimators are not widely used.

10 Conclusion

This work presents an introductory analysis of the effect of non-cooperative behavior on the average treatment effect (ATE) estimate in networks. Non-cooperative behavior adds a layer of complexity over peer effects in network experiments and is a recognized issue in the literature not considered in standard experimental analysis for the relational setting. Our work demonstrates a vulnerability in cluster-randomized network A/B testing to manipulation under non-cooperative behavior, particularly for networks with long-tailed degree distributions. We have shown that networks with strong peer effects are susceptible to ATE bias from non-cooperative behavior and identified forest-fire models and SBMs as network structures vulnerable to non-cooperative spillover effects. Our experiments using a real-world network show results consistent with our findings in synthetic networks.

Availability

Code to support graph generation, outcome simulation models, and empirical measurements of non-cooperative bias used in the experiments and simulation studies presented in this work is available at https://github.com/KDL-umass/Non-cooperative-spillover.

Acknowledgements

We thank the anonymous reviewers for their careful review and thoughtful consideration of this paper, which has greatly benefited from the reviewer suggestions. We would like to thank Dan Corkill and Andrew McGregor for initial discussions of non-cooperative behavior under cluster-randomized treatment designs for network A/B testing. We also thank Amanda Gentzel, David Arbour, Katerina Marazopoulou, and other members of the Knowledge Discovery Laboratory at University of Massachusetts Amherst for their suggestions and comments on an early version of this work.

This research was sponsored by the Defense Advanced Research Projects Agency (DARPA), the Army Research Office (ARO), and the United States Air Force (USAF), and was accomplished under Cooperative Agreement Number W911NF-20-2-0005 and Contract Number FA8750-17-C-0120. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies or views, either expressed or implied, of the DARPA, ARO, USAF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright notation herein.

References

[1] Norah Abokhodair, Daisy Yoo, and David W. McDonald. Dissecting a social botnet: Growth, content and influence in Twitter. In *Proc. of the ACM Conf. on Computer Supported Cooperative Work Social Computing (CSCW)*, 2015.

[2] Kiyan Ahmadizadeh, Bistra Dilkina, Carla P. Gomes, and Ashish Sabharwal. An empirical study of optimization for maximizing diffusion in networks. In *Proc. of the 16th Int'l. Conf. on Principles and Practice of Constraint Programming (CP)*, 2010.

[3] Luca Maria Aiello, Martina Deplano, Rossano Schifanella, and Giancarlo Ruffo. People are strange when you're a stranger: Impact and influence of bots on social networks. In *Proc. of the 6th Int'l. Conf. on Weblogs and Social Media (ICWSM)*, 2012.

[4] Leman Akoglu, Rishi Chandy, and Christos Faloutsos. Opinion fraud detection in online reviews by network

effects. In *Proc. of the 7th Int'l. Conf. on Weblogs and Social Media (ICWSM)*, 2013.

[5] Kendra Albert, Jon Penney, Bruce Schneier, and Ram Shankar Siva Kumar. Politics of adversarial machine learning. In *Towards Trustworthy ML: Rethinking Security and Privacy for ML Workshop, Int'l. Conf. on Learning Representations (ICLR)*, 2020.

[6] Peter M Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947, 2017.

[7] Susan Athey, Dean Eckles, and Guido W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association (JASA)*, 113(521), 2018.

[8] Mahmoudreza Babaei, Przemyslaw Grabowicz, Isabel Valera, Krishna P Gummadi, and Manuel Gomez-Rodriguez. On the efficiency of the information networks in social media. In *Proc. of the ACM Int'l. Conf. on Web Search and Data Mining (WSDM)*, 2016.

[9] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proc. of the Int'l. Conf. on World Wide Web (WWW)*, 2012.

[10] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[11] Abeba Birhane and Fred Cummins. Algorithmic injustices: Towards a relational ethics. *NeurIPS Black in AI Workshop*, 2019.

[12] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 2012.

[13] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. The socialbot network: when bots socialize for fame and money. In *Proc. of the Annual Computer Security Applications Conf. (ACSAC)*, 2011.

[14] Finn Brunton and Helen Nissenbaum. Political and ethical perspectives on data obfuscation. In M. Hildebrandt and K. de Vries, editors, *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*. Taylor & Francis, 2013.

[15] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Limiting the spread of misinformation in social networks. In *Proc. of the 20th Int'l. Conf. on World Wide Web (WWW)*, 2011.

[16] Qiang Cao, Xiaowei Yang, Jieqi Yu, and Christopher Palow. Uncovering large groups of active malicious accounts in online social networks. In *Proc. of the ACM SIGSAC Conf. on Computer and Communications Security (CCS)*, 2014.

[17] Robert Carlson. Rob's Giant BonusCard Swap Meet, October 2010. This page is now defunct as of access in September 2021. We retain the original access date from the reference in Brunton and Nissenbaum [14].

[18] Jusing Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. 2017.

[19] Vasek Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 1979.

[20] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proc. of the 7th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2001.

[21] John R. Douceur. The Sybil attack. In *Revised Papers from the First Int'l. Workshop on Peer-to-Peer Systems (IPTPS)*, page 251–260, 2002.

[22] Dean Eckles and Eytan Bakshy. Bias and high-dimensional adjustment in observational studies of peer effects. *Journal of the American Statistical Association (JASA)*, 2021.

[23] Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 2014.

[24] Chris Emmery, Ben Verhoeven, Guy De Pauw, Gilles Jacobs, Cynthia Van Hee, Els Lefever, Bart Desmet, Véronique Hoste, and Walter Daelemans. Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. In *Language Resources and Evaluation*, pages 597–633, 2021.

[25] Casey Fiesler and Nicholas Proferes. "Participant" perceptions of Twitter research ethics. *Social Media+ Society*, 4(1), 2018.

[26] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75 – 174, 2010.

[27] Jason Gaitonde, Jon Kleinberg, and Eva Tardos. Adversarial perturbations of opinion dynamics in networks. In *Proc. of the ACM Conf. on Economics and Computation*, 2020.

[28] Amanda Gentzel, Dan Garant, and David Jensen. The case for evaluating causal models using interventional

measures and empirical data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[29] Naman Goel, Alfonso Amayuelas, Amit Deshpande, and Ajay Sharma. The importance of modeling data missingness in algorithmic fairness: A causal perspective. In *AAAI Conf. on Artificial Intelligence*, 2021.

[30] Huan Gui, Ya Xu, Anmol Bhasin, and Jiawei Han. Network A/B testing: From sampling to estimation. In *Proc. of the 24th Int'l. Conf. on World Wide Web (WWW)*, 2015.

[31] Jochen Harant, Anja Pruchnewski, and Margit Voigt. On dominating sets and independent sets of graphs. *Comb. Probab. Comput.*, 8(6):547–553, November 1999.

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

[33] James M Hudson and Amy Bruckman. "Go away": participant objections to being studied and the ethics of chatroom research. *The Information Society*, 2004.

[34] Luke Hutton and Tristan Henderson. "I didn't sign up for this!": Informed consent in social network research. In *Proc. of the Int'l. AAAI Conf. on Web and Social Media*, 2015.

[35] Samuel D Johnson, Jemin George, and Raissa M D'Souza. Strategic seeding of rival opinions. In *Int'l. Conf. on Game Theory for Networks*, 2016.

[36] Jason J Jones, Robert M Bond, Eytan Bakshy, Dean Eckles, and James H Fowler. Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 US presidential election. *PloS one*, 12(4), 2017.

[37] Hyunseung Kang and Guido Imbens. Peer encouragement designs in causal inference with partial interference and identification of local average network effects. *arXiv preprint*, 2016.

[38] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proc. of the 9th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 137–146, 2003.

[39] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. of the Nat'l. Academy of Sciences (PNAS)*, 111(24), 2014.

[40] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78, Oct 2008.

[41] Myoung-jae Lee. Median treatment effect in randomized trials. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):595–604, 2000.

[42] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proc. of the 11th ACM SIGKDD Int'l. Conf. on Knowledge Discovery in Data Mining (KDD)*, pages 177–187, 2005.

[43] Jure Leskovec and Julian J Mcauley. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

[44] Jure Leskovec and Rok Sosič. SNAP: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(1):1, 2016.

[45] Maggie Makar, John V. Guttag, and Jenna Wiens. Learning the probability of activation in the presence of latent spreaders. In *AAAI Conf. on Artificial Intelligence*, 2018.

[46] J Nathan Matias. *Governing human and machine behavior in an experimenting society*. PhD thesis, Massachusetts Institute of Technology, 2017.

[47] J Nathan Matias, Allan Ko, and Merry Mou. The obligation to experiment, Dec 2016.

[48] Michelle N Meyer. Two cheers for corporate experimentation: The A/B illusion and the virtues of data-driven innovation. *Colo. Tech. LJ*, 13:273, 2015.

[49] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1), Dec 1978.

[50] M. Newman. The structure and function of complex networks. *SIAM Rev.*, 45:167–256, 2003.

[51] Alfred Ng. Teens have figured out how to mess with instagram's tracking algorithm: Teenagers are using group accounts to flood instagram with random user data that can't be tied to a single person. *CNet*, Feb 2020.

[52] Panagiotis Papadopoulos, Antonis Papadogiannakis, Michalis Polychronakis, Apostolis Zarras, Thorsten Holz, and Evangelos P. Markatos. K-subscription: Privacy-preserving microblogging browsing through obfuscation. In *Proc. of the Annual Computer Security Applications Conf. (ACSAC)*, 2013.

[53] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[54] Jacob Ratkiewicz, Michael Conover, Mark R. Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Detecting and tracking political abuse in social media. In *Proc. of the 5th Int'l. Conf. on Weblogs and Social Media (ICWSM)*, 2011.

[55] Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proc. of the ACM SIGKDD Int'l. Conf. on Knowledge Discovery in Data Mining (KDD)*, 2015.

[56] Manuel Gomez Rodriguez, Krishna Gummadi, and Bernhard Schoelkopf. Quantifying information overload in social media and its impact on social contagions. In *Int'l. AAAI Conf. on Weblogs and Social Media (ICWSM)*, 2014.

[57] Paul R Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association (JASA)*, 102(477), 2007.

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

[59] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proc. of the Nat'l. Academy of Sciences (PNAS)*, 105(4), 2008.

[60] Yoel Roth and Nick Pickles. Updating our approach to misleading information, May 2020.

[61] Donald B Rubin. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association (JASA)*, 75(371):591, September 1980.

[62] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association (JASA)*, 2005.

[63] Martin Saveski, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M. Airoldi. Detecting network effects: Randomizing over randomized experiments. In *Proc. of the ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2017.

[64] Scott Shane. To sway vote, Russia used army of fake Americans. *New York Times*, Sep 2017.

[65] Daniel L Sussman and Edoardo M Airoldi. Elements of estimation theory for causal effects in the presence of network interference. *arXiv preprint*, 2017.

[66] Emma Tosch, Eytan Bakshy, Emery D Berger, David D Jensen, and J Eliot B Moss. PlanAlyzer: Assessing threats to the validity of online experiments. *Proc. of the ACM on Programming Languages*, (OOPSLA), 2019.

[67] Nguyen Tran, Bonan Min, Jinyang Li, and Lakshminarayanan Subramanian. Sybil-resilient online content voting. In *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2009.

[68] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proc. of the ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2013.

[69] Gang Wang, Tianyi Wang, Haitao Zheng, and Ben Y Zhao. Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers. In *USENIX Security Symposium (USENIX Security)*, 2014.

[70] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:409–10, 1998.

[71] Douglas Guilbeault and Samuel Woolley. How Twitter bots are shaping the election. *The Atlantic*, Nov 2016.

[72] Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y. Zhao. Automated crowdturfing attacks and defenses in online review systems. In *Proc. of the ACM SIGSAC Conf. on Computer and Communications Security (CCS)*, 2017.

[73] Ercan Yildiz, Asuman Ozdaglar, Daron Acemoglu, Amin Saberi, and Anna Scaglione. Binary opinion dynamics with stubborn agents. *ACM Transactions on Economics and Computation (TEAC)*, 1(4), 2013.

[74] Xiao Zhang, Raj Rao Nadakuditi, and M. E. J. Newman. Spectra of random graphs with community structure and arbitrary degrees. *Physical Review E*, 89, 2014.