Strategic Classification in the Dark

Ganesh Ghalme [*1] **Vineet Nair** [*1] **Itay Eilat** [1] **Inbal Talgam-Cohen** [1] **Nir Rosenfeld** [1]

Abstract

Strategic classification studies the interaction between a classification rule and the strategic agents it governs. Under the assumption that the classifier is *known*, rational agents respond to it by manipulating their features. However, in many real-life scenarios of high-stake classification (e.g., credit scoring), the classifier is not revealed to the agents, which leads agents to attempt to learn the classifier and game it too. In this paper we generalize the strategic classification model to such scenarios. We define the "price of opacity" as the difference in prediction error between opaque and transparent strategy-robust classifiers, characterize it, and give a sufficient condition for this price to be strictly positive, in which case transparency is the recommended policy. Our experiments show how Hardt et al.'s robust classifier is affected by keeping agents in the dark.

1. Introduction

The increasing role of machine learning in society has led to much recent interest in learning methods that explicitly consider the involvement of human agents. A canonical instance is *strategic classification*—the study of classifiers that are robust to gaming by self-interested agents. This line of research was initiated by (Brückner & Scheffer, 2011; Hardt et al., 2016), and has quickly amassed a multitude of insightful follow-up works (see below). As a running example, consider a firm (e.g., bank) classifying users (e.g., loan applicants) into two classes (e.g., "approve" or "deny"), based on their attributes or features. Since applicants would like their request to be approved (regardless of their true solvency), they might be inclined to spend efforts on modifying their features to align with the bank's decision rule. This can lead to gaming behaviors like holding multiple credit cards, which have nothing to do with the true chances of paying back the loan. The goal of the strategic classification literature is to show when and how the firm can learn a robust classifier, which is guaranteed to perform well even when users strategically modify their features in response to it. But to be realistically applicable, and to ensure safe deployment in settings involving consequential decision-making, care must be taken as to what assumptions underlie the theoretical guarantees.

Currently known results for strategic classification rely on a key assumption: that users have *full knowledge of the classifier* and respond accordingly, or in other words, the classifier is completely transparent. Under this assumption, the strategic classification literature aims for classifiers with performance "comparable to ones that rely on secrecy." The literature thus treats the transparency/opaqueness of a classifier as a dichotomy, where only the former is assumed to require a strategy-robust learning process. However, in many real-life scenarios of high-stake classification, the situation is more nuanced. This point was made already in the original strategic classification paper: "*Secrecy is not a robust solution to the problem; information about a classifier may leak, and it is often possible for an outsider to learn such information from classification outcomes*" (Hardt et al., 2016). So often we have a hybrid situation – the users know something, not everything, about the classifier, and use their partial knowledge to strategically respond to it. This partial knowledge is a result of a second, smaller-scale learning process, this time of the users themselves, based on information aggregated by media platforms or by their own social network. As anecdotal evidence, consider the secret credit-scoring algorithms applied in the US and elsewhere, coupled with the plethora of online information on how to improve credit scores. In the case of SCHUFA, the German credit-scoring algorithm, there was even a recent crowdsourced effort in which a data donation platform gathered samples from the public with the goal of "reverse-engineering" the secret algorithm (`https://openschufa.de/english/`).

The above scenarios demonstrate that rather than preventing gaming, opacity leads users to attempt to learn the classifier and game it too. How relevant is strategic classification to such hybrid situations? What policy recommendations can the theory make for firms whose classifier is being learned? Our goal in this work is to adapt the strategic classification framework to such settings, in order to examine both the-

[*] Equal contribution [1] Technion - Israel Institute of Technology. Correspondence to: Ganesh Ghalme <ganeshg@campus.technion.ac.il>, Vineet Nair <vineet@cs.technion.ac.il>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

oretically and empirically how keeping users in the dark affects the guarantees of robust classifiers. We introduce the "optimistic firm": It applies a strategy-robust classifier while keeping classification details proprietary. The firm is optimistic in the following sense—it assumes that while gaming on behalf of the users is with respect not to f but to its learned counterpart $\widehat{f}$, the robustness of f still protects it from strategic behavior. We compare how well such a firm fares in comparison to the alternative of simply revealing the classifier. Is the firm justified in its optimistic adoption of strategic classification methods? Our results give a largely negative answer to this question.

Our Results. To address the above questions, we compare the prediction error of a robust classifier f when the strategic users must learn a version of it $\widehat{f}$, and when f is transparent to the users. We term the difference between these errors "Price of OPacity (POP)" (Sec. 3.1). Notice that whenever POP is > 0, the policy implication is that the transparent policy prevails. Our main set of theoretical results (Sec. 3.2) shows the following: We show that even if the estimate $\widehat{f}$ of f is quite good, such that the population mass on which these classifiers disagree is small, these small errors can potentially be enlarged by the strategic behavior of the user population. Indeed, if the small differences in classification incentivize the users to modify their features differently under f and $\widehat{f}$, that means that a much larger mass of users may ultimately be classified differently. We call this enlarged population subset the "enlargement region".

From the users' perspective, the enlargement region is undesirable: we show the population mass in this region will be classified negatively whereas under transparency it would have been classified positively (Thm. 3). Thus, opaqueness harms those in the region who are truly qualified. From the firm's perspective, the connection between the enlargement set and POP is not immediately clear – perhaps $\widehat{f}$ is inadvertently fixing the classification errors of f, making POP negative? We show a sufficient condition on the mass of the enlargement set for POP to be positive (Thm. 2). We demonstrate the usefulness of these results by analyzing a normally-distributed population classified linearly (Sec. 3.3). In this setting, we show via a combination of theory and experiments that POP can become very large (Prop. 6, Sec. 4). Finally, we formalize the intuition that the problem with keeping users in the dark is that the firm in effect is keeping *itself* in the dark (as to how users will strategically react). We show in Sec. 5 that if $\widehat{f}$ can be anticipated, a natural generalization of strategic classification holds (Zhang & Conitzer, 2021; Sundaram et al., 2021).

We complement our theoretical results with experiments on synthetic data as well as on a large dataset of loan requests. The results reinforce our theoretical findings, showing that POP can be quite large in practice. We use the loans dataset to further explore the implications of an opaque policy on users, showing that it can disproportionately harm users having few social connections.

More Related Work. There is a growing literature on the strategic classification model; to the best of our knowledge, this model has only been studied so far under the assumption of the classifier being known to the agent. Dong et al. (2018) study an online model where agents appear sequentially and the learner does not know the agents' utility functions. Chen et al. (2020) design efficient learning algorithms that are robust to manipulation in a ball of radius δ from agents' real positions (see also Ahmadi et al. (2020)). Milli et al. (2019) consider the the social impacts of strategic classification, and the tradeoffs between predictive accuracy and the social burden it imposes (a raised bar for agents who naturally are qualified). Hu et al. (2019) focus on a fairness objective and raise the issue that different populations of agents may have different manipulation costs. Braverman & Garg (2020) study classifiers with randomness. Kleinberg & Raghavan (2019) study a variant of strategic classification where the agent can change the ground truth by investing effort. Alon et al. (2020); Haghtalab et al. (2020) generalize their setting to multiple agents. Perdomo et al. (2020); Bechavod et al. (2021a) study causal reasoning in strategic classification. Bechavod et al. (2021a) find that strategic manipulation may help a learner recover the causal variables. Rosenfeld et al. (2020) provide a framework for learning predictors that are both accurate and that promote good actions. Levanon & Rosenfeld (2021) provide a practical, differentiable learning framework for learning in diverse strategic settings.

In concurrent and independent work, (Bechavod et al., 2021b) also consider the problem of strategic learning where the classifier is not revealed to the users. In their work, the users belong to distinct population subgroups from which they learn versions of the scoring rule and Jury is aware of these versions (close to our all-powerful Jury setting in Sec. 5). We view their work as complementary to ours in that they focus on incentivizing users to put in genuine efforts to self-improve, as advocated by (Kleinberg & Raghavan, 2019), rather than on robustness of the classifier to unwanted gaming efforts as in (Hardt et al., 2016). One finding of ours that is somewhat related is the difference in how opaque learning affects users who are more vs. less socially well-connected.

2. Preliminaries and Our Learning Setup

Classification. Let $x \in \mathcal{X} \subseteq \mathbb{R}^d$ denote a d-dimensional feature vector describing a user (e.g., loan applicant), and let D be a distribution over user population $\mathcal{X}$. Label $y \in \mathcal{Y} = \{\pm 1\}$ is binary (e.g., loan returned or not), and for simplicity we assume true labels are generated by (an unknown) ground-truth function $h(x) = y$.[1] Given access to a sample set $T = \{(x_i, y_i)\}_{i=1}^n$ with $x_i \overset{\text{iid}}{\sim} D$, $y_i = h(x_i)$, the standard goal in classification is to find a classifier $f : \mathcal{X} \to \mathcal{Y}$ from a class $\mathcal{H}$ that predicts well on D.

Strategic Classification ((Hardt et al., 2016)). In this model, users are assumed to know f, and to strategically and rationally manipulate their features in order to be classified positively by f. Performance of a classifier f is therefore evaluated on *manipulated data*, in contrast to standard supervised learning, where train and test data are assumed to be drawn from the same distribution. In the remainder of this subsection we formally describe the setup of Hardt et al. (2016). Users gain 1 from a positive classification and -1 from a negative one, and feature modification is costly. Every user x modifies her features to maximize utility. Let:

$$\Delta_f(x) = \operatorname*{argmax}_{u \in \mathcal{X}} \{f(u) - c(x, u)\} \tag{1}$$

be the utility maximizing feature vector of user x, where $c(\cdot, \cdot)$ is a publicly-known, non-negative *cost function* quantifying the cost of feature modification from x to u.[2] It is convenient to think of the mapping $\Delta_f(\cdot)$ as "moving" points in $\mathcal{X}$-space. The goal in strategic classification is to find f minimizing the *strategic error* $\texttt{err}(f) := \mathbb{P}_{x \sim D}\{h(x) \neq f(\Delta_f(x))\}$, which is the probability of the classification error by f when every user x responds to f by modifying her features to $\Delta_f(x)$.

Strategic classification can be formulated as a Stackelberg game between two players: *Jury*, representing the learner, and *Contestant*, representing the user population. The game proceeds as follows: Jury plays first and commits to a classifier $f \in \mathcal{H}$. Contestant responds by $\Delta_f(\cdot)$, manipulating feature x to modified feature $\Delta_f(x)$. The payoff to Jury is $1 - \texttt{err}(f)$. To choose f, Jury can make use of the *unmodified* sample set $T = \{(x_i, y_i)\}_{i=1}^n$, as well as knowledge of Δ_f (i.e., Jury can anticipate how contestant will play for any choice of f). The payoff to Contestant is the expected utility of a user x sampled from D and applying modification $\Delta_f(x)$, i.e., $\mathbb{E}_{x \sim D}[f(\Delta_f(x)) - c(x, \Delta_f(x))]$. Observe that Contestant's best response to Jury's choice f is thus $\Delta_f(\cdot)$ as defined in Eq. (1). We remark that Contestant's utility (and therefore Δ_f) does not directly depend on h. It will be convenient to rewrite $\Delta_f(\cdot)$ as follows: denoting by $C_f(x) = \{u \mid c(x, u) < 2, f(u) = 1\}$ the set of "feasible" modifications for x under f (where the condition that the cost < 2 comes from $f(x) \in \{-1, 1\}$),

$$\Delta_f(x) := \begin{cases} \operatorname*{argmin}_{u \in C_f(x)} c(x, u) & f(x) = -1 \wedge C_f(x) \neq \emptyset; \\ x & \text{otherwise.} \end{cases}$$

In words, x "moves" to the lowest-cost feasible point in $C_f(x)$ rather than staying put if x is (i) classified negatively, and (ii) has a non-empty feasible set.

Strategic Classification in the Dark. A key assumption made throughout the strategic classification literature is that *Contestant knows classifier f exactly*, i.e., that f is public knowledge. In this case, we say Jury implements a *transparent policy*. Our goal in this paper is to explore the case where this assumption does not hold, keeping Contestant "in the dark". We refer to this as an *opaque policy*.

We consider a *two-sided statistical Stackelberg game*, where both Jury *and* Contestant obtain information through samples.[3] In particular, we assume that since Jury does not publish f, Contestant is coerced to estimate f from data available to her (e.g., by observing the features of friends and the associated predicted outcomes). In our theoretical analysis we make this concrete by assuming that Contestant observes m samples also drawn iid from D but *classified by f*, and denote this set $T_C = \{(x_i, f(x_i))\}_{i=1}^m$. In our experiments in Sec. 4 we consider more elaborate forms of information that Contestant may have access to. Contestant uses T_C to learn a classifier $\hat{f} \in \mathcal{H}_C$ serving as an estimate of f, which she then uses to respond by playing $\Delta_{\hat{f}}$.[4] To allow $\widehat{f}$ to be arbitrarily close to f we will assume throughout that $\mathcal{H}_C = \mathcal{H}$ (otherwise the optimal error of $\widehat{f}$ can be strictly positive). Intuitively, if $\widehat{f}$ is a good estimate of f (e.g., when m is sufficiently large), then we might expect things to proceed as well as if Contestant had known f. However, as we show in Sec. 3, this is not always the case.

We have defined what Contestant knows, but to proceed, we must be precise about what Jury is assumed to know. We see two natural alternatives: (a) Jury has full knowledge of $\widehat{f}$. We refer to this setting as the *All-Powerful Jury*. (b) Jury is unaware of $\widehat{f}$, which we refer to as *Jury in the Dark*. The latter is our primary setting of interest, as we believe it to be more realistic and more relevant in terms of policy implications. It is important to stress that Jury is in the dark on account of his own decision to be opaque; any Jury that chooses transparency becomes cognizant of Contestant's strategy by definition. We present an analysis of the Jury in the Dark setting in Secs. 3,4), and return to the All-Powerful Jury in Sec. 5.

3. Jury in the Dark

When Jury is in the dark, his knowledge of how Contestant will respond is incomplete, but he must nonetheless commit

[1] The main results of our work in Sec. 3.2 hold even when D is a joint distribution over $\mathcal{X} \times \mathcal{Y}$.

[2] Ties are broken lexicographically.

[3] We call the conventional strategic learning described above as a *one-sided statistical Stackelberg game*.

[4] Our results also hold in expectation, when the sample sets for the contestant are different across the users.

to a strategy by choosing f. We analyze a Jury who assumes Contestant will play Δ_f. Indeed, if Jury believes that Contestant's estimated classifier $\widehat{f}$ is a good approximation of f, then replacing the unobserved $\widehat{f}$ with the known and similar f is a natural—if slightly optimistic—approach. The optimistic Jury implicitly assumes that small discrepancies between f and $\widehat{f}$ will not harm his accuracy too much. This approach also has a practical appeal: In practice, many important classifiers are kept opaque, yet firms are becoming more aware to strategic behavior as well as to the possibility of information leakage. A firm interested in protecting itself by learning in a manner that is strategic-aware while remaining opaque may choose to apply one of the cutting-edge robust classifiers, and virtually all current methods for strategic classification assume Contestant plays Δ_f. Despite its appeal, our results indicate that the above intuition can be misleading. By carefully tracing how errors propagate, we show that small discrepancies between f and $\widehat{f}$ can 'blow up' the error in a way that leads to complete failure of the optimistic approach.

3.1. Price of Opacity and Related Definitions

We are interested in studying the effects of Jury committing to an opaque policy, as it compares to a transparent policy. To this end we extend the definition of predictive error:

Definition 1. *The strategic error when Jury plays f and Contestant responds to $\widehat{f}$ is given by:*

$$\texttt{err}(f,\hat{f}) = \mathbb{P}_{x\sim D}\{h(x) \neq f(\Delta_{\widehat{f}}(x))\}. \tag{2}$$

Note that $\texttt{err}(f,f) = \texttt{err}(f)$; we mostly use $\texttt{err}(f,f)$ as a reminder that in principle Jury and Contestant may use different classifiers. Our main quantity of interest is:

Definition 2 (Price of Opacity)**.** *When Jury plays f and Contestant responds to $\widehat{f}$, the* ***Price of Opacity*** *(*POP*) equals to $\texttt{err}(f,\widehat{f}) - \texttt{err}(f,f)$.*

The price of opacity describes the relative loss in accuracy an optimistic Jury suffers by being opaque, i.e., holding f private. Note that POP implicitly depends on h. Our main results in Sec. 3.2 show that POP can be strictly positive, and in some cases, quite large.

Enlargement Set. The two terms in Def. 2 differ in the way Contestant plays—either using f or using $\widehat{f}$—and any difference in the errors can be attributed to discrepancies caused by this difference. This provides us a handle to trace the opaque strategic error and determine price of opacity. Define the *disagreement region* of f and $\widehat{f}$ as (see Fig. 1):

$$S := \{x \mid f(x) \neq \widehat{f}(x)\}. \tag{3}$$

Disagreement stems from errors in Contestant's estimation of f. Such errors become significant if they affect if and how points move (i.e., $\Delta_f(x)$ vs. $\Delta_{\widehat{f}}(x)$), as these differences can lead to classification errors for Contestant. We refer to the set of all such points as the *enlargement set*:

$$E := \{x \mid f(\Delta_f(x)) \neq f(\Delta_{\widehat{f}}(x))\} \tag{4}$$

The enlargement set E includes all points x which Jury classifies differently under the different strategies of Contestant (see Fig. 1). In particular, E tells us which points *prior to moving* will be disagreed on *after moving*. Because all moved points tend to concentrate around the classification boundary (since they maximize utility), we think of E as 'enlarging' S by including points that strategically moved to S. This entails a subtle but important point: even if $\widehat{f}$ is a good approximation of f and the mass on S is small, the mass of E, which includes points that *map to* S via Δ_f, can be large. In Sec. 3.2 we characterize the points in E.

Errors ϵ_1, ϵ_2 in Strategic Learning. Let ϵ_1, ϵ_2 be such that the learning error for Jury is: $\texttt{err}(f,f) = \texttt{err}(f^\star, f^\star) + \epsilon_1$, where $f^\star = \text{argmin}_{f\in\mathcal{H}} \texttt{err}(f,f)$ is the optimal strategy-aware classifier, and the learning error for the Contestant is: $\mathbb{P}_{x\in D}\{x \in S\} = \epsilon_2$. In the next sections we use ϵ_1 and ϵ_2 to reason about POP. These errors will tend to zero (with rates polynomial in the number of samples): for ϵ_1 when f is strategically-PAC learnable (see (Zhang & Conitzer, 2021; Sundaram et al., 2021)), and for ϵ_2 when $\widehat{f}$ is learnable in the standard, non-strategic PAC sense (recall that $\mathcal{H}_C = \mathcal{H}$).

Relations Between Errors. We note that the relations between $\texttt{err}(f,\widehat{f})$, $\texttt{err}(f,f)$, and $\texttt{err}(f^\star, f^\star)$ are not immediate. First, we observe that there can be instances in which $\texttt{err}(f,\widehat{f}) < \texttt{err}(f^\star, f^\star)$, i.e, the utility gained by a naïve and opaque Jury could be *higher* than that of the optimal transparent Jury. Second, we note that POP can be *negative* as there exist instances where $\texttt{err}(f,\widehat{f}) < \texttt{err}(f,f)$. We demonstrate this via experiments in App. B.

Intuitively, we expect that if n is large (ϵ_1 is small) and m is small (ϵ_2 can be large), then estimation errors of $\widehat{f}$ would harm Jury's ability to predict Contestant's moves, so much that it would lead to a strictly positive POP. This is indeed the intuition we build on for our next result which determines when POP > 0. Additionally, even if ϵ_2 is small, the probability on E can be large.

3.2. Main Results

We partition the points in E depending on how h behaves on it. This partition is used to first determine POP and then give a sufficient condition for positive POP (see Thm. 2). Next, in Thm. 3 we give an exact characterization of E in terms of S, f, $\widehat{f}$ (independent of h). Together, these results are useful in analysis, when one needs to assess whether the price of opacity exists under a reasonable distribution model D, as we demonstrate in Sec. 3.3.

Partition of E. The points in E can be partitioned as:

$$E^+ = \{x \mid h(x) = f(\Delta_f(x)) \ \wedge \ h(x) \neq f(\Delta_{\widehat{f}}(x))\};$$
$$E^- = \{x \mid h(x) \neq f(\Delta_f(x)) \ \wedge \ h(x) = f(\Delta_{\widehat{f}}(x))\}.$$

To make concrete the relation between POP, E and h define:

$$\text{POP}^+ := \mathbb{P}_{x\sim D}\{x \in E^+\}, \quad \text{POP}^- := \mathbb{P}_{x\sim D}\{x \in E^-\}.$$

Notice that, in Fig. 1, the points in region POP$^+$ increase the price of opacity, and the points in region POP$^-$ decrease the price of opacity. Further, note that since E^+ and E^- form a disjoint partition of E, we have

$$\mathbb{P}_{x\sim D}\{x \in E\} = \text{POP}^+ + \text{POP}^-. \qquad (5)$$

The next lemma follows from the definition of POP.

Lemma 1. POP $=$ POP$^+$ $-$ POP$^-$.

Note that since POP is a difference between two probabilities, its range is POP $\in [-1, 1]$.

We can now state our main result, proved using Lem. 1. The theorem provides a sufficient condition for POP > 0: if the probability mass on E exceeds the base error rates of f and $f^\star$, then POP is strictly positive. This provides a handle for reasoning about the price of opacity through the analysis of the enlargement set E. We remark that even though POP depends on the behaviour of h on E, it is interesting that the sufficiency condition in Thm. 2 depends only on f and $\widehat{f}$ which determine E. We also show in Sec. 3.3 (see Cor. 7) that there are instances where the sufficiency condition in the theorem below is indeed necessary.

Theorem 2. *The price of opacity is strictly positive, i.e.,* POP > 0*, if the following condition holds:*

$$\mathbb{P}_{x\sim D}\{x \in E\} > 2\texttt{err}(f^\star, f^\star) + 2\epsilon_1. \qquad (6)$$

The mass of the enlargement set is potentially large if the distribution D has a large mass on the "cost boundary" of f (see Fig. 1). Intuitively, enlargement can cause a "blow-up" in POP because both learning objectives—of f by Jury, and of $\widehat{f}$ by Contestant—are oblivious to the significance of this mass under strategic behavior.

Characterizing the Enlargement Set E. We identify the points in the enlargement set E when only f and $\widehat{f}$ are known without invoking an explicit dependence on the (often unknown) ground truth function h. First partition S as:

$$S_{-1,1} = \{x \mid f(x) = -1 \wedge \widehat{f}(x) = 1\};$$
$$S_{1,-1} = \{x \mid f(x) = 1 \wedge \widehat{f}(x) = -1\}.$$

Define the following sets derived from $S_{-1,1}$ and $S_{1,-1}$:

$$E_{-1,1} = \{x \mid \Delta_{\widehat{f}}(x) \in S_{-1,1}, \ f(\Delta_f(x)) = 1\};$$
$$E_{1,-1} = \{x \mid \Delta_f(x) \in S_{1,-1} \setminus \{x\}, \ \Delta_{\widehat{f}}(x) = x\}.$$

In Thm. 3 (proof in App. A) we show that $E_{-1,1}$ and $E_{1,-1}$ partitions E. We remark that this partition is different from E^+ and E^- as defined before (see Fig. 1).

Theorem 3. *The enlargement set can be partitioned as* $E = E_{-1,1} \uplus E_{1,-1}$.

One consequence of Thm. 3 is that while a transparent Jury would have classified users in E positively, an opaque Jury classifies them negatively. This disproportionately affects qualified users.

Corollary 4. *For all $x \in E$, it holds that $f(\Delta_f(x)) = 1$ and $f(\Delta_{\widehat{f}}(x)) = -1$.*

3.3. Price of Opacity for Linear Classifiers

In this section, we use results from Sec. 3.2 to exemplify how the strategic behaviour by Contestant according to her learnt classifier $\widehat{f}$ (instead of f) can harm Jury and lead to positive POP. Throughout this section, we make the simplifying assumption that $d = 1$ and that both Jury and Contestant play linear (threshold) classifiers, so as to make the computations simpler to follow and highlight Contestant's strategic behaviour. Our main findings on POP for linear classifiers, Gaussian distributions over $\mathcal{X}$, and a large class of cost functions (as in Def. 3) are the following. First, using Thm. 3 and Cor. 4 we determine E (applying knowledge of f and $\widehat{f}$), and use Thm. 2 to claim that POP > 0 is inevitable (for any h) if the probability mass on E is large (Prop. 5). In Prop. 6 we show the realizablility of h enables us to use Lem. 1 to determine the expression for POP in this setting.[5] Next, we show that the sufficiency condition in Thm. 2 is tight by giving a class of h (realizable) for which this condition is also necessary (Cor. 7). This establishes that without further assumptions on h it is impossible to derive a better condition for POP > 0. Finally in Cor. 8, with explicit knowledge of h we show POP can be made arbitrarily large (i.e., close to 1) with an appropriate choice of distribution parameters.

Our results here apply to cost functions that are *instance-invariant* (Sundaram et al., 2021),[6] with an additional feasibility requirement which ensures that each user has a feasible modification. We refer to these as *admissible costs*.

Definition 3 (Admissible Cost)**.** *A cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is* admissible *if:*
(i) $c(x, x+\delta) = c(y, y+\delta)$ for all $x, y, \delta \in \mathcal{X}$, and
(ii) for all $x \in \mathcal{X}$ there exists $\delta > 0$ s.t. $c(x, x+\delta) \leq 2$

[5] Note that such knowledge regarding h is necessary as POP depends on the behaviour of h on E.

[6] Our focus on instance-invariant costs is due to a result from Sundaram et al. (2021) showing that *non*-invaraint costs (referred to as*instance-wise* costs) are intractable even in simple cases. For example, the Strategic-VC of linear classifiers is linear for instance-invariant costs but unbounded for instance-wise costs.

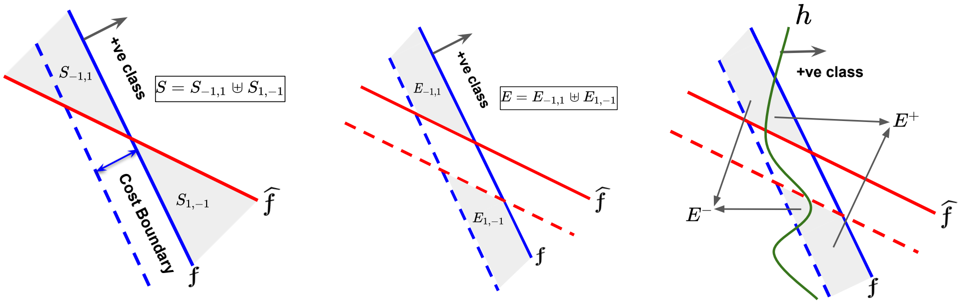

Figure 1. **(Left)** The disagreement set S between f and $\widehat{f}$, and the cost boundary of f. **(Center)** The enlargement set E derived from S. **(Right)** The partition E^+ and E^- of E. Note the following: a) E is a subset of the cost boundary of f, and b) E^+ and E^- depend on h.

In the above, condition (i) is equivalent to instance-invariance, and condition (ii) ensures feasibility.

Let c be an admissible cost function c, and define $t := \sup_{\delta>0}\{c(x, x+\delta) \le 2\}$. Let $D = \mathcal{N}(\alpha, \sigma)$ be a Gaussian distribution over $\mathcal{X}$ with $\Phi_\sigma(\cdot)$ as its CDF. The thresholds corresponding to Jury and Contestant are t_f and $t_{\widehat{f}}$ respectively, i.e., Jury's classifier is $f(x) = \text{sign}(x \ge t_f)$ and Contestant's classifier is $\widehat{f}(x) = \text{sign}(x \ge t_{\widehat{f}})$. We assume without loss of generality that $\epsilon_2 > 0$ and $t_f \ne t_{\widehat{f}}$. In Prop. 6 and Cor. 7, when we say h is realizable, we assume that $h(x) = 1$ for $x \ge \alpha$ and -1 otherwise. We now present the first result which gives the sufficient condition for POP > 0. In Prop. 5 let $\ell := 2 \cdot \texttt{err}(f^\star, f^\star) + 2\epsilon_1$.

Proposition 5. *The following are sufficient for* POP > 0*:*
(a) $\Phi_\sigma(t_f) - \Phi_\sigma(t_f - t) > \ell$ *when* $t_f > t_{\widehat{f}}$
(b) $\Phi_\sigma(t_{\widehat{f}} - t) - \Phi_\sigma(t_f - t) > \ell$ *when* $t_f < t_{\widehat{f}}$

Note that $t_f > t_{\widehat{f}} \Rightarrow E = [t_f - t, t_f]$ and $t_f < t_{\widehat{f}} \Rightarrow E = [t_f - t, t_{\widehat{f}} - t]$, and that if h is realizable, then $\texttt{err}(f^\star, f^\star) = 0$. Using this fact and assuming a reasonable relation between ϵ_2 and ϵ_1 we show there exists σ_0 such that POP > 0 for all $\sigma < \sigma_0$, and that POP can be made arbitrarily close to 1 by reducing σ.

Proposition 6. *Let h be realizable and $\epsilon_2 > 2\epsilon_1$. Then, there exists $\sigma_0 > 0$ such that for all $\sigma < \sigma_0$ it holds that:*

$$\text{POP} = \begin{cases} \Phi_\sigma(t_{\widehat{f}} - t) - \Phi_\sigma(|t - t_f|) & \text{if } t_f < t_{\widehat{f}} \\ \Phi_\sigma(t_f) - \Phi_\sigma(|t - t_f|) & \text{if } t_f > t_{\widehat{f}} \end{cases} \tag{7}$$

We use this to demonstrate instances where the sufficiency condition in Thm. 2 is also necessary.

Corollary 7. *Suppose h is realizable, $t_f < t$, and $\epsilon_2 > 2\epsilon_1$. Then there exists $\sigma_0 > 0$ such that for all $\sigma < \sigma_0$,* POP > 0 *if and only if* $\mathbb{P}_{x\sim D}\{E\} > 2\epsilon_1$.

Recall, ϵ_2 is the probability mass of the disagreement set, and intuitively, ϵ_2 is large if m is small. Moreover, the strategic VC dimension of linear classifiers in $\mathbb{R}^d$ is at most $d+1$ for admissible cost functions (Sundaram et al. (2021)) and hence threshold classifiers are strategic PAC learnable. In particular, with probability at least $1-\delta$ the following holds, which we use for Cor. 8: $\epsilon_1 \le \sqrt{\log\frac{4}{\delta}/n}$.

Corollary 8. *Suppose $\epsilon_2 > 2\sqrt{\log\frac{4}{\delta}/n}$, where $\delta \in (0,1)$. Then with probability at least $1-\delta$,* POP *is as in Eq. (7).*

Discussion. As a take-away from our analysis in this section, say the number of Contestant's samples m is fixed, and the number of Jury's samples n grows. Then the following trade-off occurs: On one hand, as Jury acquires more training samples, his strategic error $\texttt{err}(f, f)$ decreases (since $\epsilon_1 \le \sqrt{\log\frac{4}{\delta}/n}$). On the other hand, as n grows, the sufficient condition on ϵ_2 becomes weaker (so positive POP occurs more), and the expression for POP grows. So increasing the relative gap between n and m is likely to increase POP, but discarding some of Jury's training points to decrease POP is likely to increase the strategic error.

4. Experiments

In this section we complement our theoretical results with an experimental evaluation of POP. We begin with a synthetic experiment that validates our results from Sec. 3.2 and 3.3. We then proceed to analyzing a real dataset of loan applications in the peer-to-peer lending platform Prosper (`http://www.prosper.com`). This dataset is unique in that it includes: (i) logged records of loan applications, (ii) corresponding system-evaluated risk scores, and (iii) a social network connecting users of the platform, with links portraying social (rather than financial) relations (see Krumme & Herrero, 2009). We use loan details as user features and risk scores as labels to study POP behavior, and social network connections to perform an in-depth analysis of the effects of an opaque policy. Code is publicly available at `https://github.com/staretgicclfdark/strategic_rep`.

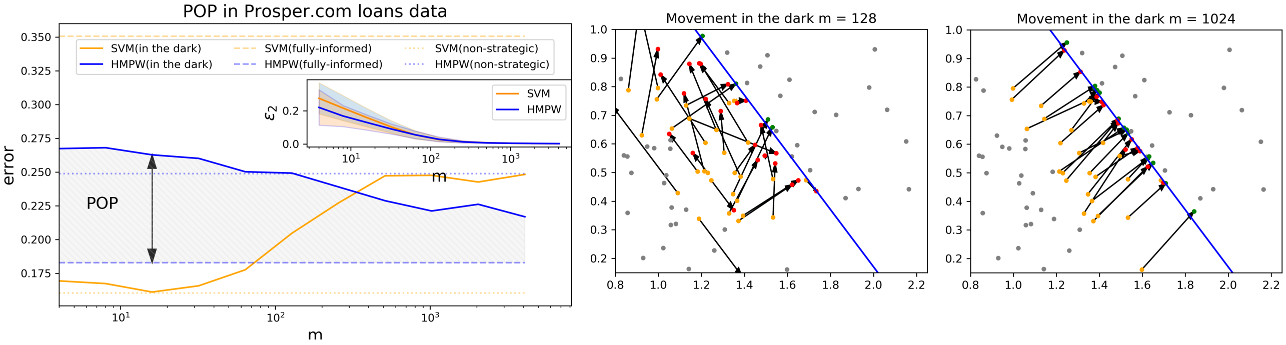

Figure 2. **(Left)** Prediction errors on loans data for three user types: *non-strategic*, *fully-informed*, and *in the dark*. POP decreases as m increases but remains considerable even for large m. **(Center+Right)** Illustrations of feature modifications for $m = 128$ and $m = 1024$, projected to 2D. Blue line is the (projected) classifier threshold, with points to its left classified as positive. Features that were not modified are in grey, and features that were modified are in orange, with arrows mapping to the modified features, with green and red indicating positive and negative outcomes, respectively.

Experimental Setting. In all experiments we use the algorithm of Hardt et al. (2016) as the strategic learner of Jury and consider settings for which this algorithm provides learning guarantees (i.e., a separable cost function and deterministic labeling). Our main experiments model Contestant as inferring $\widehat{f}$ using a simple ERM approach (in practice, SVM). We explore other risk-averse strategies for contestant in Appendix B.3. Although our theoretical analysis considered a single $\widehat{f}$ for Contestant, as we note, our results hold in expectation for the more realistic case in which each x is associated with a user, and each such user has access to her own sample set $T_C(x)$ of size m, on which she trains an individualized $\widehat{f}_x$. This is the setting we pursue in our experiments.

Because we will consider very small values for m, there is a non-negligible chance that $T_C(x)$ will include only one class of labels. A plausible modeling assumption in this case is to assert that points with one-sided information do not move. However, to prevent the attribution of POP to this behavior, we instead choose to ensure that each of Contestant's sample sets includes at least one example from each class, for which we use rejection sampling. App. B includes additional information on our setup.

4.1. Split Gaussian

Following the setup in Sec. 3.3, we generate data using $D = \mathcal{N}(0, 1)$ and $h(x) = \text{sign}(x \geq 0)$. We set $c(x, x') = \max\{0, x' - x\}$. We first sample $n = 5000$ labeled points and split them 80-20 into a train set T and a held-out test set S, and use T to train Jury's classifier f. For every $x \in S$, we repeatedly sample an additional m points labeled by f to construct training sets $T_C(x)$, and use these to train $\widehat{f}_x$ for every $x \in S$ to be used by Contestant. We repeat this for $m \in [4, 4096]$. Fig. 3 shows how POP varies with m (log scale), along with corresponding estimation error of $\widehat{f}$ w.r.t. f (i.e., ϵ_2). For small m the price of opacity is extremely large (~ 0.5). As m increases, POP decreases, but quickly plateaus at ~ 0.3 for $m \approx 100$. Note that the estimation error of f is small even for small m, and becomes (presumably) negligible at $m \approx 30$, before POP plateaus.

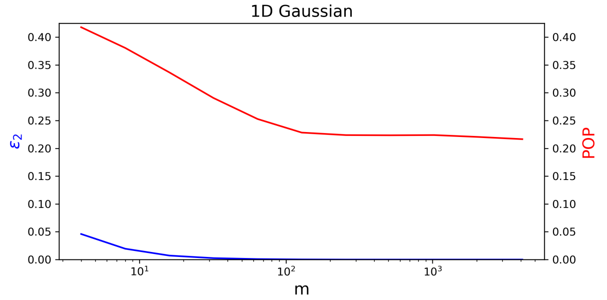

Figure 3. POP and estimation error for $\widehat{f}$ (ϵ_2) on a split 1D Gaussian. POP is large even when $\widehat{f}$ closely estimates f.

4.2. Loan Data

We now turn to studying the Prosper loans dataset. We begin by analyzing how POP varies as m increases, sampling uniformly from the data. The data includes $n = 20,222$ examples, which we partition $70-15-15$ into three sets: a training set T for Jury, a held-out test-set S, and a pool of samples from which we sample points for each $T_C(x), x \in S$. We set labels according to (binarized) system-provided risk scores, and focus on six features that are meaningful and that are amenable to modification: available credit, amount to loan, % trades not delinquent, bank card utilization, total number of credit inquiries, credit history length. A simple non-strategic linear baseline using these features achieves 84% accuracy on non-strategic data. We compare performance of the algorithm of Hardt et al. (2016) (`HMPW`) to a non-strategic linear SVM baseline. Both models are eval-

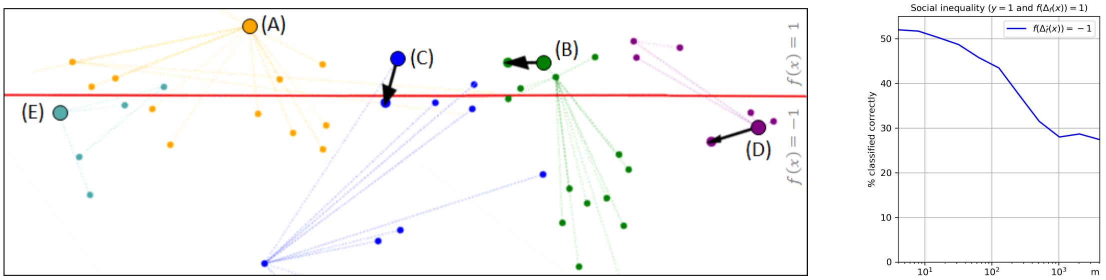

Figure 4. **(Left)** 'Stories' describing outcomes for individual users in the Prosper.com social network. Users are positioned according to their features, with points above the line approved for loan. Each user is shown along with her network neighbors, from whom she learns $\widehat{f}$. Arrows depict if and how features have been modified. **(Right)** Inequity in wrongful loan denial caused by an opaque policy.

uated under three Contestant types: *Non-Strategic*, where points do note move; *Fully-Informed*, where Contestant plays Δ_f; and *In the Dark*, where Contestant plays $\Delta_{\widehat{f}}$.

Price of Opacity. Fig. 2 (left) shows how predictive performance varies for increasing m. POP is the difference between lines corresponding to the In the Dark and Fully-Informed types (shaded area for `HMPW`). In the fully-informed case (Δ_f), the performance of `HMPW` closely matches the benchmark of SVM on non-strategic data, showing that in this case, by correctly anticipating user behavior, `HMPW` is able to a account for gaming in learning. However, when Contestant is in the dark ($\Delta_{\widehat{f}}$), for small m the performance of `HMPW` is as bad as when Contestant does not move at all, and the price of opacity is large ($\sim 9\%$). As m increases, Contestant better estimates $\widehat{f}$, and the performance of `HMPW` increases. However, even for very large m (e.g., $m = 4096$), POP remains substantial ($\sim 4\%$).

Fig. 2 (center; right) visualizes how points move under `HMPW` when Contestant is *In the Dark*, for medium-sized sample sets ($m = 128$) and large sample sets ($m = 1024$). For visualization we project points down to $\mathbb{R}^2$ in a way that aligns with the predictive model f, ensuring that points are classified as positive iff they appearing above the line. The figures demonstrate which points move and where. For $m = 128$, some points move in the right direction, others move in seemingly arbitrary directions. For $m = 1024$, most points move in the right direction, but in many cases fall short of the classification boundary.

Opacity and Social Inequity. Our analysis thus far has been focused on the effects of an opaque policy to the payoff of Jury. But an opaque policy primarily keeps *users* in the dark, and here we analyze the effects of opacity on the payoff to users. When Jury anticipates strategic behavior, he must 'raise the bar' for positive classification to prevent negative-labeled points from gaming. But due to this, positive-labeled points must also make effort to be classified as such. Ideally, Jury can set up f in a way that positive-labeled users are classified appropriately if they invest effort correctly. This, however, may require exact knowledge of f, and when positive-labeled users are in the dark, the lack of access to f may cause them to be classified negatively despite their effort in modification (see Cor. 4).

Fig. 4 (right) shows for increasing m the percentage of positive-labeled users that are classified correctly under full information (i.e., $f(\Delta_f(x)) = 1$), but are classified incorrectly when in the dark (i.e., $f(\Delta_{\widehat{f}}(x)) = -1$); this is precisely the positive enlargement set E^+ in Cor. 4. As can be seen, for low values of m, roughly 50% of these users are classified incorrectly. When considering a heterogeneous population of users varying in the amount of information they have access to (i.e., varying in m), our results indicate a clear form of inequity towards individuals who are not well-connected. This is a direct consequence of the opaque policy employed by Jury.

Stories from the Network. We now dive into the Prosper.com social network and portray stories describing individual users and how their position within the network affects how they are ultimately classified—and hence whether their loan is approved or denied—under an opaque policy. Fig. 4 (left) presents five such cases. The figure shows for each user her initial features x and her modified features $\Delta_{\widehat{f}}(x)$ (if they differ). We assume users observe the features and predictions of their neighbors in the network up to two hops away, which they use to learn $\widehat{f}$. As before, points are projected down to $\mathbb{R}^2$, so that points above the line are approved and below it are denied (to reduce cluttering, projections slightly vary across users). The plot also shows network connections between users (superimposed).

User A is approved a loan. She is classified for approval on her true features. She is well-connected and has both neighbors that have been denied and approved for loans, and her estimated $\widehat{f}$ is good. She therefore correctly infers her predicted approval and does not modify her features.

User B is also approved a loan, but at an unnecessary cost. She would have been approved on her true features, but most of her neighbors were denied loans, leading to errors in estimating $\widehat{f}$. These have led her to falsely believe that she must modify her features to be approved. The cost of modification would have been eliminated had she known f.

User C would have been approved for a loan—had she reported her true features. However, all but one of her neighbors were denied loans. This caused a bias in her estimate of $\widehat{f}$, so large that modifying her features on this account resulted in her loan being denied (unjustifiably).

User D is slightly below the threshold for approval. With the correct f, she would have been able to modify her features to receive approval. However, despite having neighbors who have been approved, her estimated $\widehat{f}$ is not exact. As a result, even after modifying her features she remains to be denied (despite believing she will be approved).

User E would not have been approved. But approval is within her reach, and had she known f, the cost of modification would have been small enough to make it worthwhile. However, she is not well connected, and errors in her estimated $\widehat{f}$ have led her to (falsely) believe that modifications for approval are too costly, and so she remains denied.

5. All-Powerful Jury

In this section, we study the setting where Jury *exactly knows* Contestant's response to his classifier f, despite f's opacity. As mentioned in Sec. 2, this setting is less realistic, but we study it to show that this knowledge is precisely what Jury is missing to successfully learn a strategic classifier when users are in the dark. We introduce the notion of a *response function* for the Contestant, denoted by $R : \mathcal{H} \to \mathcal{H}$. When Jury plays f, Contestant responds by moving the points with respect to classifier $R(f)$. For example, in the setting of Jury in the dark (Sec. 3), $R(f) = \widehat{f}$ and R is unknown to Jury. In contrast, an all-powerful Jury is assumed to know the response function R. Note that, when R is the identity map ($R(f) = f$), the all-powerful jury setting reduces to the strategic classification setting.

We now give an agnostic PAC learnability result for R-strategic classification. This generalizes a recent result of Zhang & Conitzer (2021); Sundaram et al. (2021), who defined the strategic ERM rule and strategic VC dimension, using similar techniques. Recall that $\texttt{err}(f, R(f))$ denotes the error when Jury plays the classifier f and Contestant responds to $R(f)$. In this setting, define R-optimal classifier for Jury is $f^\star_R = \operatorname{argmin}_{f\in\mathcal{H}} \text{err}(f, R(f))$. We drop the subscript R from $f^\star_R$, when it is immediate from context.

First, we introduce R-strategic ERM rule which minimizes the training error with an explicit knowledge of Contestants response function R. Given a training set $T_J = \{x_i, h(x_i)\}_{i\in[n]}$ define $\widehat{\texttt{err}}(f, R(f)) = 1/n \sum_{i\in[n]} \mathbb{1}\{f(\Delta_{R(f)}(x)) \neq h(x)\}$. The R-strategic ERM returns $\operatorname{argmin}_{f\in\mathcal{H}}\ \widehat{\texttt{err}}(f, R(f))$.

Definition 4 (R-strategic VC dimension). *Let $\mathcal{H}$ be a hypothesis class, and let $\mathcal{H}' = \{f' | \exists f \in \mathcal{H}$ such that $f'(x) = f(\Delta_{R(f)}(x))\}$. Then $\text{SVC}(\mathcal{H}) = \text{VC}(\mathcal{H}')$.*

Proposition 9. *For a given hypothesis class $\mathcal{H}$, Jury with sample set T_J containing n iid samples, using R-strategic ERM rules computes an f such that with probability $1-\delta$ the following holds: $\texttt{err}(f, R(f)) \leq \texttt{err}(f^\star, R(f^\star)) + \epsilon$, where $\epsilon \leq \sqrt{C(d\log(d/\epsilon) + \log(\frac{1}{\delta})/n}$, C is an absolute constant, and d is the R-strategic VC dimension of $\mathcal{H}$.*

Proof. For each f, let f' be such that $f'(x) = f(\Delta_{R(f)}(x))$. It is easy to see that $\mathbb{P}_{x\sim D}\{h(x) \neq f(\Delta_{R(f)}(x))\} = \mathbb{P}_{x\sim D}\{h(x) \neq f'(x)\}$. Hence, the agnostic-PAC learnability in the all powerful Jury setting can be given in terms of the VC dimension of $\mathcal{H}' = \{f' | \exists f \in \mathcal{H}$ such that $f'(x) = f(\Delta_{R(f)}(x))\}$. It follows from Def. 4 that R-strategic VC of $\mathcal{H}$ is equal to the VC dimension of $\mathcal{H}'$. The proof follows by noting that the bounds in the theorem statement are the agnostic-PAC generalization bounds for standard classification setting with the VC dimension of the hypothesis class replaced by R-strategic VC dimension of $\mathcal{H}$. □

6. Discussion

The theoretical and empirical findings in this paper show that the "optimistic firm" is wrong in hoping that small errors in estimating f by users in the dark will remain small; as we demonstrate, small estimation errors can grow into large errors in the performance of f. While there is a much-debated public discussion regarding the legal and ethical right of users to understand the reasoning behind algorithmic decision-making, in practice firms have not necessarily been propelled so far to publicize their models. Our work provides formal incentives for firms to adopt transparency as a policy. Our experimental findings provide empirical support for these claims, albeit in a simplified setting, and we view the study of more elaborate user strategies (e.g., Bayesian models, bounded-rational models, or more nuanced risk-averse behavior) as interesting future work. Finally, our work highlights an interesting open direction in the study of strategic classification—to determine the optimal policy of a firm who wishes to maintain (a level of) opaqueness alongside robustness to strategic users.

Acknowledgement: GG is thankful for the financial support from the Israeli Ministry of Science and Technology grant 19400214. VN is thankful to be supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 682203 -ERC-[Inf-Speed-Tradeoff]. This research was supported by the Israel Science Foundation (grant 336/18). Finally, we thank anonymous reviewers for their helpful comments.

References

Ahmadi, S., Beyhaghi, H., Blum, A., and Naggita, K. The strategic perceptron. Available at https://arxiv.org/abs/2008.01710, 2020.

Alon, T., Dobson, M., Procaccia, A. D., Talgam-Cohen, I., and Tucker-Foltz, J. Multiagent evaluation mechanisms. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pp. 1774–1781, 2020.

Bechavod, Y., Ligett, K., Wu, S., and Ziani, J. Gaming helps! learning from strategic interactions in natural dynamics. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1234–1242. PMLR, 13–15 Apr 2021a.

Bechavod, Y., Podimata, C., Wu, Z. S., and Ziani, J. Information discrepancy in strategic learning. Available at https://arxiv.org/pdf/2103.01028.pdf, 2021b.

Braverman, M. and Garg, S. The role of randomness and noise in strategic classification. In *1st Symposium on Foundations of Responsible Computing (FORC)*, pp. 9:1–9:20, 2020.

Brückner, M. and Scheffer, T. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 547–555, 2011.

Chen, Y., Liu, Y., and Podimata, C. Learning strategy-aware linear classifiers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC)*, pp. 55–70, 2018.

Haghtalab, N., Immorlica, N., Lucier, B., and Wang, J. Z. Maximizing welfare with incentive-aware evaluation mechanisms. In Bessiere, C. (ed.), *The Twenty-Ninth International Joint Conference on Artificial Intelligence IJCAI*, pp. 160–166, 2020.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, pp. 111–122, 2016.

Hu, L., Immorlica, N., and Vaughan, J. W. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 259–268, 2019.

Kleinberg, J. M. and Raghavan, M. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation, (EC)*, pp. 825–844, 2019.

Krumme, K. A. and Herrero, S. Lending behavior and community structure in an online peer-to-peer economic network. In *2009 International Conference on Computational Science and Engineering*, volume 4, pp. 613–618. IEEE, 2009.

Kumar, A. and Zinger, M. Network analysis of peer-to-peer lending networks cs224w project final paper. 2014.

Levanon, S. and Rosenfeld, N. Strategic classification made practical. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Milli, S., Miller, J., Dragan, A. D., and Hardt, M. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 230–239, 2019.

Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Rosenfeld, N., Hilgard, A., Ravindranath, S. S., and Parkes, D. C. From predictions to decisions: Using lookahead regularization. *Advances in Neural Information Processing Systems*, 33, 2020.

Sundaram, R., Vullikanti, A., Xu, H., and Yao, F. PAC-learning for strategic classification. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Zhang, H. and Conitzer, V. Incentive-aware pac learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.