Regulating algorithmic filtering on social media

Sarah H. Cen
Massachusetts Institute of Technology
Cambridge, MA 02139
shcen@mit.edu

Devavrat Shah
Massachusetts Institute of Technology
Cambridge, MA 02139
devavrat@mit.edu

Abstract

Social media platforms moderate content using a process known as algorithmic filtering (AF). While AF has the potential to greatly improve the user experience, it has also drawn scrutiny for its roles in spreading fake news, amplifying hate speech, facilitating digital redlining, and more. Although these externalities have sparked calls for regulations, some worry that regulations will interfere with personalization, lower profits, or restrict free speech. In this work, we are interested in whether it is possible to regulate without hurting performance or removing content. We build on the notion of regulating with respect to an implicit (social) contract that we term the *consumer-provider agreement*. We show that this contract-based regulation can be formalized as a hypothesis test and prove that this framework has desirable statistical guarantees on the platform's behavior and user's decision-making. We then show that there are conditions under which the regulation imposes little to no long-term cost on the platform. This surprising result is due to the fact that, in the presence of uncertainty, the objectives of the regulator and the platform can be partially aligned. Although not explicit objectives of the regulation, two additional benefits are that it encourages content diversity and does not remove content.

1 Introduction

Social media platforms moderate the content that users see on their feeds using a process known as *algorithmic filtering* (AF). While AF is a powerful tool with the potential to greatly improve the user experience, it has also begun to draw intense scrutiny. In particular, due to the growing popularity of social media and the ability of algorithms to moderate content on large scales, AF has played an active role in contemporary sociopolitical issues. Recently, it has been linked to the spread of fake news via algorithmic favoritism [17, 19], the amplification of hate speech via content ranking [49], and the facilitation of digital redlining via discriminatory targeted advertising [12]. These unintended side effects—or externalities—have sparked calls for social media regulations [29].

However, many fear that regulations may be harmful to the social media ecosystem. For example, some believe that the only remedy to misinformation or hate speech is content removal but that such measures restrict free speech [15]. Others believe that regulations hinder system performance [24], thereby leading to worse personalization for the user and lower revenue for the platform.

In turn, we ask: (a) Is it possible to filter content in a way that balances these concerns? (b) If so, what is the appropriate regulatory framework? (c) How does it affect the user's content and platform's revenue? To answer these questions, we build on the notion of an implicit (social) contract, as follows.

Consumer-provider agreement. In contrast to approaches in which the regulation reflects a notion of right versus wrong, we consider the relationship between the user and platform as one between a consumer and service provider. Drawing from a field of economics known as *implicit contract theory* [4, 46], two parties that voluntarily exchange goods and services are governed by an implicit contract. This contract is used to explain behaviors that cannot be predicted by competitive market theory and

Machine Learning for Economic Policy Workshop at NeurIPS 2020, Virtual.

typically reflect unspoken social norms. For example, it has been used to explain why a firm wishing to lower costs would layoff workers instead of lowering the wages of all workers.

When one party fails to uphold its end of the contract—which we call the *consumer-provider agreement*—the other is, in theory, free to terminate the relationship. However, this freedom does not exist when the cost of terminating is much higher for one party due to a power imbalance. In this work, the role of a regulation is to *correct this power imbalance by enforcing the consumer-provider agreement*. In law, implied contracts often hold the same legal force as express ones [48, 26].

Implicit contract theory can be applied to social media by observing that the platform provides the user with access to goods (e.g., social connections, news, recommendations) in exchange for the user's membership and data. While the user upholds her end of the contract simply by using the platform, there is no appraisal for whether the platform upholds its end (especially when there is a lack of competition). Therefore, the objective of a regulation would be to determine whether the platform honors the consumer-provider agreement. When the platform upholds the agreement, we call the platform's behaviors *responsible*.

Contributions. In this work, we use the consumer-provider agreement to address the three questions above. Our main contributions are theoretical: to establish that AF can be both *responsible* and *high-reward* without removing content, as follows:

1. We propose a statistical framework for enforcing the consumer -provider agreement. We call this framework a *regulation*, and we call feeds that satisfy regulation *feasible*.
2. We prove that the regulatory procedure is equivalent to hypothesis testing and that it has desirable properties, including holding the platform accountable to users and moderating the impact of filtering on user decision-making.
3. We show that, surprisingly, the regulation can impose little to no long-term cost on the platform and does so without removing content. One of the main mechanisms that the platform can use to lower cost is increasing content diversity. In other words, under regulation, the platform is incentivized to filter feeds with sufficiently high content diversity.

This work is relevant to the Machine Learning for Economic Policy NeurIPS 2020 workshop, as it uses tools from machine learning and statistical inference to study socio-economic issues stemming from AF on social media. We propose a regulatory framework that draws inspiration from concepts in economics (e.g., implicit contracts, the use of regulatory measures to address market failures, incentive alignment). A discussion of the related work and all proofs can be found in the Appendix.

2 Problem statement

Consider a specific user and platform (this statement holds for the remainder of the work). At each time step t, the platform shows the user a collection of content known as the *filtered feed*, which we denote by $Z_F^{(t)} \in \mathcal{F}$. The platform chooses the filtered feed based on the platform's objective function $\text{Rew} : \mathcal{F} \times \mathcal{X} \to \mathbb{R}$, such that $Z_F^{(t)} = \arg\max_{Z \in \mathcal{F}} \text{Rew}(Z, X^{(t)})$, where $X^{(t)} \in \mathcal{X}$ contains all other information that affects the platform's decision at time t. The reward function Rew reflects the platform's objective. For example, it may balance factors like personalization, user engagement (e.g., clicks), advertising revenue, and cost of operations.

Regulator's objective. The regulator's goal is to determine whether the platform upholds the consumer-provider agreement by observing the filtered feeds $Z_F^{(t)}$ over some period $t = 1, \dots, T$. To do so, the consumer-provider agreement must be translated into a usable form. In this work, this translational step takes the form of the *natural feed*, which we denote by $Z_N^{(t)}$. The natural feed is an alternative feed that the platform could have filtered if it adhered strictly to the consumer-provider agreement. Consider an example.

Example. In this *simplified* example, suppose that the platform's reward function has three terms: $\text{Rew}(\cdot) = \text{Rew}_{\text{per}}(\cdot) + \text{Rew}_{\text{ad}}(\cdot) + \text{Rew}_{\text{exp}}(\cdot)$, where $\text{Rew}_{\text{per}}(Z, X^{(t)})$ predicts how well Z is personalized to the user, $\text{Rew}_{\text{ad}}(Z, X^{(t)})$ predicts the advertising revenue Z would accrue, and $\text{Rew}_{\text{exp}}(\cdot)$ predicts the reward associated with the information the platform would gain from running a social experiment on the user [30]. Suppose, solely for this example, that personalization and advertising revenue are consistent with the consumer-provider agreement but the social experiment is not. Then, one way of constructing the natural feed is by solving: $Z_N^{(t)} = \arg\max_{Z \in \mathcal{F}} \text{Rew}_{\text{per}}(Z, X^{(t)}) + \text{Rew}_{\text{ad}}(Z, X^{(t)})$. The pair $(Z_F^{(t)}, Z_N^{(t)})$ is then given to the regulator.

Example 2 gives one of many possibles method for constructing $Z_N^{(t)}$. Scholars in law and philosophy, among other fields, have begun to specify the terms of the implicit social contracts between users and platforms [47, 43, 37]. Such a task is nuanced because the contract is context-specific (e.g., depends on jurisprudence and social norms of the society under consideration). In this work, we leave the task of devising the terms of the contract to these scholars and instead demonstrate how, given such terms, a regulation that balances several concerns of the user, platform, and public could be carried out.

The regulator does *not* require the platform to filter only natural feeds, as such a regulation would be too restrictive. Rather, the role of the regulator can be summarized as follows.

Definition 1 (Regulator's objective)**.** The regulator seeks to determine if the platform honors the consumer-provider agreement by testing whether the filtered feeds $Z_F^{(t)}$ are indistinguishable from a set of natural feeds $Z_N^{(t)}$ over period $t \in [T]$ with high certainty.

Meeting regulation. Suppose that the regulator's rule in Definition 1 confines the platform's choice of feeds at time t to some feasible set $\bar{\mathcal{F}}^{(t)} \subset \mathcal{F}$. Then, the platform's revised objective from above is to maximize reward subject to regulation: $Z_F^{(t)} \in \arg\max_{Z \in \bar{\mathcal{F}}^{(t)}} \text{Rew}(Z, X^{(t)})$.

The two overarching questions of this work are: (1) How does regulation affect the platform's reward (i.e., what is the cost of regulation)? (2) How does regulation affect the user's filtered feed?

Our approach. We approach these questions by establishing the existence of a statistical test that enforces the consumer-provider agreement and can be low-cost. We study the regulation's effect on the user's content by analyzing how it influences the platform's filtering. Although proposing a regulation is part of our work, our main contributions are theoretical.

3 Regulatory framework

In this section, we present a statistical procedure for testing whether the platform's AF upholds the consumer-provider agreement, then prove that the regulation has two statistical guarantees. These properties imply that the regulation can be interpreted in two ways: as enforcing the consumer-provider agreement and as moderating the effect of AF on user decision-making.

3.1 Regulatory procedure

Let $Z^{(t)} = \{\mathbf{z}_1^{(t)}, \ldots, \mathbf{z}_m^{(t)}\} \in \mathcal{Z}^m$ denote a feed at time t, where $\mathbf{z}_i^{(t)} \in \mathcal{Z} \subset \mathbb{R}^\ell$ is a feature vector describing the i-th piece of content. We assume that $\mathbf{z}_i^{(t)} \overset{\text{i.i.d.}}{\sim} q_{\mathbf{Z}}^{(t)}(\cdot)$. Let $\mathcal{P} = \{p_{\mathbf{z}}(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ for $\Theta \in \mathbb{R}^n$ be the model family containing possible feeds. In other words, $q_{\mathbf{Z}}^{(t)}(\cdot) = p_{\mathbf{z}}(\cdot; \boldsymbol{\theta}^{(t)}) \in \mathcal{P}$ for some $\boldsymbol{\theta}^{(t)} \in \Theta$. Note that $\boldsymbol{\theta}^{(t)}$ can change with time (in fact, the more quickly it changes, the easier it is to meet some of the conditions in our results). We assume that $\mathcal{P}$ satisfies common regularity conditions, as specified in Appendix D.1.

Let $Z_F^{(t)}$ and $Z_N^{(t)}$ denote the filtered and natural feeds at time t, as described in Section 2. Let $\boldsymbol{\theta}_F^{(t)}, \boldsymbol{\theta}_N^{(t)} \in \Theta$ denote the latent (unknown) parameters of the respective feeds. Let $\tilde{\boldsymbol{\theta}} : \mathcal{Z}^m \to \Theta$ denote the maximum likelihood estimator (MLE).

Let $\epsilon \in [0, 1]$ be the regulation parameter. This parameter governs the regulation strictness, where higher values indicate greater strictness. Let T be the time horizon, which determines how far into the past the regulator scrutinizes the platform's behavior. Then, based on the content over period T, the regulator (or self-regulator) decides $\tilde{H} = H_1$ if the platform should be investigated and $\tilde{H} = H_0$, otherwise. The corresponding (ϵ, T)-regulatory procedure is given in Algorithm 1.

This algorithm can be summarized by the decision rule:

$$\tilde{H} = H_1 \iff g(\tilde{M}_N, \tilde{M}_F) > \delta(\epsilon, n, T), \tag{1}$$

where $\tilde{M} := \{\tilde{\boldsymbol{\theta}}_\tau(Z^{(t)})\}_{t=1}^T$, $g(\tilde{M}_N, \tilde{M}_F) := \tilde{\mathbf{m}}^\top W^+ \tilde{\mathbf{m}}$ captures lines 2-6 in Algorithm 1, $\delta(\epsilon, n, T) := n/(T-n) \cdot F_\epsilon(n, T-n)$, and $F_\epsilon(n, T-n)$ is defined such that $\mathbb{P}(w \leq F_\epsilon(n, T-n)) = 1 - \epsilon$, where w is a random variable distributed according to the F-distribution with parameters n and $T - n$. Note that $\delta(\epsilon, n, T)$ is decreasing in ϵ.

Algorithm 1: (ϵ, T)-regulatory procedure

```
Input: Regulation parameter $\epsilon \in [0,1]$; time horizon $T < \infty$; model family $\Theta \subset \mathbb{R}^n$; filtered
       and natural feeds $\{Z_F^{(t)}\}_{t=1}^T$ and $\{Z_N^{(t)}\}_{t=1}^T$, where $|Z_F^{(t)}| = |Z_N^{(t)}| = m \forall t$ and $m, T > n$.
Result: Decision $\tilde{H} \in \{H_1, H_0\}$ corresponding to the decisions to and not to investigate the
        platform for a possible AF regulation violation, respectively.
Determine the MLE $\tilde{\boldsymbol{\theta}}(\cdot)$;
for $t = 1, \ldots, T$ do
    $\tilde{\mathbf{s}}^{(t)} = \tilde{\boldsymbol{\theta}}(Z_N^{(t)}) - \tilde{\boldsymbol{\theta}}(Z_F^{(t)})$;
end
Compute mean: $\tilde{\mathbf{m}} = \frac{1}{T} \sum_{t=1}^T \tilde{\mathbf{s}}^{(t)}$ ;
Compute covariance: $W = \frac{1}{T} \sum_{t=1}^T (\tilde{\mathbf{s}}^{(t)} - \tilde{\mathbf{m}})(\tilde{\mathbf{s}}^{(t)} - \tilde{\mathbf{m}})^\top$;
if $\tilde{\mathbf{m}}^\top W^+ \tilde{\mathbf{m}} \leq n/(T-n) \cdot F_\epsilon(n, T-n)$ then
    $\tilde{H} = H_0$;
else
    $\tilde{H} = H_1$;
end
```

3.2 Statistical guarantees of regulation

In this section, we prove that the when the (ϵ, T)-regulation returns $\tilde{H} = H_0$, then over the last T time steps, the results can be interpreted in two ways: (a) The regulator cannot determine with $(1-\epsilon)$-confidence that the platform has broken the consumer-provider agreement; or (b) The user's decision-making under the filtered feed is ϵ-close to her decision-making under the natural feed.

Interpretation (a): *Enforcing the consumer-provider agreement.* This interpretation takes the perspective of the regulator (or self-regulator). Let the regulator decide between two hypotheses:

$$H_0 : \boldsymbol{\theta}_F^{(1:T)} = \boldsymbol{\theta}_N^{(1:T)} \qquad H_1 : \boldsymbol{\theta}_F^{(1:T)} \neq \boldsymbol{\theta}_N^{(1:T)} \tag{2}$$

where H_0 is the null hypothesis, H_1 is the alternate hypothesis, and the regulator assumes that $(\boldsymbol{\theta}_N^{(t)} - \boldsymbol{\theta}_F^{(t)})$ are i.i.d. in time t. Let $\hat{H} \in \{H_0, H_1\}$ denote a decision rule for (2). In practice, the regulator first observes the filtered and natural feeds, then makes a decision $\hat{H}$. The regulator only rejects the null hypothesis (i.e., $\hat{H} = H_1$) with high confidence when there is enough evidence that the filtered content comes from a different generative model than the natural content. This outcome indicates that the platform has deviated significantly from the consumer-provider agreement.

Note that (2) is invariant to one-to-one transformations $g : \Theta \to \Theta$. An invariant hypothesis test for (2) is a test that is indifferent to (achieves the same performance under) such transformations. Under (2), we have the following result.

Proposition 1. Consider the decision $\tilde{H}$ returned by Algorithm 1. Among all invariant level-ϵ tests for (2), $\tilde{H}$ is the uniformly most powerful as $m, T \to \infty$. Formally, for any invariant test $\hat{H}$ of (2) for which the false positive rate (FPR) is $\mathbb{P}(\hat{H} = H_1 | H = H_0) \leq \epsilon$,

$$\mathbb{P}(\hat{H} = H_1 | H = H_1) \leq \mathbb{P}(\tilde{H} = H_1 | H = H_1) \text{ as } m, T \to \infty. \tag{3}$$

This result says that, if Algorithm 1 returns $\tilde{H} = H_0$, then the regulator can be $(1-\epsilon)$-confident the platform's filtering is responsible (i.e., upholds the consumer-provider agreement). In other words, the filtered feed is ϵ-indistinguishable from a natural feed. Specifically, (3) says that Algorithm 1 is the asymptotically most powerful among all invariant tests with a FPR of at most ϵ. The most powerful test is the one that is mostly likely to detect a filtering violation if there is one. Restricting tests to have a FPR no greater than ϵ mirrors the idea of having enough evidence before prosecuting.

Interpretation (b): *Moderating the influence of filtering on user decision-making*. Let $\hat{A}^{(t)} \in \{A_0, A_1\}$ denote an individual's decision between two possible actions at time t. Consider the general form of an individual's (i.e., user's) decision-making process:

$$\text{If } v_0(\hat{M}) \geq v_1(\hat{M}),\ \hat{A}^{(t)} = A_0. \text{ Else, } \hat{A}^{(t)} = A_1. \tag{4}$$

where $\hat{M} = \{\hat{\boldsymbol{\theta}}_\tau(Z^{(\tau)})\}_{\tau=1}^{t}$. Intuitively, the individual forms a belief $\hat{\boldsymbol{\theta}}_\tau(Z^{(\tau)})$ after observing information $Z^{(\tau)}$, where $\hat{\boldsymbol{\theta}}_\tau : \mathcal{Z}^m \to \Theta$ captures the user's learning behavior at time τ. If, according to her cumulative beliefs at time t, the value (or utility) v_0 of taking action A_0 is greater the value v_1 of A_1, then $\hat{A}^{(t)} = A_0$, and vice versa.

This decision-making model is very general for the following reasons: (1) the decisions are not limited to the social media context; (2) the individual's belief could depend on other factors, which are suppressed into $\hat{\boldsymbol{\theta}}_\tau(\cdot)$; (3) $v_0, v_1 : \Theta \to \mathbb{R}$ are arbitrary and could represent any value assignment function; and (4) all decisions between a finite number of choices can be written as a sequence of binary decisions. It is important to note that our result *do not assume that we observe the user's decision-making process*, simply that it exists in the form (4).

Example. Let $v_1 = r_1$ and $v_0 = r_0 + a$, where r_i is the individual's estimate of the reward associated with action i and $a >> 0$. Here, the individual is strongly predisposed to action A_0, taking action A_1 only if its proves to be much more rewarding.

Example. Suppose an individual is deciding what to order at a restaurant. Her current favorite dish is entree E, which she has ordered many times, but there are other dishes on the menu that she has not yet tried. Since she is unsure whether E is better than all other dishes, she occasionally orders something new. This strategy captures the explore-exploit trade-off of reinforcement learning. In (4), it can be modeled by letting v_i be the upper confidence bound (UCB) [38] associated with action i.

We are interested in whether a user's decisions under her filtered feeds are consistently different from her decisions under the natural feeds. In other words, the regulation does not penalize the platform for shaping a user's behaviors as such an influence is inevitable. Rather, the regulation requires that the platform does not use its role as an information gatekeeper to significantly change the user's behavior in a way that breaks the consumer-provider agreement. Let $\hat{A}_{F}^{(t)}$ denote the action that the individual chooses after observing the filtered feeds $Z_{F}^{(1:t)}$. Let $\hat{A}_{N}^{(t)}$ be defined analogously. Then, the filtered feeds induce a different decision from that induced by the natural feeds if $\hat{A}_{F}^{(t)} \neq \hat{A}_{N}^{(t)}$.

Proposition 2. Consider any decision of the form (4) at time T. Define $f(\cdot) = v_0(\cdot) - v_1(\cdot)$. Under the assumption that $\exists L \in \mathbb{R}_{\geq 0}$ such that $|f(M_1) - f(M_2)| \leq Lg(M_1, M_2)$ for all $M_1, M_2 \in \Theta^T$, then for any estimator $\hat{\boldsymbol{\theta}} : \mathcal{Z}^m \to \Theta$ and as $m, T \to \infty$, the regulation's decision $\tilde{H} = H_0$ implies that: $\mathbb{P}\big(-1 \leq \frac{f(\hat{M}_F) - f(\hat{M}_N)}{L\delta(\epsilon,n,T)} \leq 1\big) \to 1$.

This result states that, if the platform passes regulation ($\tilde{H} = H_0$), then the user's decisions under the filtered feeds are asymptotically ϵ-similar to the decisions the user would have made if the platform upheld the consumer-provider agreement (i.e., had she observed the natural feeds instead). Notably, this guarantee holds irrespective of the user's learning behaviors $\hat{\boldsymbol{\theta}}_\tau$.

Corollary 3. Consider the same setup as Proposition 2. If $|f(\hat{M}_N)| > L\delta(\epsilon, n, T)$, then the regulation's decision $\tilde{H} = H_0$ implies that $\mathbb{P}(\hat{A}_{N}^{(T)} = \hat{A}_{F}^{(T)}) \to 1$ as $m, T \to \infty$ for all $\{\hat{\boldsymbol{\theta}}_t\}_{t=1}^{T}$.

This results says that, if the user strongly favors action A_i under the natural feed, then the regulation ensures that the user takes action A_i under the filtered feed as $m, T \to \infty$.

4 Effect of regulation of the feed

In this section, we study how the regulation affects the user's feed. We show that:

- The regulation requires that the filtered feed is close in distribution to the natural feed, where it is possible to interpret closeness in terms of the information (Kullback-Leibler) divergence between the feeds. Since the natural feed is derived from the consumer-provider agreement, this implies that the platform is held more accountable to the user.
- The platform is more likely to pass regulation when it is difficult to distinguish the natural and filtered feeds with high certainty. One way the platform can induce this outcome is by increasing content diversity, meaning the regulation (inadvertently) incentivizes diversity.

We express the remaining results in terms of the generative parameters $\boldsymbol{\theta}_N$ and $\boldsymbol{\theta}_F$. As discussed in Section 3.1, $\boldsymbol{\theta}_N$ and $\boldsymbol{\theta}_F$ need not exist in practice because the platform can directly tune the feed Z_F. However, in our analysis, $\boldsymbol{\theta}_N$ and $\boldsymbol{\theta}_F$ are helpful analytical tools that illustrate how filtering schemes fare under regulation. For more intuition on the parameters $\boldsymbol{\theta} \in \Theta$, see Appendix D.1.

4.1 Feasible feed

A *feasible* feed at time step t is a filtered feed $Z_F^{(t)}$ that passes regulation. Note that setting $Z_F^{(t)} = Z_N^{(t)}$ for all time steps t always passes regulation. We call this choice of feeds the *trivial* solution. However, the platform usually prefers to have the flexibility of constructing $Z_F^{(t)}$ to be different from $Z_N^{(t)}$. In this section, we examine how the regulation limits the platform's choices of non-trivial feeds. As discussed at the start of Section 4, we turn our attention to $\boldsymbol{\theta}_N$ and $\boldsymbol{\theta}_F$. The corresponding notion of feasibility is captured in the following definition.

Definition 2. Given natural feeds $\boldsymbol{\theta}_N^{(1:T)}$ and past filtered feeds $\boldsymbol{\theta}_F^{(1:T-1)}$, the *$\alpha$-feasible set* at time T is: $\Omega_\alpha^{(T)} = \{\boldsymbol{\theta}_F^{(T)} \in \Theta : \mathbb{P}(g(\tilde{M}_N, \tilde{M}_F) \leq \delta(\epsilon, n, T)) \geq 1 - \alpha\}$.

If the platform chooses the filtered feed parameters $\boldsymbol{\theta}_F^{(T)}$ from the feasible set $\Omega_\alpha^{(T)}$, then the probability of passing regulation is high. The platform should choose $\alpha \leq \epsilon$ in order to be confident in passing regulation. We drop the subscript and superscript when the meaning of Ω is clear.

The next results study the random quantity $g(\tilde{M}_N, \tilde{M}_F)$. First, we define: $\bar{\boldsymbol{\theta}}_D = \sum_{t=1}^T \boldsymbol{\theta}_D^{(t)}/T$ for $D \in \{F, N\}$, $\mathbf{s} = \bar{\boldsymbol{\theta}}_N - \bar{\boldsymbol{\theta}}_F$, $\Sigma = \sum_{t=1}^T (I^{-1}(\boldsymbol{\theta}_N^{(t)}) + I^{-1}(\boldsymbol{\theta}_F^{(t)}))/(mT^2)$, and $\bar{g} = \mathbf{s}^\top \Sigma^+ \mathbf{s}$, where $I(\boldsymbol{\theta})$ is the Fisher information matrix about $\boldsymbol{\theta}$. Recall that $I(\boldsymbol{\theta})$ can be viewed as measuring the speed with which one can learn $\boldsymbol{\theta}$ with high certainty from i.i.d. samples drawn from $p_{\mathbf{z}}(\cdot; \boldsymbol{\theta})$.

Proposition 4. Let $T - 1 > n \geq 3$ and Σ be positive-definite. Then, as $m \to \infty$,

$$g(\tilde{M}_N, \tilde{M}_F) \overset{d}{=} Q(\bar{g}) := \frac{n}{T-n} u_1 + \frac{\sqrt{\bar{g}}}{u_2} \left(\sqrt{\bar{g}} + 2u_3 \sqrt{1 + \frac{n-1}{T-n+1} u_4} \right),$$

where $u_1 \sim F(n, T-n)$, $u_2 \sim \chi^2(T-n)$, $u_3 \sim \mathcal{N}(F0, 1)$, and $u_4 \sim F(\frac{n-1}{2}, \frac{T+1-n}{2})$ are mutually independent. As a result, the feasible set can alternatively be defined as: $\Omega_\alpha^{(T)} = \{\boldsymbol{\theta}_F^{(T)} \in \Theta : \bar{g} \leq G_\alpha\}$, where $\mathbb{P}(Q(G_\alpha) > \delta(\epsilon, n, T)) = \alpha$.

Proposition 4 reduces the analysis of the feasible set to the quantity $\bar{g}$. The smaller its value, the more likely regulation is met. The next result illustrates when $\bar{g}$ is small. Let $||A||_2$ denote the spectral norm of matrix A. Let $d_{\text{KL}}(\cdot||\cdot)$ be the Kullback-Leibler (KL) divergence.

Corollary 5. Under the same conditions as in Proposition 4,

(i) $\bar{g}$ is upper bounded by: $\bar{g} \leq G := \frac{mT^2 ||\mathbf{s}||_2^2}{\sum_{t=1}^T (||I(\boldsymbol{\theta}_N^{(t)})||_2^{-1} + ||I(\boldsymbol{\theta}_F^{(t)})||_2^{-1})}$.

(ii) Suppose the Fisher information matrix is continuous in $\boldsymbol{\theta}$. Then, as $\boldsymbol{\theta}_F^{(t)}, \boldsymbol{\theta}_F^{(t)} \to \boldsymbol{\theta}$ for all $t \in [T]$, $|\bar{g} - mTd_{\text{KL}}(p_{\mathbf{z}}(\cdot; \bar{\boldsymbol{\theta}}_N)||p_{\mathbf{z}}(\cdot; \bar{\boldsymbol{\theta}}_F))|, |\bar{g} - mTd_{\text{KL}}(p_{\mathbf{z}}(\cdot; \bar{\boldsymbol{\theta}}_F)||p_{\mathbf{z}}(\cdot; \bar{\boldsymbol{\theta}}_N))| \to 0$.

Since regulation is met when $\bar{g} \leq G_\alpha$, Corollary 5(i) implies that the platform can ensure that its filtered feed is feasible by setting $G \leq G_\alpha$. For this reason, we call G the *design parameter*. In summary, the α-feasible set is the set of filtered feed parameters for which $\bar{g} \leq G_\alpha$. By Corollary 5, $\bar{g}$ is small (and therefore regulation is met) when:

- $||\mathbf{s}||_2^2$ is small: $\boldsymbol{\theta}_F^{(t)}$ is close to $\boldsymbol{\theta}_N^{(t)}$ on average.
- $||I(\boldsymbol{\theta}_N^{(t)})||_2$ and $||I(\boldsymbol{\theta}_F^{(t)})||_2$ are small: The content is varied enough such that the regulator cannot select $\tilde{H} = H_1$ with high certainty until given a sufficiently many observations. We will see in the next section that this corresponds to having high content diversity.
- KL divergence between the feeds is small.

These results imply that the regulation requires sufficient distributional overlap between the user's natural and filtered feeds. Because the natural feed is the feed that the platform would have shown under the strict consumer-provider agreement, these results show that the *regulation holds the platform more accountable to the user and gives the user more agency over her feed.*

4.2 Content diversity

In this section, we study how the regulation affects content diversity. Although diversity is not one of the regulation's explicit objectives, we show in Section 5 that the regulation inadvertently incentivizes the platform to filter content with sufficiently high diversity. We define content diversity as follows.

Definition 3. The *content diversity* of the generative model $p_{\mathbf{z}}(\cdot; \boldsymbol{\theta})$ is $D(\boldsymbol{\theta}) = ||I(\boldsymbol{\theta})||_2^{-1}$.

This definition can be interpreted in two ways. First, the content diversity of a feed generated by $\boldsymbol{\theta}$ can be viewed as how much content that the regulator must see from that feed in order to distinguish it from other feeds $\boldsymbol{\theta}' \neq \boldsymbol{\theta}$. Alternatively, it can be viewed as the amount of content the user must see in order to develop a strong opinion (i.e., to believe in a model $\boldsymbol{\theta}$ with high certainty).

Let $H(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = -\mathbb{E}_{p(\cdot;\boldsymbol{\theta}_1)}[\log p(\cdot; \boldsymbol{\theta}_2)]$ denote the cross entropy between the generative models parameterized by $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$. Let the information entropy be denoted by $H(\boldsymbol{\theta}) = H(\boldsymbol{\theta}, \boldsymbol{\theta})$.

Lemma 6. Let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$. Then, $D(\boldsymbol{\theta}_1) \leq \frac{1}{2}(H(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) - H(\boldsymbol{\theta}_1) + O(||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^3)^{-1}$.

In other words, content diversity can be interpreted as a local measure of information entropy. With slight abuse of notation, let $D(\boldsymbol{\theta}^{(1:T)}) = \sum_{t=1}^{T} D(\boldsymbol{\theta}^{(t)})$.

Corollary 7. If $D(\boldsymbol{\theta}_F^{(1:T)}) \geq \frac{mT^2||\mathbf{s}||_2^2}{G_\alpha} - D(\boldsymbol{\theta}_N^{(1:T)})$, then $\boldsymbol{\theta}_F^{(T)} \in \Omega_\alpha^{(T)}$.

This result says that meeting regulation can be framed as having sufficiently high content diversity. In other words, content diversity is a mechanism for meeting regulation. Even so, there is a limit to how much diversity can be increased because changes in $\boldsymbol{\theta}_F^{(1:T)}$ also affect $||\mathbf{s}||_2^2$. The platform must balance both increases in diversity $D(\boldsymbol{\theta}_F^{(1:T)})$ and increases in $||\mathbf{s}||_2^2$ in order to meet Corollary 7. This balancing act is one of the main factors that determines the cost of regulation, as examined next.

5 The cost of regulation

In this section, we turn from the regulator's perspective to the platform's perspective by discussing the effect of regulation on the platform's reward (e.g., profit or performance). We show that:

- The platform can lower the cost of regulation by adding enough content diversity. In other words, under regulation, the platform is incentivized to add diversity to the filtered feed.
- There are conditions under which the platform can incur little to no cost under regulation. Intuitively, requiring that the platform respects the user's expressed preferences (via closeness to the natural feed) and encouraging the platform to explore the user's interests (via sufficiently high content diversity) allows the platform to perform well.

5.1 Reward maximization

Recall from Section 2 that the platform's objective can be formulated as a constrained optimization problem. As done in Section 4, we turn our attention to $\boldsymbol{\theta}_N$ and $\boldsymbol{\theta}_F$. In accordance, the platform's short-term objective at time t is given by:

$$\boldsymbol{\theta}_F^{(t)} \in \arg\max_{\boldsymbol{\theta} \in \Omega} R(\boldsymbol{\theta}, \boldsymbol{\theta}_N^{(t)}), \tag{5}$$

where $R : \Theta \times \Theta \to \mathbb{R}$ is the platform's reward function. There may be other information that factors into the platform's decision at time t, and we suppress this information into the function R to keep notation concise. Under this definition, the cost of regulation is defined as follows.

Definition 4. Under (5), the *cost of regulation* is: $C_\Omega^{(t)} = \max_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}, \boldsymbol{\theta}_N^{(t)}) - \max_{\boldsymbol{\theta} \in \Omega} R(\boldsymbol{\theta}, \boldsymbol{\theta}_N^{(t)})$.

The cost of regulation is the amount of short-term reward that is lost when the platform follows regulation. Increasing the size of the feasible set $\Omega_\alpha^{(T)}$ cannot increase the cost of regulation, which implies that the cost is generally low when the feasible set is large. We are interested in the cost of regulation because regulations can be difficult to pass and enforce when they impose a high cost.

5.2 Relationship between reward and diversity

In this section, we show that there are mild conditions under which the platform can increase its reward by increasing content diversity. Put differently, under regulation, the platform is incentivized to filter feeds with a sufficiently high level of content diversity. We begin with a definition.

Definition 5. Given natural feed $\boldsymbol{\theta}_N$, $\bar{\mathbf{s}}$ is a β-local direction of ascent (LDA) at $\boldsymbol{\theta}_F \in \Theta$ if $R(\boldsymbol{\theta}_F + \gamma\bar{\mathbf{s}}, \boldsymbol{\theta}_N) > R(\boldsymbol{\theta}_F, \boldsymbol{\theta}_N)$ for all $\gamma \in (0, \beta)$, $||\bar{\mathbf{s}}||_2^2 = 1$, and $\boldsymbol{\theta}_F + \gamma\bar{\mathbf{s}} \in \Theta$.

Intuitively, for a given natural feed $\boldsymbol{\theta}_N$, starting at the filtered feed $\boldsymbol{\theta}_F$ and traveling along the LDA for some interval always brings the platform to a feed that has higher reward. However, there is no guarantee that this new feed is feasible. The next result shows that, if the platform can maintain feasibility while increasing reward, it does so by adding content diversity.

Proposition 8. Consider (5) and natural feed $\boldsymbol{\theta}_N^{(T)}$ at time T. Let $\boldsymbol{\theta}_{F'}$ be a feasible feed at time T with the corresponding design parameter $G' \leq G_\alpha$. Suppose that there exists a β-LDA $\bar{\mathbf{s}}$ at $\boldsymbol{\theta}_{F'}$. Then, the following statements hold true:

(a) If $G' < G_\alpha$, then there exists $\gamma > 0$ such that $\boldsymbol{\theta}_{F''} = \boldsymbol{\theta}_{F'} + \gamma\bar{\mathbf{s}}$, $G'' \leq G_\alpha$, and $R(\boldsymbol{\theta}_{F''}, \boldsymbol{\theta}_N^{(T)}) > R(\boldsymbol{\theta}_{F'}, \boldsymbol{\theta}_N^{(T)})$.

(b) If $G' = G_\alpha$ and $\bar{\mathbf{s}}^\top \mathbf{s}' > 0$, then unless there exists $\gamma \in (0, \beta)$ such that $\boldsymbol{\theta}_{F''} = \boldsymbol{\theta}_{F'} + \gamma\bar{\mathbf{s}}$ and $D(\boldsymbol{\theta}_{F''}) > D(\boldsymbol{\theta}_{F'})$, the platform cannot increase its reward along the LDA $\bar{\mathbf{s}}$ while ensuring that new design parameter $G'' \leq G_\alpha$.

(c) Suppose $G' = G_\alpha$. If there exists $\gamma \in (0, \beta)$ and $\boldsymbol{\theta}_{F''} = \boldsymbol{\theta}_{F'} + \gamma\bar{\mathbf{s}}$ such that $D(\boldsymbol{\theta}_{F''}) > D(\boldsymbol{\theta}_{F'})$ and $D(\boldsymbol{\theta}_{F''}) > \gamma m T - D(\boldsymbol{\theta}_N^{(1:T)}) - D(\boldsymbol{\theta}_F^{(1:T-1)})$, then $G'' \leq G_\alpha$.

This result says that, given a feasible feed at time T, the platform's main mechanism for increasing reward (or lowering the cost of regulation) is adding content diversity. The first result (a) states that, if the design parameter G' is below the regulation threshold G_α, then the platform is free to increase its reward by traveling along the LDA. As long as an LDA exists at this new feed, then the platform can repeat the process in (a) until the design parameter reaches the regulation threshold G_α, which brings us to statements (b) and (c). Therefore, we can restrict our attention to the latter two results. The result in (b) states that, when $G' = G_\alpha$, the only way for the platform to increase its reward along LDA $\bar{\mathbf{s}}$ while meeting regulation is if it can also increase content diversity. The result in (c) refines (b) by stating explicit conditions under which this higher-reward feed is reachable. Intuitively, (c) says that it is easier for the platform to reach a higher reward at time T when the natural feeds and previous filtered feeds have sufficiently high content diversity.

The following corollary shows that it is possible for there to be no cost of regulation.

Corollary 9. Suppose that R is a strictly concave function of $||\mathbf{s}||_2^2$ such that the reward function can be written as $R(||\mathbf{s}||_2^2)$, where $\mathbf{s}$ is as defined in Section 4.1. Suppose $\boldsymbol{\theta}_F^* \in \arg\max_{\boldsymbol{\theta} \in \Theta} R(||\mathbf{s}||_2^2)$ and $||\mathbf{s}^*||_2^2 = \zeta < \infty$. If there exists $\zeta > 0$ and $\bar{\mathbf{s}}$ such that $\boldsymbol{\theta}_N^{(T)} + \zeta\bar{\mathbf{s}} \in \Theta$ and:

$$D(\boldsymbol{\theta}_N^{(T)} + \zeta\bar{\mathbf{s}}) \geq \frac{mT^2\zeta}{G_\alpha} - D(\boldsymbol{\theta}_N^{(1:T)}) - D(\boldsymbol{\theta}_F^{(1:T-1)}), \tag{6}$$

then there is no cost of regulation: $C_\Omega^{(T)} = 0$.

It is natural to ask whether the conditions of Corollary 9 are reasonable. To understand the condition under which there is no cost of regulation (6), consider the following example.

Example. Suppose that $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{>0}\}$ represents the family of univariate Gaussians. For simplicity, suppose that $mT^2/G_\alpha = 1$ and that the natural feeds are standard Gaussians: $\boldsymbol{\theta}_N^{(t)} = (0, 1)^\top$ for all $t \in [T]$. Recall that $D(\cdot) \geq 0$. Then, for any choice of filtered feed $\boldsymbol{\theta}_F = (0, \sigma_F^2)^\top$ where $\sigma^2 > 1$, $D(\boldsymbol{\theta}_F) = \sigma_F^2 \geq |1 - \sigma_F^2|$, which implies that (6) must hold true.

This example, though highly simplified, shows that one can find conditions under which the cost of regulation is low. Intuitively, there are two reasons the platform can achieve high reward under regulation. First, recall that the natural feed represents the consumer-provider agreement and, as a result, reflects user-centric performance metrics. Therefore, in order to accrue high reward for the given user, the platform's optimal feed should not be too far from the natural feed, which is consistent with the regulation. Second, the regulation allows the platform to deviate from the natural feed given enough content diversity. Using the understanding that showing the user a piece of content is akin to implicitly querying the user for her interest in that type of content by observing how she interacts with it, content diversity can be viewed as a mechanism for *exploration*: a way for the platform to learn the user's interests as they evolve in time. We illustrate this intuition in Appendix C.

6 Conclusion

In conclusion, we demonstrate that a regulatory framework for AF can be low-cost without removing content. Our results suggest that regulating with respect to an implicit (social) contract serves as a good starting point because it balances the interests of several stakeholders while acknowledging the responsibility that each member of the system has to the other. Several works have already begun to study and understand the implicit contracts between users and digital platforms [47, 43, 37].

Statement of original work

This work has not been published and is original work.

Broader impact

The aim of this work is to study the effects of algorithmic content filtering by social media platforms. In light of both the problems and benefits of social media, this work seeks to make two contributions that may have broader impact. Our first contribution is to propose a framework for regulating algorithmic filtering (AF) based on an implicit contract between the user and the platform. Because regulations have unanticipated effects on the system stakeholders, our second contribution is the study of this regulation on the user's content and the social media platform.

Our hope is that this work can contribute to ongoing conversations on social media platforms and provide a potential regulatory (or self-regulatory framework) for AF. We hope that, by showing that regulatory measures are not necessarily in conflict with the platform's performative objectives, the undesirable externalities caused by AF can be mitigated while preserving the benefits offered by social media. In this same spirit, we also discuss the implications of the proposed regulation on other concerns in social media, such as the preservation of free speech, the health of public discourse, and the amplification of filter bubbles.

Although a regulatory procedure is part of our proposal, our main contributions are theoretical: to establish the existence of a regulation that preserves the benefits of social media for multiple stakeholders. As such, we envision that others will be able to refine and improve upon our regulatory proposal, and we are open to further discussions on the ethical implications of our work.

Acknowledgments and Disclosure of Funding

This work is supported by a fellowship from the Massachusetts Institute of Technology (MIT) Institute for Data, Systems, and Society (IDSS) via the OCP group, the Chyn Duog Shiah Fellowship, and the Hugh Hampton Young Memorial Fellowship.

We would also like to thank Martin Copenhaver, Hussein Mozannar, Cynthia Dwork, and Christina Ilvento for their feedback and guidance during the development of this work.

References

[1] Daron Acemoglu, Munther A. Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian Learning in Social Networks. *The Review of Economic Studies*, 78(4):1201–1236, 2011.

[2] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[3] Julia Angwin, Ariana Tobin, and Madeleine Varner. Facebook (Still) Letting Housing Advertisers Exclude Users by Race, 2017.

[4] Costas Azariadis and Joseph E. Stiglitz. Implicit Contracts and Fixed Price Equilibria. *The Quarterly Journal of Economics*, pages 2–22, 1983.

[5] R. R. Bahadur. On Fisher's Bound for Asymptotic Variances. *The Annals of Mathematical Statistics*, 35(4):1545–1552, 1964.

[6] Abhijit Banerjee. A Simple Model of Herd Behavior. *The Quarterly Journal of Economics*, 107(3):797–817, 1992.

[7] BBC News. Social Media: How Can Governments Regulate It? `https://www.bbc.com/news/technology-47135058`, April 2019.

[8] Hal Berghel. Lies, Damn Lies, and Fake News. *Computer*, 50(2):80–85, 2017.

[9] Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and Applications of Robust Optimization. *SIAM Review*, 53(3):464–501, 2011.

[10] Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. The Price of Fairness. *Operations Research*, 59(1):17–31, 2011.

[11] Dimitris Bertsimas and Melvyn Sim. The Price of Robustness. *Operations Research*, 52(1):35–53, 2004.

[12] Jacob Parker Black. Facebook and the Future of Fair Housing Online. *Oklahoma Law Review*, 72:711–729, 2019.

[13] Taras Bodnar and Yarema Okhrin. On the product of inverse Wishart and normal distributions with applications to discriminant analysis and portfolio theory. *Scandinavian Journal of Statistics*, 38(2):311–331, 2011.

[14] Engin Bozdag and Jeroen van den Hoven. Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, 17(4):249–265, 2015.

[15] Valerie C. Brannon. Free Speech and the Regulation of Social Media Content, 2019.

[16] James Campbell, Avi Goldfarb, and Catherine Tucker. Privacy Regulation and Market Structure. *Journal of Economics & Management Strategy*, 24(1):47–73, 2015.

[17] Bobby Chesney and Danielle Citron. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107:1753–1820, 2019.

[18] Alexandra Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, 2017.

[19] Damian Collins, Clive Efford, Julie Elliott, Paul Farrelly, Simon Hart, Julian Knight, Ian C. Lucas, Brendan O'Hara, Rebecca Pow, Jo Stevens, and Giles Watling. Disinformation and 'fake news': Final Report, February 2019.

[20] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. AAAI, 2017.

[21] Morris H. DeGroot. Reaching a Consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[22] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284. Springer, 2006.

[23] Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2013.

[24] Benedict Evans. Regulating technology. `https://www.ben-evans.com/benedictevans/2020/7/23/regulating-technology`, July 2020.

[25] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268. Association for Computing Machinery, 2015.

[26] Steven W. Feldman. Statutes and Rules of Law as Implied Contract Terms: The Divergent Approaches and a Proposed Solution. *University of Pennsylvania Journal of Business Law*, 19:809–869, 2016.

[27] Seth Flaxman, Sharad Goel, and Justin M. Rao. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80(S1):298–320, March 2016.

[28] Joseph R. Gabriel and Steven M. Kay. On the Relationship Between the GLRT and UMPI Tests for the Detection of Signals With Unknown Parameters. *IEEE transactions on signal processing*, 53(11):4194–4203, 2005.

[29] Dipayan Ghosh. A New Digital Social Contract Is Coming for Silicon Valley. `https://hbr.org/2019/03/a-new-digital-social-contract-is-coming-for-silicon-valley`, March 2019.

[30] Vindu Goel. Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry. June 2014.

[31] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably Efficient Maximum Entropy Exploration. In *International Conference on Machine Learning*, pages 2681–2691, 2019.

[32] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, 21(2):191–207, 2018.

[33] Dennis D. Hirsch. The Law and Policy of Online Privacy: Regulation, Self-Regulation, or Co-rRegulation. *Seattle University Law Review*, 34:439, January 2011.

[34] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware Learning through Regularization Approach. In *Proceedings of 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

[35] Pauline T. Kim and Sharion Scott. Discrimination in Online Employment Recruiting. *St. Louis University Law Journal*, 63:93, 2018.

[36] Kate Klonick. The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review*, 131:1598–1670, 2017.

[37] Sanne Kruikemeier, Sophie C. Boerman, and Nadine Bol. Breaching the contract? Using social contract theory to explain individuals' online behavior to safeguard privacy. *Media Psychology*, 23(2):269–292, 2020.

[38] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[39] E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer-Verlag, New York, NY, USA, 2nd edition, 1998.

[40] E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Science & Business Media, 3rd edition, 2005.

[41] Becca Lewis and Alice E. Marwick. Media Manipulation and Disinformation Online. *New York: Data & Society Research Institute*, 2017.

[42] Alexander Ly, Maarten Marsman, Josine Verhagen, Raoul Grasman, and Eric-Jan Wagenmakers. A Tutorial on Fisher Information. *Journal of Mathematical Psychology*, 80:40–55, 2017.

[43] Kirsten Martin. Understanding Privacy Online: Development of a Social Contract Approach to Privacy. *Journal of Business Ethics*, 137(3):551–569, 2016.

[44] Pooya Molavi, Alireza Tahbaz-Salehi, and Ali Jadbabaie. A Theory of Non-Bayesian Social Learning. *Econometrica*, 86(2):445–490, March 2018.

[45] Jonathan A. Obar and Steven S. Wildman. Social Media Definition and the Governance Challenge. *Telecommunications Policy*, 39(9):745–750, 2015.

[46] Arthur M Okun. *Prices and Quantities: A Macroeconomic Analysis*. Brookings Institution Press, 2011.

[47] Iyad Rahwan. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, 20(1):5–14, 2018.

[48] Warren L. Shattuck. Contracts in Washington, 1937-1957. *Washington Law Review & State Bar Journal*, 34:24–77, 1959.

[49] Stefan Siersdorfer, Sergiu Chelaru, Jose San Pedro, Ismail Sengor Altingovde, and Wolfgang Nejdl. Analyzing and Mining Comments and Comment Ratings on the Social Web. *ACM Transactions on the Web*, 8(3), July 2014.

[50] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. Potential for Discrimination in Online Targeted Advertising. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 5–19, New York, NY, USA, 23–24 Feb 2018. PMLR.

[51] Latanya Sweeney. Discrimination in Online Ad Delivery. *Queue*, 11(3):10–29, March 2013.

[52] Larry Wasserman and Shuheng Zhou. A Statistical Framework for Differential Privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[53] Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally Robust Convex Optimization. *Operations Research*, 62(6):1358–1376, 2014.

[54] Bob Zimmer. Democracy Under Threat: Risk and Solutions in the Era of Disinformation and Data Monopoly. *Canada House of Commons*, December 2018. Report of the Standing Committee on Access to Information, Privacy and Ethics.

Appendix

A Consumer-provider agreement

In this section, we make some concluding remarks about the con-sumer-provider agreement.

Implicit contract theory. As explained in Section 1, the consumer-provider agreement captures the notion that, when two parties enter into a voluntary exchange of goods or services, there is an implicit contract between them. Historically, implicit contract theory has been used in economics to explain behaviors that cannot be predicted by competitive market theory and that typically reflect unspoken social norms.

When one party fails to uphold its end of the agreement, then the exchange of goods and services may cease. Because this outcome is generally undesirable for both parties, upholding the implicit contract can improve a firm's robustness. For example, although a firm may wish to choose policies (e.g., filter content) based on an optimization analysis, the firm's choices may be more robust when it also heeds unwritten social rules. This arises from the fact that an optimization problem cannot account for all possible factors in the system, and many of the unknown or unquantifiable factors are instead captured in social norms.

Consumer-provider agreement. In our setting, the social media platform provides the user with access to social connections, recommendations, news, and other information goods in exchange for the user's membership and data. This exchange gives rise to an implicit contract that we call the consumer-provider agreement.

The user upholds her end of the agreement simply by using the platform, but there is no appraisal for whether the platform upholds its end. The role of regulation would be to perform this appraisal. Although the user could leave the exchange, leaving often places a greater burden on the user than the platform. Especially in the absence of competition, leaving could mean giving up one's social connections. In this way, the purpose of regulation would be to correct this power imbalance by holding the platform more accountable to users according to the consumer-provider agreement.

Discussion of results. In this work, we translate the consumer-provider agreement to the social media setting using the *natural feed*. As shown in Section 3, the regulation requires that the platform's filtered feeds cannot be determined not to be natural with high certainty. In other words, that the filtered feed is indistinguishable from a natural feed with some confidence. In this way, the regulation holds the platform more accountable to the user because the content it filters must respect its implicit contract with the user.

However, we also showed in Section 5 that regulating with respect to the consumer-provider agreement is not too restrictive. In particular, there are conditions under which the cost of regulation is low, and there may even be long-term benefits to filtering close to the natural feed (as required under regulation).

One may wonder if the indistinguishability property required by regulation is too weak since the platform has enough freedom to achieve high reward. To understand why this property may be appropriate, we studied how the filtered content would affect the user's decision-making behavior. In short, the power of AF derives from the fact that it shapes the information that a user uses to make decisions. In accordance with this observation, we showed in Section 3.2 that the indistinguishability requirement implies that the user's decisions under the filtered feed are close to the decisions she would have made under the natural feed, no matter what heuristic the user uses to make decisions.

Finally, although content diversity is not an explicit objective of the regulation, we showed in Sections 4.2 and 5.2 that it incentivizes the platform to add content diversity to the user's feed. This side effect of the regulation implies that the platform is disincentivized from creating filter bubbles. Another benefit of content diversity is that it can achieve the same benefits as removing content (e.g., instead of removing articles that may contain misinformation on topic X, increasing content diversity would dilute misinformation by providing a more well-rounded representation of topic X), but it does not raise issues of restricting free speech.

Verdict. In conclusion, the consumer-provider agreement may be a good starting point for regulation because it balances the interests of the system stakeholders while acknowledging the responsibility that each parties has to the other. Several works have already begun to provide insights into implicit contracts between users and digital platforms [47, 43, 37].

We do not specify the terms of the consumer-provider agreement in this work. The terms of the consumer-provider agreement are context-specific in that they depend on the regulatory system, jurisprudence, and norms of the society in which the regulation would be deployed. The agreement also depends on the social media service under consideration, as the interactions (e.g., clicks or swipes) between users and platforms vary across social media.

For these reasons, it is out of the scope of this work to define the terms of the consumer-provider agreement, but we feel that it is an important and open problem for future study.

Another direction of future work would be to expand on the intuition in Section C. In particular, it may be possible to show that increasing accountability improves the user's trust in the platform, which ultimately leads to better performance for the platform. Such a result would show that the performative and regulatory interests can be (partially) aligned.

B Related work

Social media platforms are largely self-regulated [54]. However, in recent years, social media platforms have come under increasing scrutiny, prompting a call for regulations. Current efforts to regulate content moderation generally focus on specific issues, such as whether content is inappropriate (e.g., hate speech [7, 20]); discriminatory (e.g., race-based advertising [3, 50, 51, 35]); divisive (e.g., rankings that favors polarizing comments [49]); insulating (e.g., filter bubbles [27]), or false (e.g., fake news [41, 17, 19]).

These works typically adopt one of the following strategies: increasing content diversity (e.g., add heterogeneity to recommendations [14, 32]); drawing a global line (e.g., determining whether discrimination has occurred based on a threshold [18]); or focusing on the origin of the content (e.g., reducing fake news by whitelisting news sources [8]). There are legal and social barriers to many of these approaches [36, 15, 8, 45], including concerns that regulations might damage free speech or public discourse; violate personal rights or privacy; transfer agency away from users to big tech or government; draw highly subjective lines between acceptable and unacceptable behavior; or set precedents that are difficult to reverse. Moreover, there are concerns that regulations may lead to worse personalization and lower profits, which hurts both the user and platform [24].

We depart from the literature by studying a question at the heart of this debate: whether it is possible to regulate in a way that preserves the benefits of the social media ecosystem by allowing platforms to remain profitable while holding them accountable to users. In this way, our work does not focus on specific issues (e.g., hate speech or fake news), and our main contributions are mathematical rather than empirical.

Our focus on the consumer-provider agreement draws from implicit contract theory [4, 46], which is used by economists to explain behaviors that are observed but not justified by competitive market theory. The notion of an implicit contract between users and digital platforms is not new [47]. It has been used to explain why users stay on social media despite data privacy infractions [43, 37]. It has also been proposed as a starting point for regulation [29] because it balances the interests of both parties.

In our analysis, the social media platform is interested in how to filter a feed in order to maximize its reward (i.e., performance metric) subject to regulation. Our formulation can be viewed as an instance of robust optimization [9, 53], where our definition of the cost of regulation mirrors the "price" of robustness studied in other such works [11, 10]. Here, our study of the trade-off (or lack thereof) between regulation and reward bears resemblance to the study of optimization under fairness [10, 25, 34] and privacy [16, 33] constraints. It is worth a remark that our analysis has parallels with the field of differential privacy [22, 23] in that it compares distributions under different interventions [52]. In addition, similarly to social learning [21, 6, 1, 44], we study how information exchange—via content filtering—affects long-term outcomes, including what the user learns and how the platform performs.

C Long-term reward maximization

In this section, we follow up on the discussion in Section 5 to show that there can be benefits to following regulation in the long term.

Observation 1: Recall that the natural feed represents the con-sumer-provider agreement. Suppose that one of the terms of the agreement is that the platform provide personalized content based on the user's interests. Then, some portion of the content on the natural feed reflects the user's expressed preferences. which depend on the user's true preferences as well as whether the user has been given the opportunity to express those preferences.

Observation 2: Recall that the filtered feed is the feed that is shown to the user. One key observation is that showing a user a type of content is an implicit query of the user's interest in that content. Therefore, the ability of the user to express her preferences at time t depends on the content that is filtered at t. If the user is never shown a certain type of content, then the user's interest in that content remains unqueried and unknown.

Definition 6 (Oracle feed)**.** Suppose that, if the platform were omniscient, it would filter the oracle feed

$$\boldsymbol{\theta}_R^{(t)} = \arg\max_{\boldsymbol{\theta}\in\Theta} \text{OracleRew}(\boldsymbol{\theta}),$$

where OracleRew : $\Theta \to \mathbb{R}$ is an oracle, and OracleRew$(\boldsymbol{\theta})$ is the actual reward that the platform would receive from filtered feed $\boldsymbol{\theta}$.

The oracle feed is distinct from the optimal (unregulated) feed. The former is the ideal feed that is obtained under omniscience. The latter is computed based on R, which is a performance metric used by the platform.

Let $p_{R,t} = p_{\mathbf{z}}(\cdot; \boldsymbol{\theta}_R^{(t)})$ and similarly for $p_{N,t}$ and $p_{F,t}$. Then, using the observations above, we adopt the following model:

$$p_{N,t} \propto \left(p_{R,t-1}^a \cdot p_{F,t-1}^b\right)^{\frac{1}{a+b}}, \tag{7}$$

where $a, b \in \mathbb{R}_{\geq 0}$. In words, the user's natural feed (i.e., her expressed preferences) at time t is a function of the oracle feed (i.e., the content that would receive highest reward for the given user) and the previous filtered feed (i.e., the content that the user engaged with at the last time step).

We are interested in whether the platform can learn the user's true preferences and the optimal feed $p_{R,t}$ in the long term, as characterized in the following result.

Proposition 10. Suppose $p_{R,t} = p_R$ and $|\mathcal{Z}| < \infty$. Then, there exist a sequence of feasible filtered feeds such that $p_{F,t} \to p_R$ as $t \to \infty$. The higher the value a, the more quickly $p_{F,t}$ converges to p_R.

This result demonstrates that the long-term cost of regulation can be low. Intuitively, although the regulation places a cap on the platform's short-term reward, as long as the platform remains close to the natural feed, its long-term reward is high.

The model (7) on which this result is based assumes that the user's expressed preferences have some relation to the oracle feed, and the strength of this relation is captured by a. This is a strong condition and may not hold true in practice.

However, we posit that, if the platform respects the consumer-provider agreement by following regulation, the user's trust in the platform will increase, which would encourage them to truthfully report their preferences (e.g., interest in different types of content) to the platform. As a result, a would increase. Put differently, with better information, the platform could get closer to the oracle feed.

Remark. One of the observations of this section is that showing the user a piece of content is an implicit query of the user's interest in that content. Therefore, increasing content diversity, as studied in Sections 4.2 and 5.2, is a way for the platform to learn the user's preferences. In other words, it is a mechanism for *exploration*.

Exploration is a key concept in decision theory and machine learning. It captures the idea that, in online settings, an agent cannot simply maximize reward. He must occasionally sacrifice short-term reward in order to gain information about his environment.

In our setting, the platform must occasionally sacrifice reward in order to learn the user's preferences. Since the user's true preferences are changing and not fully observable, the platform must continually explore. Our proposed regulation naturally encourages content diversity (see Section 5.2) and therefore exploration.

This notion of diversity for exploration is not new. It is the same reason why some reinforcement learning algorithms use maximum entropy (MaxEnt) regularizers [31].

D Proofs for Section 3

D.1 Preliminaries

Consider a specific user and platform (this statement holds for the remainder of the work, unless otherwise indicated). Let the feed at time step t be given by $Z^{(t)} = \{\mathbf{z}_1^{(t)}, \dots, \mathbf{z}_m^{(t)}\} \in \mathcal{Z}^m$, where $\mathbf{z}_i^{(t)} \in \mathcal{Z} \subset \mathbb{R}^\ell$ is a feature vector describing the i-th piece of content. Let $\mathcal{P} = \{p_{\mathbf{z}}(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ for $\Theta \in \mathbb{R}$ be the model family containing the possible feeds. We refrain from specifying Θ in order to provide results that are fairly general. When implementing the regulation, the choice of Θ should become relatively clear.

For example, suppose that the platform characterizes its content by topic and sentiment, and the platform's filtering scheme depends on what topics and what corresponding sentiments it believes should be placed on the user's feed. Specifically, let $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, where $\mathbf{x}_i$ is a feature vector describing the topic of the i-th piece of content and y_i is the sentiment portrayed by that piece of content on topic $\mathbf{x}_i$. When a user interacts with content, she implicitly learns a model based on the content. Likewise, then the platform algorithmically filters content, it implicitly models the feed as a distribution over content. Suppose, purely for this example, that: $y_i = \sum_{j=1}^{\ell-1} \theta_j x_j + \theta_\ell v$, where $v \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. As mentioned above, there are two ways to interpret this model. One way to interpret it is: the user implicitly learns an approximately linear model between topics and sentiments from her feed, and her internal model is affected by additive Gaussian noise. A second interpretation is: the feed itself can be approximately represented by a set of topics, sentiments that are linear functions of those topics, and zero-mean additive Gaussian noise. As such, $\boldsymbol{\theta}$ specifies the linear function and scale of the additive noise. We make two comments. First, while this example may seem limiting, Θ could be much more sophisticated than the set of possible linear functions affected by Gaussian noise. Second, one may wonder whether such numeric features $\mathbf{x}_i$ and y_i exist. The answer is yes. Indeed, if the features did not exist, then there would be no algorithmic filtering to be regulated.

Recall the definitions:

$$\begin{aligned}
\tilde{\mathbf{s}}^{(t)} &= \tilde{\boldsymbol{\theta}}(Z_N^{(t)}) - \tilde{\boldsymbol{\theta}}(Z_F^{(t)}) \\
\tilde{\mathbf{m}}_D &= \sum_{t=1}^{T} \frac{1}{T} \tilde{\boldsymbol{\theta}}(Z_D^{(t)}) \quad \forall D \in \{F, N\} \\
\bar{\boldsymbol{\theta}}_D &= \sum_{t=1}^{T} \frac{1}{T} \boldsymbol{\theta}_D^{(t)} \quad \forall D \in \{F, N\} \\
\tilde{\mathbf{m}} &= \sum_{t=1}^{T} \frac{1}{T} \tilde{\mathbf{s}}^{(t)} = \tilde{\mathbf{m}}_N - \tilde{\mathbf{m}}_F \\
W &= \frac{1}{T} \sum_{t=1}^{T} (\tilde{\mathbf{s}}^{(t)} - \tilde{\mathbf{m}})(\tilde{\mathbf{s}}^{(t)} - \tilde{\mathbf{m}})^\top \\
\mathbf{s} &= \frac{1}{T} \sum_{t=1}^{T} (\boldsymbol{\theta}_N^{(t)} - \boldsymbol{\theta}_F^{(t)}) = \bar{\boldsymbol{\theta}}_N - \bar{\boldsymbol{\theta}}_F \\
\Sigma &= \frac{1}{mT^2} \sum_{t=1}^{T} \left(I^{-1}\left(\boldsymbol{\theta}_N^{(t)}\right) + I^{-1}\left(\boldsymbol{\theta}_F^{(t)}\right) \right).
\end{aligned}$$

Further, recall that the Fisher information matrix $I(\boldsymbol{\theta}) \in \mathbb{R}^{n \times n}$ is a positive semi-definite matrix, where the (i, j)-th entry is given by:

$$[I(\boldsymbol{\theta})]_{ij} = \mathbb{E}_{\mathbf{z} \sim p(\cdot;\boldsymbol{\theta})} \left[\frac{\partial}{\partial \theta_i} \log p(\mathbf{z}; \boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} \log p(\mathbf{z}; \boldsymbol{\theta}) \right]$$

We assume that the following regularity conditions on $\mathcal{P}$ hold:

1. Θ is an open set of $\mathbb{R}^n$.
2. Identifiability: $\mathbf{z} \overset{\text{i.i.d.}}{\sim} p_{\mathbf{z}}(\cdot; \boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ implies $p_{\mathbf{z}}(\cdot; \boldsymbol{\theta}_1)$ and $p_{\mathbf{z}}(\cdot; \boldsymbol{\theta}_2)$ are distinct.
3. Common support: The support of $p(\cdot; \boldsymbol{\theta})$ is independent of $\boldsymbol{\theta} \in \Theta$.
4. Differentiability: All the second-order partial deriviates of $\log p(\mathbf{z}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist and are continuous in $\boldsymbol{\theta}$.
5. For any $\boldsymbol{\theta}_0 \in \Theta$, there exists a neighborhood of $\boldsymbol{\theta}_0$ and a function $\Pi(\mathbf{z})$, where $\mathbb{E}_{\mathbf{z} \sim p(\cdot;\boldsymbol{\theta}_0)}[\Pi(\mathbf{z})] < \infty$ and

$$\left| \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(\mathbf{z}; \boldsymbol{\theta}) \right| \leq \Pi(\mathbf{z}),$$

   for all $\mathbf{z} \in \mathcal{Z}$, all $\boldsymbol{\theta}$ in the neighborhood of $\boldsymbol{\theta}_0$, and $i, j \in [n]$.
6. If $\boldsymbol{\theta}^*$ is the data generating process:
   (a) $\frac{\partial}{\partial \theta_i} \log p(\mathbf{z}; \boldsymbol{\theta}^*)$ is square integrable for all $i \in [n]$.
   (b) $\mathbb{E}_{\mathbf{z} \sim p(\cdot;\boldsymbol{\theta}^*)} \left[\frac{\partial}{\partial \theta_i} \log p(\mathbf{z}; \boldsymbol{\theta}^*) \right] = 0$
   (c) The Fisher information at $\boldsymbol{\theta}^*$ satisfies:

$$\begin{aligned} [I(\boldsymbol{\theta}^*)]_{ij} &= \mathbb{E}_{\mathbf{z} \sim p(\cdot;\boldsymbol{\theta}^*)} \left[\frac{\partial}{\partial \theta_i} \log p(\mathbf{z}; \boldsymbol{\theta}^*) \frac{\partial}{\partial \theta_j} \log p(\mathbf{z}; \boldsymbol{\theta}^*) \right] \\ &= -\mathbb{E}_{\mathbf{z} \sim p(\cdot;\boldsymbol{\theta}^*)} \left[\frac{\partial^2}{\partial \theta_i \theta_j} \log p(\mathbf{z}; \boldsymbol{\theta}^*) \right] \end{aligned}$$

   (d) Invertibility: Fisher information $I(\boldsymbol{\theta}^*)$ at $\boldsymbol{\theta}^*$ is positive-definite and invertible.

There are variations on these regularity conditions, and we refer the reader to other works for further details [5, 39, 42]. These regularity conditions are required for our results that use the asymptotic normality of the maximum likelihood estimator (MLE). Formally, suppose the data generating distribution is $p(\cdot; \boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$ and recall that $\tilde{\boldsymbol{\theta}} : \mathcal{Z}^m \to \Theta$ denotes the MLE. Then, under the regularity conditions above, the MLE is asymptotically normal:

$$\sqrt{m} \left(\tilde{\boldsymbol{\theta}}(Z) - \boldsymbol{\theta} \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}_n, I^{-1}(\boldsymbol{\theta})), \tag{8}$$

as $m \to \infty$, where $I^{-1}(\boldsymbol{\theta})$ denotes the inverse of the Fisher information at $\boldsymbol{\theta}$.

In this work, we adopt the shorthand notation. Suppose that $Y = y_1, \ldots, y_w$ are drawn i.i.d. from some generating distribution. We say that:

$$h(Y) \overset{d}{\approx} \mathcal{N}(h^*, S/w) \quad \text{as} \quad m \to \infty, \tag{9}$$

if

$$\sqrt{w} \left(h(Y) - h^* \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, S) \quad \text{as} \quad m \to \infty.$$

D.2 Proposition 1

Proof. Consider Algorithm 1 and (2) and recall the notation in Appendix D.1. Under the regularity conditions given in Appendix D.1, recall from (8) that the MLE is asymptotically normal. Adopting our notation (9),

$$\tilde{\boldsymbol{\theta}}(Z^{(t)}) \overset{d}{\approx} \mathcal{N} \left(\boldsymbol{\theta}^{(t)}, \frac{1}{m} I^{-1}(\boldsymbol{\theta}^{(t)}) \right), \tag{10}$$

as $m \to \infty$. Recall that $Z_N^{(t)}$ and $Z_F^{(t)}$ are independent for all $t \in [T]$. Then, as $m \to \infty$,

$$\tilde{\mathbf{s}}^{(t)} \overset{d}{\approx} \mathcal{N}\left(\boldsymbol{\theta}_N^{(t)} - \boldsymbol{\theta}_F^{(t)}, \frac{I^{-1}(\boldsymbol{\theta}_N^{(t)}) + I^{-1}(\boldsymbol{\theta}_F^{(t)})}{m}\right),$$

and, therefore, $\tilde{\mathbf{m}} \overset{d}{\approx} \mathcal{N}\left(\mathbf{s}, \Sigma\right)$ as $m \to \infty$. If $H = H_0$, then:

$$\tilde{\mathbf{m}} \overset{d}{\approx} \mathcal{N}\left(\mathbf{0}, \frac{2}{mT^2} \sum_{t=1}^{T} I^{-1}\left(\boldsymbol{\theta}_N^{(t)}\right)\right),$$

as $m \to \infty$.

Then, under the assumption $H = H_0$ and as $m \to \infty$, the statistic $\tilde{\mathbf{m}}^\top W^{-1} \tilde{\mathbf{m}} \cdot (T-n)/n$ follows an F-distribution with parameters n and $T-n$. Then, by definition, the test in line 7 is a level-ϵ test (i.e., a test for which the false positive rate $\mathbb{P}(\tilde{H} = H_1 | H = H_0)$ is ϵ) for (2) as $m \to \infty$.

Lastly, recall that generalized likelihood ratio test (GLRT) is the likelihood ratio test for which the unknown parameters are estimated using the MLE and notice that Algorithm 1 therefore implements the GLRT for (2) under the assumption that $(\boldsymbol{\theta}_N^{(t)} - \boldsymbol{\theta}_F^{(t)})$ are i.i.d. in time t. Then, by the well-known result [40, 28] that the GLRT converges to the uniformly most powerful invariant (UMPI), we have that for any invariant test $\hat{H}$ for (2) with level at most ϵ, $\mathbb{P}(\tilde{H} = H_1 | H = H_1) \geq \mathbb{P}(\hat{H} = H_1 | H = H_1)$ at $m, T \to \infty$, which concludes the proof. □

D.3 Proposition 2

Proof. Consider Algorithm 1 and (4). Recall from the proof of Proposition 1 that Algorithm 1 is equivalent to performing the GLRT and UMPI level-ϵ test as $m, T \to \infty$. By definition of the level-ϵ test, $\mathbb{P}(\tilde{H} = H_1 | H = H_0) \to \epsilon$ as $m, T \to \infty$.

By the result in Proposition 1, Algorithm 1's decision rule $\tilde{H}$ satisfies $\mathbb{P}(\tilde{H} = H_1 | H = H_1) \geq \mathbb{P}(\hat{H} = H_1 | H = H_1)$ for all invariant level-ϵ tests $\hat{H}$ for (2) as $m, T \to \infty$. It follows that, asymptotically, the level-ϵ test $\tilde{H}$ satisfies: $\mathbb{P}(\tilde{H} = H_1) \geq \mathbb{P}(\hat{H} = H_1)$ among all invariant level-ϵ tests $\hat{H}$.

Then, for any estimator $\hat{\boldsymbol{\theta}} : \mathcal{Z}^m \to \Theta$ and as $m, T \to \infty$, the regulation's decision $\tilde{H} = H_0$ implies that:

As stated in Proposition 2, let $f(\boldsymbol{\theta}) = v_0(\boldsymbol{\theta}) - v_1(\boldsymbol{\theta})$, where $\exists L \in \mathbb{R}_{\geq 0}$ such that $|f(M_1) - f(M_2)| \leq Lg(M_1, M_2)$ for all $M_1, M_2 \in \Theta^T$. Recall from (1) that:

$$\tilde{H} = H_0 \iff g(\tilde{M}_N, \tilde{M}_F) \leq \delta(\epsilon, n, T). \tag{11}$$

Combining these results and the fact that probabilities are upper bounded by 1, we have that: as $m, T \to \infty$, if $\tilde{H} = H_0$, then:

$$\begin{aligned}
&\mathbb{P}\left(g(\hat{M}_N, \hat{M}_F) \leq g(\tilde{M}_N, \tilde{M}_F)\right) \to 1 \\
&\implies \mathbb{P}\left(g(\hat{M}_N, \hat{M}_F) \leq \delta(\epsilon, n, T)\right) \to 1 \\
&\implies \mathbb{P}\left(|f(\hat{M}_F) - f(\hat{M}_N)| \leq L\delta(\epsilon, n, T)\right) \to 1,
\end{aligned} \tag{12}$$

where the first line follows from $\tilde{H}$ being the UMPI level-ϵ test, the second line follows from (11), and the last line follows by assumption. The last line is equivalent to the result statement, which concludes the proof. □

D.4 Corollary 3

Proof. Recall the setting in Proposition 2. If $|f(\hat{M}_F) - f(\hat{M}_N)| \leq L\delta(\epsilon, n, T)$ and $|f(\hat{M}_N)| > L\delta(\epsilon, n, T)$, then $f(\hat{M}_F)$ must have the same sign as $f(\hat{M}_N)$. This fact together with the last line (12) of the proof of Proposition 2 implies that $\mathbb{P}(\hat{A}_N^{(T)} = \hat{A}_F^{(T)}) \to 1$ as $m, T \to \infty$. □

E Proofs for Section 4

E.1 Proposition 4

Proof. $\tilde{\mathbf{m}}^\top W^{-1}\tilde{\mathbf{m}}$ can be decomposed as follows:

$$\begin{aligned}\tilde{\mathbf{m}}^\top W^{-1}\tilde{\mathbf{m}} &= (\tilde{\mathbf{m}}-\mathbf{s})^\top W^{-1}(\tilde{\mathbf{m}}-\mathbf{s}) - \mathbf{s}^\top W^{-1}\mathbf{s} + 2\tilde{\mathbf{m}}^\top W^{-1}\mathbf{s} \\ &= (\tilde{\mathbf{m}}-\mathbf{s})^\top W^{-1}(\tilde{\mathbf{m}}-\mathbf{s}) + (2\tilde{\mathbf{m}}-\mathbf{s})^\top W^{-1}\mathbf{s}. \end{aligned} \tag{13}$$

Recall from the Proof of Proposition 1 that

$$\begin{aligned}\tilde{\boldsymbol{\theta}}(Z^{(t)}) &\overset{d}{\approx} \mathcal{N}\left(\boldsymbol{\theta}^{(t)}, \frac{1}{m}I^{-1}(\boldsymbol{\theta}^{(t)})\right) \\ \tilde{\mathbf{s}}^{(t)} &\overset{d}{\approx} \mathcal{N}\left(\boldsymbol{\theta}_{N}^{(t)} - \boldsymbol{\theta}_{F}^{(t)}, \frac{I^{-1}(\boldsymbol{\theta}_{N}^{(t)}) + I^{-1}(\boldsymbol{\theta}_{F}^{(t)})}{m}\right) \\ \tilde{\mathbf{m}} &\overset{d}{\approx} \mathcal{N}\left(\mathbf{s}, \Sigma\right)\end{aligned}$$

as $m \to \infty$.

Recalling the definitions from Appendix D.1, these asymptotic normality results imply that W follows a Wishart distribution and $2\tilde{\mathbf{m}} - \mathbf{s}$ follows a normal distribution as $m \to \infty$. In addition, relative to the hypothesis test in (2), the first term in (13) is an ancillary statistic that follows a scaled F-distribution. Therefore, as $m \to \infty$,

$$\tilde{\mathbf{m}}^\top W^{+}\tilde{\mathbf{m}} \overset{d}{=} \frac{n}{T-n}u_1 + u_6^\top U_5^{-1}\mathbf{s}, \tag{14}$$

where $u_1 \sim F(n, T-n)$, $U_5 \sim \mathcal{W}(\Sigma, T-1, n)$, and $u_6 \sim \mathcal{N}(\mathbf{s}, \Sigma)$.

Since $T - 1 > n \geq 3$ and Σ is positive-definite, we can directly apply Corollary 1 from Bodnar and Okhrin [13] (who examined similar hypothesis testing problems) to obtain:

$$u_6^\top U_5^{-1}\mathbf{s} \overset{d}{=} \frac{\sqrt{\mathbf{s}^\top\Sigma^+\mathbf{s}}}{u_2}\left(\sqrt{\mathbf{s}^\top\Sigma^+\mathbf{s}} + 2u_3\sqrt{1 + \frac{n-1}{T-n+1}u_4}\right), \tag{15}$$

where $u_2 \sim \chi^2(T-n)$, $u_3 \sim \mathcal{N}(0,1)$, and $u_4 \sim F(\frac{n-1}{2}, \frac{T+1-n}{2})$ are mutually independent. Substituting (15) into (14) and recalling the definition $\bar{g} = \mathbf{s}^\top\Sigma^+\mathbf{s}$ gives the final result. □

E.2 Corollary 5

Proof. To prove (i), recall that the spectral norm of a symmetric matrix is the absolute value of its largest eigenvalue and that covariance matrices are positive semi-definite (i.e., have non-negative eigenvalues). Let $\lambda_i(A)$ be the i-th largest eigenvalue of A. Then,

$$\begin{aligned}\bar{g} &= \mathbf{s}^\top\Sigma^+\mathbf{s} \\ &\leq ||\mathbf{s}||_2^2\lambda_1(\Sigma^+) \\ &= \frac{||\mathbf{s}||_2^2}{\lambda_n(\Sigma)} \\ &= \frac{||\mathbf{s}||_2^2}{\lambda_n\left(\frac{1}{mT^2}\sum_{t=1}^T\left(I^{-1}(\boldsymbol{\theta}_{N}^{(t)}) + I^{-1}(\boldsymbol{\theta}_{F}^{(t)})\right)\right)} \\ &= \frac{mT^2||\mathbf{s}||_2^2}{\lambda_n\left(\sum_{t=1}^T\left(I^{-1}(\boldsymbol{\theta}_{N}^{(t)}) + I^{-1}(\boldsymbol{\theta}_{F}^{(t)})\right)\right)} \\ &\leq \frac{mT^2||\mathbf{s}||_2^2}{\sum_{t=1}^T(\lambda_n(I^{-1}(\boldsymbol{\theta}_{N}^{(t)})) + \lambda_n(I^{-1}(\boldsymbol{\theta}_{F}^{(t)})))},\end{aligned}$$

which gives the first result.

For the second result, let the entries of the Fisher information matrix be continuous in $\boldsymbol{\theta}$. Then, as $\boldsymbol{\theta}_F^{(t)}, \boldsymbol{\theta}_N^{(t)} \to \boldsymbol{\theta}$ for all $t \in [T]$, we have that $I^{-1}(\boldsymbol{\theta}_F^{(t)}), I^{-1}(\boldsymbol{\theta}_N^{(t)}) \to I^{-1}(\boldsymbol{\theta})$, which implies that $\Sigma \to \frac{2}{mT} I^{-1}(\boldsymbol{\theta})$ and $\Sigma^{-1} \to \frac{mT}{2} I(\boldsymbol{\theta})$.

Next, we use the result that:

$$d_{\mathrm{KL}}(\boldsymbol{\theta}||\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) = \frac{1}{2} \sum_{i,j \in [n]} [I(\boldsymbol{\theta})]_{ij} \Delta\boldsymbol{\theta}_i \Delta\boldsymbol{\theta}_j + O(|\Delta\boldsymbol{\theta}|^3)$$

(see (1.24) and (3.68) of Amari [2]), which is equivalent to:

$$d_{\mathrm{KL}}(\boldsymbol{\theta}||\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) = \frac{1}{2} \Delta\boldsymbol{\theta}^\top I(\boldsymbol{\theta}) \Delta\boldsymbol{\theta} + O(|\Delta\boldsymbol{\theta}|^3)$$

Therefore,

$$\begin{aligned} d_{\mathrm{KL}}(\bar{\boldsymbol{\theta}}_N||\bar{\boldsymbol{\theta}}_F) &= d_{\mathrm{KL}}(\bar{\boldsymbol{\theta}}_N||\bar{\boldsymbol{\theta}}_N - \mathbf{s}) \\ &= \frac{1}{2} \mathbf{s}^\top I(\boldsymbol{\theta}_N) \mathbf{s} + O(||\mathbf{s}||^3) \end{aligned}$$

Combining these results and re-arranging, we get that:

$$\left| \, d_{\mathrm{KL}}(\bar{\boldsymbol{\theta}}_N||\bar{\boldsymbol{\theta}}_F) - \frac{1}{mT} \mathbf{s}^\top \Sigma^{-1} \mathbf{s} \, \right| \to 0,$$

as $||\mathbf{s}||^3 \to 0$, which, from the definition $\bar{g} = \mathbf{s}^\top \Sigma^{-1} \mathbf{s}$, implies:

$$\left| \, mT d_{\mathrm{KL}}(\bar{\boldsymbol{\theta}}_N||\bar{\boldsymbol{\theta}}_F) - \bar{g} \, \right| \to 0,$$

as $||\mathbf{s}||^3 \to 0$ or, equivalently, $\boldsymbol{\theta}_F^{(t)}, \boldsymbol{\theta}_N^{(t)} \to \boldsymbol{\theta}$ for all $t \in [T]$. The last line follows from the observation that the ordering of $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_N$ can be switched without affecting the KL divergence result. □

E.3 Lemma 6

Proof. Let $\mathbf{v}$ be the principal eigenvector of $I(\boldsymbol{\theta})$ such that $||\mathbf{v}||_2^2 = 1$. Then, by the same reasoning in the proof of Corollary 5 (see Amari [2] for reference),

$$\begin{aligned} ||I(\boldsymbol{\theta}_1)||_2 &= \mathbf{v}^\top I(\boldsymbol{\theta}_1) \mathbf{v} \\ &\geq \mathbf{s}^\top I(\boldsymbol{\theta}_1) \mathbf{s} \\ &= 2 d_{\mathrm{KL}}(\boldsymbol{\theta}_1||\boldsymbol{\theta}_2) + O(||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^3) \\ &= 2(H(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) - H(\boldsymbol{\theta}_1)) + O(||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^3), \end{aligned}$$

which gives the result. □

E.4 Corollary 7

Proof. Since $\boldsymbol{\theta}_F^{(T)} \in \Omega_\alpha^{(T)}$ when $\bar{g} \leq G_\alpha$, the result follows by upper bounding the right-hand side of the expression in Corollary 5(i). by G_α. □

F Proofs for Section 5

F.1 Proposition 8

Proof. Let $\bar{\mathrm{s}}$ be as defined in Proposition statement. Let $\boldsymbol{\theta}_N^{(1:T)}$ be the natural feeds until time step T. Let $\boldsymbol{\theta}_F^{(1:T-1)}$ be the filtered feeds up until time step $T-1$. Note that, when the platform is performing the optimization at time step T in (5), these feeds cannot be changed. Let:

$$\mathbf{s}' = \frac{1}{T} \left(\boldsymbol{\theta}_N^{(T)} - \boldsymbol{\theta}_{F'}^{(T)} + \sum_{t=1}^{T-1} (\boldsymbol{\theta}_N^{(t)} - \boldsymbol{\theta}_F^{(t)}) \right), \tag{16}$$

Then (a) follows immediately from the definition of a β-LDA. By assumption,

$$G' = \frac{mT^2||\mathbf{s}'||_2^2}{D(\boldsymbol{\theta}_{F'}^{(1:T)}) + D(\boldsymbol{\theta}_{N}^{(1:T)})} < G_\alpha.$$

Let $\boldsymbol{\theta}_{F''}^{(T)} = \boldsymbol{\theta}_{F'}^{(T)} + \gamma\bar{\mathbf{s}}$ and $\boldsymbol{\theta}_{F''}^{(1:T)} = (\boldsymbol{\theta}_{F}^{(1)}, \ldots, \boldsymbol{\theta}_{F}^{(T-1)}, \boldsymbol{\theta}_{F''}^{(T)})$. Since $||\mathbf{s}' + \gamma\bar{\mathbf{s}}/T||_2^2 \leq ||\mathbf{s}'||_2^2 + \gamma/T$ and $D(\cdot) < \infty$ (recall that the Fisher information matrix is positive-definite by the regularity conditions), there exists choice of $\gamma > 0$ that is small enough such that:

$$G'' = \frac{mT^2||\mathbf{s}' + \gamma\bar{\mathbf{s}}/T||_2^2}{D(\boldsymbol{\theta}_{F''}^{(1:T)}) + D(\boldsymbol{\theta}_{N}^{(1:T)})} \leq G_\alpha.$$

Therefore, $\boldsymbol{\theta}_{F''}^{(T)}$ is a feasible feed at time T and, by definition of an LDA, $R(\boldsymbol{\theta}_{F''}, \boldsymbol{\theta}_{N}^{(T)}) > R(\boldsymbol{\theta}_{F'}, \boldsymbol{\theta}_{N}^{(T)})$.

The result in (b) follows from a similar logic. By assumption,

$$G' = \frac{mT^2||\mathbf{s}'||_2^2}{D(\boldsymbol{\theta}_{F'}^{(1:T)}) + D(\boldsymbol{\theta}_{N}^{(1:T)})} = G_\alpha.$$

Since $\bar{\mathbf{s}}^\top \mathbf{s}' > 0$, $||\mathbf{s}' + \gamma\bar{\mathbf{s}}/T||_2^2 > ||\mathbf{s}'/T||_2^2$ for any $\gamma > 0$. Therefore, the only way that:

$$\begin{aligned} G'' &= \frac{mT^2||\mathbf{s}' + \gamma\bar{\mathbf{s}}/T||_2^2}{D(\boldsymbol{\theta}_{F''}^{(1:T)}) + D(\boldsymbol{\theta}_{N}^{(1:T)})} \\ &= \frac{mT^2||\mathbf{s}' + \gamma\bar{\mathbf{s}}/T||_2^2}{D(\boldsymbol{\theta}_{F''}^{(T)}T) + D(\boldsymbol{\theta}_{F}^{(1:T-1)}) + D(\boldsymbol{\theta}_{N}^{(1:T)})} \end{aligned}$$

satisfies $G'' \leq G_\alpha$ is if $D(\boldsymbol{\theta}_{F''}) > D(\boldsymbol{\theta}_{F'})$. In this way, $D(\boldsymbol{\theta}_{F''}) > D(\boldsymbol{\theta}_{F'})$ can be interpreted as a necessary but not sufficient condition for $G'' \leq G_\alpha$.

The result in (c) refines the argument in (b). By assumption,

$$G' = \frac{mT^2||\mathbf{s}'||_2^2}{D(\boldsymbol{\theta}_{F'}^{(1:T)}) + D(\boldsymbol{\theta}_{N}^{(1:T)})} = G_\alpha,$$

and $\bar{D}_\gamma := D(\boldsymbol{\theta}_{F''}^{(T)}) - D(\boldsymbol{\theta}_{F'}^{(T)}) > 0$. Using these facts:

$$\begin{aligned} G'' &= \frac{mT^2||\mathbf{s}' + \gamma\bar{\mathbf{s}}/T||_2^2}{D(\boldsymbol{\theta}_{N}^{(1:T)}) + D(\boldsymbol{\theta}_{F'}^{(1:T)}) + \bar{D}_\gamma} \\ &\leq \frac{mT^2(||\mathbf{s}'||_2^2 + \gamma/T)}{D(\boldsymbol{\theta}_{N}^{(1:T)}) + D(\boldsymbol{\theta}_{F'}^{(1:T)}) + \bar{D}_\gamma} \\ &< \frac{mT^2||\mathbf{s}'||_2^2}{D(\boldsymbol{\theta}_{N}^{(1:T)}) + D(\boldsymbol{\theta}_{F'}^{(1:T)})} + \frac{mT\gamma}{D(\boldsymbol{\theta}_{N}^{(1:T)}) + D(\boldsymbol{\theta}_{F'}^{(1:T)}) + \bar{D}_\gamma} \\ &= G_\alpha + \frac{mT\gamma}{D(\boldsymbol{\theta}_{N}^{(1:T)}) + D(\boldsymbol{\theta}_{F'}^{(1:T)}) + \bar{D}_\gamma}. \end{aligned}$$

Because the second inequality is strict, as long as:

$$\frac{mT\gamma}{D(\boldsymbol{\theta}_{N}^{(1:T)}) + D(\boldsymbol{\theta}_{F}^{(1:T)})} < 1,$$

then it is possible to choose $\gamma > 0$ close enough to zero such that $G'' \leq G_\alpha$. We note that re-arranging this inequality yields the condition in statement (c), which concludes the proof. □

F.2 Corollary 9

Proof. The filtered feed $\boldsymbol{\theta}_{F}^{(T)} = \boldsymbol{\theta}_{N}^{(T)} + \zeta\bar{\mathbf{s}} \in \Theta$ maximizes reward because R is strictly concave in $||\mathbf{s}||_2^2$ and achieves maximum reward at $||\mathbf{s}||_2^2 = \zeta$. Therefore, the cost of regulation is 0 if $\boldsymbol{\theta}_{F}^{(T)}$ is also feasible, which is true by Corollary 7. □

F.3 Proposition 10

Proof. Suppose $p_{R,t} = p_R$. Let $c, d \in \mathbb{R}$ and $|\mathcal{Z}| < \infty$. Let $p_U = \text{Unif}(\mathcal{Z})$. We prove this result by construction. Let

$$p_{F,t} \propto \left(p_{N,t}^c \cdot p_U^d\right)^{\frac{1}{c+d}} .$$

Then,

$$p_{F,t} \propto p_{F,0}^{\left(\frac{bc}{(a+b)(c+d)}\right)^t} \cdot p_U^{\frac{td}{c+d}} \cdot \left(\Pi_{\tau=1}^{t-1} p_{R,\tau}\right)^{\frac{ac}{(a+b)(c+d)}}$$

As long as:

$$\begin{aligned} \frac{ac}{(a+b)(c+d)} &> \frac{d}{c+d} \\ a &> \frac{db}{c-d}, \end{aligned} \tag{17}$$

and $p_{F,0}(\mathbf{z}) > 0 \forall z \in \mathcal{Z}$, then as $t \to \infty$,

$$p_{F,t} \to p_R, \tag{18}$$

which is the desired result. It remains to show that the requirement (17) can be achieved under regulation.

It should be clear that d reflects a measure of the "distance" between the natural and filtered feeds. Indeed, as $d \to 0$, $p_{F,t} \to p_{N,t}$, which is always feasible. As such, for a choice of d that is small enough, $p_{F,t}$ meets regulation. Therefore, there always exists a d close enough to 0 such that it satisfies regulation and (17). □