An Adversarial Model of Network Disruption: Maximizing Disagreement and Polarization in Social Networks

Mayee F. Chen and Miklós Z. Rácz

***Abstract*—The spread of misinformation has increased markedly in recent years, a phenomenon which has been accelerated and amplified by social media such as Facebook and Twitter. While some actors spread misinformation to push a specific agenda, it has also been widely documented that others aim to simply disrupt the network by increasing disagreement and polarization across the network, thereby destabilizing society. Popular social networks are also vulnerable to large-scale attacks. Motivated by this reality, we introduce a simple model of network disruption to capture this phenomenon, where an adversary can take over a limited number of user profiles in a social network with the aim of maximizing disagreement and/or polarization in the network. We investigate this model both theoretically and empirically. We show that the adversary will always change the opinion of a taken-over profile to an extreme in order to maximize disruption. We also prove that an adversary can increase disagreement/polarization at most linearly in the number of user profiles it takes over. Furthermore, we present a detailed empirical study of several natural algorithms for the adversary on both synthetic networks and real world (Reddit and Twitter) data sets. These show that even simple, unsophisticated heuristics, such as targeting centrists, can disrupt a network effectively, causing a large increase in disagreement / polarization. Studying the problem of network disruption through the lens of an adversary thus highlights the severity of the problem.**

***Index Terms*—social networks, polarization, misinformation.**

I. Introduction

RECENT years have seen a significant increase in the spread of misinformation, a phenomenon which has been accelerated and amplified by social media such as Facebook and Twitter. This problem has been widely studied empirically [1]–[5]. By and large, the main solution proposed to tackle the spread of misinformation is to develop automated fake news detection tools (e.g., [6]). However, there are huge challenges to overcome to make this viable. To start, simply defining what is false vs. true is often controversial and by now has been hugely politicized. Moreover, rapid advances in machine learning have made possible the creation of fake audio and video that are convincingly realistic, hence the problem of detection will only become worse in the coming years.

Here we consider a completely different angle. While some actors spread misinformation to push a specific agenda, it has also been widely documented [7], [8] that others aim to simply disrupt the network by increasing disagreement and polarization across the network, thereby destabilizing society. Popular social networks are also vulnerable to large-scale attacks in which attackers have the ability to take over accounts—in September 2018 it was revealed that nearly 50 million Facebook users were compromised in a data breach this way [9]. Motivated by this reality, we introduce a simple model of *network disruption* to capture this phenomenon, where an adversary can take over some user profiles in a social network with the aim of maximizing disagreement and/or polarization in the network.

How should adversaries choose profiles, and how much disruption can this cause to the network? Does the adversary have to be sophisticated to cause significant disruption? Or can they achieve their goal via simple, unsophisticated heuristics? How do the answers to these questions depend on properties of the underlying social network? We answer these questions in this paper, and the results highlight the importance of considering an adversarial perspective in the ultimate goal of counteracting the harmful effects of malicious actors.

A. Modeling Network Disruption

Our key conceptual contribution is the introduction of an adversarial model of network disruption, based on the concepts of polarization and disagreement. First, we describe how the network evolves over time and how the final expressed opinions are used to compute polarization and disagreement. Then, we present the adversarial model that sets up how an adversary is allowed to choose profiles to maximize disruption.

1) Measuring Disruption: We model the underlying social network as a weighted graph $G = (V, E, w)$, where V is the set of vertices, corresponding to the users of the social network; E is the set of edges, connecting users who know each other; and $w : E \to [0, 1]$ is a weight function on the edges that describes the strength of the ties between users. Now consider a topic that everyone has an opinion about—gun

Manuscript received April 20, 2021; revised November 11, 2021; accepted November 20, 2021. Date of publication November 30, 2021; date of current version March 23, 2022. This work of Miklós Z. Rácz was supported in part by NSF grant DMS 1811724 and in part by a Princeton SEAS Innovation Award. Recommended for acceptance by Dr. Yan Chen. *(Corresponding author: Miklos Z. Racz.)*

Mayee F. Chen is with Stanford University, Stanford, CA 94305 USA (e-mail: mfchen@stanford.edu).

Miklós Z. Rácz is with Princeton University, Princeton, NJ 08544 USA (e-mail: mracz@princeton.edu).

Digital Object Identifier 10.1109/TNSE.2021.3131416

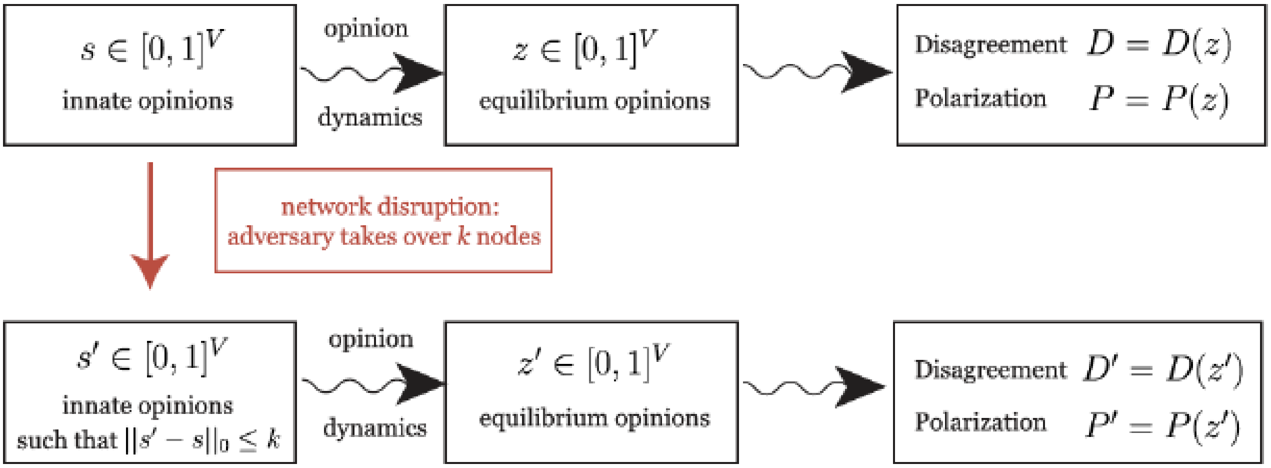

Fig. 1. *Schematic of the adversarial model of network disruption*. *Top:* On a particular topic everyone has an innate opinion, resulting in the innate opinion vector $s \in [0,1]^V$. These are mapped to equilibrium opinions $z \in [0,1]^V$ via the opinion dynamics. The equilibrium opinions give rise to natural quantities: disagreement D and polarization P. *Bottom:* The adversary can take over at most k nodes in the network and change their innate opinions, resulting in the new innate opinion vector $s' \in [0,1]^V$. The opinion dynamics are unchanged, resulting in new equilibrium opinions $z' \in [0,1]^V$, and subsequently new values of disagreement D' and polarization P'. The goal of the adversary is to *maximize* disagreement and/or polarization.

ownership, the amount of taxation, or your favorite controversial topic. We assume that everyone has an *innate opinion* about this topic and that this opinion can be quantified by a number in the interval [0,1]; for instance, 0 corresponds to strict gun control while 1 corresponds to no gun control. The innate opinions are denoted by $s = \{s_v\}_{v \in V} \in [0,\ 1]^V$.

People interact with their acquaintances on the social network and exchange opinions. As a result, their *expressed opinions* evolve and finally reach an *equilibrium*, which we denote by $z = \{z_v\}_{v \in V} \in [0,1]^V$. To be specific, in this paper we consider a simple model of opinion dynamics—known as the Friedkin-Johnsen model [10]—where users iteratively update their expressed opinions by taking a weighted average of the opinions of their friends and their own innate opinion. This results in the equilibrium opinions being $z = (I + L)^{-1} s$, where I is the identity matrix and L is the (weighted) Laplacian matrix. We emphasize that, while we focus on the Friedkin-Johnsen model, the questions that we consider about adversarial network disruption can be studied for other opinion dynamics models as well.

The equilibrium opinions z have various properties that we care about. Following [11], we introduce the following two important quantities. *Disagreement* is defined as

$$D \equiv D(z) := \sum_{(u,v) \in E} w_{u,v} (z_u - z_v)^2; \tag{1}$$

this measures how much acquaintances disagree in their opinions, globally across the network. *Polarization* is defined as

$$P \equiv P(z) := \sum_{v \in V} (z_v - \overline{z})^2, \tag{2}$$

where $\overline{z} := \frac{1}{|V|} \sum_{v \in V} z_v$ is the mean opinion; in other words, P is the variance of the opinions, multiplied by the number of vertices. We refer to [11] for a detailed discussion of these quantities, as well as related ones.

2) Modeling the Adversary: We now turn to modeling an adversarial perspective on network disruption, which is the key new idea introduced in the paper. Motivated by practical examples of hackers taking over a set of accounts, we consider an adversary that has a budget of k nodes it can control. We additionally must factor in how real users react to malicious accounts in order to delineate the adversary's capabilities as to not raise suspicion. For this purpose, we do not allow the adversary to change the graph structure (such as suddenly adding many new friends or shifting target audience) or interfere with the opinion dynamics (such as having the hacked nodes remain stubborn and ignore their friends' opinions). We thus assume the adversary can only change the innate opinions of k nodes. Since the adversary is not directly changing the expressed opinion, even extreme innate opinions are expressed subtly, raising less suspicion. One real world example motivating our adversarial model is ISIS's use of Twitter accounts to recruit new members—on the surface, the tweets and accounts did not appear out of the normal even though they pushed a malicious agenda [12].

Formally, we consider an adversary who can take over k nodes of the network and modify the innate opinions of these nodes arbitrarily. That is, the adversary can select $s' \in [0,1]^V$ such that $\|s' - s\|_0 \leq k$. Therefore, assuming the Friedkin-Johnsen model, the resulting equilibrium opinions will be $z' = (I + L)^{-1} s'$ and these will result in new values of disagreement D' and polarization P'. The goal of the adversary is to pick s' in such a way that *maximizes* disagreement D' or polarization P'. See Fig. 1 for an illustration.[1]

[1] Throughout the paper we only ever use the function $D(\cdot)$ with two arguments: z and z'. Thus, for simplicity, we denote the corresponding function values by D and D'. We hope the reader forgives the slight abuse of notation, in exchange for simplicity. (The same goes for polarization.)

B. Questions and Challenges

We open with a general question: what is the optimal solution for the adversary? That is, how should they pick the set of k vertices to hack, and how should they set the innate opinions of hacked vertices? We show that any optimal solution will set the innate opinions to an extreme; that is, if $s'_v \neq s_v$ then $s'_v \in \{0, 1\}$. Thus a brute force approach can find an optimal solution by checking all $\binom{n}{k}2^k$ possibilities, where n denotes the number of vertices. This is not feasible when k is large—so is there an efficient (polynomial time in n) algorithm to find an optimal solution? The function that we are maximizing is not submodular (see Section V) and hence off-the-shelf greedy algorithms and their guarantees do not apply directly.

Regardless if they can efficiently find an optimum or not, it is important to understand the limits of an adversarial attack under our model, prompting the next question: what is an upper bound on the amount of disruption an adversary can cause? We find that our measurement of disruption scales at most linearly in the number of profiles taken over. However, this scaling may only hold for sophisticated, knowledgeable algorithms, while it may be argued that in most cases knowing all the innate opinions exactly is unrealistic, and in other cases knowing the entire social network structure is difficult. Therefore, can the adversary cause this significant extent of disruption knowing only the network structure and nothing (or close to nothing) about the innate opinions, and vice versa? Can simple heuristics perform well, and how does performance depend on properties of the underlying social network? We investigate these questions.

II. Results

A. Theoretical Results

We analyze characteristics of the optimal solution and how polarization and disagreement scale with it. Our first result is intuitive: no matter which set of vertices the adversary chooses, the optimal way to modify the innate opinions of these nodes is to set them to one of the two extremes: 0 or 1. That is, radicalizing the taken-over account rather than giving them neutral opinions is more effective, although we again note that extreme innate opinions still give way to more subtle expressed opinions. In particular, we have the following result.

Theorem 1 (The adversary chooses extreme opinions): Consider the problem setup as above, with the adversary maximizing either disagreement, polarization, or a conical combination of these two (i.e., a linear combination with nonnegative coefficients). Assume that G has no isolated vertices. Let s' be an optimum vector of innate opinions, given the constraints (formalized in Section IV). For every $v \in V$, if $s'_v \neq s_v$, then $s'_v \in \{0, 1\}$.

This result follows from the convexity of the objective functions, together with the fact that the adversary is maximizing the objective function (see Section V for the proof). This implies that if the adversary has a budget of k (i.e., it can take over at most k nodes), then a brute force approach can find an optimal solution by checking all $\binom{n}{k}2^k$ possibilities, where n denotes the number of nodes. For constant k this gives a polynomial-time algorithm, but it performs poorly as k grows. In fact, we conjecture that solving the optimization problem of the adversary is computationally hard when k is large (e.g., $k = n^{\varepsilon}$ for constant $\varepsilon \in (0, 1)$), which is a direction for future work.

Next, we examine quantitatively the effect that the adversary can have on disagreement and polarization. First, we prove that the adversary can only increase disruption linearly in k. Specifically, for the polarization objective we show that the increase is always bounded above by $3k$; this is the content of the following theorem.

Theorem 2 (Upper bound on the increase in polarization): Let G be a weighted graph and s a vector of innate opinions such that the resulting equilibrium opinion vector z has polarization P. Suppose that the adversary has a budget of k; that is, the adversary may select $s' \in [0, 1]^V$ such that $\|s' - s\|_0 \leq k$. Let P' be the polarization of the resulting equilibrium opinion vector $z' = (I + L)^{-1}s'$. Then

$$P' \leq P + 3k.$$

For the disagreement objective, our result gives a bound of $8d_{\max}k$, where $d_{\max}$ is the (weighted) maximum degree. Thus for bounded-degree graphs this is still $O(k)$.

Theorem 3 (Upper bound on the increase in disagreement): Let G be a weighted graph and s a vector of innate opinions such that the resulting equilibrium opinion vector z has disagreement D. Suppose that the adversary has a budget of k; that is, the adversary may select $s' \in [0, 1]^V$ such that $\|s' - s\|_0 \leq k$. Let D' be the disagreement of the resulting equilibrium opinion vector $z' = (I + L)^{-1}s'$. Then

$$D' \leq D + 8d_{\max}k,$$

where $d_{\max} := \max_{v \in V} \sum_{u \in V} w_{v,u}$ is the (weighted) maximum degree.

B. Empirical Results

The theoretical results above lead to a natural question: can the adversary achieve an increase in these objective functions that grows linearly with k? We show empirically, on both synthetic and real data sets, that this is indeed the case for a range of heuristics.

We first consider a greedy algorithm, where the adversary iteratively selects nodes to take over, in each iteration choosing the node, together with one of the two extreme opinions, that maximizes the objective function. While this greedy algorithm is natural, it also uses detailed information: specifically, it assumes perfect knowledge of the network G and the innate opinions s. Since this may be unrealistic in practice, we also consider simpler heuristics for the adversary.

One such heuristic, which we term the "mean opinion" heuristic, is to choose the node whose (innate) opinion is closest to the mean and change it to one of the two extremes (either by optimizing this choice or just randomly). Such a heuristic can easily be implemented approximately by an adversary, since often it is possible to deduce whether someone has a centrist opinion by using extra information available about the node. Furthermore, it may be the case that the adversary has only approximate information about the (innate) opinions, for instance, perhaps they are "rounded" to the set $\{0, 1/2, 1\}$ (which corresponds to the two extremes and the center); in such a scenario, this heuristic is natural.

Another heuristic, which we term the "max degree" heuristic, focuses on a simple function of the underlying graph structure: iteratively choosing the largest degree nodes (in either a weighted or unweighted sense) and changing their opinion to one of the two extremes. This is motivated by practical scenarios where the network topology is only partially known; for instance, if only the node degrees are known, then this heuristic is natural.

We also compare all the algorithms to a random baseline, where the adversary selects nodes randomly and changes their opinions to random extremes—note that this approach is information agnostic.

We evaluate these algorithms on both synthetic and real data sets. For synthetic networks we use three common probabilistic generative models: Erdős-Rényi random graphs [13], [14], the preferential attachment model [15], and the stochastic block model [16], [17]. We also study Reddit and Twitter data sets that were collected in [18] and subsequently studied in [11].

Our main empirical finding is that in almost all settings—meaning, a network (synthetic or real, as above), an algorithm (from the ones described above), and an objective function (disagreement, polarization, or a conical combination)—the adversary succeeds in increasing its objective function linearly in k. The rate of increase depends on the details: the greedy algorithm performs best among these options, but the mean opinion heuristic is often not far behind. Even the random baseline gives a linear increase in k in several (though not all) settings.

We note that our empirical results only consider *iterative* algorithms. In principle, algorithms that are not iterative (e.g., inefficient algorithms such as brute force) could do much better than iterative ones. However, the upper bounds of Theorems 2 and 3 show that this is not possible: no matter the algorithm, only at most linear increase in k is possible for polarization/disagreement.

The details of all our empirical results are in Section VII below. All code and data has been posted to a public GitHub repository, available at https://github.com/mayeechen/network-disruption.

III. Related Work

The diffusion of information through networks is an important phenomenon in many disciplines. One common problem related to ours is Influence Maximization (IM), where one must select a subset of nodes to inject information into in order to maximize the number of influenced nodes by the end of the diffusion process [19]. Much work has been done on analyzing the performance of greedy algorithms for this problem [20], [21] (which, unlike ours, is submodular), and variants of it have further been studied (e.g., see [22] for a survey). Our problem instead focuses on opinion dynamics and considers a different objective of maximizing disruption. This involves not only the diffusion process but also the value of the innate and expressed opinions and where they are in the network.

Opinion dynamics have been used in various disciplines to model social learning (see, e.g., [23]). In seminal work, the DeGroot model describes how individuals reach a consensus through stochastic interactions [24]. Friedkin and Johnsen extended this model to incorporate individuals' intrinsic beliefs and prejudices [10]. In the Friedkin-Johnsen model, all agents have individual innate opinion values, and as time goes on, agents interact with each other, updating their opinions to be a weighted average of their innate opinion and the neighboring agents' opinions. Eventually, opinions converge to an equilibrium, which is a non-constant function of the innate opinions. This latter property is an important reason why we use the Friedkin-Johnsen model for opinion dynamics in this paper, in addition to its simplicity. The Friedkin-Johnsen model can be extended in a variety of ways, for instance to incorporate stubbornness and susceptibility to persuasion [25].

Several recent works have studied various network interventions to influence opinions in certain ways. Gionis, Terzi, and Tsaparas [26] studied opinion maximization in social networks, which corresponds to pushing a specific agenda. Abebe *et al.* [25] study a similar problem (opinion maximization or minimization), but where interventions happen at the level of susceptibility to persuasion. Bimpikis, Ozdaglar, and Yildiz [27] study a game-theoretic model of targeted advertising in networks, which is again a similar objective; see also the work of Lever on strategic competitions over networks [28]. Recent works of Mao *et al.* [29], [30] study competitive information spread, with a focus on understanding effects of confirmation bias.

In contrast, the work of Musco, Musco, and Tsourakakis [11]—which serves as the starting point of our work—studies polarization and disagreement, which are quite different objectives. Even though one of their settings is a slightly similar optimization problem with variable innate opinions, their technical approach and motivation are very different from our work since the goal of their work is to *minimize* polarization and disagreement.

Our key conceptual contribution is to study the *opposite* objective: *maximizing* polarization and disagreement. This corresponds to an adversarial perspective, which is motivated by recent developments over the past few years: malicious actors have increasingly been working towards disrupting networks by increasing disagreement and polarization, thereby destabilizing society [7]–[9], [31]. Also, the specific intervention we consider is taking over nodes of a network and modifying their (innate) opinions.

The recent paper [32] contains similar ideas to our work. However, their focus is on the special case when society initially has a consensus (i.e., $s = 0$), and this is perturbed by an adversary that can modify the entire innate opinion vector. They formalize the constraint on the adversary as an L_2-norm bound, whereas we use the constraint $\|s' - s\|_0 \leq k$, which has a clear interpretation in the adversary taking over at most k nodes of the network. After our work appeared on the arxiv, [32] was updated to consider L_p constraints; in particular, they prove a bound that slightly improves upon our Theorem 3. We focus on understanding the vulnerability of innate opinions and provide extensive empirical work demonstrating that simple adversarial heuristics can cause significant disruption, while [32] focuses more on a theoretical understanding of the network structure.

Finally, we note that there is a huge literature on understanding polarization in social networks, a complete overview of which is beyond the scope of this article; we refer the reader to [33], [34] and the references therein.

IV. Problem Setup

In this section we detail the problem setup for clarity. We fix an undirected weighted graph $G = (V, E, w)$ which represents the social network. Let $n = |V|$ denote the number of vertices (we often write $[n]$ for the vertex set) and let $m = |E|$ denote the number of edges. For convenience we define the weight function on all pairs of nodes, with $0 < w_{i,j} \leq 1$ if $(i, j) \in E$ and $w_{i,j} = 0$ otherwise. We also set $w_{i,i} = 0$ for all $i \in V$.

Let $d_i = \sum_{j \in V} w_{i,j}$ denote the (weighted) degree of node i and let D be the diagonal matrix with entries $d_1, \ldots, d_n$ on the diagonal. Let A denote the (weighted) adjacency matrix of G, with $A_{i,j} := w_{i,j}$ for $i, j \in V$. Let $L = D - A$ denote the weighted combinatorial Laplacian of G, which we refer to just as the Laplacian of G. Finally, let $\vec{1}$ denote the all-ones vector.

Opinion dynamics: Let $s = (s_1, \ldots, s_n) \in [0,1]^n$ denote the vector of innate opinions. In the Friedkin-Johnsen model of opinion dynamics [10], agents interact with each other as time goes on, updating their opinions to be a weighted average of their innate opinion and the neighboring agents' opinions. Formally, if $z_i^{(t)}$ denotes the expressed opinion of node i at time t (where $t \in \{0, 1, 2, \ldots\}$), then initially $z_i^{(0)} = s_i$ and the update for $t \geq 0$ is given by

$$z_i^{(t+1)} = \frac{s_i + \sum_{j \in V} w_{i,j} z_j^{(t)}}{1 + \sum_{j \in V} w_{i,j}}.$$

As $t \to \infty$, the vector of opinions converges to an equilibrium vector z that satisfies

$$z = (I + L)^{-1} s, \tag{3}$$

where I is the $n \times n$ identity matrix.

Disagreement and polarization. Following [11], we study the disagreement $D(z)$ and the polarization $P(z)$ of a vector of opinions z; see (1) and (2) for the definitions. Note that since the equilibrium opinion vector z is a function of the innate opinion vector s, disagreement D and polarization P can be considered functions of s as well, in which case we will denote them by $D(s)$ and $P(s)$, respectively. When clear from the context, we may denote these by just D and P. We also study linear combinations of these two quantities.

The objectives of the adversary. We are now ready to mathematically formulate our original questions as three optimization problems with varying objective functions. For any weighted graph G, innate opinions s, and budget $k \in \mathbb{N}$, the adversary aims to determine the optimal modified innate opinion vector s' according to the following.

- *Problem 1: Disagreement*

$$\begin{aligned} \text{maximize} \quad & D(z') \\ \text{subject to} \quad & z' = (I + L)^{-1} s,' \\ & s' \in [0,1]^n, \\ & \|s' - s\|_0 \leq k. \end{aligned} \tag{4}$$

- *Problem 2: Polarization*

$$\begin{aligned} \text{maximize} \quad & P(z') \\ \text{subject to} \quad & z' = (I + L)^{-1} s,' \\ & s' \in [0,1]^n, \\ & \|s' - s\|_0 \leq k. \end{aligned} \tag{5}$$

- *Problem 3: Weighted Sum*

$$\begin{aligned} \text{maximize} \quad & P(z') + \lambda \frac{n}{m} D(z') \\ \text{subject to} \quad & z' = (I + L)^{-1} s,' \\ & s' \in [0,1]^n, \\ & \|s' - s\|_0 \leq k. \end{aligned} \tag{6}$$

Note that in (6), we introduce λ as a parameter to describe the relative importance of disagreement versus polarization to the adversary. For this weighted sum index, we have scaled disagreement by $\frac{|V|}{|E|} = \frac{n}{m}$ so that the two terms have the same order of magnitude when $\lambda = 1$.

V. Convexity and Choosing Extreme Opinions

In this section we prove Theorem 1 and also demonstrate the lack of submodularity of the described problems. For all three optimization problems, the set of constraints do not form a convex set due to the constraint $\|s' - s\|_0 \leq k$. However, we prove that all of the objective functions are convex in s', which implies that $s'_i \in \{0, 1\}$ for all vertices i where $s'_i \neq s_i$.

Lemma 4: Disagreement is convex in s'. That is, the function $s' \mapsto D(s')$ is convex.

Proof: Disagreement can be written in quadratic form as $z'^T L z'$. Noting that $I + L$ is symmetric and using (3), D can be expressed as

$$\begin{aligned} D(s') = z'^T L z' &= ((I + L)^{-1} s')^T L ((I + L)^{-1} s') \\ &= s'^T (I + L)^{-1} L (I + L)^{-1} s'. \end{aligned}$$

The Laplacian matrix L is positive semidefinite and symmetric, so L can be written as $L = B^T B$ for some matrix $B \in \mathbb{R}^{n \times n}$. Therefore, $(I + L)^{-1} L (I + L)^{-1} = (I + L)^{-1} B^T B \quad (I + L)^{-1} = (B(I + L)^{-1})^T (B(I + L)^{-1})$, so $(I + L)^{-1} L (I + L)^{-1}$ is also positive semidefinite. Thus we can write D as a quadratic form in terms of s', with a positive semidefinite matrix, so $D(s')$ is convex in s'. ■

Lemma 5: Polarization is convex in s'. That is, the function $s' \mapsto P(s')$ is convex.

Proof: For notational convenience we drop all apostrophes from the notation. For a vector $x \in \mathbb{R}^n$ let $\widetilde{x} := x - \overline{x}\vec{1}$ denote the centered vector. With this notation we have $P(z) = \widetilde{z}^T \widetilde{z}$.

Observe that $L\vec{1} = 0$, and so $(I+L)\vec{1} = \vec{1}$ and $(I+L)^{-1}\vec{1} = \vec{1}$. Using (3) this implies that $\overline{z} = \frac{1}{n} z^T \vec{1} = \frac{1}{n} z^T (I+L)^{-1} \vec{1} = \frac{1}{n} s^T \vec{1} = \overline{s}$. In words, the mean equilibrium opinion is the same as the mean innate opinion. This, in turn, implies that $\widetilde{z} = (I+L)^{-1}\widetilde{s}$. With this notation we have that

$$P(z) = \widetilde{z}^T \widetilde{z} = \widetilde{s}^T \left((I+L)^{-1} \right)^2 \widetilde{s}.$$

For a vector $x \in \mathbb{R}^n$ define $f(x) := x^T((I+L)^{-1})^2 x$ and $g(x) := x - \overline{x}\vec{1} = \widetilde{x}$. Note that $((I+L)^{-1})^2$ is positive semidefinite, since it is the square of $(I+L)^{-1}$, which is positive semidefinite and symmetric. This implies that f is convex, since it is a quadratic form with a positive semidefinite matrix. Note also that for any two vectors $x, y \in \mathbb{R}^n$ and $\alpha \in [0,1]$ we have that $g(\alpha x + (1-\alpha)y) = \alpha g(x) + (1-\alpha) g(y)$. Therefore the convexity of $P = f \circ g$ follows directly from the convexity of f. ■

An immediate consequence of Lemmas 4 and 5 is that any conical combination of disagreement and polarization is convex in s'. This is because convexity is preserved by scaling with a positive constant, as well as across addition. Therefore our conclusions extend to the objective function of Problem 3 (see (6)) as well.

Proof of Theorem 1: Lemmas 4 and 5 show that the adversary's optimization problem is a convex maximization problem in s'. Moreover, if G has no isolated vertices then this is a strictly convex maximization problem. Therefore any coordinate of s that is changed in s' must be changed to an extreme: 0 or 1. ■

We conclude this section by a simple example that shows that the objective functions we are considering are not submodular. First, recall that a set function $f : \{0,1\}^V \to \mathbb{R}$ is *submodular* if for every $S, T \subseteq V$ with $S \subseteq T$ and for every $v \in V \setminus T$ we have that $f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$. In words, submodular functions have a diminishing returns property.

Example 1 (A single edge): Consider a graph with two nodes, denoted 1 and 2, with an edge between them with weight $w_{1,2} = 1$. Suppose that the innate opinions are initially centrist: $s_1 = s_2 = 1/2$. In this case the equilibrium opinions are also centrist: $z_1 = z_2 = 1/2$, leading to no disagreement or polarization: $D(z) = P(z) = 0$.

If an adversary has a budget of $k = 1$, they will change the innate opinion of a(n arbitrary) node to an (arbitrary) extreme: $s'_1 = 0$, $s'_2 = 1/2$. This results in the equilibrium opinions $z'_1 = 1/6$ and $z'_2 = 1/3$, giving disagreement $D(z') = 1/36$ and polarization $P(z') = 1/72$.

If an adversary has a budget of $k = 2$, they will change the innate opinions to opposite extremes: $s''_1 = 0$, $s''_2 = 1$. This results in the equilibrium opinions $z''_1 = 1/3$ and $z''_2 = 2/3$, giving disagreement $D(z'') = 1/9$ and polarization $P(z'') = 1/18$.

For both disagreement and polarization the increase in the second step is greater than the increase in the first step, and hence these objective functions are not submodular.

Because all three objective functions are not submodular, we are unable to apply the theoretical guarantees of greedy algorithms for submodular maximization (e.g., [35]). We instead focus directly on bounding the extent of disruption an adversary can cause, independent of the algorithm, and then conduct an empirical study to evaluate the performance of greedy algorithms with respect to this bound.

VI. Bounds on Network Disruption

In this section we prove Theorems 2 and 3. We start with a preliminary lemma which gives a bound on the L_1-norm of the difference between the modified equilibrium opinion vector z' and the original equilibrium opinion vector z.

Lemma 6: Let s be the original innate opinion vector and let s' be the modified innate opinion vector, satisfying $\|s' - s\|_0 \leq k$. Let z and z' be the respective equilibrium opinion vectors. Then

$$\|z' - z\|_1 \leq k.$$

Proof: Since $z = (I+L)^{-1}s$, we have that

$$\begin{aligned} \|z' - z\|_1 &= \left\|(I+L)^{-1}(s' - s)\right\|_1 \\ &\leq \sum_{i=1}^{n} \sum_{a=1}^{n} \left|(s'_a - s_a)(I+L)^{-1}_{ia}\right| \\ &= \sum_{i=1}^{n} \sum_{a=1}^{n} \left|s'_a - s_a\right| (I+L)^{-1}_{ia}, \end{aligned}$$

where the inequality is due to the triangle inequality and the final equality is because the entries of $(I+L)^{-1}$ are nonnegative. Without loss of generality, assume that nodes $1, \ldots, k$ comprise the set of nodes taken over by the adversary. Since $s_i \in [0,1]$, we must have $|s'_i - s_i| \leq 1$ for all i. Thus

$$\|z' - z\|_1 \leq \sum_{i=1}^{n} \sum_{a=1}^{k} \left|s'_a - s_a\right| (I+L)^{-1}_{ia} \leq \sum_{i=1}^{n} \sum_{a=1}^{k} (I+L)^{-1}_{ia}.$$

Now interchanging the order of summation we have that

$$\sum_{i=1}^{n} \sum_{a=1}^{k} (I+L)^{-1}_{ia} = \sum_{a=1}^{k} \sum_{i=1}^{n} (I+L)^{-1}_{ia} = \sum_{a=1}^{k} 1 = k.$$

Here we used the fact that the column sums of $(I+L)^{-1}$ are all equal to 1, which follows from the fact that $(I+L)^{-1}\vec{1} = \vec{1}$ (shown in Section V) and that $(I+L)^{-1}$ is symmetric. ■

A. Bound on the Increase in Polarization

Proof of Theorem 2: We first rewrite P' in a way to make P appear. This can be done by adding and subtracting under the square, and then expanding the square:

$$
\begin{aligned}
P' &= \sum_{i=1}^{n}\left(z'_i - \overline{z}'\right)^2 = \sum_{i=1}^{n}\left(z'_i - z_i + z_i - \overline{z} + \overline{z} - \overline{z}'\right)^2 \\
&= P + \sum_{i=1}^{n}\left(z'_i - z_i\right)^2 + n\left(\overline{z} - \overline{z}'\right)^2 + 2\sum_{i=1}^{n}\left(z'_i - z_i\right)\left(z_i - \overline{z}\right) \\
&\quad + 2\sum_{i=1}^{n}\left(z'_i - z_i\right)\left(\overline{z} - \overline{z}'\right) + 2\sum_{i=1}^{n}\left(z_i - \overline{z}\right)\left(\overline{z} - \overline{z}'\right). \qquad (7)
\end{aligned}
$$

Since $\sum_{i=1}^{n}(z_i - \overline{z}) = 0$, the last term in (7) is zero. The first term in (7) is equal to $-2n(\overline{z} - \overline{z}')^2$, because $\sum_{i=1}^{n}(z'_i - z_i) = n(\overline{z}' - \overline{z})$. Plugging this back into the display above we obtain that

$$
P' = P + \sum_{i=1}^{n}\left(z'_i - z_i\right)^2 + 2\sum_{i=1}^{n}\left(z'_i - z_i\right)\left(z_i - \overline{z}\right) - n\left(\overline{z} - \overline{z}'\right)^2. \qquad (8)
$$

The last term in (8) is nonpositive, so we may drop it. For the first sum in (8), note that $z_i \in [0,1]$ for every $i \in [n]$, so $(z'_i - z_i)^2 \leq |z'_i - z_i|$. Together with Lemma 6 this shows that $\sum_{i=1}^{n}(z'_i - z_i)^2 \leq k$. Finally, for the other sum in (8), using the bound $|z_i - \overline{z}| \leq 1$ we have that $\sum_{i=1}^{n}(z'_i - z_i)(z_i - \overline{z}) \leq \sum_{i=1}^{n}|z'_i - z_i| \leq k$. Altogether this shows that $P' \leq P + 3k$ as desired. ■

B. Bound on the Increase in Disagreement

Proof of Theorem 3: We start by rewriting D' in a way to make D appear. This can be done by adding and subtracting under the square, and then expanding the square. In the following all summations over i and j go from 1 to n, so we do not write this out further.

$$
\begin{aligned}
D' &= \sum_{i,j} w_{i,j}\left(z'_i - z'_j\right)^2 \\
&= \sum_{i,j} w_{i,j}\left(z'_i - z_i + z_i - z_j + z_j - z'_j\right)^2 \\
&= D + \sum_{i,j} w_{i,j}\Big\{\left(z'_i - z_i\right)^2 + \left(z'_j - z_j\right)^2 \\
&\qquad + 2\left(z'_i - z_i\right)\left(z_j - z'_j\right) \\
&\qquad + 2\left(z_i - z_j\right)\left(z'_i - z_i + z_j - z'_j\right)\Big\}.
\end{aligned}
$$

We now bound the four sums above. The first two sums are equal by symmetry, and we have that

$$
\begin{aligned}
\sum_{i,j} w_{i,j}\left\{\left(z'_i - z_i\right)^2 + \left(z'_j - z_j\right)^2\right\} &= 2\sum_{i,j} w_{i,j}\left(z'_i - z_i\right)^2 \\
= 2\sum_{i} d_i\left(z'_i - z_i\right)^2 &\leq 2d_{\max}\sum_{i}\left|z'_i - z_i\right| \leq 2d_{\max}k,
\end{aligned}
$$

where we used Lemma 6 for the last inequality and the fact that $|z'_i - z_i| \in [0,1]$ in the inequality before that. Next, using the inequality $(z'_i - z_i)(z_j - z'_j) \leq |z'_i - z_i|$ we have that

$$
2\sum_{i,j} w_{i,j}\left(z'_i - z_i\right)\left(z_j - z'_j\right) \leq 2\sum_{i,j} w_{i,j}\left|z'_i - z_i\right| \leq 2d_{\max}k.
$$

Finally, we use the bound $(z_i - z_j)(z'_i - z_i + z_j - z'_j) \leq |z'_i - z_i| + |z'_j - z_j|$ to obtain that

$$
\begin{aligned}
&2\sum_{i,j} w_{i,j}\left(z_i - z_j\right)\left(z'_i - z_i + z_j - z'_j\right) \\
&\leq 2\sum_{i,j} w_{i,j}\left(\left|z'_i - z_i\right| + \left|z_j - z'_j\right|\right) = 4\sum_{i,j} w_{i,j}\left|z'_i - z_i\right| \\
&\leq 4d_{\max}k.
\end{aligned}
$$

Putting everything together we obtain that $D' \leq D + 8d_{\max}k$ as desired. ■

VII. Algorithms and Experiments

We analyze the performance of the different heuristics across our three objectives and comment on how factors in the underlying social network—such as the degrees of the vertices and the distribution of innate opinion vectors—play a role. In particular, in our experiments we consider maximizing disagreement D, polarization P, and a weighted sum $P + \frac{n}{m}D$ (i.e., $\lambda = 1$). In the descriptions of the algorithms, we refer to the adversary's objective as f.

A. Algorithms for the Adversary

We present six adversarial heuristics that are designed under varying levels of information available about the network structure and opinions. We start with a natural greedy algorithm and then turn to other simpler heuristics. All algorithms below are iterative, picking vertices one at a time until at most k vertices have been selected. We denote by Ω the set of vertices that have already been selected by the adversary; initially $\Omega = \emptyset$.

Greedy. In each iteration i, we select a vertex and set its opinion to 0 or 1 to result in the greatest increase in the objective function $f((I + L)^{-1}s')$, given that $i - 1$ opinions have already been picked and modified according to this algorithm. We then add this vertex to Ω, update s', and repeat k times. If no modification results in an increase in the objective function at the ith iteration, with $i < k$, then we stop.

Mean Opinion. First, we select the index j^* such that

$$
j^* := \arg\max_{j \notin \Omega}\left|s'_j - \frac{1}{n}\sum_{i=1}^{n} s'_i\right|.
$$

In words: among opinions that have not been changed yet, we choose the vertex whose opinion is closest to the current network's average opinion to be j^*. Second, we must change the opinion s'_{j^*} to 0 or 1. To do this, we optimize and set $s'_{j^*} = a^* := \arg\max_{a \in \{0,1\}}\{f((I + L)^{-1}s') : s'_{j^*} = a\}$.

Note that the first step of this heuristic does not require any knowledge of the underlying graph structure, which can be the case in practice when edges are unknown; for instance, when hiding a list of followers or friends. Furthermore, if the adversary only has a rough idea of the nodes' opinions, this heuristic is intuitive and implementable approximately: pick a "centrist" node with the most neutral opinion.

Mean Opinion (Randomized). This algorithm is similar to the Mean Opinion algorithm, except the second step is replaced with randomly picking s'_{j^*} to be equal to 0 or 1 with equal probability. This algorithm can thus be entirely performed without knowledge of the underlying graph.

Max Degree. First, we select the index j^* such that

$$j^* := \arg\max_{j\in\Omega} \sum_{i\in V} \mathbf{1}_{\{w_{i,j}>0\}}.$$

In words: we choose our vertex to be the one that is connected to the most other vertices in the network. Second, we optimize the opinion s'_{j^*} as in Mean Opinion.

Max Weighted Degree. This algorithm is similar to Max Degree, except in the first step we choose j^* by maximizing the weighted degree: $j^* := \arg\max_{j\in\Omega} \sum_{i\in V} w_{i,j}$.

The latter two algorithms exploit the network structure in a simple way and so they may be practical for an adversary that has access to the underlying graph but may not have the means or data necessary to deduce what the opinions are.

Random. First, select a vertex $j^* \notin \Omega$ uniformly at random. Second, set s'_{j^*} to either 0 or 1 with equal probability. This completely random algorithm offers a natural baseline to compare against.

B. Synthetic Experiments

We evaluate the algorithms described above on synthetic networks generated using three probabilistic models: the Erdős-Rényi model, the preferential attachment model, and the stochastic block model. In all three cases, our results suggest that Greedy, Mean Opinion, and Mean Opinion (Randomized) cause disruption that scales linearly in k.

For each of the models, we generate a random graph with $n = 1000$ vertices. Weights on the edges are chosen independently and uniformly at random from (0,1) (and nonedges have zero weight). We experiment with k in the range $0 \le k \le n/2$. For each iteration until $n/2$, we plot the disagreement, polarization, and weighted sum when the adversary disrupts the network according to the six algorithms presented.[2]

Erdős-Rényi model. In the Erdős-Rényi model [13], [14] every pair of nodes is connected independently with some probability $p \in [0, 1]$. This model serves as a natural null model for random graphs, with no underlying structure. In Fig. 2 we take $p = 0.2$; other values of p show qualitatively similar behavior. We set the innate opinion vector s to have i.i.d. values which are uniformly distributed in $[0, 1]$.

The simulated performance of the six algorithms are shown in Fig. 2 (top row). We observe that all three objective functions are increasing roughly linearly in k, for all six algorithms, with the Greedy algorithm performing the best. We also see that Mean Opinion and Mean Opinion (Randomized), the two heuristics that exploit the innate opinion vector, perform better than Max Degree and Max Weighted Degree, which exploit network structure. In fact, the latter two heuristics appear to be only slightly better than the Random baseline for all three objectives.

Preferential attachment model. Compared to Erdős-Rényi random graphs, more realistic graphs can be constructed with the preferential attachment process [15]. While the Erdős-Rényi random graph serves as a natural null model for a network with no structure, the preferential attachment process instead follows the natural concept that vertices that are more connected will receive more edges in the future. This is often true in social networks; for instance, new accounts on a social media platform are perhaps more likely to follow a popular account rather than a less known one. We choose to generate a network using a preferential attachment process with parameter $m = 5$, meaning that at each time step, a new vertex is connected to m existing nodes, choosing each existing node with probability proportional to its degree. We again set the innate opinion vector s to have i.i.d. values which are uniformly distributed in $[0, 1]$.

The simulated performance of the six algorithms are shown in Fig. 2 (middle row). Relatively, the greedy algorithm still has the best performance, followed by Mean Opinion and Mean Opinion (Randomized) for the k defined in the synthetic experiments, and the performance of all algorithms seems to increase linearly in this range of k. We observe, however, that while Max Degree and Max Weighted Degree start out worse than pure randomization, they appear eventually to surpass Random and increase at a rate faster than other algorithms. Lastly, we observe that the scale of the objectives is significantly larger than is observed for the Erdő s-Rényi model (see the top row of Fig. 2); perhaps this is due to the Erdős-Rényi graph being much denser and thus "averaging" all the opinions more.

Stochastic block model. The stochastic block model [16], [17] is able to represent planted clusters, unlike the other models we consider. These sorts of communities often arise in social networks, as seen in the Twitter data set which we discuss below. We consider the simplest version of the stochastic block model, with two communities C_1 and C_2 each of size $n/2$. Let the connectivity within both communities have parameters $p_{11} = p_{22} = 0.7$, that is, pairs of vertices within the communities share an edge with probability 0.7 (independently across pairs), and let the connectivity between the two communities have parameter $p_{12} = 0.1$. Moreover, different communities often have different opinion distributions. Therefore, in our experiments we set the innate opinions s_v for $v \in C_1$ to be independent draws from the $\mathrm{Beta}(5, 2)$ distribution, while the opinions of s_v for $v \in C_2$ are i.i.d. $\mathrm{Beta}(2, 5)$. This means that opinions in C_1 are biased towards 1, and opinions in C_2 are biased towards 0. Experiments with different parameters show similar qualitative behavior.

The simulated performance of the six algorithms are shown in Fig. 2 (bottom row). Similar to the other synthetic

[2] To clarify, each figure presents results for a single realization of the random setup (random graph and innate opinions) and, for randomized algorithms, a single realization of randomness in the algorithm. We do this so that curves represent the trajectory of a single adversary's action.

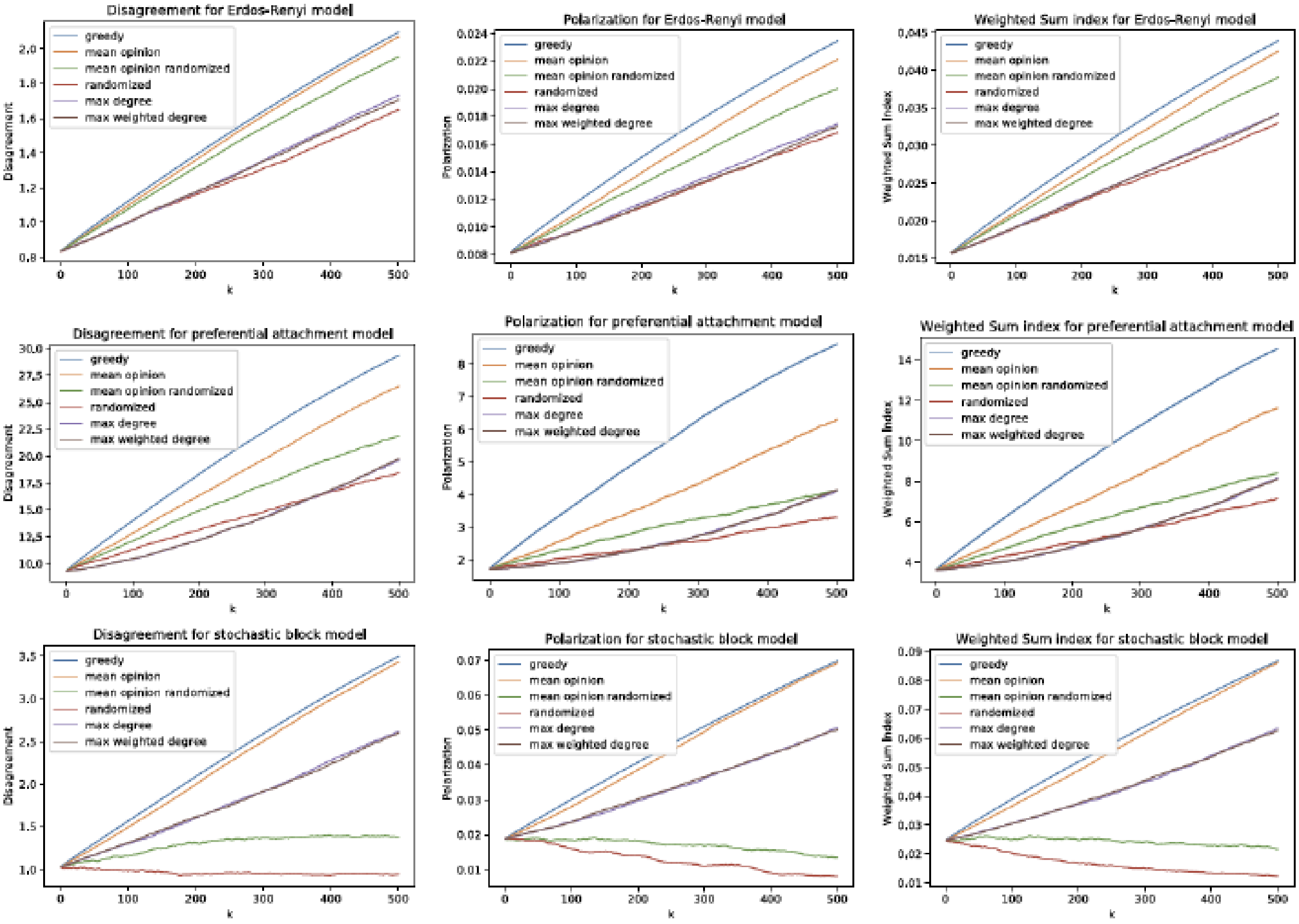

Fig. 2. *Performance of network disruption algorithms in synthetic experiments. Top row:* under the Erdős-Rényi model with $p = 0.2$ and opinions distributed according to $\mathrm{Uni}(0, 1)$. *Middle row:* under the preferential attachment model with $m = 5$ and opinions distributed according to $\mathrm{Uni}(0, 1)$. *Bottom row:* under the stochastic block model with $p_{11} = p_{22} = 0.7$, $p_{12} = 0.1$, and opinions distributed according to $\mathrm{Beta}(5, 2)$ and $\mathrm{Beta}(2, 5)$.

experiments, the Greedy and Mean Opinion algorithms perform the best across the three objectives, increasing linearly in k. However, in this case the Random baseline actually decreases the value of all three objectives, while Mean Opinion (Randomized) decreases for polarization and the weighted sum. We conjecture that this is because choosing between 0 and 1 heavily depends on which community j^* is in due to how the innate opinions are generated using two beta distributions rather than just a uniform distribution over $[0, 1]$.

C. Analysis of Reddit and Twitter Data Sets

We also evaluate the six algorithms for the adversary on two real data sets, finding that polarization and disagreement can increase one order of magnitude when an adversary takes over just 10% of the accounts. These data sets, one on Twitter and one on Reddit, contain the edge set for the social networks as well as the list of opinions of the users over time. They were originally collected by [18] by tracking interactions between users and using natural language processing techniques to map text to opinions; they were subsequently studied in [11]. We pick the innate opinion vector to be the most recently recorded opinion vector, which is also how [11] chooses innate opinion vectors.

Twitter. This network has $n = 548$ vertices and $m = 3638$ edges, where the vertices represent the individuals tweeting over a certain time period about a debate on the Delhi legislative assembly elections of 2013 (identified by a set of hashtags), and their opinions correspond to the sentiment of the tweets. Each edge is an undirected interaction between users.

The simulated performance of the six algorithms on the Twitter data set are shown in Fig. 3 (top row).[3] The Greedy and Mean Opinion algorithms still have the largest increases in all three objectives for this data set, with the Greedy algorithm performing best. On the other hand, Mean Opinion (Randomized) and Random perform relatively poorly, with Max Degree and Max Weighted Degree eventually outperforming the former two for all three objectives. This relative ordering

[3] Again, each figure presents results for a single realization of the randomness (for those algorithms that involve randomness). The same applies to the bottom row of Fig. 3.

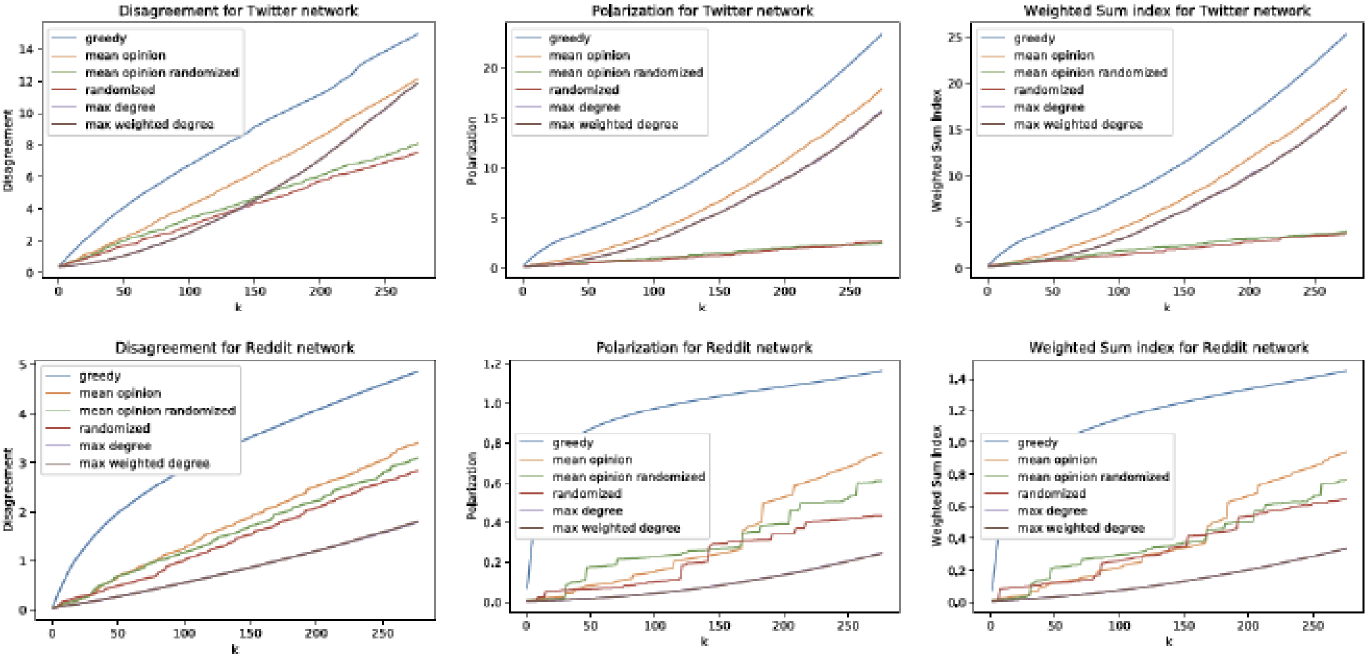

Fig. 3. *Performance of network disruption algorithms on real data sets. Top row:* on a Twitter data set. *Bottom row:* on a Reddit data set.

of the performance of different algorithms is similar to that of the stochastic block model discussed previously (Fig. 2, bottom row). In fact, when the Twitter network is visualized, we can see that there are two main communities, and a third smaller and less dense community. Therefore, we can attribute a lot of the performance results to the underlying community structure. However, the distribution of innate opinions does not follow two beta distributions, but instead is approximately Gaussian with mean 0.602 and standard deviation 0.08, which mitigates the decrease in performance that results from randomly setting a^* amid beta-distributed opinions.

In Table I, we list the exact values for disagreement, polarization, and their weighted sum of the Twitter network when the adversary uses the greedy algorithm, at the start of the algorithm ($k = 0$) and when k is equal to $20, 50, 100, 200$. This table suggests that, even if the adversary can only change the opinions on 20 accounts (approximately 3.6% of the nodes), the disagreement in the network increases by over 4 times, while the polarization and weighted sum increase by over 7 times. This quantitatively demonstrates the significant amount of disruption—increase in disagreement and polarization—that a malicious actor may inflict upon a social network.

Reddit. This network has $n = 556$ vertices and $m = 8969$ edges, where the vertices represent individuals who have posted in a politics subreddit, and their opinions correspond to the sentiment in this subreddit over a certain time period. There is an edge between users if they both post in at least two other same subreddits. We also discard three vertices from this graph that are not connected to any other vertices, as keeping these vertices implies that algorithms can simply change these opinions to yield large increases in polarization without any consequences for the opinion dynamics.

TABLE I
VALUES OF OBJECTIVE FUNCTIONS FOR THE TWITTER DATA SET. THESE ARE UNDER THE GREEDY ALGORITHM AT $k = 0$ (original), 20, 50, 100, AND 200

Objective	$k=0$	$k=20$	$k=50$	$k=100$	$k=200$
Disagreement	0.48	2.12	4.17	6.81	11.20
Polarization	0.29	2.34	3.89	6.70	15.05
Weighted Sum	0.37	2.66	4.48	7.59	16.54

The simulated performance of the six algorithms on the Reddit data set are shown in Fig. 3 (bottom row).[4] Again, the greedy algorithm performs best, with a large increase especially for small k. While the graphs for polarization and for the weighted sum have very noticeable jumps, for all three objectives MEAN OPINION, MEAN OPINION (RANDOMIZED), and RANDOM perform similarly. We conjecture that random is not the worst in this case for two reasons: firstly, the Reddit data set's opinions roughly follow a Gaussian distribution with mean 0.498 and standard deviation 0.04, meaning that the values are more tightly concentrated around a very neutral opinion than the Twitter data set. Moreover, the distribution of degrees of the vertices is more uniform than that of the Twitter data set (which appears to follow a power law instead), suggesting that arbitrarily choosing a vertex and then randomly setting its opinion can still result in good performance.

In Table II, we list the exact values for disagreement, polarization, and their weighted sum of the Reddit network when the adversary uses the greedy algorithm at the start of the

[4] We remark that some of the curves in these figures have relatively big jumps, a phenomenon that is not present in the figures about other networks. It turns out that the Reddit data set has many nodes with degree 1, and these larger jumps happen when the algorithm happens to pick these nodes to change. All other networks appearing in the paper do not have nodes of degree 1, which is why this phenomenon does not appear in the other figures.

TABLE II
Values of Objective Functions for the Reddit Data Set. These Are Under the Greedy Algorithm at $k = 0$ (original), 20, 50, 100, and 200

Objective	$k = 0$	$k = 20$	$k = 50$	$k = 100$	$k = 200$
Disagreement	0.09	1.14	2.00	2.88	4.09
Polarization	0.07	0.72	0.88	0.98	1.09
Weighted Sum	0.08	0.79	0.99	1.15	1.33

algorithm ($k = 0$) and when k is equal to 20, 50, 100, 200. This table suggests that, even if the adversary can only change the opinions on 20 accounts (approximately 3.6% of the nodes), all objectives are able to increase roughly tenfold. Just like the corresponding results for the Twitter data set, this quantitatively demonstrates the significant amount of disruption—increase in disagreement and polarization—that a malicious actor may inflict upon a social network.

VIII. Conclusion and Discussion

Our primary conceptual contribution is the introduction of an adversarial model of network disruption. This presents an important lens through which to study the unfortunate recent trend of malicious actors interfering in social networks in order to destabilize society.

The key conclusion from our results is that an adversary can significantly disrupt a network—in particular, increasing disagreement and polarization—using simple, unsophisticated methods. This mirrors recent findings analyzing real-world data; for instance, the authors in [31] conclude that the Internet Research Agency's operations to interfere with the 2016 U.S. presidential election "were largely unsophisticated". This adversarial approach thus highlights the severity of the problem, and we hope this motivates further research into addressing it via strategies for defending against network disruption.

We list several avenues for further study in the following bullet points, ranging from specific questions concerning the model we studied to broad questions concerning adversarial models on social networks.

- *Hardness of optimal network disruption:* As mentioned in Section II, we conjecture that solving the optimization problem of the adversary is computationally hard when k is large. Recent work of Gionis, Terzi, and Tsaparas [26] on a related opinion maximization problem uses a reduction to vertex cover to show hardness; see also [25] where this proof is adapted to another setting. Adapting this proof to our setting is challenging due to the different nature of our objective function, coupled with the opinion dynamics whose effect is difficult to isolate.
- *Performance guarantees for the adversary:* In Section VII we investigated empirically the performance of several natural algorithms for the adversary, on several different random graphs, as well as on Reddit and Twitter data sets. While performances varied, depending on the algorithm and the underlying social network, one thing that most had in common was a linear growth in the objective function, as a function of the budget k. Is it possible to prove such a performance guarantee (at least for some heuristic)?
- *Other opinion dynamics:* We focused here on the Friedkin-Johnsen model of opinion dynamics, but everything we discussed can be studied under other models. How robust are the results to such changes?
- *Other adversarial disruption models:* We have considered a setup where an adversary can change the innate opinions of k nodes in a network. While we discuss the motivation behind our model in Section I-A2, we may want to model how an adversary disrupts a network in more nuanced ways. For instance, we can combine our model with that of [25], which uses a susceptibility parameter to reflect how easily users are influenced by (adversarial) opinions. We may also want to consider a setting where adversaries can create new accounts (i.e., bots), for which we would need to both set opinions and add edges. Nonetheless, we hope that the simplicity of our model, while it may not completely represent the complex realities of social networks, provides a first step into understanding adversarial disruption.
- *Defense strategies:* Our empirical results show that the adversary does not have to be sophisticated in order to significantly disrupt the network. This highlights the need to think critically about defense strategies that can counteract network disruption. For instance, is it possible to tackle network disruption by modifying the network itself (e.g., by carefully suggesting new edges to add)?

Ultimately, we hope that considering an adversarial viewpoint will better equip us to minimize the deleterious effects of malicious actors.

Acknowledgments

We thank Cameron Musco, Christopher Musco, and Charalampos Tsourakakis for sharing the Reddit data set used in [11] and for clarifying details of their data analysis. We also thank anonymous reviewers for their careful reading of the paper and their helpful questions and suggestions that helped improve the paper.

References

[1] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *J. Econ. Perspectives*, vol. 31, no. 2, pp. 211–236, May 2017.

[2] A. Fourney, M. Z. Racz, G. Ranade, M. Mobius, and E. Horvitz, "Geographic and temporal trends in fake news consumption during the 2016 US presidential election," in *Proc. ACM Conf. Inf. Knowl. Manag.*, 2017, pp. 2071–2074.

[3] A. Guess, B. Nyhan, and J. Reifler, "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign," 2018. [Online]. Available: http://www.dartmouth.edu/ nyhan/fake-news-2016.pdf

[4] A. Spangher, G. Ranade, B. Nushi, A. Fourney, and E. Horvitz, "Analysis of strategy and spread of Russia-sponsored content in the US in 2017," 2018. [Online]. Available: https://arxiv.org/abs/1810.10033

[5] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[6] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some like it hoax: Automated fake news detection in social networks," in *Proc. 2nd Workshop Data Sci. Social Good*, 2017, pp. 1–15.

[7] Facebook, "An update on information operations on facebook," Sep. 2017. [Online]. Available: https://about.fb.com/news/2017/09/information-operations-update/

[8] U.S.H.P.S.C. on Intelligence (USHPSCI), "Exposing Russia's effort to sow discord online: The internet research agency and advertisements," 2018. [Online]. Available: https://intelligence.house.gov/social-media-content/default.aspx

[9] J. C. Wong, "Facebook says nearly 50 m users compromised in huge security breach," The Guardian. Sep. 2018. [Online]. Available: https://www.theguardian.com/technology/2018/sep/28/facebook-50-million-user-accounts-security-berach

[10] N. E. Friedkin and E. C. Johnsen, "Social influence and opinions," *J. Math. Sociol.*, vol. 15, no. 3-4, pp. 193–206, 1990.

[11] C. Musco, C. Musco, and C. E. Tsourakakis, "Minimizing polarization and disagreement in social networks," in *Proc. World Wide Web Conf.*, 2018, pp. 369–378.

[12] L. Blaker, "The islamic state's use of online social media," *Military Cyber Affairs*, vol. 1, no. 1, pp. 1–9, 2015, Art. no. 4.

[13] P. Erdös and A. Rényi, "On random graphs, I," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[14] P. Erdös and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci.*, vol. V.A, no. 1-2, pp. 17–60, 1960.

[15] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[16] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Netw.*, vol. 5, no. 2, pp. 109–137, 1983.

[17] E. Abbe, "Community detection and stochastic block models: Recent developments," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 1–86, 2017.

[18] A. De, S. Bhattacharya, P. Bhattacharya, N. Ganguly, and S. Chakrabarti, "Learning a linear influence model from transient opinion dynamics," in *Proc. 23rd ACM Int. Conf. Inf. Knowl. Manag.*, 2014, pp. 401–410. [Online]. Available: https://doi.org/10.1145/2661829.2662064

[19] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2003, pp. 137–146. [Online]. Available: https://doi.org/10.1145/956750.956769

[20] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2007, pp. 420–429.

[21] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2009, pp. 199–208.

[22] S. Banerjee, M. Jenamani, and D. K. Pratihar, "A survey on influence maximization in a social network," *Knowl. Inf. Syst.*, vol. 62, no. 9, pp. 3417–3455, Mar. 2020. [Online]. Available: http://dx.doi.org/10.1007/s10115-020-01461-4

[23] E. Mossel and O. Tamuz, "Opinion exchange dynamics," *Probability Surv.*, vol. 14, pp. 155–204, 2017.

[24] M. H. DeGroot, "Reaching a consensus," *J. Amer. Stat. Assoc.*, vol. 69, no. 345, pp. 118–121, 1974.

[25] R. Abebe, J. Kleinberg, D. Parkes, and C. E. Tsourakakis, "Opinion dynamics with varying susceptibility to persuasion," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 1089–1098.

[26] A. Gionis, E. Terzi, and P. Tsaparas, "Opinion maximization in social networks," in *Proc. SIAM Int. Conf. Data Mining*, 2013, pp. 387–395.

[27] K. Bimpikis, A. Ozdaglar, and E. Yildiz, "Competitive targeted advertising over networks," *Operations Res.*, vol. 64, no. 3, pp. 705–720, 2016.

[28] C. R. Lever, "Strategic competitions over networks," Ph.D. dissertation, Stanford University, 2010.

[29] Y. Mao, S. Bolouki, and E. Akyol, "Spread of information with confirmation bias in cyber-social networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 2, pp. 688–700, Apr.–Jun. 2020.

[30] Y. Mao, E. Akyol, and N. Hovakimyan, "Impact of confirmation bias on competitive information spread in social networks," *IEEE Control Netw. Syst.*, vol. 8, no. 2, pp. 816–827, Jun. 2021.

[31] R. L. Boyd *et al.*, "Characterizing the internet research agency's social media operations during the 2016 U.S. presidential election using linguistic analyses," 2018. [Online]. Available: https://psyarxiv.com/ajh2q/

[32] J. Gaitonde, J. Kleinberg, and E. Tardos, "Adversarial perturbations of opinion dynamics in networks," 2020. [Online]. Available: https://arxiv.org/abs/2003.07010

[33] E. Pariser, *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think.* Penguin, U.K., 2011. [Online]. Available: https://www.amazon.com/Filter-Bubble-Personalized-Changing-Think/dp/0143121235

[34] J. Hazla, Y. Jin, E. Mossel, and G. Ramnarayan, "A geometric model of opinion polarization," 2019. [Online]. Available: https://arxiv.org/abs/1910.05274

[35] A. Krause and D. Golovin, "Submodular function maximization,*" Tractability*, vol. 3, pp. 71–104, 2014.

Mayee F. Chen is a third year Ph.D. student with Stanford University in the Computer Science Department advised by Christopher Ré. She completed the undergraduate studies with Princeton University, majoring in ORFE. Mayee's current research interests revolve around how to encode and evaluate sources of supervision and side information throughout the ML pipeline (e.g. weakly/semi/self-supervised) through both information-theoretic and geometric lenses.

Miklós Z. Rácz is an Assistant Professor with Princeton University in the ORFE Department, as well as an affiliated faculty member with the Center for Statistics and Machine Learning (CSML). Before coming to Princeton, he received the Ph.D. in statistics from UC Berkeley and was then a Postdoc in the Theory Group with Microsoft Research, Redmond. Miki's research focuses on probability, statistics, and their applications, and he is particularly interested in network science.