Performative Power

Moritz Hardt*
Max-Planck Institute for Intelligent Systems, Tübingen
hardt@is.mpg.de

Meena Jagadeesan*
UC Berkeley
mjagadeesan@berkeley.edu

Celestine Mendler-Dünner*
Max-Planck Institute for Intelligent Systems, Tübingen
cmendler@tuebingen.mpg.de

Abstract

We introduce the notion of performative power, which measures the ability of a firm operating an algorithmic system, such as a digital content recommendation platform, to cause change in a population of participants. We relate performative power to the economic study of competition in digital economies. Traditional economic concepts struggle with identifying anti-competitive patterns in digital platforms not least due to the complexity of market definition. In contrast, performative power is a causal notion that is identifiable with minimal knowledge of the market, its internals, participants, products, or prices. Low performative power implies that a firm can do no better than to optimize their objective on current data. In contrast, firms of high performative power stand to benefit from steering the population towards more profitable behavior. We confirm in a simple theoretical model that monopolies maximize performative power. A firm's ability to personalize increases performative power, while competition and outside options decrease performative power. On the empirical side, we propose an observational causal design to identify performative power from discontinuities in how digital platforms display content. This allows to repurpose causal effects from various studies about digital platforms as lower bounds on performative power. Finally, we speculate about the role that performative power might play in competition policy and antitrust enforcement in digital marketplaces.

1 Introduction

Digital platforms pose a well-recognized challenge for antitrust enforcement. Traditional market definitions, along with associated notions of competition and market power, map poorly onto digital platforms. A core challenge is the difficulty of precisely modeling the interactions between the market participants, products, and prices. An authoritative report, published by the Stigler Committee [2019], details the many challenges associated with digital platforms, among them: *"Pinpointing the locus of competition can also be challenging because the markets are multisided and often ones with which economists and lawyers have little experience. This complexity can make market definition another hurdle to effective enforcement."* Published the same year, a comprehensive report from the European Commission calls for *"less emphasis on analysis of market definition, and more emphasis on theories of harm and identification of anti-competitive strategies."* [Crémer et al., 2019]

Our work responds to this call by developing a normative and technical proposal for reasoning about power in digital economies, while relaxing the reliance on market definition. Our running example is a digital content recommendation platform. The platform connects content creators with viewers, while monetizing views through digital advertisement. Key to the business strategy of a

*Authors listed in alphabetical order.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

firm operating a digital content recommendation platform is its ability to predict revenue for content that it recommends. Often framed as a supervised learning task, the firm trains a statistical model on observed data to predict some proxy of revenue, such as clicks, views, or engagement. Better predictions enable the firm to more accurately identify content of interest and thus increase profit.

A second way of increasing profit is more subtle. The platform can use its predictions to steer participants towards modes of consumption and production that are easier to predict and monetize. For example, the platform could reward consistency in the videos created by content creators, so that the popularity of their videos becomes more predictable. Similarly, the platform could recommend addictive content to viewers, appealing to behavioral weaknesses to drive up viewer engagement. How potent such a strategy is depends on the extent to which the firm is able to steer participants, which we argue reveals a salient power relationship between the platform and its participants.

1.1 Our contribution

We introduce the notion of *performative power* that quantifies a firm's ability to steer a population of participants. We argue that the sensitivity of participant behavior to algorithmic changes in the platform provides an important indicator of the firm's power. Performative power is a causal statistical notion that directly quantifies how much participants change in response to actions by the platform, such as updating a predictive model. In doing so it avoids market specifics, such as the number of firms involved, products, and monetary prices. Neither does it require a competitive equilibrium notion as a reference point. Instead, it focuses on where rubber meets the road: the algorithmic actions of the platform and their causal powers.

We build on recent developments in performative prediction [Perdomo et al., 2020] to articulate the difference between learning and steering in prediction. We show that under low performative power, a firm cannot do better than standard supervised learning on observed data. Intuitively, this means the firm optimizes its loss function *ex-ante* on data it observes. We interpret this optimization strategy as analogous to the firm being a price-taker, an economic condition that arises under perfect competition in classical market models. We contrast this optimization strategy with a firm that performs *ex-post* optimization and takes advantage of performative power to achieve lower expected risk.

To better understand our definition, we study performative power in the concrete algorithmic market model of strategic classification. Strategic classification models participants as best-responding agents that change their features rationally in response to a predictor to achieve a better prediction outcome. We study the role of different economic factors by extending the standard model to incorporate competing firms and outside options. Our first observation is that a *monopoly* firm can derive significant performative power because participants are willing to incur a cost up to the utility of using the service in order to adjust to the firm's predictor. Moreover, performative power is maximized if the firm can personalize decisions to individual users. Our second observation is that performative power decreases in the presence of *competition* and *outside options*. When firms compete for participants, offering services that are perfect substitutes for each other, then even two firms can lead to zero performative power. This result stands in analogy with classical Bertrand competition.

On the empirical side, we propose a causal design to identify performative power in the context of a recommender system arranging content into display slots. This design, we call *discrete display design (DDD)*, establishes a connection between performative power and the causal effect of display position on consumption. To derive a lower bound on performative power, DDD constructs a hypothetical algorithmic action that aggregates the causal effects of display position across the population. This allows us to repurpose reported causal effects of display position as lower bounds on performative power. It also charts out a concrete empirical strategy for measuring power in digital economies.

Finally, we examine the role of performative power in competition policy. We contrast performative power with traditional measures of market power, describe how performative power can capture complex behavioral patterns, and discuss the potential role of performative power in antitrust debates.

1.2 Related work

Our notion of performative power builds on the development of performativity in prediction by Perdomo et al. [2020]. Performativity captures that the predictor can influence the data-generating process, a dependency ruled out by the traditional theory of supervised learning. A growing line of work on performative prediction, e.g., [Mendler-Dünner et al., 2020; Drusvyatskiy and Xiao,

2022; Izzo et al., 2021; Dong and Ratliff, 2021; Miller et al., 2021; Brown et al., 2022; Li and Wai, 2021; Ray et al., 2022; Jagadeesan et al., 2022; Wood et al., 2022], has studied different optimization challenges and solution concepts in performative prediction. Rather than viewing performative effects as a challenge for the learning algorithm, we argue that performativity reveals a salient power relationship between the decision maker and the population. From an optimization perspective, our work demonstrates that sufficiently high performative power is necessary for performative optimization approaches to be beneficial compared with standard supervised learning.

The *strategic classification* setup we use for our case study was proposed in [Brückner et al., 2012; Hardt et al., 2016] and is closely related to a line of work in the economics community [Frankel and Kartik, 2022; Ball, 2020; Hennessy and Goodhart, 2020; Frankel and Kartik, 2019]. A line of work on strategic classification makes the assumption that performative effects are the result of individuals manipulating their features so as to best respond to the deployment of a predictive model. The focus has been on describing participant behavior in response to a single firm acting in isolation. Our extensions incorporate additional market factors into the model, such as outside options or the choice between competing firms, which we believe are helpful for gaining a better understanding of interactions in digital economies. Beyond the case of a single classifier, recently, Narang et al. [2022] and Piliouras and Yu [2022] consider multiple firms simultaneously applying retraining algorithms in performative environments and analyze convergence to equilibrium solutions. Ginart et al. [2021] study another model of feedback loops arising from competition between machine learning models.

There is extensive literature on the topic of competition on digital platforms that we do not attempt to survey here. For starting points, see, for example, recent work by Bergemann and Bonatti [2022], a survey by Calvano and Polo [2021], a discussion by Parker et al. [2020], the reports already mentioned [Stigler Committee, 2019; Crémer et al., 2019], as well as a macroeconomic perspective on the topic [Syverson, 2019].

2 Performative power

Fix a set $\mathcal{U}$ of participants interacting with a designated firm, where each $u \in \mathcal{U}$ is associated with a data point $z(u)$. Fix a metric $\mathrm{dist}(z, z')$ over the space of data points. Let $\mathcal{F}$ denote the set of actions a firm can take. We think of an action $f \in \mathcal{F}$ as a predictor that the firm can deploy at a fixed point in time. For each participant $u \in \mathcal{U}$ and action $f \in \mathcal{F}$, we denote by $z_f(u)$ the potential outcome random variable representing the counterfactual data of participant u if the firm were to take action f.

Definition 1 (Performative Power)**.** *Given a population $\mathcal{U}$, an action set $\mathcal{F}$, potential outcome pairs $(z(u), z_f(u))$ for each unit $u \in \mathcal{U}$ and action $f \in \mathcal{F}$, and a metric* dist *over the space of data points, we define the* performative power *of the firm as*

$$\mathrm{P} := \sup_{f \in \mathcal{F}} \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}\left[\mathrm{dist}\left(z(u), z_f(u)\right)\right] ,$$

where the expectation is over the randomness in the potential outcomes.

The expression inside the supremum generalizes an average treatment effect, corresponding to scalar valued potential outcomes and the absolute value as metric. We could generalize other causal quantities such as heterogeneous treatment effects, but this avenue is not subject of our paper. The definition takes a supremum over possible actions a firm can take at a specific point in time. We can therefore lower bound performative power by estimating the causal effect of any given action $f \in \mathcal{F}$.

Having specified the sets $\mathcal{F}$ and $\mathcal{U}$, estimating performative power amounts to causal inference involving the potential outcome variables $z_f(u)$ for unit $u \in \mathcal{U}$ and action $f \in \mathcal{F}$. In an observational design, an investigator is able to identify performative power without an experimental intervention on the platform. We propose and apply one such observational design in Section 5. In an experimental design, the investigator deploys a suitably chosen action to estimate the effect. Neither route requires understanding the specifics of the market in which the firm operates. It is not even necessary to know the firm's objective function, how it optimizes its objective, and whether it successfully achieves its objective. In practice, the dynamic process that generates the potential outcome $z_f(u)$ may be highly complex, but this complexity does not enter the definition. Consequently, the definition applies to complex multisided digital economies that defy mathematical specification. To make this abstract concept of performative power more concrete, let us instantiate it in a concrete example.

Example: Digital content recommendation. Consider a digital content recommendation platform, such as the video sharing services YouTube or Twitch. The platform aims to recommend channels that generate high revenue, personalized to each viewer. Towards this goal, the platform collects data to build a predictor f for the value of a channel c to a viewer with preferences p. Let $x = (x_c, x_p)$ be the features used for the prediction task that capture attributes x_c of the channel and the attributes x_p of the viewer preferences. Let y be the target variable, such as *watch time*, that acts as a proxy for the monetary value of showing a channel to a specific viewer. For concreteness, take the supervised learning loss $\ell(f(x), y)$ incurred by a predictor f to be the squared loss $(f(x) - y)^2$.

When defining performative power, participants could either be viewers or content creators. The definition is flexible and applies to both. By selecting the units $\mathcal{U}$, which features to include in the data point z, and how to specify the distance metric dist, we can pinpoint specific power relationships.

Content creators. The predictor f can affect the type of videos that content creators stream on their channels. For example, content creators might strategically adjust various features of their content relevant for the predicted outcome, such as the length, type or description of their videos, to improve their ranking. Thus, by changing how it predicts the monetary value of a channel, the platform can induce changes in the content on the channel. To measure this source of power, we let the participants $\mathcal{U}$ be content creators and suppose that each content creator $u \in \mathcal{U}$ maintains a channel of videos. Let the data point $z(u)$ correspond to features x_c characterizing the channel c created by content creator u. Let dist be a metric over features of content. The resulting instantiation of performative power measures the changes in content induced by potential implementations $\mathcal{F}$ of the prediction function. In Section 4, we investigate this form of performative power from a theoretical perspective by building on the setup of *strategic classification*.

Viewers. The predictor f can shape the consumption patterns of viewers, since viewers tend to follow recommendations when deciding what content to consume (e.g. [Ursu, 2018]). Thus, by changing which content it recommends to a user, the platform can induce changes in the target variable: how much time the users spends watching content on a given channel. Let's suppose that we wish to investigate the effect of the predictor on viewer consumption of a certain genre of content (e.g. radical content). To formalize this source of power, we let the units $\mathcal{U}$ be viewers. Let the data point $z(u)$ correspond to how long the viewer u spends watching content in the genre of interest. More formally, let $z(u)$ for a user u with preferences p be equal to the cumulative watch time over pairs (x_c, x_p) where c is a channel within the genre of interest. Let $\mathrm{dist}(z, z') = |z - z'|$ capture the difference in watch time. The resulting instantiation of performative power measures the changes in consumption of a given genre of content induced by a set of prediction functions $\mathcal{F}$. In Section 5, we propose an *observational design* to identify this quantity.

3 Learning versus steering

Performative power enters the firm's optimization problem and has direct consequences for their optimization strategies. Instead of identifying the best action f while treating data as fixed, high performative power enables the firm to *steer* the population towards data that it prefers. In the following, we elucidate the role of performative power in optimization and study the equilibria attained in an economy of performative predictors.

3.1 Optimization strategies

Let us focus on predictive accuracy as the optimization objective of the firm. Hence, the goal of the firm is to choose a predictive model f that suffers small loss $\ell(f(x), y)$ measured over instances (x, y). To make the role of steering explicit we distinguish between the *ex-ante* loss $\ell(f(x(u)), y(u))$ and the *ex-post* loss $\ell(f(x_f(u)), y_f(u))$. The former describes the loss that the firm can optimize when building the predictor. The latter describes the loss that the firm observes after deploying f. More formally, the *ex-post risk* that the firm suffers after deploying f on a population $\mathcal{U}$ is given by

$$\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \ell(f(x_f(u)), y_f(u)) \,. \tag{1}$$

Expression (1) is an instance of what Perdomo et al. [2020] call *performative risk* of a predictor. That is the loss a predictor incurs on the distribution over instances it induces. To simplify notation we adopt their conceptual device of a distribution map: $\mathcal{D}(\theta)$ maps a predictive model, characterized by model parameters θ, to a distribution over data instances.

To express the firm's optimization problem within the framework of performative prediction, we define a data instance as $z(u) = (x(u), y(u))$ for $u \in \mathcal{U}$ so we can capture performativity in the features as well as in the labels. Let the firm's action correspond to choosing a parameter vector θ for its predictor f_θ from an action set $\mathcal{F} = \Theta$. Then, the *aggregate distribution* over data $\mathcal{D}(\theta)$ corresponds to the distribution over the potential outcome variable $z_\theta(u)$ after the firm takes action $\theta \in \Theta$, where the randomness comes from u being uniformly drawn from $\mathcal{U}$ as well as randomness in the potential outcomes. The firm's ex-post risk (1) for deploying f_θ is the performative risk:

$$\mathrm{PR}(\theta) := \mathop{\mathbb{E}}_{z \sim \mathcal{D}(\theta)}[\ell(\theta;\, z)],$$

where the loss typically corresponds to the mismatch between the predicted label and the true label: $\ell(\theta;\, z) = \ell(f_\theta(x), y)$ for $z = (x, y)$.

Observe that θ arises in two places in the objective: in the distribution $\mathcal{D}(\theta)$ and in the loss $\ell(\theta;\, z)$. Thus, for any choice of model ϕ, we can decompose the performative risk $\mathrm{PR}(\theta)$ as:

$$\mathrm{PR}(\theta) = \mathrm{R}(\phi, \theta) + (\mathrm{R}(\theta, \theta) - \mathrm{R}(\phi, \theta)) \tag{2}$$

where $\mathrm{R}(\phi, \theta) := \mathbb{E}_{z \sim \mathcal{D}(\phi)}\, \ell(\theta;\, z)$ denotes the loss of a model θ on the distribution $\mathcal{D}(\phi)$. This tautology highlights the difference between learning and steering and we differentiate between the following two optimization approaches:

Ex-ante optimization. Ex-ante optimization focuses on optimizing the first term in the decomposition (2). For any ϕ, the resulting minimizer can be computed statistically and corresponds to $\theta_{\mathrm{SL}} := \arg\min_{\theta \in \Theta} \mathrm{R}(\phi, \theta)$. Let f_ϕ be any previously chosen model, then employing supervised learning on historical data sampled from $\mathcal{D}(\phi)$ corresponds to what we call ex-ante optimization.

Ex-post optimization. In contrast to ex-ante optimization, *ex-post optimization* accounts for the impact of the model on the distribution. It trades-off the two terms in (2), and directly optimizes the performative risk $\theta_{\mathrm{PO}} := \arg\min_{\theta \in \Theta} \mathrm{PR}(\theta)$. Solving this problem exactly, and finding the performative optimum θ_{PO} requires optimization over the distribution map $\mathcal{D}(\theta)$.

In the context of digital content recommendation, ex-ante optimization corresponds to training θ on historical data collected by the platform, whereas ex-post optimization selects θ based on randomized experiments, A/B testing or explicit modeling of $\mathcal{D}(\theta)$. It holds that $\mathrm{PR}(\theta_{\mathrm{PO}}) \leq \mathrm{PR}(\theta_{\mathrm{SL}})$, as in ex-post optimization the firm can choose to steer the population towards more predictable behavior[2]– an objective worth pursuing as argued in [Shmueli and Tafti, 2020; Ward, 2022].

3.2 Gain of ex-post optimization is bounded by a firm's performative power

We show that the gain of ex-post optimization over ex-ante optimization can be bounded by the firm's performative power with respect to the set of actions Θ and the data vector $z = (x, y)$. Intuitively, if the firm's performative power is low, then the distributions $\mathcal{D}(\theta)$ and $\mathcal{D}(\phi)$ for any $\theta, \phi \in \Theta$ are close to one another. Coupled with a regularity assumption on the loss, this means that the second term in (2) should be small. Thus, using the ex-ante approach of minimizing the first term produces a near-optimal ex-post solution, as we demonstrate in the following result:

Proposition 1. *Let* P *be the performative power of a firm with respect to the action set* Θ*. Let* L_z *be the Lipschitzness of the loss in* z *with respect to the metric* dist*. Let* θ_{PO} *be the ex-post solution and* θ_{SL} *be the ex-ante solution computed from* $\mathcal{D}(\phi)$ *for any past deployment* $\phi \in \Theta$*. Then, we have that*

$$\mathrm{PR}(\theta_{\mathrm{SL}}) \leq \mathrm{PR}(\theta_{\mathrm{PO}}) + 4L_z\mathrm{P}.$$

The gain achievable through ex-post optimization is bounded by performative power. Hence, a firm with small performative power cannot do much better than ex-ante optimization and might be better off sticking to classical supervised learning practices instead of engaging with ex-post optimization.

3.3 Ex-post optimization in an economy of predictors

The result in Proposition 1 studies the optimization strategy of a single firm in isolation. In this section, we investigate the interaction between the strategies of multiple firms that optimize simultaneously over the same population. We consider an idealized marketplace where C firms

[2]The conceptual distinction between learning and steering applies to *general optimization objectives*. Ex-ante optimization corresponds to optimizing on historical data, whereas ex-post optimization corresponds to optimizing over the counterfactuals

all engage in ex-post optimization and we assume all exogenous factors remain constant. Let $\mathcal{D}^C(\theta^1, \ldots, \theta^{i-1}, \theta^i, \theta^{i+1}, \ldots, \theta^C)$ be the distribution over $z(u)$ induced by each firm $i \in [C]$ deploying model f_{θ^i}. Let ℓ_i denotes the loss function chosen by firm i. We say a set of predictors $[f_{\theta^1}, \ldots, f_{\theta^C}]$ is a *Nash equilibrium* if and only if no firm has an incentive to unilaterally deviate from their predictor using ex-post optimization:

$$\theta^i \in \arg\min_{\theta \in \Theta} \mathop{\mathbb{E}}_{z \sim \mathcal{D}^C(\theta^1, \ldots, \theta^{i-1}, \theta, \theta^{i+1}, \ldots, \theta^C)} [\ell_i(\theta;\, z)].$$

First, we show that at the Nash equilibrium, the suboptimality of each predictor f_{θ^i} on the induced distribution depends on the performative power of the respective firm.

Proposition 2. *Suppose that the economy is in a Nash equilibrium $(\theta^1, \ldots, \theta^C)$, and firm i has performative power P_i with respect to the action set Θ. Let L_z be the Lipschitzness of the loss ℓ_i in z with respect to the metric* dist*. Then, it holds that $\mathbb{E}_{z \sim \mathcal{D}}[\ell_i(\theta^i;\, z)] \leq \min_\theta \; \mathbb{E}_{z \sim \mathcal{D}}[\ell_i(\theta;\, z)] + L_z \mathrm{P}_i$, where $\mathcal{D} = \mathcal{D}^C(\theta^1, \ldots, \theta^C)$ is the distribution induced at the equilibrium.*

Proposition 2 implies that if the performative power of all firms is small ($\mathrm{P}_i \to 0 \; \forall i$), then the equilibrium becomes indistinguishable from that of a static, non-performative economy with distribution $\mathcal{D}$ over content. Interestingly, there is an important distinction with the static setting: if the firms were to collude—for example, because of common ownership [Azar et al., 2018] —then they would be able to significantly shift the distribution. In particular, even if the performative power P_i of any given firm i is small, the aggregate performative power of a set of firms S can be much larger.

To illustrate this, we consider a *mixture economy*, where all of the firms share a common loss function ℓ and performative power is uniformly distributed across firms (we formalize the construction of a mixture economy in Appendix B.1). We can apply Proposition 2 to analyze the equilibria as $C \to \infty$.

Corollary 1. *Suppose that all firms $i \in [C]$ share the same loss function $\ell_i = \ell$. Let θ^* be a symmetric Nash equilibrium in the mixture economy with C platforms. As $C \to \infty$, it holds that:*

$$\mathop{\mathbb{E}}_{z \sim \mathcal{D}^C(\theta^*, \ldots, \theta^*)} [\ell(\theta^*;\, z)] \;\to\; \min_\theta \mathop{\mathbb{E}}_{z \sim \mathcal{D}^{C=1}(\theta^*)} [\ell(\theta;\, z)].$$

Corollary 1 demonstrates that a symmetric equilibrium approaches a *performatively stable point* of the monopoly economy where $C = 1$ as the number of firms in the economy grows and the performative power of each individual firm diminishes. In contrast, if all C firms collude, they would obtain the performative power of a monopoly platform and instead choose a *performatively optimal point* of $\mathcal{D}^{C=1}$. A competitive economy can thus exhibit a very different equilibrium from that of the monopoly or collusive economy.

4 Performative power in strategic classification

We now turn to a stylized market model and investigate how performative power depends on the economy in which the firm operates. Specifically, we use *strategic classification* [Hardt et al., 2016] as a test case for our definition. In strategic classification, participants strategically adapt their features with the goal of achieving a favorable classification outcome. Hence, performative power is determined by the degree to which a firm's classifier can impact participant features. We use this concrete market setting to examine the qualitative behavior of performative power in the presence of competition and outside options.

4.1 Strategic classification setup

Let $x(u)$ be the *features* and $y(u)$ the *binary label* describing a participant $u \in \mathcal{U}$. A firm chooses a binary predictor $f : \mathbb{R}^m \to \{0, 1\}$ and incurs loss $\ell(f(x), y) = |f(x) - y|$. Let $\mathcal{D}_{\text{orig}}$ denote the base distribution over features and labels $(x_{\text{orig}}(u), y_{\text{orig}}(u))$ absent any strategic adaptation, which we assume is continuous and supported everywhere. Let $\mathcal{D}(f)$ be the distribution over potential outcomes $(x_f(u), y_f(u))$ that arises from the response of participant u to the deployment of a model f. We assume that participant u incurs a cost $c(x_{\text{orig}}(u), x')$ for changing their features to x'. In line with the standard setup, the cost for feature changes is measured relative to the *original* features. We assume c is a metric and any feature change that deviates from the original features results in nonnegative cost for participants. Further, we assume the label does not change, i.e., $y_f(u) = y_{\text{orig}}(u)$.

Since performative effects surface as changes in participant features, we measure performative power over $z(u) = x(u)$. In Appendix B.2, we describe how this instantiation of performative power can be interpreted in the context of content recommendation platforms and hiring platforms.

4.2 Performative power in the monopoly setting

Consider an economy with a single firm that offers utility $\gamma > 0$ to its participants for a positive prediction. Participants want to use the service regardless of what classifier the firm chooses. In addition, assume there is an *outside options* at utility level $\beta > 0$ that does not demand any extra effort from participants. This decreases the budget they are willing to invest to their *surplus utility* $\Delta\gamma = \max(0, \gamma - \beta)$. We adopt the following standard rationality assumption on participant behavior.

Assumption 1 (Participant Behavior Specification)**.** *Let $\Delta\gamma \geq 0$ be the surplus utility that a participant can expect from a positive classification outcome from classifier f over any outside option. Then, a participant $u \in \mathcal{U}$ with original features $x_{\text{orig}}(u)$ will change their features according to $x_f(u) = \arg\max_{x'} \left(\Delta\gamma f(x') - c(x_{\text{orig}}(u), x')\right)$.*

As a result of Assumption 1 a participant will change their features if and only if the cost of a feature change is no larger than $\Delta\gamma$. Furthermore, if participants change their features, then they will expend the minimal cost required to achieve a positive outcome. This specification of participant behavior allows us to bound the firm's performative power in terms of the cost function c and the distance function dist. Namely, the set of potential values $x_f(u)$ can take on is restricted to $\mathcal{X}_{\Delta\gamma}(u) := \{x : c(x_{\text{orig}}(u), x) \leq \Delta\gamma\}$. Thus, the effect of a change in the decision rule on participant u can be upper bounded by the distance between $x(u)$ and the most distant point in $\mathcal{X}_{\Delta\gamma}$. Aggregating these unilateral effects yields a bound on performative power:

Lemma 1. *The performative power P of the firm with respect to any set of predictors $\mathcal{F}$ and a population $\mathcal{U}$ can be upper bounded as*

$$\mathrm{P} \leq \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sup_{x' \in \mathcal{X}_{\Delta\gamma}(u)} \mathrm{dist}(x(u), x'). \tag{3}$$

If the firm action space $\mathcal{F}$ is restricted to a parametric family of predictors, the upper bound in Lemma 1 need not be tight. A typical decision rule, such as a linear threshold classifier, does not impact all participants $u \in \mathcal{U}$ equally. More specifically, the amount of change that the firm can induce with a decision rule f on an individual u depends the relative position of their features $x_{\text{orig}}(u)$ to the decision boundary. (In Appendix B.3 we provide a precise bound for threshold classifiers).

Interestingly, the ability to fully *personalize* decisions to each user maximizes a firm's performative power and allows it to extract the maximum utility from each individual participant simultaneously.

Proposition 3. *Consider a population $\mathcal{U}$ of users. Let $\mathcal{F}$ be the set of all functions from $\mathbb{R}^m$ to $\{0, 1\}$. Suppose that the first coordinate of $x(u)$ is immutable, and uniquely identifies user $u \in \mathcal{U}$: $x_{\text{orig}}(u)[1] = x_f(u)[1] = i$ for all $f \in \mathcal{F}$. Then, the firm can personalize the decision rule so that the performative power of the firm matches the upper bound in* (3)*.*

Let us investigate the role of $\Delta\gamma$ in the upperbound (3). Recall that user behavior is determined by the cost c of changing features relative to $x_{\text{orig}}(u)$, performative power is measured as the distance from the current state $x(u)$ with respect to dist. The Lipschitz constant $L := \sup_{x,x'} \frac{\mathrm{dist}(x,x')}{c(x,x')}$ allows us to relate the two metrics and derive a simpler upper bound:

Corollary 2. *The performative power of a firm in the monopoly setup can be bounded as $\mathrm{P} \leq 2L\,\Delta\gamma$, where $\Delta\gamma$ measures the surplus utility offered by the service of the firm over outside options.*

Corollary 2 makes explicit that $\Delta\gamma > 0$ is a prerequisite for a firm to have any performative power. This qualitative behavior is in line with common intuition in economics that monopoly power relies on the firm offering a service that is superior to existing options (i.e. $\gamma > \beta$).

4.3 Firms competing for participants

We next consider a dynamic model of *competition* between two firms where participants always choose the firm that offers higher utility. In this model of perfectly elastic demand, we demonstrate how the presence of competition reduces the performative power of a firm. In particular, we will show

that for a natural constraint on the firm's action set each firm's performative power can drop to zero at equilibrium, even when $\gamma > \beta$ and the firm offers participants a nonzero utility for their service.

To model competition in strategic classification, we specify participant behavior as follows: Given that the first firm deploys f_1 and the second firm deploys f_2, then participant u will choose f_1 if $\max_{x'} (f_1(x') - c(x_{\text{orig}}(u), x')) > \max_{x'} (f_2(x') - c(x_{\text{orig}}(u), x'))$, and choose f_2 otherwise. We assume that a participant tie-breaks in favor of the lower threshold, randomizing if the two thresholds are equal. After choosing firm $i \in \{1, 2\}$, they change their features according to Assumption 1 as $x_f(u) = \arg\max_{x'} (\gamma f_i(x') - c(x_{\text{orig}}(u), x'))$, where γ is the utility of a positive outcome.

For the firm we assume their classifier is chosen based on the following utility function: For a rejected participant the firm receives utility 0 and for an accepted participant the firm receives utility $\alpha > 0$ if they have a positive label and utility $-\alpha$ if they have a negative label. Given this utility function we assume the firm's action set is constraint to models $\mathcal{F}^+(\theta)$ that yield non-negative utility for the firm given that the other firm deploys θ.

To formally study competition, we focus on a 1-dimensional setup where $\mathcal{F}$ is the set of threshold functions. We assume that the cost function $c(x, x')$ is continuous in both of its arguments, strictly increasing in x' for $x' > x$, strictly decreasing for $x' < x$, and satisfies $\lim_{x' \to \infty} c(x, x') = \infty$. Furthermore, we assume that the posterior $p(x) = \mathbb{P}_{\mathcal{D}_{\text{orig}}}[Y = 1 \mid X = x]$ is strictly increasing in x with $\lim_{x \to -\infty} p(x) = 0$, and $\lim_{x \to \infty} p(x) = 1$.

Given this market specification, our next result shows that the presence of competition can drive the performative power of each competing firm to zero. This qualitative behavior bears an intriguing resemblance to a state of *perfect competition* in classical pricing economies. A well-known results on market power under Bertrand competition in economics, see e.g., [Baye and Kovenock, 2008], show how a state of zero power is reached with only two competing firms.

Proposition 4. *Consider the 1-dimensional setup with two competing firms specified above. Suppose that the economy is at a symmetric Nash equilibrium (θ^*, θ^*). If $L < \infty$, then the performative power of either firm with respect to the actions $\mathcal{F}^+(\theta^*)$ of non-negative utility solutions is* $\mathrm{P} = 0$.

The intuition behind Proposition 4 is that performative power of a firm purely arises from how much larger the current threshold is than the minimum threshold a firm can deploy within their action set $\mathcal{F}^+(\theta)$. At the Nash equilibrium (where both firms best-respond with respect to their utility functions taking their own performative effects into account) this set of non-utility decreasing actions becomes a singleton for both firms. The formal proof of the result can be found in Appendix C.9.

5 Discrete display design

Now that we have examined the theoretical properties of performative power, we turn to the question of *measuring* performative power from observational data. We focus on our running example of digital content recommendation and we propose an observational design to measure the recommender system's ability to shape consumption patterns.

5.1 The causal effect of position

We assume that there are C pieces of content $\mathcal{C} = \{0, 1, 2, \ldots, C-1\}$ that the platform can present in m display slots. We make the convention that item 0 corresponds to leaving the display slot empty. The case $m = 2$ already encapsulates the main idea so that we restrict our attention to the case of two display slots. The first display slot is more desirable as it is more likely to catch the viewer's attention. For example, the first display slot could be the first ad slot on a page of search results. Researchers have studied the causal effect of position on consumption, often via quasi-experimental methods such as regression discontinuity designs, but also through experimentation in the form of A/B tests.

Definition 2 (Causal effect of position)**.** *Let the treatment $T \in \{0, 1\}$ be the action of flipping the content in the first and second display slots for a viewer u, and let the potential outcome variable $Y_t(u)$ indicate whether viewer u consumes the content that is initially in the first display slot under the treatment $T = t$. We call the corresponding average treatment effect*

$$\beta = \left| \mathbb{E}\left[Y_1 - Y_0\right] \right|$$

the causal effect of position, where the expectation is taken over the population of viewers and the randomness in the potential outcomes.

For example, Narayanan and Kalyanam [2015] estimate the causal effect of position in search advertising, where advertisements are displayed across a number of ordered slots whenever a keyword is searched. They measured the causal effect of position on click-through rate of participants.

5.2 From causal effect of position to performative power

The identification strategy we propose, called *discrete display design (DDD)*, derives a lowerbound on performative power by repurposing existing measures of the causal effect of position. Note that we focus on content recommendation in this section, the design however can be generalized to other settings where the firm's action corresponds to a discrete decision of how to display content. Setting up the DDD involves two steps: First, we need to instantiate the definition of performative power with a suitable action set so that the firm's actions result in swapping the position of content items, and second, we plug in the causal effect of position to lower bound performative power.

While the first step is mostly a technical exercise, the second step relies on a crucial assumption. In particular, it involves relating the unilateral causal effect of position to performative power that quantifies the effect of an action on the entire population of viewers. Thus, for being able to extrapolate the effect from a single viewer to the population DDD relies on a non-interference assumption. In the advertising example, this means that the ads shown to one viewer do not influence the consumption behavior of another viewer. Let us investigate the two steps in greater detail:

Step 1: Instantiating performative power. Let the units $\mathcal{U}$ be the set of viewers. For each viewer $u \in \mathcal{U}$ let $z(u) \in \mathbb{R}^C$ be the distribution over content items $\mathcal{C}$ consumed by viewer u, represented as a histogram. More formally, let $z(u)$ be a vector in the C-dimensional probability simplex where the ith coordinate is the probability that viewer u consumes content item i. The metric $\mathrm{dist}(z, z')$ is the ℓ_1-distance $\mathrm{dist}(z, z') = \sum_{i=0}^{C-1} |z[i] - z'[i]|$.

The decision space $\mathcal{F}$ of the firm corresponds to its decisions of how to arrange content in the $m = 2$ display slots. It is natural to decompose this decision into a continuous score function s followed by a discrete conversion function κ that maps scores into an allocation. The score function $s\colon \mathcal{U} \to \mathbb{R}^C$ maps the viewer to a vector of scores, where each coordinate is an estimate of the quality of the match between the viewer and the corresponding piece of content. The conversion function $\kappa\colon \mathbb{R}^C \to \mathcal{C}^2$ takes as input the vector of scores and outputs an ordered list of items with the top 2 scores. We assume the platform displays these 2 items in order and the conversion function κ is fixed. Hence, we identify the firm's action space with the set of feasible score functions $\mathcal{S} \subseteq \mathcal{U} \to \mathbb{R}^C$.

To define the reference state $z(u)$ we think of s_{curr} as being the score function currently deployed by the platform. Let δ be the maximum difference in the highest score and second highest score for any user under s_{curr}. Consider the set $\mathcal{S}$ of scoring functions defined as

$$\mathcal{S} := \left\{ s\colon \mathcal{U} \to \mathbb{R}^C \mid \forall u \in \mathcal{U}\colon \|s(u) - s_{\text{curr}}(u)\|_\infty \leq \delta \right\}.$$

Notably, there exists an $s_{\text{swap}} \in \mathcal{S}$ that is capable of swapping the order of the first and second highest scoring item under s_{curr} for any user $u \in \mathcal{U}$ simultaneously. We denote the counterfactual variable corresponding to a score function $s \in \mathcal{S}$ as $z_s(u)$. Given this specification, performative power with respect to the action set $\mathcal{S}$ can be bounded by the causal effect of s_{swap} as follow

$$\mathrm{P} = \sup_{s \in \mathcal{S}} \frac{1}{\mathcal{U}} \sum_{u \in \mathcal{U}} \|z_{s_{\text{curr}}}(u) - z_s(u)\|_1 \geq \frac{1}{\mathcal{U}} \sum_{u \in \mathcal{U}} \|z_{s_{\text{curr}}}(u) - z_{s_{\text{swap}}}(u)\|_1. \tag{4}$$

Step 2: Lower bounding performative power. To relate the lower bound on performative power from (4) to the causal effect of position, let $i_{\text{top}}(u) = \kappa \circ s_{\text{curr}}(u)[1]$ denote the coordinate of the item displayed to user u in the first display slot under s_{curr}. Then, we can lower bound each term in the sum (4) as $\|z_{s_{\text{curr}}}(u) - z_{s_{\text{swap}}}(u)\|_1 \geq |z_{s_{\text{curr}}}(u)[i_{\text{top}}(u)] - z_{s_{\text{swap}}}(u)[i_{\text{top}}(u)]|$. Now to enable us to study the effect of changing s_{curr} to s_{swap} independently for each user we place the following non-interference assumption on the counterfactual variables which closely relates to the stable unit treatment value assumption (SUTVA) [Imbens and Rubin, 2015] prevalent in causal inference.[3]

Assumption 2 (No interference across units)**.** *For any $u \in \mathcal{U}$ and any pair of scoring functions $s_1, s_2 \in \mathcal{S}$, if $\kappa(s_1(u)) = \kappa(s_2(u))$, it also holds that $z_{s_1}(u) = z_{s_2}(u)$.*

[3] If an approximate version of Assumption 2 holds, i.e., if $|z_{s_1}(u) - z_{s_2}(u)| < \delta'$, then our lower bound for performative power will apply up to a δ' additive error.

The assumption requires that there are no spill-over or peer effects and the content a viewer consumes only depends on the content recommended to them and not the content recommended to other viewers. The effect of a unilateral change to the consumption of item $i_{\text{top}}(u)$ under s_{swap} exactly corresponds to what we defined as the causal effect of position. Aggregating these unilateral causal effects across all viewers in the population we obtain a lowerbound on performative power. The proof of Theorem 1 can be found in Appendix C.12.

Theorem 1. *Let* P *be performative power as instantiated above. If Assumption 2 holds, then performative power is at least as large as the causal effect of position*

$$\mathrm{P} \geq \beta.$$

Let us return to the search advertisement marketplace study of Narayanan and Kalyanam [2015] to demonstrate how we can leverage Theorem 1 to relate the findings of their observational causal design to performative power. In particular, they examine position effects in search advertising, where ads are displayed across a number of ordered slots whenever a keyword is searched. They found that the effect of showing an ad in display slot 1 versus display slot 2 corresponds to 0.0048 clicks per impression (see Table 2 in their paper). By treating each incoming keyword query as a distinct "viewer" u, this number exactly corresponds to what we defined as the causal effect of position. Thus, we can apply Theorem 1 to get $P \geq 0.0048$. Putting this into context; the mean click-through rate in display slot 2 is 0.023260. Hence, the lower bound 0.0048 is a 21% percent increase relative to the baseline. The firm thus has a substantial ability to shape what advertisements users click on.

6 Discussion

Performative power is a conceptually new notion of power. We dedicate this discussion to the potential role we envision performative power can play in competition policy and antitrust enforcement.

There is a widely recognized need for new investigative tools in the context of digital economies. A challenge with traditional measures of market power, such as the Lerner Index [Lerner, 1934], or the Herfindahl–Hirschman Index (HHI) is that they rely on an idealized model for how the market works. But these established models of classical pricing economies often map poorly to digital economies. Reasons include the complexity and multi-sided nature of digital economies that pose a major hurdle for market definition [Stigler Committee, 2019; Crémer et al., 2019]. But also the important role of behavioral weaknesses of consumers —such as tendencies for single-homing, vulnerability to addiction, and the impact of framing and nudging on participant behavior [e.g. Thaler and Sunstein, 2008; Fogg, 2002]– these can be exploited by firms and often bear unrecognized market power Stigler Committee [2019]. We further expand on this discussion in Appendix A

By focusing on directly observable statistics performative power could be helpful to deal with markets that resist a clean specification. Performative power is sensitive to the market nuances without modeling them. For example, suppose that as a result of uncertainty about market boundaries, a regulator failed to account for a substantial competitor in a marketplace. The presence of this competitor nonetheless decreases the performative power of a firm, indicating that the market is more competitive than suspected.

However performative power also put a heavy burden on the investigator. A major step in using perfomrative power as a tool is its instantiation; the choice of action set $\mathcal{F}$, the definition of a population $\mathcal{U}$, and the specification of what to include in the features z. Without carefully arguing about the interpretation of these variables, the measure of performative power as a measure of *'potential for harm'* might represent a very loose upperbound on what manifests in practice. As a result, going from performative power to consumer harm requires additional substantive arguments. We show in Appendix B.4 how performative power directly translates into consumer harm for strategic classification. However, in general, harm and power are two distinct normative concepts. Performative power should be understood as a tool to flag market situations that merit further investigations.

Acknowledgments

We're grateful to Guy Aridor, Dirk Bergemann, Emilio Calvano, Gabriele Carovano, and Martin C. Schmalz for helpful feedback and pointers.

References

Jose Azar, Martin C. Schmalz, and Isabel Tecu. Anticompetitive effects of common ownership. *The Journal of Finance*, 73(4):1513–1565, 2018.

Ian Ball. Scoring strategic agents. *ArXiv:1909.01888*, 2020.

Michael R. Baye and Dan Kovenock. *Bertrand competition*. Palgrave Macmillan UK, London, 2008.

Dirk Bergemann and Alessandro Bonatti. Data, competition, and digital platforms. 2022.

Jean Magnan de Bornier. The "Cournot-Bertrand Debate": A Historical Perspective. *History of Political Economy*, 24(3):623–656, 09 1992.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 6045–6061. PMLR, 2022.

Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *JMLR*, 13(1):2617–2654, September 2012.

Emilio Calvano and Michele Polo. Market power, competition and innovation in digital markets: A survey. *Information Economics and Policy*, 54:100853, 2021.

Jacques Crémer, Yves-Alexandre de Montjoye, and Heike Schweitzer. *Competition Policy for the digital era : Final report*. Publications Office of the European Union, 2019.

Roy Dong and Lillian J Ratliff. Approximate regions of attraction in learning with decision-dependent distributions. *arXiv preprint arXiv:2107.00055*, 2021.

Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 2022.

B. J. Fogg. Persuasive technology: Using computers to change what we think and do. *Ubiquity*, dec 2002. doi: 10.1145/764008.763957.

Alex Frankel and Navin Kartik. Muddled Information. *Journal of Political Economy*, 127(4): 1739–1776, 2019.

Alex Frankel and Navin Kartik. Improving information from manipulable data. *Journal of the European Economic Association*, 20(1):79–115, 2022.

Tony Ginart, Eva Zhang, Yongchan Kwon, and James Zou. Competing AI: how does competition feedback affect machine learning? In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130, pages 1693–1701. PMLR, 2021.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Innovations in Theoretical Computer Science (ITCS)*, 2016.

Christopher Hennessy and Charles Goodhart. Goodhart's law and machine learning. *SSRN*, 2020.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: Performative gradient descent. In *International Conference on Machine Learning (ICML)*, 2021.

Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In *International Conference on Machine Learning (ICML)*. PMLR, 2022.

A. P. Lerner. The concept of monopoly and the measurement of monopoly power. *The Review of Economic Studies*, 1(3):157–175, 1934.

Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. *ArXiv:2110.00800*, 2021.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In *Advances in Neural Information Processing Systems*, volume 33, pages 4929–4939, 2020.

John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning (ICML)*, pages 7710–7720. PMLR, 2021.

Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Fairness, Accountability, and Transparency (FAccT)*, pages 230–239, 2019.

Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J. Ratliff. Multiplayer performative prediction: Learning in decision-dependent games. *ArXiv:2201.03398*, 2022.

Sridhar Narayanan and Kirthi Kalyanam. Position effects in search advertising and their moderators: A regression discontinuity approach. *Marketing Science*, 34(3):388–407, may 2015. ISSN 1526-548X.

Geoffrey Parker, Georgios Petropoulos, and Marshall W Van Alstyne. Digital platforms and antitrust. 2020.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning (ICML)*, volume 119, pages 7599–7609. PMLR, 2020.

Georgios Piliouras and Fang-Yi Yu. Multi-agent performative prediction: From global stability and optimality to chaos. *Arxiv:2201.10483*, 2022.

Mitas Ray, Lillian J Ratliff, Dmitriy Drusvyatskiy, and Maryam Fazel. Decision-dependent risk minimization in geometrically decaying dynamic environments. 2022.

Galit Shmueli and Ali Tafti. "Improving" prediction of human behavior using behavior modification. *Arxiv:2008.12138*, 2020.

Stigler Committee. Final report: Stigler committee on digital platforms. available at https://research.chicagobooth.edu/stigler/media/news/committee-on-digitalplatforms-final-report, September 2019.

Chad Syverson. Macroeconomics and market power: Context, implications, and open questions. *Journal of Economic Perspectives*, 33(3):23–43, August 2019.

Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, 2008.

Raluca M. Ursu. The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science*, 37(4):530–552, 2018.

Jacob Ward. *The Loop: How Technology is Creating a World without Choices and How to Fight Back*. Hachette Books, 2022.

Killian Wood, Gianluca Bianchin, and Emiliano Dall'Anese. Online projected gradient descent for stochastic optimization with decision-dependent distributions. *IEEE Control Systems Letters*, 6: 1646–1651, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Performative power in competition policy

We expand on the discussion from Section 6 about the potential role of performative power in competition policy and antitrust enforcement as a tool complementary to existing machinery. Performative power differs from traditional measures of market power. It's a causal notion of influence that does not require a precise specification of the market. What it needs is only a specification of a population, the firm's action space, and an outcome variable of interest. The definition therefore applies to markets that defy a clean mathematical definition. We do not need to know or model all the possible mechanisms that relate the firm's action to the participant outcome. To give an example, suppose that as a result of uncertainty about market boundaries we failed to acknowledge a substantial competitor in a marketplace. The presence of this competitor nonetheless decreases the performative power of a firm, indicating that the market is more competitive than suspected.

A.1 Measures of market power in economics

Typical measures of market power in economic theory focus on classical pricing markets of homogeneous goods, where a firm's primary action is choosing a price to sell the good or the quantity of the good to sell. The scalar nature of these quantities enables them to be easily compared across different market contexts and firms. In addition, the utility of the firm and the utility of participants are inversely related: a higher price yields greater utility for the firm and lower utility for all participants. This simple relationship allows one to directly reason about participant welfare and profit of firms. However, the situation of a digital economy is more complex [Stigler Committee, 2019; Crémer et al., 2019]. As a result, classical measures from competition theory that primarily focused on price effects struggle to accurately characterize these economies. Let us illustrate these challenges using two textbook definitions of market power.

Lerner index. The Lerner index [Lerner, 1934] quantifies the pricing power of a firm, measuring by how much the firm can raise the price above marginal costs. Marginal costs reflect the price that would arise in a perfectly competitive market. A major issue of applying this standard definition of market power in the context of digital economies is that it is not clear what the competitive reference state should look like, "We have lost the competitive benchmark,"[4] as Jacques Crémer said. Thus, measures based on profit margin cannot directly be adopted as a proxy for market power in these settings.

Market share. Measures such as the Herfindahl–Hirschman index (HHI), which is used by the US federal trade commission[5] to measure market competitiveness, is based on *market share*: the fraction of participants who participate in a given firm. However, the validity of market share as a proxy for power relies on a specific model of competition where the elasticity of demand is low. This model is challenging to justify[6] in the context of digital economies where opening an account on a platform is very simple and usually free of charge. In addition, not all participants with accounts on a digital platform are equally active and inactive participants should not factor into the market power of a firm the same way active participants do. Market share is not sufficiently expressive to make this distinction.

A.2 Complex consumer behavior

Participant behavior plays a critical role in digital market places. As outlined by the Stigler Committee [2019], "the findings from behavioral economics demonstrate an under-recognized market power held by incumbent digital platforms." In particular, behavioral aspects of consumers—such as tendencies for single-homing, vulnerability to addiction, and as well as the impact of framing and nudging on participant behavior [e.g. Thaler and Sunstein, 2008; Fogg, 2002]—can be exploited by firms in digital economies, but do not factor into traditional measures of market power. By focusing on changes in participant features, performative power has the potential to capture the effects of these behavioral patterns while again sidestepping the challenges of explicitly modeling them.

[4] Opening statement at the 2019 Antitrust and competition conference – digital platforms, markets, and democracy

[5] See `https://www.justice.gov/atr/herfindahl-hirschman-index` (retrieved January, 2022).

[6] This critique is similar to the disconnect between the Cournot model and the Bertrand model in classical economics [Bornier, 1992]. E.g., "concentration is worse than just a noisy barometer of market power" [Syverson, 2019].

A.3 From performative power to consumer harm

Performative power focuses on measuring power rather than harm. The relationship to harm depends on the choice and interpretation of the outcome variable and requires additional substantive arguments.

High performative power and the ability of a firm to steer participants naturally bears the potential for user harm, whenever there exists a misalignment between the utilities of the platform and the participants. Recent work by Shmueli and Tafti [2020] backs a similar concern with empirical evidence. However, in general, harm and power are two fundamentally distinct normative concepts. Performative power does not necessarily imply harm, but it can serve as an indicator of *potential* harm that can help flag market situations that merit further investigations by regulators. In fact, the report of the European Commission on competition policy for the digital era [Crémer et al., 2019] suggests in several contexts where regulators should be suspicious of power and suggests putting burden of proof on the firm rather than the regulators.

In specific contexts, we can establish an exact correspondence between performative power and harm. By carefully choosing the parameters in the definition, performative power can be instantiated to more closely capture manifested harm to a particular (sub-)population of participants we wish to investigate. We implement this principle in Section 4 where we establish a connection between performative power and the utility of users. More generally, this connection can be achieved if the attributes $z(u)$ consists of the sensitive features that are impacted by the firm, the distance function is aligned with the utility function of participants, and the set $\mathcal{F}$ reflects actions that are taken by the firm.

B Additional results and discussions for Section 3 and Section 4

B.1 Formalization of a mixture economy

We formalize the construction of a mixture economy. Let $z(u), z_\theta^{C=1}(u)$ denote the pair of counterfactual outcomes before and after the deployment of θ in a hypothetical monopoly economy where a single firm holds all the performative power. Let $\mathcal{D}^{C=1}(\theta)$ be the distribution map associated with the variables $z_\theta^{C=1}(u)$ for $u \in \mathcal{U}$. In a uniform mixture economy, we assume that each participant $u \in \mathcal{U}$ uniformly chooses between the C firms. The counterfactual $z_\theta(u)$ associated with one firm changing its predictor to θ is equal to $z(u)$ with probability $1 - 1/C$ and $z_\theta^{C=1}(u)$ otherwise.

B.2 Interpretation of performative power in setup in Section 4

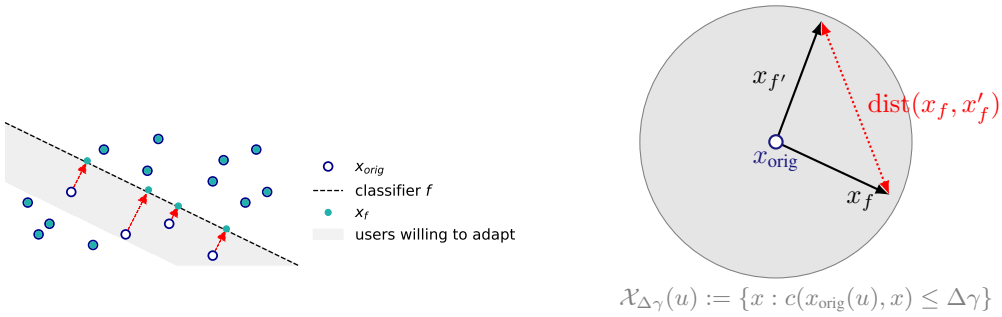
In our running example of digital content recommendations where participants correspond to content creators, the instantiation of performative power in Section 4 measures how much the content of each channel changes with changes in the recommendation algorithm. As a different example, suppose that firm is a hiring platform such as HireVue that uses video data from interviews to a compute a performance prediction for an applicant for a job. In this case, the participants would correspond to applicants, and the data vector would correspond to the applicant's voice patterns and interview responses.

When instantiating the definition of performative power, the choice of distance metric enables us to define how to weight specific feature changes. In our content recommendation example, if we are interested in the burden on *content creators*, we choose the distance metric to be aligned with the cost function c of producing a piece of content. However, if we are interested in measuring the impact of changes in content on viewers, a distance metric that reflects harm to viewers might be more appropriate. We keep this distance metric abstract in our analysis.

B.3 Example: Monopoly power in heterogeneous setting

Different participants are typically impacted differently by a classifier, depending on their relative position to the decision boundary, as visualized in Figure 1(a). As a result of this heterogeneity, the upper bound in Lemma 1 is not necessarily tight, because the firm can not extract the full utility from all participants simultaneously.

Let us investigate the effect of heterogeneity in a concrete 1-dimensional setting where $\mathrm{dist}_X(x, x') = c(x, x') = |x - x'|$. Consider a set of actions $\mathcal{F}$ that corresponds to the set of all threshold functions.



(a) participants adapting to classifier f

(b) participant expenditor constraint

Figure 1: Illustrations for 2-dimensional strategic classification example

Suppose that the posterior $p(x) = \mathbb{P}[Y = 1 \mid X = x]$ satisfies the following regularity assumptions: $p(x)$ is strictly increasing in x with $\lim_{x\to-\infty} p(x) = 0$, and $\lim_{x\to\infty} p(x) = 1$. Now, let θ_{SL} be the supervised learning threshold, which is the unique value where $p(\theta_{\mathrm{SL}}) = 0.5$. We can then obtain the following bound on the performative power P with respect to any $\mathcal{F}$ assuming the firm's classifier is θ_{SL} in the current economy (see Proposition 5):

$$0.5\Delta\gamma \mathop{\mathbb{P}}_{\mathcal{D}_{\mathrm{orig}}} \left[x \in [\theta_{\mathrm{SL}}, \theta_{\mathrm{SL}} + 0.5\Delta\gamma]\right] \leq \mathrm{P} \leq \Delta\gamma\,. \tag{5}$$

This bound illustrates how performative power in strategic classification depends on the fraction of participants that fall in between the old and the new threshold. As long as the density in this region is non-zero, a platform that offers $\Delta\gamma > 0$ utility will also have strictly positive performative power, providing a lower bound on P.

Proposition 5. *Suppose that* $\mathrm{dist}(x, x') = c(x, x') = |x - x'|$. *Let us consider a set of actions* $\mathcal{F}$ *that corresponds to the set of all threshold functions. Suppose that the posterior* $p(x) = \mathbb{P}[Y = 1 \mid X = x]$ *satisfies the following regularity assumptions:* $p(x)$ *is strictly increasing in* x *with* $\lim_{x\to-\infty} p(x) = 0$, *and* $\lim_{x\to\infty} p(x) = 1$. *Now, let* θ_{SL} *be the supervised learning threshold, which is the unique value where* $p(\theta_{\mathrm{SL}}) = 0.5$. *If the firm's classifier is* θ_{SL} *in the current economy, then performative power* P *with* $\mathcal{F}$ *can be bounded as:*

$$0.5\gamma \mathop{\mathbb{P}}_{\mathcal{D}_{orig}} \left[x \in [\theta_{\mathrm{SL}}, \theta_{\mathrm{SL}} + 0.5\Delta\gamma]\right] \leq \mathrm{P} \leq 2\Delta\gamma\,. \tag{6}$$

B.4 Optimization strategies, performative power, and user harm

The fact that a monopoly firm has nonzero performative power has consequences for the optimization strategies that it would use, as we discussed in Section 3. To make this explicit, let's contrast the solutions of ex-ante and ex-post optimization in a simple one-dimensional setting.

Example 1 (1-dimensional setting)**.** *Consider a 1-dimensional feature vector* $x \in \mathrm{R}$ *and suppose that the posterior* $p(x) = \mathbb{P}[Y = 1 \mid X = x]$ *is strictly increasing in* x *with* $\lim_{x\to-\infty} p(x) = 0$, *and* $\lim_{x\to\infty} p(x) = 1$. *Now consider a set of actions* $\mathcal{F}$ *that corresponds to the set of all threshold functions and set* $\mathrm{dist}(x, x') = c(x, x') = |x - x'|$. *Let* θ_{SL} *be the supervised learning threshold from ex-ante optimization, which is the unique value where* $p(\theta_{\mathrm{SL}}) = 0.5$. *Then, the ex-post threshold lies at* $\theta_{\mathrm{PO}} = \theta_{\mathrm{SL}} + \Delta\gamma$.

In Example 1 ex-post optimization leads to a higher acceptance threshold than ex-ante optimization. Thus, for any setting where the participants utility is decreasing in the threshold (e.g., the class of utility functions that Milli et al. [2019] call *outcome monotonic*), this implies that ex-post optimization creates stronger negative externalities for participants than ex-ante optimization. Furthermore, the effect grows with the performative power of the firm. In the extreme case of the monopoly setting with no outside options, ex-post optimization can leave certain participants with a net utility of 0 and thus can transfer the entire utility from these participants to the firm.

C Proofs

C.1 Auxiliary results

The proofs for Section 3 leverage the following lemma, which bounds the diameter of Θ with respect to Wasserstein distance in distribution map.

Lemma 2. *Let* P *be the performative power with respect to* Θ*. For any* $\theta, \theta' \in \Theta$*, it holds that* $\mathcal{W}(\mathcal{D}(\theta), \mathcal{D}(\theta')) \leq 2\mathrm{P}$.

Proof. Let θ_{curr} be the current classifier weights. We use the fact that for any weights $\theta'' \in \Theta$, it holds that $\mathcal{W}(\mathcal{D}(\theta_{\text{curr}}), \mathcal{D}(\theta'')) \leq \frac{1}{|U|}\sum_{u\in\mathcal{U}} \mathbb{E}[\mathrm{dist}(z(u), z_{\theta''}(u))]$ where the expectation is over randomness in the potential outcomes. This follows from the definition of Wasserstein distance—in particular that we can instantiate the mass-moving function by mapping each participant to themselves. Thus, we see that:

$$\begin{aligned}
\mathcal{W}(\mathcal{D}(\theta), \mathcal{D}(\theta')) &\leq \mathcal{W}(\mathcal{D}(\theta), \mathcal{D}(\theta_{\text{curr}})) + \mathcal{W}(\mathcal{D}(\theta_{\text{curr}}), \mathcal{D}(\theta')) \\
&\leq \frac{1}{|U|}\sum_{u\in\mathcal{U}} \mathbb{E}[\mathrm{dist}(z(u), z_{\theta}(u))] + \frac{1}{|U|}\sum_{u\in\mathcal{U}} \mathbb{E}[\mathrm{dist}(z(u), z_{\theta'}(u))] \\
&\leq 2 \sup_{\theta''\in\Theta} \frac{1}{|U|}\sum_{u\in\mathcal{U}} \mathbb{E}[\mathrm{dist}(z(u), z_{\theta''}(u))] \\
&\leq 2\mathrm{P},
\end{aligned}$$

where the last line uses the definition of performative power that bounds the effect of any θ in the action set Θ on the participant data z. □

C.2 Proof of Proposition 1

Let ϕ be the previous deployment inducing the distribution on which the supervised learning threshold θ_{SL} is computed. Let θ^* be an optimizer of $\min_{\theta\in\Theta} \mathrm{R}(\theta_{\mathrm{PO}}, \theta)$, where we recall the definition of the decoupled performative risk as $\mathrm{R}(\phi, \theta) := \mathbb{E}_{z\sim\mathcal{D}(\phi)}\, \ell(\theta;\, z)$. Then, we see that for any ϕ:

$$\begin{aligned}
&\mathrm{PR}(\theta^{\mathrm{SL}}) - \mathrm{PR}(\theta_{\mathrm{PO}}) \\
&= \big(\mathrm{R}(\theta^{\mathrm{SL}}, \theta^{\mathrm{SL}}) - \mathrm{R}(\phi, \theta^{\mathrm{SL}})\big) + \mathrm{R}(\phi, \theta^{\mathrm{SL}}) - \mathrm{R}(\theta_{\mathrm{PO}}, \theta_{\mathrm{PO}}) \\
&\leq \big(\mathrm{R}(\theta^{\mathrm{SL}}, \theta^{\mathrm{SL}}) - \mathrm{R}(\phi, \theta^{\mathrm{SL}})\big) + \mathrm{R}(\phi, \theta^{*}) - \mathrm{R}(\theta_{\mathrm{PO}}, \theta^{*}) \\
&\leq L_z \mathcal{W}(\mathcal{D}(\theta^{\mathrm{SL}}), \mathcal{D}(\phi)) + L_z \mathcal{W}(\mathcal{D}(\phi), \mathcal{D}(\theta_{\mathrm{PO}})) \\
&\leq 4 L_z P\,.
\end{aligned}$$

The first inequality follows because θ^* minimizes risk on the distribution $\mathcal{D}(\theta_{\mathrm{PO}})$, while θ^{SL} minimizes risk on $\mathcal{D}(\phi)$. The second inequality follows from the dual of the Wasserstein distance where L_z is the Lipschitz constant of the loss function in the data argument z. The last inequality follows from Lemma 2.

Now, suppose that ℓ is γ-strongly convex. Then we have that:

$$\mathrm{R}(\theta, \theta_{\mathrm{PO}}) - \mathrm{R}(\theta, \theta_{\mathrm{SL}}) \geq \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{SL}}\|^2$$

Again applying Lemma 2,

$$\begin{aligned}
\mathrm{PR}(\theta_{\mathrm{SL}}) &= \mathrm{R}(\theta_{\mathrm{SL}}, \theta_{\mathrm{SL}}) \\
&\leq \mathrm{R}(\theta, \theta_{\mathrm{SL}}) + L_z \mathcal{W}(\mathcal{D}(\phi), \mathcal{D}(\theta_{\mathrm{SL}})) \\
&\leq \mathrm{R}(\phi, \theta_{\mathrm{SL}}) + 2L_z P \\
&\leq \mathrm{R}(\phi, \theta_{\mathrm{PO}}) - \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{SL}}\|^2 + 2L_z P \\
&\leq \mathrm{R}(\theta_{\mathrm{PO}}, \theta_{\mathrm{PO}}) + L_z \cdot \mathcal{W}(\mathcal{D}(\phi), \mathcal{D}(\theta_{\mathrm{PO}})) - \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{SL}}\|^2 + 2L_z P \\
&\leq \mathrm{PR}(\theta_{\mathrm{PO}}) + 4L_z P - \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{SL}}\|^2.
\end{aligned}$$

Using that $\mathrm{PR}(\theta_{\mathrm{PO}}) \leq \mathrm{PR}(\theta_{\mathrm{SL}})$, we find that

$$\frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{SL}}\|^2 \leq 4L_z P\,.$$

Rearranging gives

$$\|\theta_{\mathrm{PO}} - \theta_{\mathrm{SL}}\| \leq \sqrt{\frac{8L_z P}{\gamma}}\,.$$

C.3 Proof of Proposition 2

Let's focus on firm i, fixing classifiers selected by the other firms. Let's take PR and R to be defined with respect to $\mathcal{D}(\cdot) = \mathcal{D}(\theta_1, \ldots, \theta_{i-1}, \cdot, \theta_{i+1}, \ldots, \theta_C)$. Let $\theta^* = \arg\min_\theta \mathrm{R}(\theta_i, \theta)$. We see that:

$$\begin{aligned}
\mathrm{PR}(\theta^i) &\leq \mathrm{PR}(\theta^*) \\
&\leq \mathrm{R}(\theta_i, \theta^*) + L_z \mathcal{W}(\mathcal{D}(\theta_i), \mathcal{D}(\theta^*)) \\
&\leq \min_\theta \mathrm{R}(\theta_i, \theta) + L_z \left(\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[\mathrm{dist}(z(u), z_{\theta^*}(u))] \right) \\
&\leq \min_\theta \mathrm{R}(\theta_i, \theta) + L_z P_i.
\end{aligned}$$

Rewriting this, we see that:

$$\mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell_i(\theta^i;\, z)] \leq \min_\theta \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell_i(\theta;\, z)] + L_z P_i.$$

If, in addition, ℓ_i is γ-strongly convex, then we know that:

$$L_z P_i \geq \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell_i(\theta^i;\, z)] - \min_\theta \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell(\theta;\, z)] \geq \frac{\gamma}{2}\|\theta^i - \min_\theta \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell_i(\theta;\, z)]\|^2.$$

Rearranging, we obtain that

$$\left\| \theta^i - \min_\theta \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell(\theta;\, z)] \right\|_2 \leq \sqrt{\frac{2L_z P_i}{\gamma}}\,.$$

C.4 Proof of Corollary 1

Let P be the performative power associated with the variables $z_\theta^{C=1}$. We first claim that the performative power of any firm in the mixture model is at most P/C. This follows from the fact that for a given firm the potential outcome $z_\theta(u)$ is equal to $z(u)$ with probability $1 - 1/C$ and equal to $z_\theta^{C=1}(u)$ with probability $1/C$.

Let's focus on platform i, fixing classifiers selected by the other platforms. Let's take PR and R to be with respect to $\mathcal{D}(\cdot) = \mathcal{D}^C(\theta^*, \cdots, \theta^*, \cdot, \theta^*, \ldots, \theta^*)$. Now, we can apply Proposition 2 to see that

$$\mathrm{PR}(\theta^*) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}^{C=1}(\theta^*)}[\ell(\theta^*;\, z)] \leq \min_\theta \mathop{\mathbb{E}}_{z \sim \mathcal{D}^{C=1}(\theta^*)}[\ell(\theta;\, z)] + \frac{L_z P}{C}.$$

Thus, in the limit as $C \to \infty$, it holds that

$$\mathop{\mathbb{E}}_{z \sim \mathcal{D}^{C=1}(\theta^*)}[\ell(\theta^*;\, z)] \to \min_\theta \mathop{\mathbb{E}}_{z \sim \mathcal{D}^{C=1}(\theta^*)}[\ell(\theta;\, z)]$$

as desired.

C.5 Proof of Lemma 1

Fix a classifier f and a unit u. By Assumption 1, we know that $x(u)$ and $x_f(u)$ are both in $\mathcal{X}_{\Delta_\gamma}(u)$. The claim follows from

$$\mathrm{dist}(x_{\mathrm{orig}}(u), x_f(u)) \leq \sup_{x' \in \mathcal{X}_{\Delta_\gamma}(u)} \mathrm{dist}(x_{\mathrm{orig}}(u), x').$$

C.6 Proof of Proposition 3

The first coordinate being immutable corresponds to $c(x, x') = \infty$ for $x[1] \neq x'[1]$. The proof is by construction of a classifier $f^* : \mathrm{R}^m \to \{0, 1\}$. For each individual u we define the set

$$\tilde{\mathcal{X}}(u) := \operatorname*{arg\,sup}_{x' \in \mathcal{X}_{\Delta\gamma}(u)} \operatorname{dist}(x(u), x').$$

Now let f^* be such that

$$f^*(x) = \begin{cases} 1 & x \in \tilde{\mathcal{X}}(u) \text{ with } u = x[1] \\ 0 & x \notin \tilde{\mathcal{X}}(u) \text{ with } u = x[1] \end{cases}$$

where we used that the first coordinate of the feature vector x uniquely identifies the individual. The effect of f^* on a population $\mathcal{U}$ corresponds to

$$\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \operatorname{dist}(x(u), x_{f^*}(u)) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sup_{x' \in \mathcal{X}_{\Delta\gamma}(u)} \operatorname{dist}(x(u), x') \tag{7}$$

Thus for any $\mathcal{F}$ that contains f^* the performative power is maximized.

C.7 Proof of Corollary 2

Applying Lemma 1, it suffices to show that the diameter of the set $\mathcal{X}_{\Delta\gamma}(u)$ can be upper bounded by $2L\Delta\gamma$ for any $u \in \mathcal{U}$. We see that for any $x, x' \in \mathcal{X}_{\Delta\gamma}(u)$, using that c is a metric, it holds that:

$$\begin{aligned} \operatorname{dist}(x, x') &\leq L \cdot c(x, x') \\ &\leq L \cdot (c(x_{\text{orig}}(u), x) + c(x_{\text{orig}}(u), x')) \\ &\leq 2L\Delta\gamma. \end{aligned}$$

C.8 Proof of Proposition 5

The upper bound follows from Corollary 4. For the lower bound, we take f to be the threshold classifier given by $\theta_{\text{SL}} + \Delta\gamma$. We see that for $x_{\text{orig}}(u) \in [\theta_{\text{SL}}, \theta_{\text{SL}} + \Delta\gamma]$, it holds that $x_f(u) = \theta_{\text{SL}} + \Delta\gamma$ and $x(u) = x_{\text{orig}}(u)$. This means that the performative power is at least:

$$\begin{aligned} \mathrm{P} &= \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[\operatorname{dist}(x(u), x_f(u))] \\ &= \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[|x(u) - x_f(u)|] \\ &\geq \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} I[x_{\text{orig}}(u) \in [\theta_{\text{SL}}, \theta_{\text{SL}} + \Delta\gamma]]\, \mathbb{E}[|\theta_{\text{SL}} + \Delta\gamma - x_{\text{orig}}(u)|] \\ &\geq \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} I[x_{\text{orig}}(u) \in [\theta_{\text{SL}}, \theta_{\text{SL}} + \frac{1}{2}\Delta\gamma]] \cdot \frac{1}{2}\Delta\gamma \\ &\geq \frac{1}{2}\Delta\gamma \mathop{\mathbb{P}}_{\mathcal{D}_{\text{orig}}} [x \in [\theta_{\text{SL}}, \theta_{\text{SL}} + \frac{1}{2}\Delta\gamma]], \end{aligned}$$

as desired.

C.9 Proof of Proposition 4

The proof relies on the following proposition that we will prove in the next section.

Proposition 6. *Consider the 1-dimensional setup specified in Section 4.3, and suppose that the economy is at a symmetric state where both firms choose classifier θ. For any $\mathcal{F}$, consider one of the firms, let $\mathcal{F}$ denote their action set and let $\theta_{\min}$ be the minimum threshold classifier in $\mathcal{F}$. Then, the performative power of the firm is upper bounded by:*

$$\mathrm{P} \leq L \min(c(\theta_{\min}, \theta), \gamma) + \gamma L p_{\text{reach}}([\theta_{\min}, \theta]).$$

where $p_{\text{reach}}([\theta_{\min}, \theta]) := \mathbb{P}_{\mathcal{D}_{orig}} [x \in [\xi(\theta_{\min}), \xi(\theta)]]$ with $\xi(\theta')$ being the unique value such that $\xi(\theta') < \theta'$ and $c(\xi(\theta'), \theta') = 1$.

The performative power of a firm thus purely arises from how much larger the market threshold is than the minimum threshold a firm can deploy within their action set $\mathcal{F}$.

We now apply Proposition 6 to see that the performative power for any $\mathcal{F}$ is upper bounded by

$$B := L \min(c(\theta_{\text{min}}, \theta^*), \gamma) + L\gamma p_{\text{reach}}([\theta_{\text{min}}, \theta^*])$$

where (θ^*, θ^*) is a symmetric state. Now considering the constraints on the firm's action set given by the firm's utility function, using Proposition 2, we see that $\mathcal{F}(\theta^*) = [\theta^*, \infty)$. This means that $\theta_{\text{min}} = \theta^*$, and so $\xi(\theta_{\text{min}}) = \xi(\theta^*)$. Thus, $B = 0$ which demonstrates that the performative power is upper bounded by 0, and is thus equal to 0.

C.10 Proof of Proposition 6

Proof of Proposition 6. Consider a classifier $f \in \mathcal{F}(\theta)$ with threshold θ', and suppose that a firm changes their classifier to f. It suffices to show that:

$$\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[\text{dist}(x(u), x_f(u))] \leq L \min(c(\theta_{\text{min}}, \theta), \gamma) + L\gamma p_{\text{reach}}([\theta_{\text{min}}, \theta]).$$

For technical convenience, we reformulate this in terms of the cost function c. Based on the definition of L, it suffices to show that:

$$\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[c(x(u), x_f(u))] \leq \min(c(\theta_{\text{min}}, \theta), \gamma) + \gamma p_{\text{reach}}([\theta_{\text{min}}, \theta]).$$

Case 1: $\theta' > \theta$**.** Participants either are indifferent between θ and θ' or prefer θ to θ'. Due to the tie breaking rule, the firm will thus lose all of its participants. Thus, all participants will switch to the other firm and adapt their features to that firm which has threshold θ. This is the same behavior as these participants had in the current state, so $x_f(u) = x(u)$ for all participants u. This means that

$$\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[c(x(u), x_f(u))] = 0$$

as desired.

Case 2: $\theta' < \theta$**.** Participants either are indifferent between θ and θ' or prefer θ' to θ. Due to the tie breaking rule, the firm will thus gain all of the participants. We break into several cases:

$$\begin{cases} x_f(u) = x(u) = x_{\text{orig}}(u) & \text{if } x_{\text{orig}}(u) < \xi(\theta') \\ x_f(u) = \theta', x(u) = x_{\text{orig}}(u) & \text{if } x_{\text{orig}}(u) \in [\xi(\theta'), \min(\theta', \xi(\theta)))] \\ x_f(u) = x(u) = x_{\text{orig}}(u) & \text{if } x_{\text{orig}}(u) \in (\theta', \xi(\theta)) \\ x_f(u) = \theta', x(u) = \theta & \text{if } x_{\text{orig}}(u) \in (\xi(\theta), \theta') \\ x_f(u) = x_{\text{orig}}(u), x(u) = \theta & \text{if } x_{\text{orig}}(u) \in [\max(\theta', \xi(\theta)), \theta] \\ x_f(u) = x(u) = x_{\text{orig}}(u) & \text{if } x_{\text{orig}}(u) \geq \theta. \end{cases}$$

The only cases that contribute to $\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[c(x(u), x_f(u))]$ are the second, fourth, and fifth cases. Thus, we can upper bound $\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[c(x(u), x_f(u))]$ by:

$$\underbrace{\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U} | x_{\text{orig}}(u) \in [\xi(\theta'), \min(\theta', \xi(\theta))]} \mathbb{E}[c(x(u), x_f(u))]}_{(A)} + \underbrace{\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U} | x_{\text{orig}}(u) \in (\xi(\theta), \theta')} \mathbb{E}[c(x(u), x_f(u))]}_{(B)}$$

$$+ \underbrace{\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U} | x_{\text{orig}}(u) \in [\max(\theta', \xi(\theta)), \theta]} \mathbb{E}[c(x(u), x_f(u))]}_{(C)}$$

For (A), we see that

$$\begin{aligned}(A) &= \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U} | x_{\text{orig}}(u) \in [\xi(\theta'), \min(\theta', \xi(\theta)))} \mathbb{E}[c(x_{\text{orig}}(u), \theta')] \\ &\leq \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U} | x_{\text{orig}}(u) \in [\xi(\theta'), \min(\theta', \xi(\theta)))} \mathbb{E}[c(\xi(\theta'), \theta')] \\ &= \gamma \cdot \mathbb{P}_{\mathcal{D}_{\text{orig}}}[x \in [\xi(\theta'), \min(\theta', \xi(\theta)))] \\ &\leq \gamma \cdot \mathbb{P}_{\mathcal{D}_{\text{orig}}}[x \in [\xi(\theta'), \xi(\theta))].\end{aligned}$$

For (B), we see that:

$$\begin{aligned}(B) &= \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U} | x_{\text{orig}}(u) \in (\xi(\theta), \theta')} \mathbb{E}[c(\theta, \theta')] \\ &= c(\theta, \theta') \cdot \mathbb{P}_{\mathcal{D}_{\text{orig}}}[x \in (\xi(\theta), \theta')] \\ &= \min(c(\theta, \theta'), \gamma) \cdot \mathbb{P}_{\mathcal{D}_{\text{orig}}}[x \in (\xi(\theta), \theta')].\end{aligned}$$

For (C), we see that:

$$\begin{aligned}(C) &= \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U} | x_{\text{orig}}(u) \in [\max(\theta', \xi(\theta)), \theta]} \mathbb{E}[c(x_{\text{orig}}(u), \theta)] \\ &\leq \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U} | x_{\text{orig}}(u) \in [\max(\theta', \xi(\theta)), \theta]} \mathbb{E}[\min\left(c(\theta', \theta), c(\xi(\theta), \theta)\right)] \\ &= \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U} | x_{\text{orig}}(u) \in [\max(\theta', \xi(\theta)), \theta]} \mathbb{E}[\min\left(c(\theta', \theta), \gamma\right)] \\ &= \min(c(\theta', \theta), \gamma) \cdot \mathbb{P}_{\mathcal{D}_{\text{orig}}}[x \in [\max(\theta', \xi(\theta)), \theta]]\end{aligned}$$

Putting this all together, we obtain that:

$$\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[c(x(u), x_f(u))] \leq \gamma p_{\text{reach}}([\theta', \theta]) + \min(c(\theta', \theta), \gamma)$$

for $p_{\text{reach}}([\theta', \theta]) := \mathbb{P}_{\mathcal{D}_{\text{orig}}}[x \in [\xi(\theta'), \xi(\theta)]]$ as desired. Since $\gamma p_{\text{reach}}([\theta', \theta]) + \min(c(\theta', \theta), \gamma)$ is decreasing in θ', this expression is maximized when $\theta' = \theta_{\text{min}}$. Thus we obtain an upper bound of

$$\gamma \cdot p_{\text{reach}}([\theta_{\text{min}}, \theta]) + \min(c(\theta_{\text{min}}, \theta), \gamma)\,. \qquad \square$$

C.11 Statement and proof of Proposition 7

Proposition 7. *Consider the 1-dimensional setup described in Section 4.3. Then, a symmetric solution $[\theta^*, \theta^*]$ is an equilibrium if and only if θ^* satisfies*

$$\mathbb{E}_{(x,y) \sim \mathcal{D}_{orig}}[y = 1 \mid x \geq \xi(\theta^*)] = \tfrac{1}{2}, \tag{8}$$

where $\xi(\theta^*)$ is the unique value such that $c(\xi(\theta^*), \theta^*) = \gamma$ and $\xi(\theta^*) < \theta^*$. Both firms earn zero utility at this equilibrium. Moreover, the set $\mathcal{F}^+(\theta^*)$ of actions that a firm can take at equilibrium that achieve nonnegative utility is exactly equal to $[\theta^*, \infty)$, assuming the other firm chooses the classifier θ^*.

Proof. The proof proceeds in two steps. First, we establish that $[\theta^*, \theta^*]$ is an equilibrium; next, we show that $[\theta, \theta]$ is not in equilibrium for $\theta \neq \theta^*$.

Establishing that $[\theta^*, \theta^*]$ is an equilibrium and $\mathcal{F}^+(\theta^*) = [\theta^*, \infty)$. First, we claim that $[\theta^*, \theta^*]$ is an equilibrium. At $[\theta^*, \theta^*]$, each participant chooses the first firm with $1/2$ probability. The

expected utility earned by a firm is:

$$\begin{aligned}
\frac{1}{2}\int_{\xi(\theta)}^{\infty} p_{\text{orig}}(x)(p(x)-(1-p(x))\mathrm{d}x &= \int_{\xi(\theta)}^{\infty} p_{\text{orig}}(x)(p(x)-0.5)\mathrm{d}x \\
&= \int_{\xi(\theta)}^{\infty} p_{\text{orig}}(x)p(x)\mathrm{d}x - 0.5\int_{\xi(\theta)}^{\infty} p_{\text{orig}}(x)\mathrm{d}x \\
&= \int_{\xi(\theta)}^{\infty} p_{\text{orig}}(x)\mathrm{d}x \left(\frac{\int_{\xi(\theta)}^{\infty} p_{\text{orig}}(x)p(x)\mathrm{d}x}{\int_{\xi(\theta)}^{\infty} p_{\text{orig}}(x)\mathrm{d}x} - \frac{1}{2}\right) \\
&= \left(\int_{\xi(\theta)}^{\infty} p_{\text{orig}}(x)\mathrm{d}x\right)\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\text{orig}}}[y=1 \mid x \geq \xi(\theta)] - \frac{1}{2}\right) \\
&= 0\,.
\end{aligned}$$

If the firm chooses $\theta > \theta^*$, then since the cost function is strictly monotonic in its second argument, participants either are indifferent between θ and θ^* or prefer θ to θ^*. Due to the tie breaking rule, the firm will thus lose all of its participants and incur 0 utility. Thus the firm has no incentive to switch to θ.

If the firm chooses $\theta < \theta^*$, then it will gain all of the participants. The firm's utility will be:

$$\begin{aligned}
&\int_{\xi(\theta)}^{\infty} p_{\text{orig}}(x)(p(x)-(1-p(x))\mathrm{d}x \\
&\qquad = \int_{\xi(\theta)}^{\xi(\theta^*)} p_{\text{orig}}(x)(p(x)-(1-p(x))\mathrm{d}x + \int_{\xi(\theta^*)}^{\infty} p_{\text{orig}}(x)(p(x)-(1-p(x))\mathrm{d}x \\
&\qquad = 2\int_{\xi(\theta)}^{\xi(\theta^*)} p_{\text{orig}}(x)(p(x)-0.5)\mathrm{d}x.
\end{aligned}$$

It is not difficult to see that at θ^*, it must hold that $p(\xi(\theta^*)) \leq 0.5$. Since the posterior is strictly increasing, this means that $p(\xi(\theta)) < p(\xi(\theta^*)) = 0.5$, so the above expression is negative. This means that the firm will not switch to $\xi(\theta)$.

Moreover, this establishes that $\mathcal{F}(\theta*) = [\theta^*, \infty)$.

$[\theta, \theta]$ is not in equilibrium if $\xi(\theta^*)$ does not satisfy (8). If $\theta < \theta^*$, then the firm earns utility

$$\frac{1}{2}\left(\int_{\xi\theta}^{\infty} p_{\text{orig}}(x)(p(x)-(1-p(x))\right)\mathrm{d}x\,,$$

which we already showed above was negative. Thus, the firm has incentive to change their threshold to above θ so that it loses the full participant base and gets 0 utility.

If $\theta > \theta^*$, then the firm earns utility

$$U = \frac{1}{2}\left(\int_{\xi(\theta)}^{\infty} p_{\text{orig}}(x)(p(x)-(1-p(x))\right)\mathrm{d}x\,,$$

which is strictly positive. Fix $\epsilon > 0$, and suppose that the firm changes to a threshold θ' such that $c(\theta', \theta) = \epsilon$. Then it would gain all of the participants and earn utility:

$$\begin{aligned}
\int_{\xi(\theta')}^{\infty} p_{\text{orig}}(x)(p(x)-(1-p(x))\mathrm{d}x &= \int_{\xi(\theta')}^{\xi\theta} p_{\text{orig}}(x)(p(x)-(1-p(x))\mathrm{d}x \\
&\qquad + \int_{\xi(\theta)}^{\infty} p_{\text{orig}}(x)(p(x)-(1-p(x))\mathrm{d}x \\
&= \int_{\xi(\theta')}^{\xi(\theta)} p_{\text{orig}}(x)(p(x)-(1-p(x))\mathrm{d}x + 2U.
\end{aligned}$$

We claim that this expression approaches $2U$ as $\epsilon \to 0$. To see this, note that $c(\xi(\theta'), \theta) \to \gamma$ and so $\xi(\theta') \to \xi(\theta)$ as $\epsilon \to 0$. This implies that $\int_{\xi(\theta')}^{\xi(\theta)} p_{\text{orig}}(x)(p(x)-(1-p(x))\mathrm{d}x \to 0$ as desired. Thus, the expression approaches $2U > U$ as desired. This means that there exists ϵ such that the firm changing to θ' results in a strict improvement in utility. □

C.12 Proof of Theorem 1

Proof. Recall the definition of the action set $\mathcal{S}$. We prove Theorem 1 by constructing a $s_{\text{swap}} \in \mathcal{S}$ and relating the effect of a change in the score function from s_{curr} to s_{swap} to the causal effect of position.

For $u \in \mathcal{U}$ let $i_1(u)$ and $i_2(u)$ denote the index of the content item shown to user u under s_{curr} in the first and second display slot, respectively. Now, let the score function s_{swap} be such that the content displayed in the first two display slots is swapped relative to s_{curr}, simultaneously for all users $u \in \mathcal{U}$:

$$\tilde{s}(u)[i] = \begin{cases} s_{\text{curr}}(u)[i_2(u)] & i = i_1(u) \\ s_{\text{curr}}(u)[i_1(u)] & i = i_2(u) \\ s_{\text{curr}}(u)[i] & \text{otherwise.} \end{cases} \tag{9}$$

It holds that $s_{\text{swap}} \in \mathcal{S}$, since $|s_{\text{curr}}(u)[i_1(u)] - s_{\text{curr}}(u)[i_2(u)]| \leq \delta$ for all $u \in \mathcal{U}$. We lower bound performative power as

$$\mathrm{P} = \sup_{s \in \mathcal{S}} \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}\left[\|z(u) - z_s(u)\|_1\right] \geq \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}\left[\|z(u) - z_{s_{\text{swap}}}(u)\|_1\right] \tag{10}$$

To bould the difference between the counterfactual variable $z(u)$ and $z_{_{\text{swap}}}(u)$, we decompose s_{swap} into a series of unilateral *swapped* score functions, one for each viewer. The score function s^u_{swap} associated with viewer u swaps the scores of content that currently appears in the first two display slots for viewer u and keeps the scores of the other viewers unchanged.

Assumption 2 implies that $z_{s_{\text{swap}}}(u) = z_{s^u_{\text{swap}}}(u)$, since there are no peer effects; $z_{s_{\text{swap}}}(u)$ is independent of $s_{\text{swap}}(u')$ for $u' \neq u$. Thus, we can aggregate the unilateral effects across all viewers $u \in \mathcal{U}$ to obtain the effect of s_{swap} as:

$$\mathrm{P} \geq \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}\left[\|z(u) - z_{s_{\text{swap}}}(u)\|_1\right] = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}\left[\|z(u) - z_{s^u_{\text{swap}}}(u)\|_1\right] \tag{11}$$

Reasoning about unilateral effects allows us to relate the summands in (11) to the causal effect of position. In particular, focus on coordinate $i_1(u)$ in the norm, and let $Y_0(u) = z(u)[i_1(i)]$ and $Y_1(u) = z_{s^u_{\text{swap}}}(u)[i_1(u)]$. Then, we have

$$\mathrm{P} \geq \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}\left|z(u)[i_1] - z_{\tilde{s}_u}(u)[i_1]\right| = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}\left|Y_0(u) - Y_1(u)\right| = \beta.$$

where the causal effect of position β is defined as in Definition 2. □