Defending Elections against Malicious Spread of Misinformation

Bryan Wilder,[1] **Yevgeniy Vorobeychik**[2]
[1]Center for Artificial Intelligence in Society, University of Southern California, bwilder@usc.edu
[2]Department of Computer Science & Engineering, Washington University in St. Louis, yvorobeychik@wustl.edu

Abstract

The integrity of democratic elections depends on voters' access to accurate information. However, modern media environments, which are dominated by social media, provide malicious actors with unprecedented ability to manipulate elections via misinformation, such as fake news. We study a zero-sum game between an attacker, who attempts to subvert an election by propagating a fake new story or other misinformation over a set of advertising channels, and a defender who attempts to limit the attacker's impact. Computing an equilibrium in this game is challenging as even the pure strategy sets of players are exponential. Nevertheless, we give provable polynomial-time approximation algorithms for computing the defender's minimax optimal strategy across a range of settings, encompassing different population structures as well as models of the information available to each player. Experimental results confirm that our algorithms provide near-optimal defender strategies and showcase variations in the difficulty of defending elections depending on the resources and knowledge available to the defender.

Introduction

Free and fair elections are essential to democracy. However, the integrity of elections depends on voters' access to accurate information about candidates and issues. Oftentimes, such information comes via news media or political advertising. When these information sources are accurate and transparent, they serve an important role in producing well-functioning elections. However, because of the great impact that messaging can have on voter behavior (Gerber, Karlan, and Bergan 2009; DellaVigna and Kaplan 2007; Brader 2005), such information can also subvert legitimate elections when deliberately falsified by malicious actors.

In traditional media environments, such subversion is relatively difficult because professional news organizations serve as gatekeepers to information spread. However, modern media environments are increasingly decentralized due to the importance of social networks such as Facebook or Twitter, which allow outside actors to spread political information directly amongst voters (Chi and Yang 2011; Wattal et al. 2010; Holcomb, Gottfried, and Mitchell 2013). This presents an unprecedented opportunity for malicious actors to spread deliberately falsified information – "fake news" – and in doing so, influence the results of democratic elections. Such concerns are particularly salient in light of the 2016 U.S. presidential election. Recent research shows that, on average, an American adult was exposed to at least one fake news story during the campaign (Allcott and Gentzkow 2017) and that these stories influenced voter attitudes (Pennycook, Cannon, and Rand 2017).

Prior work on election control has considered a number of mechanisms for election interference, including bribery (Faliszewski et al. 2009; Baumeister et al. 2015; Erdélyi, Reger, and Yang 2017; Yang, Shrestha, and Guo 2016), adding or deleting voters (Erdélyi, Hemaspaandra, and Hemaspaandra 2015; Loreggia et al. 2015; Faliszewski, Hemaspaandra, and Hemaspaandra 2011; Liu et al. 2009), and adding or deleting candidates (Chen et al. 2015; Liu et al. 2009). Only recently has social influence been explicitly studied as a means of election control (Sina et al. 2015; Wilder and Vorobeychik 2018; Faliszewski et al. 2018). Further, with only a few exceptions which do not consider social influence (Li, Jiang, and Wu 2017; Yin et al. 2018), election control has so far primarily been studied from the attacker's perspective (to establish the computational complexity of controlling an election when the attacker is the only actor).

We therefore ask the following natural question: *how can a defender mitigate the impact of fake news on an election?* For instance, a social media platform or a news organization may have the ability to detect and label fake news stories on a given advertising channel, or propagate a counter-message with more accurate information. We model this interaction as a zero-sum game between an attacker, attempting to influence voters by advertising on a subset of possible channels, and a defender who enacts counter-measures on a subset of channels. The goal for the attacker is to maximize the expected number of voters who switch to the attacker's preferred candidate, whereas the defender's goal is to minimize this quantity. Note that in this model the defender is neutral with respect which candidate actually wins; they focus solely on minimizing the attacker's malicious influence.

Computing equilibria is computationally challenging due to the exponential number of possible actions for each player. Complicating the problem, in practice the defender may have considerable uncertainty about which candidate

each voter prefers at the start of the game (information which is needed to effectively target limited resources). We provide efficient algorithms, backed by theoretical guarantees and empirical analysis, across a range of settings:

1. In the *disjoint* case, each voter can be reached by only one advertising channel, modeling a case where each channel corresponds to a different demographic group. We give an FPTAS for the minimax equilibrium strategies.
2. In the *nondisjoint* case, each voter can be reached by an arbitrary set of channels. We first prove that the associated computational problem is APX-hard. We then provide an algorithm with a bicriteria guarantee: it guarantees the defender a constant-factor approximation to the optimal payoff but relaxes the budget constraint.
3. We consider three models of uncertainty about voter preferences. The first is *stochastic* uncertainty where the preference profile is drawn from a distribution. The second is *asymmetric* uncertainty where the preference profile is drawn from a distribution and the attacker observes the realized draw. The third is *adversarial* uncertainty where the preference profile is chosen to be the worst possible for the defender within an uncertainty set. Collectively, these models allow us to capture a range of assumptions about the information available to each player. Surprisingly, we show that across all three models, and in both the disjoint and nondisjoint cases, the defender can obtain exactly the same approximation ratios as when preferences are known exactly.

Problem Formulation

We consider a set of voters V (with $|V| = n$) and a set of advertising channels C (with $|C| = m$). C and V form a bipartite graph; that is, each voter is reachable by one or more advertising channels. The voters participate in an election between two candidates, c_a and c_d. An attacker aims to ensure that one of these candidates, c_a, wins the election. A defender aims to protect the election against this manipulation. Each voter v has a preferred candidate who they vote for. Let $\theta_v = 1$ if v initially prefers c_d and 0 otherwise.

The attacker attempts to alter election results by spreading a message (a fake news story) amongst the voters. More precisely, the attacker has a limited advertising budget and can send the message through at most k_a channels. If channel u is chosen by the attacker, then any voter v with an edge to u switches their vote to c_a with probability p_{uv}, where all such events are independent. The defender can protect voters from the attacker's misinformation, for example by detecting and labeling falsified stories on a given advertising channel, or by attempting to propagate a counter-message of their own. If the defender protects channel v, each voter connected to v is "immunized" against the attacker's message independently with probability q_{uv}. The defender may select up to k_d channels.

We model this interaction as a zero-sum game between the attacker and defender. In this setting, equilibrium strategies are unaffected by whether one party must first commit to a strategy (formally, the Nash and Stackelberg equilibria are equivalent). Hence without loss of generality, we consider a simultaneous-move game and seek to compute a Nash equilibrium. The defender's strategy space is all subsets of k_d channels to protect, while the attacker's strategy space consists of all subsets of k_a channels to attack. Hence, each player has an exponentially large number of pure strategies, substantially complicating equilibrium computation.

We now introduce the attacker's objective, which determines the payoffs for the game. When the defender chooses a set of channels S_d and the attacker chooses S_a, let $f(S_d, S_a)$ be the expected number of voters who previously preferred c_d but switch their vote to c_a. The randomness is over which voters are reached by the attacker's message (determined by the probabilities p_{uv} and q_{uv}). Formally, we can express f as

$$f(S_d, S_a) = \sum_{v \in V} \theta_v \left(\prod_{u \in S_d} 1 - q_{uv} \right) \left(1 - \prod_{u \in S_a} 1 - p_{uv} \right)$$

where the first product is the probability that the defender fails to reach voter v and the second is the probability that the attacker succeeds. The term θ_v means that only voters who initially prefer c_d count (since they are the only ones who can switch). The attacker's payoff is simply $f(S_d, S_a)$, while the payoff for the defender is $-f(S_d, S_a)$; in words, the defender aims to minimize the spread of misinformation.

We consider two models for how the population may be structured. In the *disjoint* model, the advertising channels partition the population so that each voter has an edge to exactly one channel. This models a case where the channels represent demographic groups and the attacker is deciding which demographics to target. In the more general *nondisjoint* model, voters may be reached through multiple channels; thus, the edges can form an arbitrary bipartite graph.

We begin by considering the case where θ (the voters' initial preferences) are common knowledge. Subsequently, we consider the setting in which voter preferences are uncertain.

Related Work

We survey related work in two areas. First, recent work in social choice studies the interaction between social influence and elections. However, all such work examines the attacker's problem of manipulating the election, leaving open the question of how elections can be defended against misinformation. Most closely related is the work of Wilder and Vorobeychik (2018), who study the attacker's problem of manipulating an election in a model where social influence spreads amongst voters from an attacker's chosen "seed nodes". However, they do not study the corresponding defender problem. Our model is also somewhat different in that we consider advertising to voters across a set of channels, rather than influence among the voters themselves. The work of Berdereck et al. (2016) is also closely related. They study the attacker's problem in a bribery setting where a single action (e.g., placing an ad) can sway multiple voters. Faliszewski et al. (2018) extend this to a domain where the initially bribed agents can influence others. Bredereck and Elkind (2017) also study a problem of manipulating diffusions on social networks, though not specifically in the context of elections. Sina et al. (2015) study a different form of

Algorithm 1 FPLT(ϵ)

Arbitrarily initialize S_d^0 and S_a^0
for $t = t...T$ **do**
  Draw p_a, p_d uniformly at random from $[0, \frac{1}{\epsilon}]^m$
  //TopK returns the set consisting of the indices of the smallest k entries of the given vector
  $S_a^t = TopK(\sum_{s=1}^{t-1} \ell(S_d^s) + p_a, k_a)$
  $S_d^t = TopK(\sum_{s=1}^{t-1} \ell(S_a^s) + p_d, k_d)$
return $\{S_a^t\}$ and $\{S_d^t\}$

manipulation, where edges may be added to the graph. Together, this body of work demonstrates substantial interest in the election control literature in emerging threats such as fake news. Our contribution is the first study of these problems from the perspective of a defender.

Second, our work is related to a complementary literature on budget allocation problems. Budget allocation is the attacker's problem in our model with no defender intervention: allocating an advertising budget to maximize the number of people reached. Efficient algorithms are available for a number of variants on this model (Alon, Gamzu, and Tennenholtz 2012; Soma et al. 2014; Miyauchi et al. 2015; Staib and Jegelka 2017). None of this work studies the game-theoretic problem of a defender trying to prevent an attacker from reaching voters. Soma et al. (2014) study a game where multiple advertisers compete for consumers, but not where one advertiser solely attempts to block the other. Their game is a potential game with pure strategy equilibria; however, it is easy to give examples in our model where the zero-sum nature of the attacker-defender interaction requires randomization. This makes equilibrium computation harder because we cannot simply use the best response dynamics. Our work is also related to the influence blocking maximization (IBM) problem (He et al. 2012) where one player attempts to limit the spread of a cascade in a social network. However, in IBM the starting points of the cascade are fixed in advance; in our problem the adversary chooses a randomized strategy to evade the defender.

Disjoint populations

In this setting, the population of voters is partitioned by the channels. Let V_u denote the set of voters affiliated with channel u. Exploiting the disjoint structure of the population, we can use linearity of expectation to rewrite the utility function $f(S_d, S_a)$ as

$$\sum_{u \in S_a \setminus S_d} \sum_{v \in V_u} \theta_v p_{uv} + \sum_{u \in S_a \cap S_d} \sum_{v \in V_u} \theta_v p_{uv}(1 - q_{uv})$$
$$= \sum_{u \in S_a} \sum_{v \in V_u} \theta_v p_{uv} - \sum_{u \in S_a \cap S_d} \sum_{v \in V_u} \theta_v p_{uv} q_{uv}.$$

Importantly, this expression is *linear* in each player's decisions. More formally, let $1[S]$ denote the indicator vector of a set S. Define the loss vector $\ell(S_a)$ to have value $1[u \in S_a] \sum_{v \in V_u} \theta_v p_{uv} q_{uv}$ in coordinate u. Then, we have

Algorithm 2 OnlineGradient(η, α, T, k_a)

$x_i^0 = \frac{1}{mk_a}$ for $i = 1...m$
for $t = 1...T$ **do**
  //Greedily maximizes a function subject to budget
  S_d^t = Greedy($g(\cdot|x_a^t)$, αk_d)
  $\nabla^t = \nabla F(x^{t-1}|S_d^t)$
  x^{t+1} = Update(x_t, ∇^t)
return $\{S_d^t\}$
function EXPONENTIATEDGRADIENTUPDATE
  $y^t = \min\{x^t e^{\eta \nabla^t}, 1\}$
  $x^{t+1} = \frac{k_a y^t}{||y^t||_1}$
function EUCLIDEANUPDATE
  $x^{t+1} = \arg\min_{y \in \mathcal{X}} ||y - (x^t + \eta \nabla^t)||_2$

that $f(S_d, S_a) = \sum_{u \in S_a} \sum_{v \in V_u} \theta_v p_{uv} - 1[S_d]^\top \ell(S_a)$, where the first term is constant with respect to S_d. Similarly, we can define a loss vector $\ell(S_d)$ which encapsulates the attacker's payoff for any defender action S_d.

To exploit this structure, we employ an algorithmic strategy based on online linear optimization. In such problems, a player seeks to optimize a (possibly adversarially chosen) sequence of linear functions over a feasible set. The aim is to achieve low *regret*, which measures the gap in hindsight to the best fixed decision over T rounds. We map online linear optimization onto our problem as follows. The feasible set for each player consists of m-dimensional binary vectors, where a 1 indicates that the player has chosen the corresponding channel and a 0 indicates that they have not. A vector is feasible if it sums to at most k_d (for the defender) or k_a (for the attacker). Both the attacker and defender will choose a series of actions from the corresponding feasible sets. In iteration t, if the attacker chooses a set S_a^t, and the defender receives a loss vector $\ell(S_a^t)$ and suffers loss $1[S_d^t]\ell(S_a^t)$. The attacker's loss functions are defined similarly.

Each player will generate their actions using the classical *Follow The Perturbed Leader (FTPL)* algorithm of Kalai and Vempala (2005) (Algorithm 1). At each iteration, each player best responds to the uniform distribution over all strategies played so far by their opponent, plus a small random perturbation. Note that best response here corresponds to linear optimization over the player's feasible set. Since any budget-satisfying vector is feasible, we simply select the highest-weighted k_d elements (or k_a for the attacker). Since FTPL has a no-regret guarantee for online linear optimization neither player can gain significantly by deviating from their history of play once the number of iterations is sufficiently high. More precisely, we have the following:

Theorem 1. *With $\frac{4n^2 \max\{k_a, k_d\}}{\epsilon^2}$ iterations of FTPL, uniform distributions on $\{S_a^t\}$ and $\{S_d^t\}$ form an ϵ-equilibrium.*

Nondisjoint populations

When voters may be reachable from multiple advertising channels, the approach from the previous section breaks down because utility is no longer linear for either player: selecting one channel reaches a subset of voters and hence

reduces the gain from selecting additional channels. Indeed, we can obtain the following hardness result:

Theorem 2. *In the nondisjoint setting, computing an optimal defender mixed strategy is APX-hard.*

The intuition is that the maximum coverage problem is essentially a special case of ours. However, diminishing returns provides useful algorithmic structure. Formally, both players' best response functions are closely related to submodular optimization problems. A set function is submodular if for all $A \subseteq B$ and $u \in V \setminus B$, $f(B \cup \{u\}) - f(B) \leq f(A \cup \{u\}) - f(A)$. We will only deal with monotone functions, where $f(A \cup \{u\}) - f(A) \geq 0$ holds for all A, u.

Our overall approach is to work in the marginal space of the attacker, by keeping track of only the marginal probability that they select each channel. That is, the attacker's current mixed strategy is concisely represented by a fractional vector x, where x_u gives the probability of selecting channel u. We run an approximate no-regret learning algorithm to update x over a series of iterations. At each iteration t, x is updated via a gradient step on a reward function induced by a set S_d^t played by the defender. Specifically, we will choose S_d^t to be an approximate best response to the current attacker mixed strategy.

There are two principal challenges that must be solved to enable this approach. First, we need to design an appropriate no-regret algorithm for the attacker. This is a challenging task as the attacker's utility is no longer linear (or even concave) in the marginal vector x. Second, we need to compute approximate best responses for the defender, which itself is NP-hard.

We resolve the first challenge by running an online gradient algorithm for the attacker, where the continuous objective at each iteration is the *multilinear extension* of an objective induced by the defender's strategy S_d^t. The multilinear extension is a fractional relaxation of a submodular set function. We define the multilinear extension $F(\cdot, S_d)$ induced by a defender strategy S_d as

$$F(x|S_d) = \sum_{v \in V} \theta_v \left(\prod_{u \in S_d} 1 - q_{uv} \right) \left(1 - \prod_{u=1}^{m} 1 - x_u p_{uv} \right)$$

That is, $F(x|S_d)$ is the expected value of $f(S_d, S_a)$ when each channel u is independently included in S_a with probability x_u. This is a special case of the multilinear extension more generally defined for arbitrary submodular set functions (Calinescu et al. 2011).

While F is in general not concave, we show that gradient-ascent style algorithms enjoy a no-regret guarantee against a $\frac{1}{2}$-approximation of the optimal strategy in hindsight. Our general strategy is to analyze online mirror ascent for continuous submodular functions. By making specific choices for the mirror map, we obtain two concrete algorithms (the update rules in Algorithm 2). The first is standard online gradient ascent, which takes a gradient step followed by Euclidean projection onto the feasible set $\mathcal{X} = \{x | \sum_u x_u \leq k_a, 0 \leq x \leq 1\}$. The second is an exponentiated gradient algorithm, which scales each entry of x according to the gradient and then normalizes to enforce the budget constraint. We have the following convergence guarantees:

Theorem 3. *Suppose that we apply Algorithm 2 to a sequence of multilinear extensions $F(\cdot|S_d^1)...F(\cdot|S_d^T)$. Let $b = \max_{|S_d| \leq k_d, u \in C} f(S_d, \{u\})$. Then, after T iterations, we have that*

$$\frac{1}{2} \max_{x^* \in \mathcal{X}} \sum_{t=1}^{T} F(x^*|S_d^t) - \sum_{t=1}^{T} F(x^t|S_d^t) \leq \sqrt{2} L D_{k_a} \sqrt{T}.$$

where for the exponentiated gradient update, $L = b$ and $D_{k_a} = k_a \log(m)$ and for the Euclidean update, $L = b\sqrt{m}$ and $D_{k_a} = \sqrt{k_a}$.

Our proof builds on the fact that for any single continuous submodular function, any local optimum is a $\frac{1}{2}$-approximation to the global optimum and translates this into the online setting. We remark that a no-regret guarantee for online gradient ascent for submodular functions was recently shown in (Chen, Hassani, and Karbasi 2018). Our more general analysis based on mirror ascent gives their result as a special case, and also allows us to analyze the exponentiated gradient update. The advantage is that the theoretical convergence rate is substantially better for exponentiated gradient, reducing the dimension dependence from $O(\sqrt{m})$ to $O(\log m)$. However, we also include the result for online gradient ascent since it tends to perform better empirically.

The second challenge is computing defender best responses. We show that the defender's best response problem is also closely related to a submodular maximization problem. Accordingly, we can compute approximate best responses via a greedy algorithm. Specifically, we show that the defender can obtain an ϵ-approximation to the optimal best response when the greedy algorithm is given an expanded budget of $\ln\left(\frac{n}{\epsilon}\right) k_d$ nodes.

In more detail: fix an attacker mixed strategy, denoted as σ_a. The defender best response problem is $\min_{|S_d| \leq k_d} \mathbb{E}_{S_a \sim \sigma_a} [f(S_d, S_a)]$. That is, we wish to minimize the number of voters who switch their vote, in expectation over σ_a. We consider the following equivalent problem

$$\max_{|S_d| \leq k_d} \mathop{\mathbb{E}}_{S_a \sim \sigma_a} [f(\emptyset, S_a) - f(S_d, S_a)],$$

i.e., maximizing the number of voters who do not switch as a result of the defender's action. Define $g(S_d|\sigma_a) = \mathbb{E}_{S_a \sim \sigma_a} [f(\emptyset, S_a) - f(S_d, S_a)]$. The key observation enabling efficient best response computations is the following:

Lemma 1. *For any attacker mixed strategy σ_a, $g(\cdot|\sigma_a)$ is a monotone submodular function.*

Accordingly, we can compute ϵ-optimal best responses by running the greedy algorithm with an expanded budget:

Theorem 4. *Running the greedy algorithm on the function g with a budget of $\ln\left(\frac{n}{\epsilon}\right) k_d$ outputs a set S_d satisfying $\mathbb{E}_{S_a \sim \sigma_a} [f(S_d, S_a)] \leq \min_{|S^*| \leq k_d} \mathbb{E}_{S_a \sim \sigma_a} [f(S_d, S_a)] + \epsilon$.*

Note that running greedy with the original budget k_d would give a $(1 - 1/e)$ approximation for the function g. However, a constant factor approximation for maximizing g

may not translate into any approximation for minimizing f because of the constant term $f(\emptyset, S_a)$ in the definition of g. Expanding the budget by a logarithmic factor gives a $1-\epsilon$ approximation with respect to g, and when ϵ is small enough the guarantee can be translated back in terms of f.

Combining the no-regret guarantee for the attacker and the best response approximation guarantee for the defender yields the following guarantee for the sequence of sets S_d^t:

Theorem 5. *After T iterations, let $\hat{\sigma}_T$ be the uniform distribution on $S_d^1 ... S_d^T$ output by Algorithm 2. The defender's payoff using $\hat{\sigma}_T$ is bounded as*

$$\max_{|S_a| \leq k_a} \mathop{\mathbb{E}}_{S_d \sim \hat{\sigma}_T} [f(S_d, S_a)] \leq 2\left(\tau + \epsilon + \frac{\sqrt{2}LD}{\sqrt{T}}\right).$$

Now, if we take $T = \left(\frac{4\sqrt{2}LD}{\epsilon}\right)^2$ and run greedy with $\epsilon' = \frac{\epsilon}{4}$, we obtain that $\hat{\sigma}_T$ is a 2-approximation Nash equilibrium strategy for the defender up to additive loss ϵ, using a budget of $\left(\ln\left(\frac{n}{\epsilon}\right) + O(1)\right) k_d$. Each iteration takes time $O(nm + m \log m + mn\alpha k)$ where the first term is to compute the attacker's gradient, the second to project onto their feasible strategy set, and the third is to run greedy for the defender (see the supplement for details).

Preference uncertainty

The previous two sections showed how to compute approximately optimal equilibrium strategies for the defender when both players know the starting preferences of the voters exactly. However, in practice the preferences will be subject to uncertainty, complicating the problem of optimally targeting resources. We now explore three models of preference uncertainty, each of which makes an increasingly conservative assumption about the information available to the defender. In each case, we show how to extend our algorithmic techniques to obtain approximately optimal defender strategies.

Stochastic uncertainty

We start with the least conservative assumption that the joint preference profile of the voters is drawn from a distribution which is known to both players. Each aims to maximize their payoff in expectation over the unknown draw from this distribution. We show that in both the disjoint and nondisjoint settings, the same algorithmic techniques go through with a natural modification to account for uncertainty.

Recall that θ denotes the voter preferences. θ is now drawn from a known joint distribution D. Let $f_\theta(S_d, S_a)$ denote the expected number of voters who switch to c_a under preferences θ. The payoffs are given by $\mathbb{E}_{\theta \sim D}[f_\theta(S_d, S_a)]$. Via linearity of expectation, we can write this as

$$\sum_{v \in V} \Pr[\theta_v = 1] \prod_{u \in S_d} (1 - q_{uv}) \left(1 - \prod_{u \in S_a} 1 - p_{uv}\right).$$

Dependence on the random preferences appears only through the term $\Pr[\theta_v = 1]$. This has two important consequences. First, we can evaluate the objective and implement the corresponding algorithms using access only to the marginals of the distribution. For many distributions of interest (e.g., product distributions where each voter adopts a preference independently), these will be known explicitly, and they can in general be evaluated to arbitrary precision via random sampling. Second, since the probability term is a nonnegative constant with respect to the strategies S_d and S_a, the payoffs retain properties such as linearity (in the disjoint case) or submodularity (in the nondisjoint case). Accordingly, we can obtain exactly the same computational guarantees as in the deterministic case, merely substituting the above expression for the payoffs:

Algorithm 3 FPLT-Asymmetric(ϵ)

Arbitrarily initialize S_d^0 and $S_a^0(\theta_j)$
for $t = t...T$ **do**
  Draw p_a^j, p_d uniformly at random from $[0, \frac{1}{\epsilon}]^m$
  //TopK returns the set consisting of the indices of the smallest k entries of the given vector
  $S_a^t(\theta_j) = TopK(\sum_{s=1}^{t-1} \ell_{\theta_j}(S_d^s) + p_a^j, k_a)\ j = 1...N$
  $S_d^t = TopK(\sum_{s=1}^{t-1} \frac{1}{N} \sum_{j=1}^{N} \ell_{\theta_j}(S_a^s(\theta_j)) + p_d, k_d)$
return $\{S_a^t\}$ and $\{S_d^t\}$

Algorithm 4 OG-Asymmetric($\eta, \alpha, T, k_a N$)

Draw $\theta_1 ... \theta_N$ iid from D
$x_i^0(\theta_j) = \frac{1}{mk_a}$ for $i = 1...m,\ j = 1...N$
for $t = 1...T$ **do**
  S_d^t = Greedy$\left(\frac{1}{N} \sum_{j=1}^{N} g(\cdot | x^{t-1}(\theta_j)), \alpha k_d\right)$
  for $j = 1...N$ **do**
    $\nabla^t(\theta_j) = \nabla F(x^{t-1}(\theta_j) | S_d^t)$
    $x^{t+1}(\theta_j) = \text{Update}(x^t(\theta_j), \nabla^t(\theta_j))$
return $\{S_d^t\}$

Theorem 6. *By substituting $\Pr[\theta_v = 1]$ for θ_v in the definition of f, FTPL achieves the same guarantee for the stochastic objective as in Theorem 1. Further, making this substitution in the definition of $F(x|S_d)$ and running Algorithm 2 yields the same guarantee as in Theorem 5.*

Asymmetric uncertainty

We now consider a case where the true voter preferences are still drawn from a distribution, but the players have access to asymmetric information about the draw. Specifically, the defender knows only the prior distribution, while the attacker has access to the true realized draw. We aim to solve the defender problem:

$$\min_{\sigma_d} \mathop{\mathbb{E}}_{\theta \sim D} \left[\max_{|S_a| \leq k_a} \mathop{\mathbb{E}}_{S_d \sim \sigma_d} [f_\theta(S_a, S_d)]\right]. \quad (1)$$

Here, the defender minimizes in expectation over the distribution of voter preferences, but the attacker maximizes knowing the actual draw $\theta \sim D$. We show how to compute approximately optimal defender strategies for an arbitrary distribution D, assuming only the ability to draw i.i.d. samples. We first prove a concentration bound for the number

of samples required to approximate the true problem over defender mixed strategies with bounded support:

Lemma 2. *Draw $N = O\left(\frac{n^2 mT}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\log m\right)$ samples. With probability at least $1-\delta$, for defender mixed strategy σ_d with support size at most T,*

$$\left|\mathop{\mathbb{E}}_{\theta\sim D}\left[\max_{|S_a|\le k_a}\mathop{\mathbb{E}}_{S_d\sim\sigma_d}\left[f_\theta\left(S_a,S_d\right)\right]\right]-\frac{1}{N}\sum_{i=1}^{N}\max_{|S_a|\le k_a}\mathop{\mathbb{E}}_{S_d\sim\sigma_d}\left[f_\theta\left(S_a,S_d\right)\right]\right|\le\epsilon$$

We now give generalizations of our earlier algorithms for the disjoint and nondisjoint settings. Each algorithm first draws sufficient samples for Lemma 2 to hold. Then, it simulates a separate adversary for each of the samples, mimicking the ability of the adversary to respond to the true draw of θ. Each adversary runs a separate instance of a no-regret learning algorithm (FTPL for the disjoint case and online gradient for the nondisjoint case). In each iteration, the defender updates according to the *expectation* over all of the adversaries (since the defender does not know the true θ). More precisely, in the disjoint case, the defender's loss function in iteration t is given by the average of the loss functions generated by each of the individual adversaries. The defender takes a FTPL step according to this average loss. In the nondisjoint case, the defender computes a greedy best response where the objective is given by average influence averted over all of the current adversary strategies. We show the following approximation guarantee for each setting:

Theorem 7. *Using inputs $T = \frac{4n^2\max\{k_a,k_d\}}{\epsilon^2}$, and $N = O\left(\frac{n^2mT}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\log m\right)$ for Algorithm 3, the uniform distribution over $\{S_d^t\}$ is an ϵ-equilibrium defender strategy.*

Theorem 8. *Run Algorithm 4 with $T = \frac{2L^2D^2}{\epsilon^2}$ iterations, $\eta = \frac{1}{L\sqrt{2T}}$, $\alpha = \ln\frac{n}{\epsilon} + O(1)$, and $N = O\left(\frac{n^3T}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\log n\right)$ samples. Let $\hat{\sigma}_T$ be the uniform distribution on $S_d^1 ... S_d^T$. With probability at least $1-\delta$, the defender's payoff using $\hat{\sigma}_T$ is bounded as*

$$\mathop{\mathbb{E}}_{\theta\sim D}\left[\max_{|S_a|\le k_a}\mathop{\mathbb{E}}_{S_d\sim\hat{\sigma}_T}\left[f_\theta\left(S_a,S_d\right)\right]\right]\le 2\tau+\epsilon.$$

where τ is the optimal value for Problem 1.

That is, the defender can obtain the same approximation guarantee in the same number of iterations. Each iteration takes time $O(N(mn + m\log m))$ to update all of the adversaries, while the defender best response problem still requires one call to greedy as before.

Adversarial uncertainty

We now consider the most conservative uncertainty model, in which the voters' preferences are chosen adversarially within some uncertainty set. Specifically, there is a nominal preference profile $\hat{\theta}$ (e.g., $\hat{\theta}$ may be an estimate from historical data). We are guaranteed that the true θ lies within the uncertainty set $\mathcal{U}_\ell = \{\theta : |\{v : \theta_v \neq \hat{\theta}_v\}| \le \ell\}$. That is, the true θ may differ in up to ℓ places from our estimate. The defender solves the robust optimization problem

$$\min_{\sigma_d}\max_{\theta\in\mathcal{U}_\ell}\max_{S_a\le|k_a|}\mathop{\mathbb{E}}_{S_d\sim\sigma_d}\left[f(S_d,S_a)\right] \tag{2}$$

which optimizes against the worst case $\theta \in \mathcal{U}_\ell$. Note that Problem 2 essentially places the choice of θ under the control of the attacker (formally, we can combine the two max operations). We show that the attacker component of the algorithms when payoffs are common knowledge can be generalized to handle this expanded strategy set. Essentially, the attacker will now have two kinds of actions. First, selecting a channel for a fake news message (as before). Second, directly reaching a given voter by changing their initial preference. We equivalently simulate the second class of actions by adding a new channel v' for each voter v. The new channel has $q_{v',v} = 0$ and $p_{v',v} = 1$. That is, the attacker always succeeds in influencing v and can never be stopped by the defender. The attacker's pure strategy set now consists of all choices of k_d normal channels and ℓ of the new channels.

Our result from the disjoint case goes through essentially unchanged. Algorithm 5 runs FTPL for both players, as before. The only change is in the linear optimization step for the attacker, which now selects separately the top k_a regular channels and ℓ new channels (lines 5 and 6). We have the following guarantee:

Theorem 9. *Using $T = \frac{4n^2\max\{k_a+\ell,k_d\}}{\epsilon^2}$ for Algorithm 5, the uniform distribution over $\{S_d^t\}$ is an ϵ-equilibrium defender strategy.*

The main technical difference is in the nondisjoint case, where the attacker's problem now corresponds to submodu-

Algorithm 5 FPLT-Adversarial(ϵ)

1: Arbitrarily initialize S_d^0 and S_a^0
2: **for** $t = t...T$ **do**
3: Draw p_a, p_d uniformly at random from $[0, \frac{1}{\epsilon}]^m$
4: //TopK returns the set consisting of the indices of the smallest k entries of the given vector
5: $S_a^t = TopK\left(\left[\sum_{s=1}^{t-1}\ell(S_d^s)+p_a\right]_{1:m}, k_a\right)\cup$
6: $TopK\left(\left[\sum_{s=1}^{t-1}\ell(S_d^s)+p_a\right]_{m+1:m+n}, \ell\right)$
7: $S_d^t = TopK(\sum_{s=1}^{t-1}\ell(S_a^s)+p_d, k_d)$
8: **return** $\{S_a^t\}$ and $\{S_d^t\}$

Algorithm 6 OG-Adversarial(η, α, T, k_a)

1: $x_i^0 = \frac{1}{mk_a}$ for $i = 1...m+n$
2: **for** $t = 1...T$ **do**
3: S_d^t = Greedy(x^{t-1}, αk_d)
4: $\nabla^t = \nabla F(x^{t-1}|S_d^t)$
5: $x_{1:m}^{t+1}$ = Update($x_{1:m}^t, \nabla_{1:m}^t, k_a$)
6: $x_{m+1:m+n}^{t+1}$ = Update$\left(x_{m+1:m+n}^t, \nabla_{m+1:m+n}^t, \ell\right)$
7: **return** $\{S_d^t\}$

lar maximization over a partition matroid (since the budget constraint is now split into two categories instead of a single category as before). More general matroid constraints can complicate submodular maximization, e.g., the greedy algorithm no longer obtains the optimal approximation ratio. Fortunately, our use of a continuous relaxation and online gradient ascent for the attacker can be shown to generalize without loss to arbitrary matroid constraints:

Theorem 10. *After T iterations, let $\hat{\sigma}_T$ be the uniform distribution on $S_d^1 \ldots S_d^T$ output by Algorithm 6. The defender's payoff using $\hat{\sigma}_T$ (with respect to Problem 2) is bounded as*

$$\max_{|S_a| \leq k_a} \mathop{\mathbb{E}}_{S_d \sim \hat{\sigma}_T} [f(S_d, S_a)] \leq 2 \left(\tau + \epsilon + \frac{L^2 D_{k_a+\ell}^2}{2\sqrt{T}} \right).$$

Experiments

We now examine our algorithms' empirical performance, and what the resulting values reveal about the difficulty of defending elections across different settings. We focus on the nondisjoint setting for two reasons. First it is the more general case. Second, FTPL is guaranteed to converge to an ϵ-optimal strategy in the disjoint setting, while in the nondisjoint setting is important to empirically assess our algorithm's approximation quality. Our experiments use the Yahoo webscope dataset (Yahoo 2007). The dataset logs bids placed by advertisers on a set of phrases. We create instances where the phrases are advertising channels and the accounts are voters. To generate each instance, we sample a random subset of 100 channels and 500 voters. Each propagation probability is drawn uniformly at random from $[0, 0.2]$ for each player. Each voter's preference is also drawn uniformly at random. All results are averaged over 30 iterations.

We start with fully known preferences and examine the approximation quality of Algorithm 2. Importantly, we do not increase the defender's budget (i.e., $\alpha = 1$). Empirically, Algorithm 2 performs substantially better than its theoretical guarantee, rendering bicriteria approximation unnecessary.

We use the mixed strategies that Algorithm 2 outputs to compute upper and lower bounds on the value of the game. The upper bound b_u is the attacker's best response to the defender mixed strategy, while the lower bound b_ℓ is the defender's best response to the attacker mixed strategy. It is easy to see that the defender cannot obtain utility better than b_ℓ, and Algorithm 2's mixed strategy guarantees utility no worse than b_u. Hence, we use $\frac{b_u - b_\ell}{b_\ell}$ as an upper bound on the optimality gap. Since finding exact best responses is NP-hard, we use mixed integer programs (see the supplement).

Table 1 shows that Algorithm 2 computes highly accurate defender equilibrium strategies across a range of values for k_a and k_d. We use $T = 50$ iterations with $\eta = 0.05$. *The average optimality gap is always (provably) under 6%.* Moreover, this value is an upper bound, and the real gap may be smaller. We conclude that Algorithm 2 is highly effective at computing near-optimal defender strategies. Next, Figure 1 examines how the attacker's payoff varies as a function of k_a and k_d. Even for large k_d, the defender cannot completely erase the attacker's impact (to be expected since

k_d/k_a	5	10	20
5	0.016 ± 0.007	0.016 ± 0.010	0.026 ± 0.015
10	0.017 ± 0.008	0.020 ± 0.008	0.037 ± 0.017
20	0.014 ± 0.006	0.025 ± 0.012	0.053 ± 0.022

Table 1: Upper bound on optimality gap for Algorithm 2. Average over 30 instances; $\pm$ denotes standard deviation.

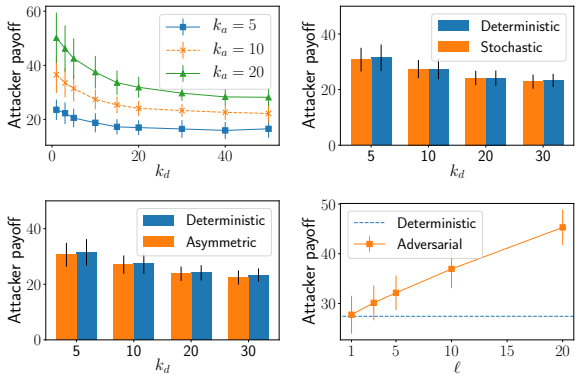

Figure 1: Top left: Attacker's payoff as the budget constraint for each player varies. Top right: attacker payoff with stochastic uncertainty. Bottom left: asymmetric uncertainty. Bottom right: adversarial uncertainty, varying the uncertainty set size ℓ.

$q_{uv} < 1$ and so the defender's message is not perfectly effective). However, the defender can obtain a large reduction in the attacker's influence when k_a is high. The empirical payoffs are convex in k_d, meaning that the defender achieves this reduction with a moderate value of k_d and sees little improvement afterwards. When k_a is low, even large defender expenditures have a relatively little impact. Intuitively, it is harder for the defender to ensure an intersection between their own strategy and the attacker's when the attacker only picks a small number of channels to begin with.

Next, we examine the impact of uncertainty. Figure 1 shows the attacker's payoff under stochastic, asymmetric, and adversarial uncertainty compared to fully known payoffs. Stochastic uncertainty leaves the attacker's payoff virtually identical. Surprisingly, this also holds for the asymmetric case. However, in the adversarial setting, the attacker's payoff scales linearly with ℓ, indicating that the defender cannot mitigate the impact of such uncertainty. Hence, the defender can benefit substantially from gathering enough information to at least estimate the distribution of θ, even if the attacker still has privileged information.

Conclusion: We introduce and study the problem of a defender mitigating the impact of adversarial misinformation on an election. Across a range of population structures and uncertainty models, we provide polynomial time approximation algorithms to compute equilibrium defender strategies, which empirically provide near-optimal payoffs. Our results show that the defender can substantially benefit from modest resource investments, and from gathering enough infor-

mation to estimate voter preferences.

Acknowledgments: This work was partially supported by the National Science Foundation (CNS-1640624, IIS-1649972, and IIS-1526860), Office of Naval Research (N00014-15-1-2621), and Army Research Office (W911NF1610069, MURI W911NF1810208).

References

Allcott, H., and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31(2):211–236.

Alon, N.; Gamzu, I.; and Tennenholtz, M. 2012. Optimizing budget allocation among channels and influencers. In *WWW*, 381–388. ACM.

Baumeister, D.; Erdélyi, G.; Erdélyi, O. J.; and Rothe, J. 2015. Complexity of manipulation and bribery in judgment aggregation for uniform premise-based quota rules. *Mathematical Social Sciences* 76:19–30.

Brader, T. 2005. Striking a responsive chord: How political ads motivate and persuade voters by appealing to emotions. *American Journal of Political Science* 49(2):388–405.

Bredereck, R., and Elkind, E. 2017. Manipulating opinion diffusion in social networks. In *IJCAI*.

Bredereck, R.; Faliszewski, P.; Niedermeier, R.; and Talmon, N. 2016. Large-scale election campaigns: Combinatorial shift bribery. *Journal of Artificial Intelligence Research* 55:603–652.

Calinescu, G.; Chekuri, C.; Pál, M.; and Vondrák, J. 2011. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing* 40(6):1740–1766.

Chen, J.; Faliszewski, P.; Niedermeier, R.; and Talmon, N. 2015. Elections with few voters: Candidate control can be easy. In *AAAI*, volume 15, 2045–2051.

Chen, L.; Hassani, H.; and Karbasi, A. 2018. Online continuous submodular maximization. In *International Conference on Artificial Intelligence and Statistics*, 1896–1905.

Chi, F., and Yang, N. 2011. Twitter adoption in congress. *Review of Network Economics* 10(1).

DellaVigna, S., and Kaplan, E. 2007. The fox news effect: Media bias and voting. *The Quarterly Journal of Economics* 122(3):1187–1234.

Erdélyi, G.; Hemaspaandra, E.; and Hemaspaandra, L. A. 2015. More natural models of electoral control by partition. In *International Conference on Algorithmic Decision-Theory*, 396–413. Springer.

Erdélyi, G.; Reger, C.; and Yang, Y. 2017. The complexity of bribery and control in group identification. In *AAMAS*, 1142–1150.

Faliszewski, P.; Hemaspaandra, E.; Hemaspaandra, L. A.; and Rothe, J. 2009. Llull and copeland voting computationally resist bribery and constructive control. *Journal of Artificial Intelligence Research* 35:275–341.

Faliszewski, P.; Gonen, R.; Koutecký, M.; and Talmon, N. 2018. Opinion diffusion and campaigning on society graphs. In *IJCAI*, 219–225.

Faliszewski, P.; Hemaspaandra, E.; and Hemaspaandra, L. A. 2011. Multimode control attacks on elections. *Journal of Artificial Intelligence Research* 40(1):305–351.

Gerber, A. S.; Karlan, D.; and Bergan, D. 2009. Does the media matter? a field experiment measuring the effect of newspapers on voting behavior and political opinions. *American Economic Journal: Applied Economics* 1(2):35–52.

He, X.; Song, G.; Chen, W.; and Jiang, Q. 2012. Influence blocking maximization in social networks under the competitive linear threshold model. In *SDM*, 463–474.

Holcomb, J.; Gottfried, J.; and Mitchell, A. 2013. News use across social media platforms. *Pew Research Journalism Project*.

Kalai, A., and Vempala, S. 2005. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences* 71(3):291–307.

Li, Y.; Jiang, Y.; and Wu, W. 2017. Protecting elections with minimal resource consumption. In *AAMAS*.

Liu, H.; Feng, H.; Zhu, D.; and Luan, J. 2009. Parameterized computational complexity of control problems in voting systems. *Theoretical Computer Science* 410(27-29):2746–2753.

Loreggia, A.; Narodytska, N.; Rossi, F.; Venable, K. B.; and Walsh, T. 2015. Controlling elections by replacing candidates or votes. In *AAMAS*.

Miyauchi, A.; Iwamasa, Y.; Fukunaga, T.; and Kakimura, N. 2015. Threshold influence model for allocating advertising budgets. In *ICML*, 1395–1404.

Pennycook, G.; Cannon, T. D.; and Rand, D. G. 2017. Prior exposure increases perceived accuracy of fake news.

Sina, S.; Hazon, N.; Hassidim, A.; and Kraus, S. 2015. Adapting the social network to affect elections. In *AAMAS*, 705–713.

Soma, T.; Kakimura, N.; Inaba, K.; and Kawarabayashi, K.-i. 2014. Optimal budget allocation: Theoretical guarantee and efficient algorithm. In *ICML*, 351–359.

Staib, M., and Jegelka, S. 2017. Robust budget allocation via continuous submodular functions. In *ICML*.

Wattal, S.; Schuff, D.; Mandviwalla, M.; and Williams, C. B. 2010. Web 2.0 and politics: the 2008 us presidential election and an e-politics research agenda. *MIS quarterly* 669–688.

Wilder, B., and Vorobeychik, Y. 2018. Controlling elections through social influence. In *AAMAS*, 265–273.

Yahoo. 2007. Yahoo! webscope dataset ydata-ysm-advertiser-bids-v1 0. http://research.yahoo.com/Academic_Relations.

Yang, Y.; Shrestha, Y. R.; and Guo, J. 2016. How hard is bribery with distance restrictions? In *ECAI*, 363–371.

Yin, Y.; An, B.; Hazon, N.; and Vorobeychik, Y. 2018. Optimal defense against election control by deleting voter groups. *Artificial Intelligence* 259:32–51.