Fuzzy Systems and Soft Computing
ISSN : 1819-4362

Framework for the Mathematical Evaluation of Social Media Dynamics

Pratyush Ranjan Mohapatra,Biswadarsi Biswal
Dept.of Computer Science And Engineering,Gandhi Institute For Technology,Bhubaneswar,752054
Email:pratyush@gift.edu.in

Abstract

Social media platforms (SMPs) leverage algorithmic filtering (AF) as a means of selecting the content that constitutes a user's feed with the aim of maximizing their rewards. Selectively choosing the contents to be shown on the user's feed may yield a certain extent of influence, either minor or major, on the user's decision-making, compared to what it would have been under a natural/fair content selection. As we have witnessed over the past decade, algorithmic filtering can cause detrimental side effects, ranging from biasing individual decisions to shaping those of society as a whole, for example, diverting users' attention from whether to get the COVID-19 vaccine or inducing the public to choose a presidential candidate. The government's constant attempts to regulate the adverse effects of AF are often complicated, due to bureaucracy, legal affairs, and financial considerations. On the other hand SMPs seek to monitor their own algorithmic activities to avoid being fined for exceeding the allowable threshold. In this paper, we mathematically formalize this framework and utilize it to construct a data-driven statistical auditing procedure to regulate AF from deflecting users' beliefs over time, along with sample complexity guaran- tees. This state-of-the-art algorithm can be used either by authorities acting as external regulators or by SMPs for self-auditing.

Keywords: Auditing, social media platforms, algorithmic filtering, distributional testing, testing Markov chains.

1. Introduction

Social media platforms (SMPs), e.g., Google, Facebook, and Twitter, are increasingly becoming the prevailing, most easily accessible, and most popular platforms for individual media consumption across the Western world (Mitchell et al., 2016). Indeed, media platforms act as intermediaries between users and the wealth of information collected from their friends, news, opinion leaders, celebrities, politicians and advertisers. So pervasive and eclectic is the stream of information content collected for each user at any given time that it compels social networks to filter out all but the most relevant information, display it in the user's news feed, and its order of appearance. To that end, in the last decade, social

.

platforms have been adopted various *algorithmic filtering* (AF) methods (Caplan, 2018) to select and sort collections of contents to be shown on their user's feed.

Notwithstanding the potential of AF to provide users with a richer, more diverse, and more engaging experience, over the past two decades, these methods have been abused by social network platforms to selectively filter user feeds in an effort to maximize their returns (including revenue, user accumulation, popularity gain, etc.). This phenomenon has brought about harmful side effects (DeVito et al., 2017; Bozdag, 2013). For example, an artificial comment ranking that encourages over-representation of one side's opinion or polarization of opinions (Siersdorfer et al., 2014), has sow hatred between groups (Lee, 2016)). Similarly, the prioritization of a specific topic contributes to the dissemination of deliberately disregarded fake news (Lewis and Marwick, 2017; Chesney and Citron, 2019; Damian et al., 2019; Pariser, 2011). Disseminating fake news may sway the presidential election results (Blake, 2018). Advertisements that promote products based on erroneous claims regarding the user's interests (Speicher et al., 2018; Sweeney, 2013), leading to some information being more (or less) visible along with many others. Intensive dietary recommendations may cause users to change their own diet (Chau et al., 2018; Jane et al., 2018), etc.

The foregoing examples, among many others, embody the potentially damaging fact that subjectively filtering the content to be shown on the user's feed might not overlap with the individual user's or the society's good as a whole, resulting in widespread adverse impacts on both individuals and society (Pariser, 2011). This, in turn, heavily impacts users' learning, shapes their thinking and decisions, and ultimately influences how they behave as individuals or as a whole society.

These negative influences have led to a number of calls for regulatory action by the authorities; however, their increasing enforcement attempts encounter multiple hurdles, such as, legal barriers, cumbersome and entangled bureaucracy, high human resource costs, which usually ends with no concrete results (Brannon, 2019; Klonick, 2017; Berghel, 2017). The legal difficulties are mainly driven by the concern that regulations might limit free speech (Klonick, 2017; Brannon, 2019), infringe on privacy by requiring content disclosure, subjectively define of what is right or wrong media behaviour (Obar and Wildman, 2015), undermine innovation or suppress jobs and revenues (e.g. through advertising restrictions).

Meanwhile, the increasing enforcement of regulations that aims at fining violations encourages the platforms to use self-regulatory methods to prevent unintentional internal activities and avoid penalties (Medzini, 2021). Among others, Twitter suspends tens of thousands accounts suspected of being involved in promoting conspiracy theories (News, January 2021). Facebook has set up an independent internal team named "Oversight Board" to foster freedom of expression by making principled, independent decisions about contents (Board, March 2020); YouTube has removed videos urging violence (Independent, January 2021).

The suggestion of the notion of an implicit agreement between users and social media platforms is far from being new (Manning, 1989). This notion draws from a general implicit contract theory (Koszegi, 2014), which economists use to explain behaviors that are observed but not justified by competitive market theory. In particular, it has been invoked to explain the reason for users to keep using social media despite data privacy infractions (Kruikemeier et al., 2020; Sarikakis and Winter, 2017; World Economic Forum, 2016). It has also been advanced as a starting point for regulation (Quinn, 2016), since it balances the interests of both parties.

As social media become increasingly popular information sources, a fundamental question remains: *Is there a systematic and responsible way to regulate the effect of social media platforms on users learning and decision-making?* Even though it may be possible to do so, due to the many issues raised above, and many other related ones, designing and reinforcing a regulation is still a notoriously difficult open problem (Kurbalija, 2016). Accordingly, the challenging quest for currently a far reaching fundamental theory for systematic regula- tory procedures that satisfy several social, legal, financial, and user related requirements, and its prospective practical ramifications, constitute the main impetus behind this paper. Motivated to guarantee compliance with a consumer-provider agreement, in this paper, we propose a data-driven statistical auditing procedure to regulate AF, which monitors ad- verse influences on the user learning (and thus on decision-making), while allowing real-time enforcement.

1.1 Related Work

Various attempts aim at regulating content moderation have been proposed over the last few years; however, all of these attempts generally focus on monitoring specific violations of the social platform-user agreement. Specifically, common methods for content moderation fall broadly into one of three categories (see, e.g., (Campbell, 2019; Mohseni et al., 2019)):

1. *Content control*, which aims at tagging or removing suspicious items. However, the ability of AI algorithms to identify such rough items grows more slowly than the ability to create them (Paschen, 2019), and objectivity of human content control is often less trusted (Anderson and Rainie, 2017). Content control strategies include: drawing a line in the sand (e.g., determining whether discrimination has occurred by thresholding the difference between two proportions (Chouldechova, 2017)); detecting hate speech (e.g., using deep learning technique Jahan and Oussalah (2021); Rodr´ıguez et al. (2019) or NLP clustering methods Davidson et al. (2017)); or finding the origin of the content (e.g., reducing fake news by whitelisting news sources Berghel (2017) or detecting the sources that generate misleading posts R ácz and Richey (2020)).
2. *Transparency*, where users are required to provide lawful identification. This approach imposes a serious toll on user privacy and anonymity, while not even necessarily stopping unintended spread of misinformation.
3. *Punishment*, where the network provider or the state impose penalties for malicious spreading of fake information. This extreme approach is clearly the least desirable from both privacy and human-rights perspectives.

Most related to our work are (Cen and Shah, 2021) and (Cen et al., 2023). Specifically, in (Cen and Shah, 2021) the concept of “counterfactual regulations” was proposed and analyzed. Counterfactual regulations deal with regulatory statements of the form: “The platform should produce similar feeds for given users who are identical except for one single property”. The users differentiating property could be, for example, gender, religion, left or right wing affiliation, age, among many others. Accordingly, (Cen and Shah, 2021) proposed an auditing procedure to test whether a counterfactual regulation statement is met or not, under a certain i.i.d. observational model. More recently, (Cen et al., 2023) introduced the notion of baseline/reference feed as “the content that a user would see without filtering”.[1] Then, they studied the problem of regulating AF with respect to this baseline, and proposed a framework and a procedure for regulating and auditing SMPs with respect to such a baseline. As we explain below in detail, our paper follows some of the general ideas in (Cen and Shah, 2020) and (Cen and Shah, 2021), but deviates in the way the setting is formulated and analyzed.

Finally, research and modeling of counterfactual regulation draw parallel ideas from the differential privacy literature (Dwork et al., 2006, 2014), as in the case of comparing outcomes under different interventions (Wasserman and Zhou, 2010). While our paper addresses questions similar to those studied in social learning and opinion dynamics, e.g. (Acemoglu et al., 2011; Molavi et al., 2018; Banerjee, 1992), it is distinct from this literature in the sense that our research focuses on the question of how the flow of information, mediated by social networks, leads to undesirable biases in the way users learn and, consequently, to a detrimental change in their decision-making and ultimately in their actions. Furthermore, this is accomplished without the need to actually access the users’ beliefs, actions, or thoughts.

1.2 Main Contributions

Our main goal is to develop an auditing procedure for content moderation over social networks. We split this subsection into two parts: the first focuses on our conceptual contributions to the general area of social media regulation, while the second discusses our technical contributions.

1.2.1 Conceptual Contributions

Unifying framework. Following the lead of (Cen and Shah, 2020) and (Cen and Shah, 2021), we formulate a statistical unifying framework for online platform auditing. This framework considers the three involved parties: platform, users, and an auditor, all interacting and evolving over time (see, Figure 1). At each time point, the platform shows its users collections of content, known as “filtered feeds.” As each user in the platform browses through his own feed, he implicitly forms a belief, and ultimately modifies his actions. The auditor’s meta-objective is to moderate the effect of socially irresponsible externalities caused by the AF’s effect on user learning and decision-making, either as individuals or as a society. To that end, the platform supplies the auditor with anonymous data of two types: filtered and reference. The latter is constructed by ignoring any aspect of a platform’s fiscal motivation, thus representing a natural/fair filtering of content rather than a subjective form of filtering (see, Section 2, for a precise definition), prioritizing the users’ experience. We show that the auditor’s task can be formulated as a certain closeness testing problem (see, e.g., (Daskalakis et al., 2018b; Canonne et al., 2022)). In addition to the filtered vs. reference approach above, similarly to (Cen and Shah, 2021), we also study counterfactual regulations.

[1] The idea of a baseline feed was originally proposed in an old version of (Cen and Shah, 2021), which can be found in (Cen and Shah, 2020).

Automatic online auditing procedure. We propose an auditing procedure that does not require any prior explicit regulation statement. The auditing procedure monitors any damaging influence on the users' decision-making over a predefined adjustable time-frame, compared to what it would have been without subjective filtering of the users' feeds, namely, under a natural/fair content filtering. This is accomplished by formulating a measure called "belief-variability", which estimates the influence of the AF on the beliefs of all the users. Using this variability, we then formulate the auditor's objective as a sequential hypothesis testing problem. As a binary hypothesis tester, the auditor examines whether the platform exceeds a tunable threshold of acceptable values of this estimated measurement of influence, doing so over a predefined time frame with a given confidence level. The auditor outputs whether or not regulation is being complied with, meaning whether public opinion is being biased or not. For example, this auditing procedure could easily detect the intensive promotion of a presidential candidate via posts, advertisements, the prioritization of related user comments, artificial adversarial users, or polarized recommendations. Finally, we propose an auditing procedure for deciding whether a platform complies with a given counterfactual regulation statement over the course of time.

We next highlight the main differences and contributions compared to (Cen and Shah, 2021) and (Cen et al., 2023). Specifically, both of these papers follow a "worst-case" approach, where auditing is designed to prevent violations associated with (a hypothetical) "most gullible" user, i.e., the user whose decisions are most influenced by AF. The idea is that if this user passes regulation, then all other users will pass regulation as well. We instead propose a "global" approach, where we average the influence of the platform's AF over a set of users. It should be clear that each approach has its own advantages and disadvantages. For example, the worst-case approach might be sensitive to adversarial users; in real-world SMPs, where any party is free to create a user without any supervision, a set of adversarial users can act as more naive/gullible than the most gullible user already in existence, and thus fool the auditor. Also, since the most gullible user is model driven (and not chosen from the data) then he/she might be unrealistically "too gullible", and then the auditor will announce false alarm violations excessively. Finally, the worst-case approach prevents all users from gradually changing their opinions. This is because, under this approach, the auditing process will immediately result in a violation when the most gullible user alters its opinion slightly. As a result, all other users will not have the opportunity to make slow and natural changes to their opinions, as they would with our average approach. In some sense, the above problematic issues are less severe/relevant in our average approach. It should be emphasized, however, that the outcome of any approach would depend on how seriously the platform engages in conversations on designing the test, model family, and reference feeds.

Another difference that we would like to emphasize is that the probabilistic setting considered in our paper is different from the one in (Cen and Shah, 2021) and (Cen et al., 2023). Specifically, in those papers, an i.i.d. time-independent generative model was assumed for the filtered (and reference) feeds. This implies that feeds are statistically independent, and excludes violations of regulation over time. In "real-world" cases, this approach may be inherently challenging. Indeed, in cases where regulations must be enforced over time, the procedures in (Cen and Shah, 2021) and (Cen et al., 2023) must be repeated endlessly. Furthermore, this allows an "uncooperative" platform to comply with regulation at a specific time when being tested, but not at any other time. In our paper, on the other hand, as an initial attempt and approximation to resolve the above issues, we follow a more complicated time-dependent Markovian model. As so, our auditing procedures, analysis, and results are inherently different from those in the aforementioned papers.

1.2.2 Technical Contributions

In addition to formulating a mathematical model for social media auditing, our paper contributes to the study of the *closeness testing problem*. The closeness testing problem have been extensively studied in the past few years (see, e.g., (Daskalakis et al., 2018b; Batu et al., 2013; Chan et al., 2014; Acharya et al., 2015)), as well as its extended version, the tolerant closeness testing problem (e.g., Daskalakis et al. (2018b); Canonne et al. (2022)). The vanilla form of the later is as follows. We are given i.i.d. sample access to distributions P and Q over $[n]$, and bounds $\varepsilon_2 > \varepsilon_1 \geq 0$, and $\delta > 0$. The task is to distinguish with probability of at least $1 - \delta$ between $\|P - Q\|_1 \leq \varepsilon_1$ and $\|P - Q\|_1 \geq \varepsilon_2$, whenever P, Q satisfy one of these two inequalities. In our setting, samples (or, feeds) are assumed to be generated from a certain Markovian probabilistic model (rather than being i.i.d.). Testing Markov chains is a new and active area of research with a number of remarkable recent results, such as testing symmetric Markov chains (Daskalakis et al., 2018a), testing Ergodic Markov chains (Wolfer and Kontorovich, 2019, 2020) or testing irreducible Markov chains (Chan et al., 2021). In this paper, we construct a method to solve a generalized form of the two problems above. Specifically, rather than a single pair of distributions, we are given samples from multiple pairs (see, Levi et al. (2011) for a related testing problem) of hidden irreducible Markov chains, and we need to decide whether the total sum of distances between these hidden pairs of chains is ε_1-close, or ε_2-far away. Similarly to majority of the papers mentioned above, we focus on the case where probabilistic distance measure is l_∞, with the understanding that other metrics can be analyzed. We propose a testing algorithm to the problem above, along with sample and complexity guarantees. It turns out that a major part of the analysis of our algorithm is related to the study of the covering time of random walks on undirected graphs (Chan et al., 2021). Specifically, we obtain an upper bound on the time it takes for multiple parallel random walks to cover each state a given number of times. Our analysis might be of independent interest.

1.3 Notations

For a positive integer m, we denote $[m] \equiv \{1, 2, \ldots, m\}$. The underlying space in the paper is $\mathbb{R}^n$, i.e., the space of all real-valued n length column vectors endowed with the dot product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}$. For $p \geq 1$, The l_p-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is given by $\|\mathbf{x}\|_p \equiv \sqrt[p]{\sum_{i=1} |x_i|^p}$. The l_∞-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is $\|\mathbf{x}\|_\infty = \max_{i=1,2,\ldots,n} |x_i|$. The p-norm of matrix $\mathbf{A}$ induced by vector p-norms is defined by $\|\mathbf{A}\|_p = \sup_{x/=0} \frac{\|\mathbf{A}x\|_p}{\|x\|_p}$. In the special cases of $p = 1, \infty$, the induced matrix norms can be computed or estimated by $\|\mathbf{A}\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|$, which is simply the maximum absolute column sum of the matrix; $\|\mathbf{A}\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|$, which is simply the maximum absolute row sum of the matrix. $\mathbf{e}$ is used to denote the vector of all ones and $\mathbf{0}$ is the vector of all zeros. We denote by $|S|$ the number of element in the set S. The function $a : (\mathrm{N} \times \mathrm{N}) \rightarrow \{\{\mathrm{N}\}, \{\mathrm{N}\}\}$ takes a pair of elements and return set containing those two element. The function a_f :

$(N \times N) \rightarrow \{N\}$ takes a pair of elements and return the first element in the pair (first coordinate). Similarly, the function $a_s : (N \times N) \rightarrow \{N\}$ takes a pair of elements and return the second element in the pair (first coordinate). Finally, we let Δn be the n-dimensional probability simplex.

2. Framework: Setup and Goal

In this section, we formalize mathematically our framework, including the setup and goals. Here, we opted to keep the exposition simple and concise, by presenting only the essential ingredients of our model which are needed for our main results. However, we refer the interested reader to the appendix, where we include a detailed and consistent construction of our framework, with deeper discussions and motivations for our definitions and assumptions.

2.1 The setup

Consider a system with the following three parties: a SMP, a *user*, and an *auditor*, as illustrated in Figure 1. At each time step $t \in N$, the platform shows each user a collection of contents (e.g., posts, videos, photos, ads, etc.) known as *filtered feeds*. We denote the filtered feed shown to user $u \in [U]$ at time $t \in N$ by $\mathbf{X}^F_u(t)$, and assume that it consists of $M \in N$ pieces of contents, namely, $\mathbf{X}^F_u(t) = \{\mathbf{x}^F_{1,u}(t), \ldots, \mathbf{x}^F_{M,u}(t)\}$, where $\mathbf{x}^F_{j,u}(t) \in X$ denotes a piece of content, for $1 \leq j \leq M$.

Generally speaking, the AF mechanism is not known and should not be disclosed to the auditor. Nonetheless, it should be clear that for the auditor to be able to inspect the SMP, something about the feeds generation process must be assumed. From the auditor's point of view, the platform is a sequential feeds generating system, relying on a probabilistic relationship of the current feed conditioned on the previous feeds. Specifically, in this paper, we assume that the feeds are generated at random according to a quasi-Markov homogeneous model; we divide the time horizon into batches, and assume that in each batch, the platform's AF process is modeled as a large probabilistic state machine. One can think of these batches as time interval where the platform collects new data to create new successive feeds.

Mathematically, let $T_{total} \in N$ denote the time horizon, which determines how far into the past the auditor scrutinizes the platform's behavior. Assume we have $B \in N$ batches each of length $T \in N$, such that in batch $b \in [B]$ we have a time sampling sequence $b \cdot T < t_{0,b} < t_{1,b} < \cdots < t_{T,b} \leq (b+1) \cdot T$. In each batch, from the auditor's point of view, the piece of content $\mathbf{x}^F_{l,u}(t_{i,b})$, at time $t_{i,b}$, for $l \in [M]$, is drawn from a first-order irreducible Markov chain, namely, $P(\mathbf{x}^F_{l,u}(t_{i,b}) | \mathbf{x}^F_{l,u}(t_{0,b}), \ldots, \mathbf{x}^F_{l,u}(t_{i-1,b})) = P(\mathbf{x}^F_{l,u}(t_{i,b}) | \mathbf{x}^F_{l,u}(t_{i-1,b}))$, and $P(\mathbf{x}^F_{l,u}(t_{i,b}) = s_2 | \mathbf{x}^F_{l,u}(t_{i-1,b}) = s_1) , Q_{u,b}(s_1, s_2)$, for any two possible states $s_1, s_2 \in X$. We denote the transition probability matrix by in batch $b \in [B]$ by $Q^{\mathbf{F}}_{u,b} = [Q_{u,b}(s_1, s_2)]_{s_1,s_2 \in F}$. We assume further that the M Markov trajectories are i.i.d. Note that over different intervals, indexed by b, the filtering process could be transformed into a new state machine subjected to a different transition probabilities. For example, this transformation may occur over time when new external data incur noticeable changes in the platform's reward. Thus, in the bth

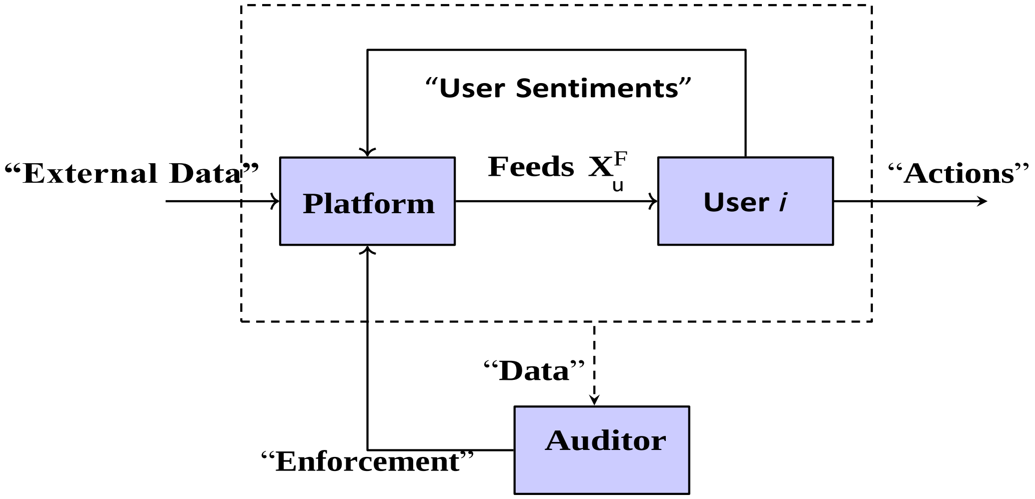

Figure 1: An illustration of the interaction between the platform, the user, and the auditor.

batch, the observed feeds are,

$$\underbrace{\left\{\mathbf{x}^F_{l,u}(t_{0,b})\right\}_{l=1}^M}_{\text{Feed 1}}, \underbrace{\left\{\mathbf{x}^F_{l,u}(t_{1,b})\right\}_{l=1}^M}_{\text{Feed 2}}, \ldots, \underbrace{\left\{\mathbf{x}^F_{l,u}(t_{T,b})\right\}_{l=1}^M}_{\text{Feed T}}.$$

Reference feeds. Following (Wachter and Mittelstadt, 2019; Ghosh, 2019; Cen and Shah, 2020; Petty, 2000), we define a reference boundary that is formed based on the users consent, and its location is determined by domain experts. While user *u*'s filtered feed $\mathbf{X}^F_u(t)$ at time *t* is chosen by the platform in a certain reward-maximizing methodology, the *reference feeds* $\mathbf{X}^R_u(t)$ could have been hypothetically selected by the platform if it strictly followed the consumer-provider agreement. These reference feeds are specific to each user $u \in [U]$ and time *t*. In this scenario, the platform would construct the feed based solely on the user's interests, without any subjective preferences influencing the content selection. Essentially, the only natural situation where the platform can filter content without introducing any subjective bias into the user's decision-making process and actions is by selecting feasible content that maximizes the user's benefit/reward. This approach ensures that the user's feed reflects their own preferences, which may align with the platform's benefits at times, but not necessarily always. Mathematically, the user's exclusive benefit is quantified by a personal reward function that encompasses only the components measuring the user's benefits. Similarly to the filtered feeds, we assume a Markovian generative model for the reference feeds, and denote by $P^{\mathbf{R}}_{u,b}$, $[P_{u,b}(s_1, s_2)]_{i,j\in F}$ the corresponding matrix transition probabilities in batch $b \in [B]$. In the appendix, we give examples of how the filtered and reference feeds are constructed by means of a certain reward function maximization, and elucidate the difference between the filtered and reference feeds. It should be emphasized that the specific reference feed construction hinted above, and described in more detail in the appendix, is just one possible example; our results and algorithms only require that there is some fixed reference feed (per user).

Counterfactual regulation. In addition to the "filtered vs. reference" approach, we will also analyze the following alternative auditing framework. Let S be a *regulatory statement* that an inspector (or, perhaps, the platform itself) wish to test. For example, S could be: "*The platform should produce similar feeds, in the course of a given time horizon* T, *for users who are identical except for property P*", where P could be ethnicity, sexual orientation, gender, a combination of these factors, etc. Let $U_P \subset [U] \times [U]$ be a subset of pairs of users that comply with P. Then, for any pair of users $(i, j) \in U_P$, the inspector's objective is to determine whether the platform's filtering algorithm cause user i's and user j's beliefs and actions to be significantly different. We formulate this objective rigorously in the next section. We mention here that a similar approach to the above was proposed recently in (Cen and Shah, 2021), assuming a time-independent i.i.d. model.

2.2 Auditor's goal

Average violation. We now define the meaning of "violation" from the auditor's perspective. Let $U \subseteq [U]$ be a certain subset of users. We define the *total filtering-variability metric* as,

$$V_{\text{filter}} = \frac{1}{|U|}\sum_{u\in U}\max_{i\in F} d_{TV}\left(P_{u,b}(i,\cdot),\ Q_{u,b}(i,\cdot)\right) \tag{1}$$

$$= \frac{1}{|U|}\sum_{u\in U}\max_{i\in F}\left\|\mathbf{P}^{\mathbf{R}}_{u,b}(i) - \mathbf{Q}^{\mathbf{F}}_{u,b}(i)\right\|_1 \tag{2}$$

$$= \frac{1}{|U|}\sum_{u\in U}\left\|P^{\mathbf{R}}_{u,b} - Q^{\mathbf{F}}_{u,b}\right\|_\infty , \tag{3}$$

where $\mathbf{P}^{\mathbf{R}}_{u,b}(i) \triangleq [P_{u,b}(i,j)]_{j\in F}$ and $\mathbf{Q}^{\mathbf{F}}_{u,b}(i) \triangleq [Q_{u,b}(i,j)]_{j\in F}$. We discuss the choice of the above metric in the appendix. Without loss of generality, in the rest of this paper, we focus on the special case where $U = [U]$. Also, we will consider a single specific interval for testing, say, $\{t_0, t_1, \ldots, t_T\} = [T]$, and therefore drop the dependency of the above notations on the batch index b. The underlying assumption here is that T is sufficiently large so as to allow for reliable testing, as dictated by our sample complexity guarantees, presented in the next section. An interesting question is to consider the case where the batch sizes are unknown, and then more sophisticated sequential/adaptive testing algorithms are needed.

Testing. Following the above, from the auditor's perspective, we define a violation event as the case where V_{filter} is "unusually large". Specifically, the audit's decision task is formulated as the following hypothesis testing problem,

$$H_0 : V_{\text{filter}} \le \varepsilon_1 \quad \text{vs.} \quad H_1 : V_{\text{filter}} \ge \varepsilon_2, \tag{4}$$

where $\varepsilon_2 > \varepsilon_1 \ge 0$ govern the auditing strictness. Devising successful statistical tests which solve (4) with high probability, guarantee that whenever the auditor decision is H_0, then the platform honors the consumer-provider agreement, since the beliefs and actions are indistinguishable under the filtered and reference feeds. Conversely, rejecting H_0 with high confidence implies that AF causes significantly different learning outcomes. Calculating V_{filter} requires knowledge of the filtering and reference distributions; a condition rarely met in practice. Accordingly, the auditor needs to solve (4) using only samples from these distributions; we assume that for $t \geq 1$ the auditor observes the filtered and reference feeds $\{\mathbf{X}_i^F(t), \mathbf{X}_i^R(t)\}$, for all users $u \in [U]$, and utilize these to test for violations. In practice, it might be challenging for the auditor to have both the reference and filtered feeds at hand. As so, it is an interesting question for future research to analyze the scenario where this full information is not available, e.g., only partial and perhaps quantized/noisy observations are given. Note that this type of hypothesis testing problem is reminiscent of the well- studied *tolerant closeness testing* problem (see, e.g., Daskalakis et al. (2018b); Canonne et al. (2022)). We are now in a position to state the testing problem faced by the auditor.

Problem 1 (Auditor testing) *Fix $\varepsilon_1, \varepsilon_2 \in (0, 1)$ and $\delta \in (0, 1)$ with $\varepsilon_1 < \varepsilon_2$. Given a set of t_T pairs of Markovian trajectories $\mathbf{X}_u^F(t_1), \mathbf{X}_u^R(t_1)$, . . . , $\mathbf{X}_u^F(t_T), \mathbf{X}_u^R(t_T)$ drawn from an* unknown *corresponding pair of Markov chains $Q_u^{\mathbf{F}}, P_u^{\mathbf{R}}$, for each user $u \in U$, an $(\varepsilon_1, \varepsilon_2, \delta)$-sum of pairs tolerant closeness testing algorithm outputs YES if $V_{\text{filter}} \leq \varepsilon_1$ and 'NO if $V_{\text{filter}} \geq \varepsilon_2$, with probability at least $1 - \delta$.*

As we mentioned earlier, the testing problem above is similar to the well-studied Markov tolerant closeness testing problem (e.g., Chan et al. (2021)). Nonetheless, the vanilla setting of this type of testing, is simpler than the one we are after, mainly because in our problem we deal with a *sum* of the distances between pairs of latent Markov chains, rather than a single distance, as it is in the standard setting. Finally, Figure 2 illustrates the filtered vs. reference testing scheme considered in this paper.

Remark 1 (Worst-case violation) *As we have mentioned in the introduction and above, in this paper we focus on a global approach by averaging the influence of the platform on the users. Here, we would like to mention that using the same techniques we develop in this paper, a worst-case approach, in the same vein as in (Cen and Shah, 2021), can be analyzed as well. Specifically, the idea in the worst-case approach is that if the platform's influence on the most gullible user's decision-making exceeds a predefined threshold, then it would mean a violation of the platform-user agreement. Now, within the proposed framework, we define this most gullible user as,*

$$u_{\text{gullible}} = \arg\max_{u\in[U]} \left\| P_{u,b}^{\mathbf{R}} - Q_{u,b}^{\mathbf{F}} \right\|_\infty ,$$

i.e., the platform's influence on his feed is the most significant. Intuitively, if the filtered feed satisfies regulation for the most gullible user, then it satisfies regulation for all other users whose learning is, by definition, less affected by the filtered feed, in the above sense. Accordingly, the auditor testing problem can be formulated as testing between

$$H_0^{\text{worst}} : \max_{u\in[U]} \left\| P_{u,b}^{\mathbf{R}} - Q_{u,b}^{\mathbf{F}} \right\|_\infty \leq \varepsilon_1 \quad \text{vs.} \quad H_1^{\text{worst}} : \max_{u\in[U]} \left\| P_{u,b}^{\mathbf{R}} - Q_{u,b}^{\mathbf{F}} \right\|_\infty \geq \varepsilon_2. \tag{5}$$

2.3 Auxiliary definitions and lemmas

This subsection is devoted to present several notations, definitions, and a lemma that will be needed to present our main results. As mentioned in the previous subsection, the problem

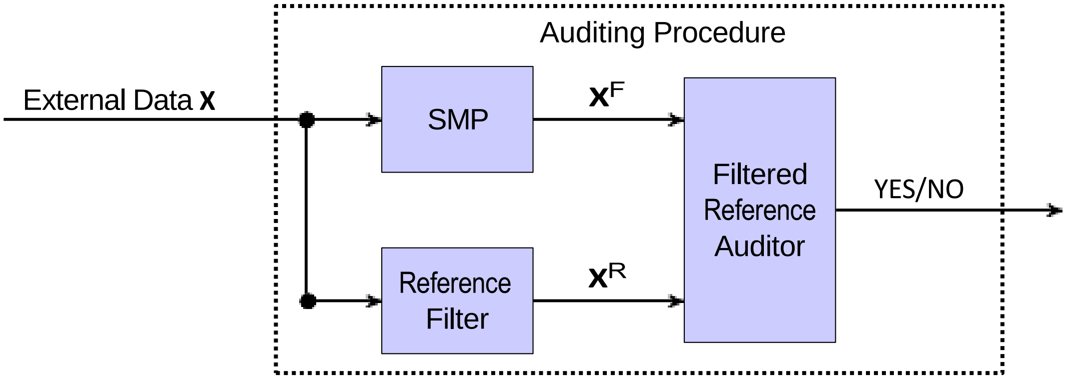

Figure 2: An illustration of the auditing procedure. The SMP and the uniformal filter get as an input the external data procedure, then output the filtered and the reference feeds, respectively. Both feeds are seen by the auditor, where the last outputs “YES” when the regulation is not violated, or “NO” otherwise.

of closeness testing of a single pair of Markov chains was considered in, for example, (Chan et al., 2021); it was shown that the testing algorithm and sample complexity depend on the *k*-cover time. The former is defined as the first time that a random walk has visited every state of the Markov chain at least *k* times, while the later is the maximization of the expectation of this random variable over all initial states. As a natural generalization, we define the $\boldsymbol{l}$–joint-*k*-cover time, as the expected time it takes for $\boldsymbol{l} \geq 1$ independent random walks to cover all states at least *k* times. In the language of our framework, an interpretation of this $\boldsymbol{l}$–joint-*k*-cover time is the expected time it takes the platform to show all users all feasible contents.

Definition 2 (*l*-joint-*k*-cover time) *Let* $Z^\infty_{1,1}, Z^\infty_{2,1}, ..., Z^\infty_{l,1}$ *be l-independent infinite trajectories drawn by the same Markov chain* M. *For* $t \geq 1$, *let* $\{N_i^{Z^j}(t), \forall i \in [n]\}$ *be the counting distribution of states* $i \in [n]$ *appearing in the subtrajectory* $Z^t_{j,1}$ *up to time t. For any* $k, \boldsymbol{l} \in \mathbb{N}$, *the random* $\boldsymbol{l}$*–joint-k-cover time* $\tau^{(k)}_{\text{cov}}(\boldsymbol{l}; M)$, *is the first time when all* $\boldsymbol{l}$ *independent random walks have jointly visited every state of* M *at least k times, i.e.,*

$$\tau^{(k)}_{\text{cov}}(\boldsymbol{l}; \boldsymbol{M}) \triangleq \inf \left\{ t \geq 0 : \forall i \in [n], \sum_{j=1}^{\boldsymbol{l}} N_i^{Z^j}(t) \geq k \right\}. \tag{6}$$

Accordingly, the l-joint-k-cover time is given by

$$t^{(k)}_{\text{cov}}(\boldsymbol{l}; \boldsymbol{M}) \triangleq \max_{\mathbf{v} \in [n]^{\boldsymbol{l}}} \mathbb{E}\left[\tau^{(k)}_{\text{cov}}(\boldsymbol{l}; \boldsymbol{M}) \mid Z_{1,1} = v_1, Z_{2,1} = v_2, ..., Z_{\boldsymbol{l},1} = v_{\boldsymbol{l}} \right], \tag{7}$$

where the coordinates of $\mathbf{v} = (v_1, v_2, \ldots, v_l) \in [n]^l$ *correspond to initial states.*

Throughout the paper, we will also use the notation $t^{(k)}_{\text{cov}}(l; \mathrm{P})$ to refer to $t^{(k)}_{\text{cov}}(l; M)$, where P denotes the matrix transition probabilities of M. For simplicity of notation, we denote $t_{\text{cov}}(M) \equiv t^{(1)}_{\text{cov}}(1; M)$. In addition, unless we explicitly deal with two different chains, we omit the dependency of $t^{(k)}_{\text{cov}}(l; M)$ on M and use $t^{(k)}_{\text{cov}}(l)$ instead. We denote by π the stationary distribution of M, and accordingly we define the *mixing time* as $t_{\text{mix}}(M) , \min \{ t \geq 1 : \max_{\mu\in\Delta_d} d_{TV}(\mu M^t, \pi) \leq 1/4 \}$. Finally, we denote the minimum stationary probability as $\pi_{\wedge} = \min_{i\in[n]} \pi_i$. Our goal is to bound $t^{(k)}_{\text{cov}}(l; M)$ in terms of $t_{\text{cov}}(M)$. To date, studies have focused on one of the following two separate cases:

1. Upper bounding the expected time required to cover all n states of an irreducible Markov chain, k times, with a single random walk, given by $t^{(k)}_{\text{cov}}(1; M)$, in the terms of $t_{\text{cov}}(M)$. Specifically, it is shown in (Chan et al., 2021) that for irreducible Markov chain,

$$t^{(k)}_{\text{cov}}(1; M) = O\left(t_{\text{cov}}(M) \log n + \frac{k \log n}{\pi_{\wedge}} \right). \tag{8}$$

2. Upper bounding the expected time required to cover all n states of some general irreducible Markov chain, with l multiple independent random walks, given by $t^{(1)}_{\text{cov}}(l; M)$, in the terms of t_{cov}. Many bounds, relying on different assumptions, exist in the literature. For example, combining Theorem 3.2, Lemma 4.3, and Theorem 4.7 in (Rivera et al., 2023), we get that for irreducible Markov chain,

$$t^{(1)}_{\text{cov}}(l; M) = O\left(t_{\text{mix}}(M) \vee \frac{t_{\text{cov}}(M) \log n}{l} \right). \tag{9}$$

For our case, we obtain the following bound on the $t^{(k)}_{\text{cov}}(l; M)$, for any $k, l \geq 1$.

Lemma 3 *For any $k, l \geq 1$ and irreducible Markov chains M,*

$$t^{(k)}_{\text{cov}}(l; M) = O\left(\left(t_{\text{mix}}(M) \vee \frac{t_{\text{cov}}(M) \log n}{l} \right) \log n + \frac{k \log n}{\pi_{\wedge}} \right). \tag{10}$$

Finally, following (Daskalakis et al., 2018a), for a length q trajectory Z_1^q of an irreducible Markov chain M, and for any state $i \in [n]$, we define the mapping $\psi^{(i)}_k(Z_1^q)$ as follows: we look at the first k visits to state i (i.e., at times $t = t_1, \ldots, t_k$ with $Z_t = i$) and write down the corresponding transitions in Z_1^q, i.e., Z_{t+1}. In other words, the mapping returns the k succeeding states of state i. We note that every state is visited almost surely, since M is an irreducible finite-state Markov chain. Therefore, the above mapping defines a proper probability distribution. Most importantly, as we will show later on, this distribution is independent across all different states and/or independent for a particular state i because of the Markov property.

3. Main Results

In this section, we present our main results. Specifically, in Subsection 3.1, we start by presenting an algorithm, along with sample complexity guarantees, for closeness testing the sum of distances of pairs of discrete distributions using i.i.d. samples. This in turn will serve as a sub-routine in the auditing procedure we propose and analyze in Subsection 3.2. Finally, in Subsection 3.3 we analyze the counterfactual regulation approach.

3.1 Warm up: Testing a family of discrete distributions

As a warm-up, we start by generalizing the vanilla i.i.d. tolerant closeness testing problem (e.g., Canonne et al. (2022)), to the case where one is given a *set* of pairs of measurements drawn from a *set* of pairs of probability distributions, and is tasked with deciding whether the total sum of distances between these pairs of distributions is close or far away. This problem is formulated mathematically as follows.

Problem 2 (Sum closeness testing) *Given sample access the pairs of distributions* (P_u, Q_u) *over* $[n]$*, for* $u \in [U]$*, and bounds* $\varepsilon_2 > \varepsilon_1 \geq 0$*, and* $\delta > 0$*, distinguish with probability of at least* $1 - \delta$ *between* $\sum_{u=1}^{U} |P_u - Q_u|_1 \leq |U| \cdot \varepsilon_1$ *and* $\sum_{u=1}^{U} |P_u - Q_u|_1 \geq |U| \cdot \varepsilon_2$*, whenever the distributions satisfy one of these two inequalities.*

The vanilla i.i.d. tolerant closeness testing corresponds to U = 1. As we will see in the following subsection, an algorithm to Problem 2 will serve as a building block to the actual testing problem we are after in Problem 1. We next propose a procedure solving the above testing problem along with sample complexity guarantees. We establish first a few notations. Let $S_{u,P}$ and $S_{u,Q}$ be two sets of $m \in N$ samples drawn from P_u and Q_u, respectively, for all $u \in [U]$, and let $S_P \triangleq \{S_{1,P}, \ldots, S_{U,P}\}$ and $S_Q \triangleq \{S_{1,Q}, \ldots, S_{U,Q}\}$. For every $u \in U$, let $V_{u,i}$ and $\tilde{V}_{u,i}$ count the number of occurrences of symbol $i \in [n]$, in the first and the second sets of m samples (each), sampled from P_u, respectively. Similarly, we denote $Y_{u,i}$ and $\tilde{Y}_{u,i}$ the corresponding samples from Q_u, for every $u \in U$. As is customary in the literature of distributional testing (e.g., Canonne et al. (2022)), we use the "Poissonization" trick, and assume that the sample sizes of P_u and Q_u, for every symbol $i \in [n]$, are Poisson-distributed with mean m, namely, $V_{u,i}, \tilde{V}_{u,i} \sim \text{Poisson}(m \cdot P_{u,i})$ and $Y_{u,i}, \tilde{Y}_{u,i} \sim \text{Poisson}(m \cdot Q_{u,i})$, where $P_{u,i}$ $(Q_{u,i})$ is the probability of symbol i under P_u (Q_u). Define,

$$f_{u,i} \triangleq \begin{cases} \max\{\sqrt{mn}|P_{u,i} - Q_{u,i}|, n(P_{u,i} + Q_{u,i}), 1\}, & \text{if } m > n, \\ \max\{m(P_{u,i} - Q_{u,i}), 1\}, & \text{otherwise}, \end{cases} \tag{11}$$

where $\tilde{V}_{u,i}, \tilde{Y}_{u,i}$ are used to estimate $f_{u,i}$ with $\hat{f}_{u,i}$, defined as,

$$\hat{f}_{u,i} \triangleq \begin{cases} \max\left\{ n\frac{|\tilde{V}_{u,i} - \tilde{Y}_{u,i}|}{\sqrt{m/n}}, \frac{\tilde{V}_{u,i} + \tilde{Y}_{u,i}}{m/n}, 1 \right\}, & \text{if } m > n, \\ \max\left\{ \tilde{V}_{u,i} + \tilde{Y}_{u,i}, 1 \right\}, & \text{otherwise}. \end{cases} \tag{12}$$

Additionally, define $G_{u,i} \triangleq (V_{u,i} - Y_{u,i})^2 - V_{u,i} - Y_{u,i}$, and finally,

$$G \triangleq \sum_{u=1}^{U} \sum_{i=1}^{R} \frac{G_{u,i}}{\hat{f}_{u,i}}. \tag{13}$$

Consider the routine IIDTESTER(S_P, S_Q, δ, ε_1, ε_2, m, n) in Algorithm 1. The constant $c > 0$ is an absolute constant determined in the course of the analysis. We have the following result.

Algorithm 1: Tolerant closeness tester for the i.i.d. pairs

```
Input: $U$, $n$, $m$, $\varepsilon_1$, $\delta$, and samples $S_P$ and $S_Q$ from $\{(P_u, Q_u)\}_{u\in[U]}$.
Set $\tau \longleftarrow - c \min\left(\frac{m^{3/2}\varepsilon_2}{n^{\frac{1}{2}}}, \frac{Um^2\varepsilon_2^2}{n}\right)$
Compute $G$ in (13).
If $G < \tau$, then Return YES
Else $G \geq \tau$, then Return NO
```

Algorithm 2: Filtered vs. reference auditing procedure

```
Input: $T$, $n$, $|X|$, $\varepsilon_1, \varepsilon_1, \delta, \bar{m}$, and feeds $\{\mathbf{X}_u^R(t), \mathbf{X}_u^F(t)\}_{t=1}^T$, for $u \in [U]$.
Output: YES if $V_{\text{filter}} \leq \varepsilon_1$ / NO if $V_{\text{filter}} \geq \varepsilon_2$.
For $i \leftarrow 1, 2 \ldots\ldots, n$
    Set $S^R \leftarrow \emptyset$ and $S^F \leftarrow \emptyset$
    For every user $u \leftarrow 1, 2 \ldots\ldots, U$
        If $\sum_{j=1}^M N_i^{\mathbf{x}_{j,u}^R} < \bar{m}$ or $\sum_{j=1}^M N_i^{\mathbf{x}_{j,u}^F} < \bar{m}$
            Return NO
        Calculate $S_u^R \leftarrow \cup_{j=1}^M \psi_{\bar{m}}^{(i)}\{\mathbf{x}_{j,u}^R(t)\}_{t=1}^T$ and $S_u^F \leftarrow \cup_{j=1}^M \psi_{\bar{m}}^{(i)}\{\mathbf{x}_{j,u}^F(t)\}_{t=1}^T$
        Do $S^R \leftarrow S^R \cup S_u^R$ and $S^F \leftarrow S^F \cup S_u^F$
    If IIDTESTER$(S^R, S^F, \delta, \varepsilon_1, \varepsilon_2, \bar{m}, n)$ = NO
        Return NO
Return YES
```

Theorem 4 (Sample complexity) *There exists an absolute constant $c > 0$ such that, for any $0 \leq \varepsilon_2 \leq 1$ and $0 \leq \varepsilon_1 \leq c\varepsilon_2$, given*

$$m = O\left(\sqrt{\frac{n}{\varepsilon^4\delta U}} + n\frac{\varepsilon_1^2}{\varepsilon_2^4} + n\frac{\varepsilon_1}{\varepsilon_2^2} + \frac{n^{2/3}}{U\varepsilon_2^{4/3}}\right), \tag{14}$$

samples from each of $\{P_u\}_{u=1}^U$ and $\{Q_u\}_{u=1}^U$, Algorithm 1 distinguish between $\sum_{u=1}^U \|P_u - Q_u\|_1 \leq U \cdot \varepsilon_1$ and $\sum_{u=1}^U \|P_u - Q_u\|_1 \geq U \cdot \varepsilon_2$, with probability at least $1 - \delta$.

3.2 Filtered vs. reference auditing

In this subsection, we present our auditing procedure for the filtered vs. reference feeds approach. We denote by $m(n, \varepsilon_1, \varepsilon_2, \delta)$ the sample complexity of the i.i.d. tester in Algorithm 1, and assume that it satisfies the condition in Theorem 4. Let $\bar{m}$, $m(n, \varepsilon_1, \varepsilon_2, \delta/4n)$. Consider the auditing procedure in Algorithm 2. We have the following result.

Theorem 5 (Sample complexity) *Given an $(\varepsilon_1, \varepsilon_2, \delta)$ i.i.d. tolerant-closeness-tester for n state distributions with the sample complexity of m $(n, \varepsilon_1, \varepsilon_2, \delta)$, then we can $(\varepsilon_1, \varepsilon_2, \delta)$ testing hypothesis* (4) *using,*

$$T = O\left(\max_{u\in[U]} \max_{W\in\{Q_u^F, P_u^R\}} t_{\text{cov}}^{\bar{m}}(M; W) \log\frac{U}{\delta}\right), \tag{15}$$

samples per user.

Note that in step 8 of Algorithm 2, feed samples S^R and S^F from the Markov chains are supplied to the i.i.d. tester in Algorithm 1. These samples are pulled using the mapping ψ, and thus are guaranteed i.i.d., as mentioned right after Lemma 3. The sample condition in (15) guarantees that all states are visited "reasonable" number of times, jointly by all the M chains, and for all users. Accordingly, we can apply an i.i.d. identity tester to each state's conditional distribution, and the auditing procedure return "YES" if this distribution passes its corresponding i.i.d. test.

At this point we would like to mention that our auditing procedure is not required to be disclosed to the internal AF mechanism used by the platform, which may not consent to be shared. This provides also a flexibility in regulating the model with no need for adaptation with respect to any future modification of the internal AF. Furthermore, our procedure can be applied using only access to users' observations (their feeds) in order to infer the influence of the platforms on their beliefs, decision-making, and ultimately on their actions, while having no access to their actual beliefs. It is clear that this way no further privacy leakage is incurred from the auditing process.[2] The bound we derived on *m*-joint *k*-cover time in Subsection 2.3 gives a simpler sample complexity bound for the auditing procedure. In particular, using Lemma 3, we get that the number of samples, per user, can be bounded as,

$$\mathrm{T} = \mathrm{O}_\delta \quad \max_{u \in [U]} \max_{W \in \{Q_u^F, P_u^R\}} \quad t_{\mathrm{mix}}(W) \vee \frac{t_{\mathrm{cov}}(W) \log |X|}{M} \quad \log |X| + \frac{\overline{m} \log |X|}{\pi_\wedge(W)} \quad , \tag{16}$$

where $\pi_\wedge(W)$ denotes the minimum stationary distribution of the Markov chain with transition probability matrix W, and O_δ hides logarithmic factors in δ.

3.3 Counterfactual regulation

Above, we have focused on the "filtered vs. reference" feeds approach. However, it is clear that other frameworks can be formulated. Consider the following as an alternative. Let S be a *regulatory statement* that an inspector (or, perhaps, the platform itself) wish to test. For example, S could be: "*The platform should produce similar articles for users who are identical except for property P*", where *P* could be ethnicity, sexual orientation, gender, a combination of these factors, etc. Let $U_P \subset V \times V$ be a subset of pairs of users that comply with *P*. Then, for any pair of users $(i, j) \in U_P$, the inspector's objective is to determine whether the platform's filtering algorithm cause user *i*'s and user *j*'s beliefs and actions be significantly different. A similar approach, was studied recently in (Cen and Shah, 2021) under a time-independent i.i.d. model. We take into account the inherent dependency on the time dimension as in "real-world" applications regulations must be enforced over time, as explained in the introduction. Similarly to Subsection 2.2, we define the notion of counterfactual violation as follows.

[2] While data hacking remains a possibility, it would not be considered a regulatory flaw, as it could occur regardless of regulation. The involved parties are the platform and auditor, where the platform possesses data access and the auditor utilizes data for testing, thus maintaining user data privacy with sensible precautions.

Definition 6 (Counterfactual total variability) *Let* $U_P \subset [U]\times[U]$ *be a subset of pairs of users that comply with P. Then, for any pair of users* $(i, j) \in U_P$*, the total variability in algorithmic filtering behavior for counterfactual users is given by*

$$\bar{V}_{cu}(S, U_P) \triangleq \frac{1}{|U_P|}\sum_{(i,j)\in U_P} \max_{l\in F} d_{TV}\left(Q_i(l,\cdot), Q_j(l,\cdot)\right) \tag{17}$$

$$= \frac{1}{|U_P|}\sum_{(i,j)\in U_P} \max_{l\in F} \left\| \mathbf{Q}_i(l) - \mathbf{Q}_j(l) \right\|_1 \tag{18}$$

$$= \frac{1}{|U_P|}\sum_{(i,j)\in U_P} \left\| Q_i^F - Q_j^F \right\|_\infty . \tag{19}$$

Then, in the same spirit of the previous subsection, we define the investigator's task to test for violations in the following sense:

$$H_0^S : \bar{V}_{cu}(S, U_P) \le \varepsilon_1 \quad \text{vs.} \quad H_1^S : \bar{V}_{cu}(S, U_P) \ge \varepsilon_2. \tag{20}$$

As before, the goal here is to construct good inspection procedures given only S and a black-box access to the filtering algorithm. Note also that U_P need not correspond to real users and could represent hypothetical users. Now, comparing (3) and (4) with (19) and (20), it is clear that the hypothesis test in (20) is the same as the one in (4), if each pair of filtered and reference distributions that correspond to some user is replaced with a pair of filtered distributions that correspond to a pair of users in U_P . Accordingly, consider the counterfactual auditing procedure that appears in Algorithm 3. It is clear that the underlying idea in Algorithm 3 is the same as the one in Algorithm 2. The following is a direct consequence of Theorem 5.

Theorem 7 (Sample complexity) *Given an* $(\varepsilon_1, \varepsilon_2, \delta)$ *i.i.d. tolerant-closeness-tester for n-state distributions with sample complexity* $m(n, \varepsilon_1, \varepsilon_2, \delta)$*, then we can* $(\varepsilon_1, \varepsilon_2, \delta)$ *testing hypothesis (20) using,*

$$T = O\left(\max_{(u,v)\in U_P}\ \max_{W\in\{Q_u^F,Q_v^F\}} t_{cov}^{\tilde{m}}(M; W)\log\frac{|U_P|}{\delta}\right), \tag{21}$$

samples for each pair of users in U_P.

It should be mentioned that the auditing procedure requires a black-box access to the filtering algorithm only, and the internal filtering mechanism is oblivious to the auditor (SMPs will not grant auditors full access to their filtering algorithm). This in turn also implies that auditing procedure can work even if the filtering algorithm changes over time. Finally, as in the previous subsection, the bounds we derived on m-joint$k$-cover time in Subsection 2.3 give simpler sample complexity bounds. Indeed, the number of samples, for each pair of users in U_P, can be written as,

$$T = O_\delta\left(\max_{(u,v)\in U_P}\ \max_{W\in\{Q_u^F,Q_v^F\}} t_{mix}(W) \vee \frac{t_{cov}(W)\log|X|}{M}\log|X| + \frac{\bar{m}\log|X|}{\pi_\wedge(W)}\right). \tag{22}$$

Algorithm 3: Counterfactual auditing procedure

```
Input: $T$, $n$, $|X|$, $\underline{\varepsilon}_1, \varepsilon_1, \delta, \overline{m}$, and feeds $\{\mathbf{X}_{\underline{u}}^F(t), \mathbf{X}_v^F(t)\}_{t=1}^T$, for every $(u,v) \in U_P$.
Output: YES if $V_{cu}(S, U_P) \leq \varepsilon_1$ / NO if $V_{cu}(S, U_P) \geq \varepsilon_2$.
For $i \leftarrow 1, 2 \ldots\ldots, n$
    Set $S \leftarrow \emptyset$ and $\tilde{S} \leftarrow \emptyset$
    For every pair $(u,v) \in U_P$
        If $\sum_{j=1}^{M} N_i^{\mathbf{x}_{j,u}^F} < \overline{m}$ or $\sum_{j=1}^{M} N_i^{\mathbf{x}_{j,v}^F} < \overline{m}$
            Return NO
        Calculate $S_u \leftarrow \cup_{j=1}^{M} \psi_{\bar{m}}^{(i)}\{\mathbf{x}_{j,u}^F(t)\}_{t=1}^T$ and $\tilde{S}_v \leftarrow \cup_{j=1}^{M} \psi_{\bar{m}}^{(i)}\{\mathbf{x}_{j,v}^F(t)\}_{t=1}^T$
        Do $S \leftarrow S \cup S_u$ and $\tilde{S} \leftarrow \tilde{S} \cup \tilde{S}_v$
    If IIDTESTER$(S, \tilde{S}, \delta, \varepsilon_1, \varepsilon_2, \overline{m}, n)$ = NO
        Return NO
Return YES
```

4. Proofs

This section is devoted to the proofs of our results.

4.1 Proof of Theorem 4

In this subsection we prove Theorem 4. To this end, we start by proving a few auxiliary results which characterize the first and second order statistics of the count in (13).

4.1.1 Auxiliary Results

Lemma 8 *Let* $\delta_1 \in (0,1)$, *and recall the definitions in* (11)–(13). *Then, there exist absolute constants* $c_1, c_2, c_3 > 0$, *such that the following hold with probability at least* $1 - \delta_1$,

$$\mathbb{E}\left[G\hat{f}_{u,i}, u \in [U], i \in [n]\right] \geq \frac{\delta_1 m^2 \left(\sum_{u=1}^{U} |P_u - Q_u|_1\right)^2}{c_1 \sum_{u=1}^{U} \sum_{i=1}^{n} f_{u,i}}, \tag{23}$$

and

$$\mathbb{E}\left[G\hat{f}_{u,i}, u \in [U], i \in [n]\right] \leq \frac{c_2}{\delta_1} \sum_{u=1}^{U} \sum_{i=1}^{n} \frac{m^2 (P_{u,i} - Q_{u,i})^2}{f_{u,i}}, \tag{24}$$

and

$$\mathrm{Var}\left[G\hat{f}_{u,i}, u \in [U], i \in [n]\right] \leq \frac{c_3}{\delta_1} \sum_{u=1}^{U} \sum_{i=1}^{n} \frac{\mathrm{Var}[G_{u,i}]}{f_{u,i}^2}. \tag{25}$$

Proof [Proof of Lemma 8] By standard properties of the Poisson distribution, the random variables in the definition of $G_{u,i}$ are statistically independent. Therefore,

$$\mathbb{E}[G_{u,i}] = \mathbb{E}[(V_{u,i} - Y_{u,i})^2 - V_{u,i} - Y_{u,i}]$$

$$
\begin{aligned}
&= \mathrm{E}[V_{u,i}^2] - 2\mathrm{E}[V_{u,i}]\mathrm{E}[Y_{u,i}] + \mathrm{E}[Y_{u,i}^2] - \mathrm{E}[V_{u,i}] - \mathrm{E}[Y_{u,i}]\\
&= (mP_{u,i})^2 + mP_{u,i} - 2m^2P_{u,i}Q_{u,i} + (mQ_{u,i})^2 + mQ_{u,i} - m^2P_{u,i} - mQ_{u,i}\\
&= (mP_{u,i})^2 - 2m^2P_{u,i}Q_{u,i} + (mQ_{u,i})^2 = m^2(P_{u,i} - Q_{u,i})^2\\
&= m^2\,|P_{u,i} - Q_{u,i}|^2\,. \qquad (26)
\end{aligned}
$$

Hence, $G_{u,i}$ is an unbiased estimator of $m^2|P_{u,i} - Q_{u,i}|^2$. Similarly,

$$
\begin{aligned}
\mathrm{Var}(G_{u,i}) &= \mathrm{Var}[(V_{u,i} - Y_{u,i})^2 - V_{u,i} - Y_{u,i}]\\
&= \mathrm{E}[((V_{u,i} - Y_{u,i})^2 - V_{u,i} - Y_{u,i})^2] - \mathrm{E}[((V_{u,i} - Y_{u,i})^2 - V_{u,i} - Y_{u,i})]^2\\
&= \mathrm{E}[((V_{u,i} - Y_{u,i})^4 - 2(V_{u,i} - Y_{u,i})^3 + (V_{u,i} - Y_{u,i})^2]\\
&\qquad - \mathrm{E}[((V_{u,i} - Y_{u,i})^2 - V_{u,i} - Y_{u,i})]^2\\
&= 4m^3(P_{u,i} - Q_{u,i})^2(P_{u,i} + Q_{u,i}) + 2m^2(P_{u,i} + Q_{u,i}). \qquad (27)
\end{aligned}
$$

Next, using the fact that $G_{u,i}$ and $\hat{f}_{u,i}$ are independent, by the linearity of the expectation, we obtain that the conditional expectation of G is,

$$
\mathrm{E}\left[G\,\hat{f}_{u,i}, u \in [U], i \in [n]\right] = \mathrm{E}\left[\sum_{u=1}^{U}\sum_{i=1}^{n}\frac{G_{u,i}}{\hat{f}_{u,i}}\,\middle|\,\hat{f}_{u,i}, u \in [U], i \in [n]\right] \qquad (28)
$$

$$
= \sum_{u=1}^{U}\sum_{i=1}^{n}\frac{\mathrm{E}\,[G_{u,i}]}{\hat{f}_{u,i}}. \qquad (29)
$$

Similarly, the conditional variance of G is,

$$
\mathrm{Var}\left[G\,\hat{f}_{u,i}, u \in [U], i \in [n]\right] = \mathrm{Var}\left[\sum_{u=1}^{U}\sum_{i=1}^{n}\frac{G_{u,i}}{\hat{f}_{u,i}}\,\middle|\,\hat{f}_{u,i}, u \in [U], i \in [n]\right] \qquad (30)
$$

$$
= \sum_{u=1}^{U}\sum_{i=1}^{n}\frac{\mathrm{Var}\,[G_{u,i}]}{\hat{f}_{u,i}^2}. \qquad (31)
$$

Combining (26) and (29) we get,

$$
\begin{aligned}
\mathrm{E}\left[G\,\hat{f}_{u,i}, u \in [U], i \in [n]\right] &= \sum_{u=1}^{U}\sum_{i=1}^{n}\frac{\mathrm{E}\,[G_{u,i}]}{\hat{f}_{u,i}} &(32)\\
&= \sum_{u=1}^{U}\sum_{i=1}^{n}\frac{m^2(P_{u,i} - Q_{u,i})^2}{\hat{f}_{u,i}} &(33)\\
&\geq \frac{m^2\left(\sum_{u=1}^{U}\sum_{i=1}^{n}|P_{u,i} - Q_{u,i}|\right)^2}{\sum_{u=1}^{U}\sum_{i=1}^{n}\hat{f}_{u,i}} &(34)\\
&\geq \frac{m^2\left(\sum_{u=1}^{U}\left|P_u - Q_u\right|_1\right)^2}{\sum_{u=1}^{U}\sum_{i=1}^{n}\hat{f}_{u,i}}, &(35)
\end{aligned}
$$

where the first inequality follows from the following fact that for any sequence of real-valued numbers $\{a_i\}_{i=1}^n$ and positive real-valued numbers $\{b_i\}_{i=1}^n$, we have,

$$\sum_{i=1}^{n} \frac{a_i^2}{b_i} \geq \frac{\left(\sum_{i=1}^{n} |a_i|\right)^2}{\sum_{i=1}^{n} b_i}, \tag{36}$$

and the last inequality follows by applying Cauchy-Schwarz to,

$$\sum_{i=1}^{n} |a_i| = \sum_{i=1}^{n} \frac{\sqrt{b_i}\,|a_i|}{\sqrt{b_i}}. \tag{37}$$

Next, Lemma 2.5 in (Canonne et al., 2022) states that there exist absolute constants $c_1, c_2, c_3 > 0$ such that, for every $u \in [U]$, $i \in [n]$, we have $\mathrm{E}[\hat{f}_{u,i}] \leq c_1 f_{u,i}$, $\mathrm{E}[\hat{f}_{u,i}^{-1}] \leq \frac{c_2}{f_{u,i}}$, and $\mathrm{E}[\hat{f}_{u,i}^{-2}] \leq \frac{c_3}{f_{u,i}^2}$. Moreover, by definition, the random random variables $\hat{f}_{u,i}$ are non-negative, and thus, applying by Markov's inequality, we obtain that,

$$\sum_{i=1}^{n} \hat{f}_{u,i} \leq \frac{1}{\delta_1} \sum_{i=1}^{n} \mathrm{E}\left[\hat{f}_{u,i}\right], \tag{38}$$

with probability at least $1 - \delta_1$, for any $u \in [U]$. Combined with Lemma 2.5 in (Canonne et al., 2022), this means that, with probability at least $1 - \delta_1$,

$$\mathrm{E}\left[G\hat{f}_{u,i}, u \in [U], i \in [n]\right] \geq \frac{\delta_1 m^2 \sum_{u=1}^{U} \left|P_u - Q_u\right|_1^2}{c_1 \sum_{u=1}^{U} \sum_{i=1}^{n} f_{u,i}}. \tag{39}$$

Next, applying Markov's inequality for the non-negative random variable $\mathrm{E}\left[G\,\hat{f}_{u,i}, u \in [U], i \in [n]\right]$, along with Lemma 2.5 in (Canonne et al., 2022), we obtain with probability at least $1 - \delta_1$,

$$\begin{aligned}
\mathrm{E}\left[G\hat{f}_{u,i}, u \in [U], i \in [n]\right] &\leq \frac{1}{\delta_1} \mathrm{E}\left[\mathrm{E}\left[G\hat{f}_{u,i}, u \in [U], i \in [n]\right]\right] && (40)\\
&= \frac{1}{\delta_1} \mathrm{E}\left[\sum_{u=1}^{U} \sum_{i=1}^{n} \frac{\mathrm{E}\left[G_{u,i}\right]}{\hat{f}_{u,i}}\right] && (41)\\
&\leq \frac{c_2}{\delta_1} \sum_{u=1}^{U} \sum_{i=1}^{n} \frac{m^2 (P_{u,i} - Q_{u,i})^2}{f_{u,i}}. && (42)
\end{aligned}$$

Similarly, with probability at least $1 - \delta_1$,

$$\begin{aligned}
\mathrm{Var}\left[G\hat{f}_{u,i}, u \in [U], i \in [n]\right] &\leq \frac{1}{\delta_1} \mathrm{E}\left[\mathrm{Var}\left[G\hat{f}_{u,i}, u \in [U], i \in [n]\right]\right]\\
&= \frac{1}{\delta_1} \mathrm{E}\left[\sum_{u=1}^{U} \sum_{i=1}^{n} \frac{\mathrm{Var}\left[G_{u,i}\right]}{\hat{f}_{u,i}^2}\right] && (43)\\
&\leq \frac{c_3}{\delta_1} \sum_{u=1}^{U} \sum_{i=1}^{n} \frac{\mathrm{Var}\left[G_{u,i}\right]}{f_{u,i}^2}. && (44)
\end{aligned}$$

This concludes the proof. ■

By the union bound, we conclude that (23)–(25), hold simultaneously with probability at least $1 - 3\delta_1$. We next bound the terms in the right-hand-side of (23)–(25), separately for $m \ge n$ and $m \le n$, respectively. We follow similar ideas as in (Canonne et al., 2022, Lemma 2.3) and (Canonne et al., 2022, Lemma 2.4). We have the following result.

Lemma 9 *For $m \ge n$, the following hold,*

$$\sum_{u=1}^{U}\sum_{i=1}^{n}\frac{\mathrm{Var}[G_{u,i}]}{f_{u,i}^{2}} \le \frac{10Um^{2}}{n}, \tag{45}$$

$$\sum_{i=1}^{n}\frac{m^{2}(P_{u,i}-Q_{u,i})^{2}}{f_{u,i}} \le \frac{m^{3/2}\left\|P_u-Q_u\right\|_1}{n^{\frac{1}{2}}}, \tag{46}$$

and

$$\frac{m^{2}\sum_{u=1}^{U}\left\|P_u-Q_u\right\|_1^{2}}{\sum_{u=1}^{U}\sum_{i=1}^{n}f_{u,i}} \ge \min\left\{\frac{m^{3/2}\sum_{u=1}^{U}\left\|P_u-Q_u\right\|_1}{2\,(Un)^{\frac{1}{2}}}, \frac{m^{2}\sum_{u=1}^{U}\left\|P_u-Q_u\right\|_1^{2}}{6Un}\right\}. \tag{47}$$

Proof [Proof of Lemma 9] We start by proving (45). From (27), we get

$$\begin{aligned}
\sum_{u=1}^{U}\sum_{i=1}^{n}\frac{\mathrm{Var}[G_{u,i}]}{f_{u,i}^{2}} &= \sum_{u=1}^{U}\sum_{i=1}^{n}\frac{4m^{3}(P_{u,i}-Q_{u,i})^{2}(P_{u,i}+Q_{u,i})+2m^{2}(P_{u,i}+Q_{u,i})^{2}}{f_{u,i}^{2}} && (48)\\
&= \sum_{u=1}^{U}\sum_{i=1}^{n}\frac{4m^{3}(P_{u,i}-Q_{u,i})^{2}(P_{u,i}+Q_{u,i})+2m^{2}(P_{u,i}+Q_{u,i})^{2}}{\left(\max\left\{\sqrt{mn}\cdot|P_{u,i}-Q_{u,i}|, n\cdot(P_{u,i}+Q_{u,i}), 1\right\}\right)^{2}} && (49)\\
&\le \sum_{u=1}^{U}\sum_{i=1}^{n}\frac{4m^{3}(P_{u,i}+Q_{u,i})}{mn} + \sum_{u=1}^{U}\sum_{i=1}^{n}\frac{2m^{2}}{n^{2}} && (50)\\
&= \frac{10Um^{2}}{n}, && (51)
\end{aligned}$$

where the inequality follows by lower bounding the denominator in (49) by $mn(P_{u,i} - Q_{u,i})^2$ for the first term in the numerator, and by $n^2(P_{u,i} + Q_{u,i})^2$ for the second term in the numerator. Next, we prove (46). We have,

$$\begin{aligned}
\sum_{i=1}^{n}\frac{m^{2}(P_{u,i}-Q_{u,i})^{2}}{f_{u,i}} &= \sum_{i=1}^{n}\frac{m^{2}|P_{u,i}-Q_{u,i}|^{2}}{\max\left\{\sqrt{mn}\cdot|P_{u,i}-Q_{u,i}|, n\cdot(P_{u,i}+Q_{u,i}), 1\right\}} && (52)\\
&\le \sum_{i=1}^{n}\frac{m^{3/2}|P_{u,i}-Q_{u,i}|}{n^{\frac{1}{2}}} && (53)
\end{aligned}$$

$$= \frac{m^{3/2} \| P_u - Q_u \|_1}{n^{\frac{1}{2}}}. \tag{54}$$

Finally, we prove (47). Note that,

$$\begin{aligned}
\frac{m^2 \left(\sum_{u=1}^{U} \| P_u - Q_u \|_1\right)^2}{\sum_{u=1}^{U} \sum_{i=1}^{n} f_{u,i}} &= \frac{m^2 \left(\sum_{u=1}^{U} \| P_u - Q_u \|_1\right)^2}{\sum_{u=1}^{U} \sum_{i=1}^{n} \max\left\{ \sqrt{mn} \cdot |P_{u,i} - Q_{u,i}|, n \cdot (P_{u,i} + Q_{u,i}), 1 \right\}} \\
&\geq \frac{m^2 \left(\sum_{u=1}^{U} \| P_u - Q_u \|_1\right)^2}{\sum_{u=1}^{U} \sum_{i=1}^{n} \left(\sqrt{mn} \cdot |P_{u,i} - Q_{u,i}| + n \cdot (P_{u,i} + Q_{u,i}) + 1 \right)} \qquad (55) \\
&= \frac{m^2 \left(\sum_{u=1}^{U} \| P_u - Q_u \|_1\right)^2}{\sqrt{mn} \cdot \sum_{u=1}^{U} \| P_u - Q_u \|_1 + 2Un + Un} \qquad (56) \\
&\geq \min\left\{ \frac{m^{3/2} \sum_{u=1}^{U} \| P_u - Q_u \|_1}{2U\sqrt{n}}, \frac{m^2 \left(\sum_{u=1}^{U} \| P_u - Q_u \|_1\right)^2}{6Un} \right\}, \qquad (57)
\end{aligned}$$

where the first inequality follows by the trivial bound $\max(a, b) \leq a + b$, for any two non-negative numbers a and b. ■

Applying Lemma 9 on (23)–(25), we obtain the following corollary for $m \geq n$.

Corollary 10 *For $m \geq n$, the following hold with probability at least $1 - \delta_1$,*

$$\begin{aligned}
&\mathrm{E}\left[G \hat{f}_{u,i}, u \in [U], i \in [n] \right] \\
&\quad \geq \frac{\delta_1}{c_1} \min\left\{ \frac{m^{3/2} \sum_{u=1}^{U} \| P_u - Q_u \|_1}{2U\sqrt{n}}, \frac{m^2 \left(\sum_{u=1}^{U} \| P_u - Q_u \|_1\right)^2}{6Un} \right\},
\end{aligned} \tag{58}$$

$$\mathrm{E}\left[G \hat{f}_{u,i}, u \in [U], i \in [n] \right] \leq \frac{c_2}{\delta_1} \sum_{u=1}^{U} \frac{m^{3/2} \| P_u - Q_u \|_1}{n^{\frac{1}{2}}}, \tag{59}$$

and

$$\mathrm{Var}\left[G \hat{f}_{u,i}, u \in [U], i \in [n] \right] \leq \frac{c_3}{\delta_1} \frac{10Um^2}{n}. \tag{60}$$

Next, we move froward to the case where $m \leq n$. We have the following result.

Lemma 11 *For $m \leq n$, the following hold,*

$$\sum_{i=1}^{n} \frac{\mathrm{Var}(G_{u,i})}{f_{u,i}^2} \leq 24m, \tag{61}$$

$$\sum_{i=1}^{n} \frac{m^2 (P_{u,i} - Q_{u,i})^2}{f_{u,i}} \leq m \| P^{u} - Q_{u} \|_1 , \tag{62}$$

and

$$\frac{m^2 \sum_{u=1}^{U} \| P_u - Q_u \|_1^2}{\sum_{u=1}^{U} \sum_{i=1}^{n} f_{u,i}} \geq \frac{m^2 \sum_{u=1}^{U} \| P_u - Q_u \|_1^2}{3n}. \tag{63}$$

Proof [Proof of Lemma 11] As before, we start by proving (61). We have,

$$\sum_{i=1}^{n} \frac{\mathrm{Var}(G_{u,i})}{f_{u,i}^2} = \sum_{i=1}^{n} \frac{4m^3 (P_{u,i} - Q_{u,i})^2 (P_{u,i} + Q_{u,i}) + 2m^2 (P_{u,i} + Q_{u,i})^2}{f_{u,i}^2} \tag{64}$$

$$= \sum_{i=1}^{n} \frac{4m^3 (P_{u,i} - Q_{u,i})^2 (P_{u,i} + Q_{u,i}) + 2m^2 (P_{u,i} + Q_{u,i})^2}{(\max\{m \cdot (P_{u,i} + Q_{u,i}), 1\})^2} \tag{65}$$

$$\leq \sum_{i=1}^{n} \frac{4m^3 (P_{u,i} - Q_{u,i})^2 (P_{u,i} + Q_{u,i}) + 4m^2 (P_{u,i} - Q_{u,i})^2 + 8m^2 Q_{u,i}^2}{\max\{m^2 \cdot (P_{u,i} + Q_{u,i})^2, 1\}} \tag{66}$$

$$\leq \sum_{i=1}^{n} \frac{4m^3 (P_{u,i} - Q_{u,i})^2 (P_{u,i} + Q_{u,i})}{m^2 \cdot (P_{u,i} + Q_{u,i})^2} + \sum_{i=1}^{n} \frac{4m^2 (P_{u,i} - Q_{u,i})^2}{m \cdot (P_{u,i} + Q_{u,i})} + \sum_{i=1}^{n} \frac{8m^2 Q_{u,i}^2}{\max\{m^2 (P_{u,i} + Q_{u,i})^2, 1\}} \tag{67}$$

$$\leq 4m \sum_{i=1}^{n} |P_{u,i} - Q_{u,i}| + 4m \sum_{i=1}^{n} |P_{u,i} - Q_{u,i}| + \sum_{i=1}^{n} \frac{8m^2 Q_{u,i}^2}{\max\{m (P_{u,i} + Q_{u,i}), 1\}} \tag{68}$$

$$\leq 8m \| P_u + Q_u \|_1 + \sum_{i=1}^{n} 8m Q_{u,i} \tag{69}$$

$$\leq 24m, \tag{70}$$

where the first inequality follows from the fact that $(a + b)^2 \leq 2(a - b)^2 + 4b^2$, for any $a, b \geq 0$, the second inequality follows by lower bounding the denominator by individual terms in the maximum, and the third inequality follows from the trivial bound $\frac{|a-b|}{a+b} \leq 1$, for $a, b \geq 0$. Next, for (62), we note that,

$$\sum_{i=1}^{n} \frac{m^2 (P_{u,i} - Q_{u,i})^2}{f_{u,i}} = \sum_{i=1}^{n} \frac{m^2 |P_{u,i} - Q_{u,i}|^2}{\max\{m \cdot (P_{u,i} + Q_{u,i}), 1\}} \tag{71}$$

$$\leq \sum_{i=1}^{n} m |P_{u,i} - Q_{u,i}| = m \| P_u - Q_u \|_1. \tag{72}$$

Finally, we prove (63). We have,

$$\frac{m^2 \sum_{u=1}^{U} \| P_u - Q_u \|_1^2}{\sum \sum} = \frac{m^2 \sum_{u=1}^{U} \| P_u - Q_u \|_1^2}{\sum \sum} \tag{73}$$

$$^{U}_{u=1} \quad ^{n}_{i=1} \; f_{u,i} \qquad\qquad ^{U}_{u=1} \quad ^{n}_{i=1} \; \max\{m \cdot (P_{u,i} + Q_{u,i}) , 1\}$$

$$\geq \frac{m^2 \left(\sum_{u=1}^{U} |P_u - Q_u|_1\right)^2}{\sum_{u=1}^{U} \left(m \cdot |P_u + Q_u|_1 + L\right)} \tag{74}$$

$$= \frac{m^2 \left(\sum_{u=1}^{U} |P_u - Q_u|_1\right)^2}{2m + n} \tag{75}$$

$$\geq \frac{m^2 \left(\sum_{u=1}^{U} |P_u - Q_u|_1\right)^2}{3n}. \tag{76}$$

This concludes the proof. ■

Applying Lemma 11 on (23)–(25), we obtain the following corollary for $m \leq n$.

Corollary 12 *For $m \leq n$, the following hold with probability at least $1 - \delta_1$,*

$$\mathrm{E}\left[G\hat{f}_{u,i}, u \in [U], i \in [n]\right] \geq \frac{\delta_1}{c1} \frac{m^2 \left(\sum_{u=1}^{U} |P_u - Q_u|_1\right)^2}{3n}, \tag{77}$$

$$\mathrm{E}\left[G\hat{f}_{u,i}, u \in [U], i \in [n]\right] \leq \frac{c_2}{\delta_1} \sum_{u=1}^{U} m |P_u - Q_u|_1, \tag{78}$$

$$\mathrm{Var}\left[G\hat{f}_{u,i}, u \in [U], i \in [n]\right] \leq \frac{24mUc_3}{\delta_1}, \tag{79}$$

and

$$\mathrm{Var}\left[G\hat{f}_{u,i}, u \in [U], i \in [n]\right] \leq \frac{1}{40} \mathrm{E}\left[G\hat{f}_{u,i}, u \in [U], i \in [n]\right]^2 + 324\mathrm{E}\left[G\hat{f}_{u,i}, u \in [U], i \in [n]\right] + 648m^2 \sum_{u=1}^{U} |Q_u|_2^2. \tag{80}$$

Proof [Proof of Corollary 12] Inequalities (77)–(79) follow almost directly from Lemma 11, and we next focus on (80). First, note that

$$\mathrm{Var}\left[G \hat{f}_{u,i}\right] = \sum_{u=1}^{U}\sum_{i=1}^{n} \frac{\mathrm{Var}(G_{u,i})}{\hat{f}_{u,i}^2} \tag{81}$$

$$= \sum_{u=1}^{U}\sum_{i=1}^{n} \frac{4m^3 (P_{u,i} - Q_{u,i})^2 (P_{u,i} + Q_{u,i}) + 2m^2 (P_{u,i} + Q_{u,i})^2}{\hat{f}_{u,i}^2} \tag{82}$$

$$\overset{(a)}{\leq} 4m^3 \left(\sum_{u=1}^{U}\sum_{i=1}^{n} \frac{(P_{u,i} - Q_{u,i})^4}{\hat{f}_{u,i}^2}\right)^{\frac{1}{2}} \left(\sum_{u=1}^{U}\sum_{i=1}^{n} \frac{(P_{u,i} + Q_{u,i})^2}{\hat{f}_{u,i}^2}\right)^{\frac{1}{2}} + \sum_{u=1}^{U}\sum_{i=1}^{n} \frac{2m^2 (P_{u,i} + Q_{u,i})^2}{\hat{f}_{u,i}^2} \tag{83}$$

$$\overset{(b)}{\leq} 4m^3 \left(\sum_{u=1}^{U}\sum_{i=1}^{n} \frac{(P_{u,i} - Q_{u,i})^2}{\hat{f}}\right) \left(\sum_{u=1}^{U}\sum_{i=1}^{n} \frac{(P_{u,i} + Q_{u,i})^2}{\hat{f}^2}\right)^{\frac{1}{2}}$$

u=1 i=1 u,i u=1 i=1 u,i

$$+\sum_{u=1}^{U}\sum_{i=1}^{n}\frac{2m^2\left(P_{u,i}+Q_{u,i}\right)^2}{\hat f^2_{u,i}} \tag{84}$$

$$4\left(m^2\sum_{u=1}^{U}\sum_{i=1}^{n}\frac{\left(P_{u,i}-Q_{u,i}\right)^2}{\hat f_{u,i}}\right)\left(m^2\sum_{u=1}^{U}\sum_{i=1}^{n}\frac{\left(P_{u,i}+Q_{u,i}\right)^2}{\hat f^2_{u,i}}\right)^{\frac{1}{2}}$$

$$+\sum_{u=1}^{U}\sum_{i=1}^{n}\frac{2m^2\left(P_{u,i}+Q_{u,i}\right)^2}{\hat f^2_{u,i}}, \tag{85}$$

where (a) follows from the Cauchy-Schwartz inequality, and (b) is due to the monotonicity of the l_p norm, i.e., for any vector u, $\|u\|_2 \leq \|u\|_1$. Then,

$$\mathrm{Var}\left[G\hat f_{u,i}\right] = 4\,\mathrm{E}\left[G\hat f_{u,i}\right]\left(\sum_{u=1}^{U}\sum_{i=1}^{n}\frac{m^2\left(P_{u,i}+Q_{u,i}\right)^2}{\hat f^2_{u,i}}\right)^{\frac{1}{2}}+\sum_{u=1}^{U}\sum_{i=1}^{n}\frac{2m^2\left(P_{u,i}+Q_{u,i}\right)^2}{\hat f^2_{u,i}} \tag{86}$$

$$\overset{(a)}{\leq}\frac{1}{40}\,\mathrm{E}\left[G\hat f_{u,i}\right]^2+(160+2)\sum_{u=1}^{U}\sum_{i=1}^{n}\frac{m^2\left(P_{u,i}+Q_{u,i}\right)^2}{\hat f_{u,i}} \tag{87}$$

$$\overset{(b)}{\leq}\frac{1}{40}\,\mathrm{E}\left[G\hat f_{u,i}\right]^2+162\sum_{u=1}^{U}\sum_{i=1}^{n}\frac{2m^2\left(P_{u,i}-Q_{u,i}\right)^2}{\hat f^2_{u,i}}+162\sum_{u=1}^{U}\sum_{i=1}^{n}\frac{4m^2Q^2_{u,i}}{\hat f^2_{u,i}} \tag{88}$$

$$\overset{(c)}{\leq}\frac{1}{40}\,\mathrm{E}\left[G\hat f_{u,i}\right]^2+324\sum_{u=1}^{U}\sum_{i=1}^{n}\frac{m^2\left(P_{u,i}-Q_{u,i}\right)^2}{\hat f_{u,i}}+648\sum_{u=1}^{U}\sum_{i=1}^{n}m^2Q^2_{u,i} \tag{89}$$

$$=\frac{1}{40}\,\mathrm{E}\left[G\hat f_{u,i}\right]^2+324\mathrm{E}\left[G\hat f_{u,i}\right]+648m^2\sum_{u=1}^{U}\|Q_u\|_2^2 \tag{90}$$

$$=\frac{1}{40}\,\mathrm{E}\left[G\left(\hat f_{u,i}, u\in[\mathrm{U}], i\in[n]\right)\right]^2+324\mathrm{E}\left[G\left(\hat f_{u,i}, u\in[\mathrm{U}], i\in[n]\right)\right]$$

$$+\,648m^2\sum_{u=1}^{U}\|Q_u\|_2^2, \tag{91}$$

where in (a) we use the fact that $2ab \leq a^2 + b^2$, (b) follows from $(a+b)^2 \leq 2(a-b)^2 + 4b^2$, and finally (c) is because $\hat f_{u,i} \geq 1$. ■

4.1.2 Proof of Theorem 4.

We start with the case where $m \geq n$. By Chebyshev's inequality,

$$\mathrm{P}\left(|G-\mu|\leq\frac{\sigma}{\sqrt{\delta}}\left(\hat f_{u,i}, u\in[\mathrm{U}], i\in[n]\right)\right)\geq 1-\delta, \tag{92}$$

where

$$\mu = \mathrm{E}\left[G\, f_{u,i} \text{ for } u \in [U], i \in [n]\right], \tag{93}$$

$$\sigma^2 = \mathrm{Var}\left[G\, \widehat{f}_{u,i} \text{ for } u \in [U], i \in [n]\right]. \tag{94}$$

Accordingly, using Corollary 10, we get that with probability at least $1-\delta$,

$$\frac{\delta_1}{c_1}\min\left(\frac{m^{3/2}\sum_{u=1}^{U}|P_u-Q_u|_1}{2U\sqrt{n}}, \frac{m^2\left(\sum_{u=1}^{U}|P_u-Q_u|_1\right)^2}{6Un}\right) - m\sqrt{\frac{10c_3U}{\delta\delta_1 n}} \leq G, \tag{95}$$

and

$$G \leq \frac{c_2}{\delta^1}\sum_{u=1}^{U}\frac{m^{3/2}|P_u-Q_u|_1}{n^{\frac{1}{2}}} + m\sqrt{\frac{10c_3U}{\delta\delta_1 n}}. \tag{96}$$

We will consider the two possible cases where either $\sum_{u=1}^{U}|P_u-Q_u|_1 \geq \varepsilon_2 U$ or $\sum_{u=1}^{U}|P_u-Q_u|_1 \leq \varepsilon_1 U$. Starting with the former, the lower bound in (95) reduces to

$$G \geq \frac{\delta_1}{c_1}\min\left(\frac{m^{3/2}\varepsilon_2 U}{2U\sqrt{n}}, \frac{m^2\varepsilon_2^2 U^2}{6Un}\right) - m\sqrt{\frac{10c_3U}{\delta\delta_1 n}} \tag{97}$$

$$\geq \frac{\delta_1}{12c_1}\min\left(\frac{m^{3/2}\varepsilon_2}{\sqrt{n}}, \frac{m^2\varepsilon_2^2 U}{n}\right), \tag{98}$$

where the second inequality holds as long as $m \geq \max\left(\frac{12^2}{5}\frac{10c_3c_1^2}{\delta_1^2}\frac{U}{\delta\varepsilon_2^2}, \frac{12c_1}{\delta_1}\sqrt{\frac{10c_3}{\delta_1}}\sqrt{\frac{n}{U\delta\varepsilon_2^4}}\right) = C\max\left(\frac{U}{\delta\varepsilon_2^2}, \sqrt{\frac{n}{U\delta\varepsilon_2^4}}\right)$ for some constant $C > 0$. Therefore, with probability at least $1-\delta$ the test announces that $\sum_{u=1}^{U}|P_u-Q_u|_1 \geq U\varepsilon_2$ correctly.

On the other hand, in the case where $\sum_{u=1}^{U}|P_u-Q_u|_1 \leq \varepsilon_1 U$, using Corollary 10 once again, we get that with probability at least $1-\delta$,

$$G \leq \frac{c_2}{\delta_1}\frac{m^{3/2}\varepsilon_1 U}{\sqrt{n}} + m\sqrt{\frac{10c_3U}{\delta\delta_1 n}} \tag{99}$$

$$\leq \frac{\delta_1}{24c_1}\min\left(\frac{m^{3/2}\varepsilon_2}{\sqrt{n}}, \frac{m^2\varepsilon_2^2 U}{n}\right), \tag{100}$$

where the second inequality holds as long as $m > \max\left(\frac{\delta_1\varepsilon_2}{c_2\varepsilon_1}\sqrt{\frac{1}{U\delta}}, \frac{c_2^2 n\varepsilon_1^2}{\delta_1^2\varepsilon_2^4}\right) = \tilde{C}\max\left(\frac{\varepsilon_2}{\varepsilon_1}\sqrt{\frac{1}{U\delta}}, \frac{n\varepsilon_1^2}{\varepsilon_2^4}\right)$, for some constant $\tilde{C} > 0$. Therefore, with probability at least $1-\delta$ the test announces that $\sum_{u=1}^{U}|P_u-Q_u|_1 \leq \varepsilon_1$ correctly.

Finally, we turn to the case $m \leq n$. Using Corollary 12 and repeating the same arguments above, we get that the test announces $\sum_{u=1}^{U}|P_u-Q_u|_1 \geq \varepsilon_2 U$ correctly if $m \geq n\frac{\varepsilon_1}{\varepsilon_2^2}$, while announces $\sum_{u=1}^{U}|P_u-Q_u|_1 \leq \varepsilon_1 U$ correctly, as long as $m \geq \frac{n^{2/3}}{U\varepsilon_2^{4/3}}$. This concludes the proof.

4.2 Proof of Theorem 5

We start with the following lemma generalizes the "exponential decay lemma" in (Chan et al., 2021) for the case where we have *m* independent Markov chains. This result will be essential in the proof of Theorem 5. For simplicity of notations, we will sometimes suppress the dependency of the covering quantities on M.

Lemma 13 (Exponential decay) *For* M *independent irreducible Markov chains* $\{Z^{\infty}_{m,1}\}^{M}_{m=1}$, *on the same state space* $[n]$, *for any* k, L $\in$ N, *and any initial distribution* **q** *over* $[n]^{M}$, *we have*

$$P \quad \tau^{(k)}_{cov}(m) \geq eLt^{(k)}_{cov}(m; M) \quad \leq e^{-L}. \tag{101}$$

Proof [Proof of Lemma 13] Consider $\tau^{(k)}_{cov}(M; M)$ with any fixed starting states $Z_{l,1} = v_l$, for $l \in [M]$. By Markov's inequality,

$$P(\tau^{(k)}_{cov}(M) \geq et^{(k)}_{cov}(M; M)) \leq P \quad \tau^{(k)}_{cov}(M) \geq eE \overset{h}{} \tau^{(k)}_{cov}(M) \quad Z_{l,1} = v_l, \ \forall j \in [M] \overset{i}{} \tag{102}$$

$$\leq e^{-1}. \tag{103}$$

Note that this inequality holds for any initial states $\mathbf{v} = (v_1, . . . , v_M) \sim \mu$, where μ is any discrete distribution over $[n]^M$. We next analyze sub-trajectories of our Markov chain of length $v \triangleq et^{(k)}_{cov}(M; M)$. Specifically, for any $1 \leq l \leq L$, we define E_l as the event that the set of M sub-trajectories of the Markov chains $\{Z^{lv}_{m,(l-1)v+1}\}^{M}_{m=1}$ jointly cover the state space k times. According to (102), we have

$$P(E^c_1) = P \quad \tau^{(k)}_{cov}(M) \geq et^{(k)}_{cov}(M; M) \quad \leq e^{-1}. \tag{104}$$

Denote the distribution of $\{Z_{m,v}\}_{m\in[M]}$ conditioned on E^c_1 by μ^r, and let $\tau^{(k)'}_{cov}(M)$ be the M-joint k-cover time of $\{Z^{\infty}_{m,v+1}\}^{M}_{m=1}$. We have,

$$P(E^c_2|E^c_1) = P \quad \tau^{(k)'}_{cov}(M) \geq et^{(k)}_{cov}(M) \ \ \tau^{(k)}_{cov}(M) \geq et^{(k)}_{cov}(M; M) \tag{105}$$

$$= P \quad \tau^{(k)'}_{cov}(M) \geq et^{(k)}_{cov}(M; M) \ \ \{Z_{m,v}\}_{m\in[M]} \sim \mu^r \ \leq e^{-1}, \tag{106}$$

where we used the fact that E_1 is determined by $\{Z^{v}_{m,1}\}_{m\in[M]}$, and that by the Markov property the event E_2 do not depend on $\{Z^{v-1}_{m,1}\}_{m\in[M]}$. Thus,

$$P(E^c_1 \cap E^c_2) = P(E^c_1)\, P(E^c_2 \mid E^c_1) \leq e^{-2}. \tag{107}$$

Using the same arguments above, by induction, we can show that P $\cap_{i\in[L]} E^c_i \leq e^{-L}$. Define E as the event that the set of M sub-trajectories of the Markov chains $\{Z^{Lv}_{m,1}\}^{M}_{m=1}$ jointly cover the state space k times. Then, it is clear that $\cup_{i\in[L]} E_i \subseteq E$, and thus, $P(E^c) \leq$ P $\cap_{i\in[L]} E^c_i \leq e^{-L}$, which concludes the proof. ■

We are now in a position to prove Theorem 5. Recall that $\overline{m} = m(n, \varepsilon_1, \varepsilon_2, \delta/4n)$ is the sample complexity guarantee associated with Algorithm 1. For $u \in [U]$ and $l \in N$ we define the events,

$$E_{R,u}(l) \triangleq \bigcap_{m=1}^{M} \left\{ N_i^{x^R_{m,u}}(l) \geq \bar{m},\ \forall i \in [n] \right\}, \tag{108}$$

$$E_{F,u}(l) \triangleq \bigcap_{m=1}^{M} \left\{ N_i^{x^F_{m,u}}(l) \geq \bar{m},\ \forall i \in [n] \right\} \cdot \tag{109}$$

Furthermore, for any $i \in [n]$, we define $\tilde{E}_i$ as the event that steps 6-8 in Algorithm 2 return "NO", namely, that the first $\bar{m}$ succeeding samples of state $i \in [n]$ in the union of the M Markov trajectories do not pass the i.i.d. tester in Algorithm 1.

To establish Theorem 5, we consider the two possible cases where $V_{\text{filter}} \leq \varepsilon_1$, and the complementary case where $V_{\text{filter}} \geq \varepsilon_2$. Starting with the former, according to Lemma 13, by taking $L = \log\frac{4U}{\delta}$, we get $P\left(\tau^{(\bar{m})}_{\text{cov}}(M; Q^F_u) \geq e t^{(\bar{m})}_{\text{cov}}(M; Q^F_u) \log\frac{4U}{\delta}\right) \leq \frac{\delta}{4U}$ and $P\left(\tau^{(\bar{m})}_{\text{cov}}(M; P^R_u) \geq e t^{(\bar{m})}_{\text{cov}}(M; P^R_u) \log\frac{4U}{\delta}\right) \leq \frac{\delta}{4U}$, for all $u \in [U]$. Thus, for a length $l = T$ trajectory in (108)–(109), where T is as stated in Theorem 5, i.e.,

$$T = e \log\frac{4U}{\delta} \max_{u \in [U]} \max_{W \in \{Q^F_u, P^R_u\}} t^{\bar{m}}_{\text{cov}}(M; W), \tag{110}$$

we will have $\bar{m}$ samples for each state in $[n]$ with probability $P(E_{R,u}(T)) \geq 1 - \frac{\delta}{4U}$, for any $u \in [U]$. Similarly, $P(E_{F,u}(T)) \geq 1 - \frac{\delta}{4U}$, for any $u \in [U]$. Thus, by a union bound over the two chains and the set all users $u \in [U]$, the probability of passing the condition in step 5 of Algorithm 2 is at least $\geq 1 - \frac{\delta}{2}$. Furthermore, by the sample complexity guarantee in Theorem 4 associated with the i.i.d. tester in Algorithm 1, we have $P(\tilde{E}_i) \leq \frac{\delta}{4n}$, implying that, $P(\cup_{u \in U} E^c_{R,u}(T) \cup E^c_{F,u}(T) \cup \tilde{E}_1 \ldots \cup \tilde{E}_n) \leq \frac{3\delta}{4}$. Thus, with probability at least $1 - \delta$, the tester will return YES.

Next, we move forward to the complementary case. Note that the only case the algorithm outputs YES is when it do not pass steps 4 and 7 in Algorithm 2 for all states, which means it will have enough samples for testing each state, and the i.i.d. tester outputs YES for all sub-tests. Since $V_{\text{filter}} \geq \varepsilon_2$ implies that there exists $i_\wedge \in [n]$ such that $\sum_{u=1}^{U} \left\| \mathbf{P}^{\mathbf{R}}_u(i_\wedge) - \mathbf{Q}^{\mathbf{F}}_u(i_\wedge) \right\|_1 \geq U\varepsilon_2$, this guarantees that the sub-test for $i_\wedge$ will return NO with probability $P(\tilde{E}_{s_i}) \geq 1 - \frac{\delta}{4n}$ again due to the sample complexity guarantees for the i.i.d. tester in Theorem 4. Thus, the probability for the whole procedure outputting YES is $P(\cap_{u \in [U]} E_{R,u}(T) \cap E_{F,u}(T) \cap \tilde{E}^c_1 \ldots \cap \tilde{E}^c_n) \leq P(\tilde{E}^c_{i_s}) \leq \delta/4n$.

Combining both cases above, it is clear that Algorithm 2 will output the correct answer with probability at least $1 - \delta$.

4.3 Proof of Lemma 3

To prove Lemma 3 we start with the following concentration bound on the random l-joint k-hitting time $\tau^{(k)}_{\text{hit}}(l; i)$, defined as the first time when a particular state $i \in [n]$ is visited k

times jointly by the l Markov chains. Mathematically, we define

$$\tau^{(k)}_{\text{hit}}(l;i) \triangleq \inf\left\{t \geq 0 : \sum_{j=1} N_i^{Z_j}(t) \geq k\right\}, \tag{111}$$

for $i \in [n]$. The l-joint k-hitting time is then defined as

$$t^{(k)}_{\text{hit}}(l) \triangleq \max_{i\in[n],\mathbf{v}\in[n]^{\ell}} \mathrm{E}\left[\tau^{(k)}_{\text{hit}}(l;i) \mid Z_{1,1}=v_1, Z_{2,1}=v_2, ..., Z_{l,1}=v_l\right]. \tag{112}$$

Finally, we define the l–joint return time. For some state $i \in [n]$, the random l–joint return time $\tau_{\text{ret}}(l;i)$ is the first time one of the l Markov chains starting at i return to i. The l–joint return time is then defined as $t_{\text{ret}}(l;i) = \mathrm{E}[\tau_{\text{ret}}(l;i)|Z_{1,1} = i, Z_{2,1} = i, ..., Z_{l,1} = i]$. It is standard result that for irreducible chains $t_{\text{ret}}(1;i) = 1/\pi_i$, where π is the stationary distribution.

Lemma 14 *Let $Z^{\infty}_{1,1}, Z^{\infty}_{2,1}, ..., Z^{\infty}_{l,1}$ be l-independent infinite trajectories drawn by the same Markov chain M. Then, for any $i \in [n]$,*

$$\mathrm{P}\left[\tau^{(k)}_{\text{hit}}(l;i) \geq t\right] \leq \exp\left(-\frac{t}{e\mu_i}\right), \tag{113}$$

for any $t \geq 0$, where $\mu_i \triangleq t^{(1)}_{\text{hit}}(l) + \frac{k}{\pi_i}$.

Proof [Proof of Lemma 14] First, we note that by the Markov property we have

$$\tau^{(k)}_{\text{hit}}(l;i) = \tau^{(1)}_{\text{hit}}(l;i) + (k-1)\cdot t_{\text{ret}}(l;i) \tag{114}$$

$$\leq \tau^{(1)}_{\text{hit}}(l;i) + \frac{k}{\pi_i} \tag{115}$$

$$\leq t^{(1)}_{\text{hit}}(l) + \frac{k}{\pi_i} = \mu_i, \tag{116}$$

where the last inequality holds by definition with probability one. Thus, by Markov inequality we get,

$$\mathrm{P}\left[\tau^{(k)}_{\text{hit}}(l;i) \geq e\mu_i\right] \leq \frac{1}{e}. \tag{117}$$

Using the same arguments as in the proof of the exponential decay result in Lemma 13, we can show that for any $l \geq 1$,

$$\mathrm{P}\left[\tau^{(k)}_{\text{hit}}(l;i) \geq en\mu_i\right] \leq e^{-\kappa}. \tag{118}$$

Thus, the result follows by taking $n = t/e\mu_i$. ■

Using the above result we are now in a position to prove Lemma 3. First, it is clear that $\tau^{(k)}_{\text{cov}}(l) \leq \max_{i\in[n]} \tau^{(k)}_{\text{hit}}(l;i)$. Thus, by the union bound and Lemma 14, we have for any $t \geq 0$,

$$\mathrm{P}\left[\tau^{(k)}_{\text{cov}}(l) \geq t\right] \leq \sum_{i\in[n]} \exp\left(-\frac{t}{e\mu_i}\right) \leq n\exp\left(-\frac{t}{e\min_{i\in[n]}\mu_i}\right). \tag{119}$$

Therefore, we have,

$$\mathrm{E}[\tau_{\mathrm{cov}}^{(k)}(l)] = \int_0^\infty \mathrm{P}\left(\tau_{\mathrm{cov}}^{(k)}(l) \geq t\right) \mathrm{d}t \tag{120}$$

$$\leq \int_0^{e \min_{i\in[n]} \gamma_i \log n} \mathrm{d}t + n \int_{e \min_{i\in[n]} \gamma_i \log n}^{\infty} \exp\left(-\frac{t}{e \min_{i\in[n]} \gamma_i}\right) \mathrm{d}t \tag{121}$$

$$= e \min_{i\in[n]} \gamma_i \log n + e^2 \min_{i\in[n]} \gamma_i. \tag{122}$$

Thus, it follows that $t_{\mathrm{cov}}^{(k)}(l) = \mathrm{O}\left(\min_{i\in[n]} \gamma_i \log n\right) = \mathrm{O}\left(t_{\mathrm{hit}}^{(1)}(l) \log n + \frac{k \log n}{\pi_s}\right)$, and therefore, $t_{\mathrm{cov}}^{(k)}(l) = \mathrm{O}\left(t_{\mathrm{cov}}^{(1)}(l) \log n + \frac{k \log n}{\pi_s}\right)$. Finally, combining Theorem 3.2, Lemma 4.3, and Theorem 4.7 in (Rivera et al., 2023), we have that,

$$t_{\mathrm{cov}}^{(1)}(l) = \mathrm{O}\left(\max\left(t_{\mathrm{mix}}, \frac{t_{\mathrm{cov}} \log n}{l}\right)\right), \tag{123}$$

which concludes the proof.

5. Conclusion and Future Research

In this paper, we modeled the relationship between the three stakeholders: the platform, the users, and the auditor. The essence of the modeling is that from the auditor's perspective the platform is a content-generating system formulated by a multidimensional first order Markov chain (as the fixed number of pieces of the content appearing on each feed), where at every time step the platform samples a new feed, according to the Markov transition- matrix (conditional probability). We developed an auditing method that tests whether there are unexpected deviations in the user's decision-making process over a predefined time horizon. Unexpected deviations in the user's decision-making process might be a result of the selective filtering of the contents to be shown on the user's feed in comparison to what would be the users' decision-making process under natural filtering. We proposed also an auditing procedure for online counterfactual regulations.

There are several exciting directions for future work, including the following. A major goal going forward is to evince our auditing procedure on real social media content. Specifically, while our work propose a theoretical framework for SMP auditing, we left several fundamental questions that revolve around implementability, such as, how do we know if the framework is effective or useful? What are the metrics that should be used? We are currently investigating these kind of questions. From the technical perspective, there are many interesting generalizations an open questions that we plan to investigate. For example, studying a sequential version of the testing problems proposed in this paper are of particular importance. Indeed, in real-world platforms decisions must be taken as quickly as possible so that proper countermeasures can be taken to suppress regulation violation. Moreover, real-world networks are gigantic and therefore it is quite important to study the performance of low-complexity algorithms. Also, it is of both theoretical and practical importance to consider more general probabilistic models which will, for example, capture the dynamic relationships between users, the varying influence of individual users within the platform, and in general weaken some of the assumptions we made about the behaviour of users, the platform/algorithm, the social relationships and dynamics. It would be interesting to consider more complicated/real-world motivated generative models, such as, higher-order Markov chains, as well as simple structured dependencies among the M contents. In this paper we considered testing against a single agree upon definition for a reference feed. However, there is more than one "natural/fair" way to filter contents. Accordingly, it would be more robust and general to require closeness to the set of "natural/fair" references, which may be very different from one another. From the conceptual perspective, while our paper propose several definition for the notions of "variability" and "violation", there probably are other possible definitions, which take into account some perspectives of responsible regulation which we ignored, and are important to investigate.

Finally, it should be clear that each approach, worst-case (Cen and Shah, 2021) or average, has its own advantages and disadvantages. For example, the worst-case approach might be sensitive to adversarial users; in real-world SMPs, where any party is free to create a user without any supervision, a set of adversarial users can act as more naive/gullible than the most gullible user already in existence, and thus fool the auditor. Also, the worst- case approach prevents all users from gradually changing their opinions. This is because, under this approach, the auditing process will immediately result in a violation when the most gullible user alters its opinion slightly. As a result, all other users will not have the opportunity to make slow and natural changes to their opinions, as they would with our average approach. In some sense, the above problematic issues are less severe/relevant in our average approach. In the average approach, the auditing procedure would not prevent the SMP from identifying a set $O(1)$-many users who, for example, are most likely to tip the outcome of an election and promoting one presidential candidate to them, while using the reference feed itself for the remaining users. However, if the set of chosen users is "large enough", coupled with good-faith effort to choose and test features, this issue can be resolved. It is clear, nonetheless, that further research of both approaches (and perhaps others) is needed so as to achieve better understanding of this complicated problem of auditing/regulating SMPs.

Acknowledgments

This work was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 1734/21).

Appendix A. Detailed Construction of Our Framework

In this appendix, we provide more detailed discussions about the framework, definitions, and assumptions presented in Section 2.

A.1 User-platform relationship

Users. We describe the *users learning and decision-making pipeline*. As users browse through their feeds, they implicitly form internal *beliefs* about the observed contents, and based on those beliefs they later take *actions/decisions*. For example, how individuals vote or the products they buy are decisions that are affected by the content they see on social media. In addition, the decisions does not have to occur on the platform. For instance, the platform could show information on COVID-19, but the decision could be whether to get the vaccine. Let us formulate this mathematically. Let Ω be a compact metrizable set of possible states; this set can be finite, countably infinite, or continuums, and its elements $\omega \in \Omega$ can be either scalars or vectors. At each time step $t \geq 1$, each user $u \in [U]$ is associated with a belief $B^F_{u,t} \in \Delta(\Omega)$, where $\Delta(\Omega)$ is the simplex of probability distributions over the state space. At start $t = 0$, without loss of essential generality, we may assume that $B^F_{u,0} = \text{Uniform}(\Omega)$, for all $u \in [U]$. The belief $B^F_{u,t}$ is a posterior distribution on Ω *conditioned* on the information available to user u at time t. This information consists of the observed feeds $\{\mathbf{X}^F_u(l)\}_{l \leq t}$. The total history information available to user u at time t by $H^F_{t,u} \triangleq \{\mathbf{X}^F_u(l) : l \leq t\}$. Accordingly, user u's belief at time t is defined as $B^F_{u,t}(d\omega, h_{t,u}) \triangleq P^F(d\omega | H^F_{t,u} = h_{t,u})$, for a given sequence of feeds $h_{t,u}$. Based on the beliefs users take decisions (or, actions); each user have a set of possible *actions* at time $t \geq 0$. For user $u \in U$, let $A_u(t)$ denote a compact metrizable action space, and $A_{u,i}(t) \in A_u(t)$ be the ith action. Also, let $U_u : \Omega \times A_u \rightarrow R$ be (possibly continuous) user u's utility function. Consequently, for any belief $B^F_{u,t} \in \Delta(\Omega)$ and a utility function U_u we define $\mathrm{br}^F_{u,t}(h_{t,u})$ as the set of actions that maximizes user u's expected utility, i.e.,

$$\mathrm{br}^F_{u,t}(h_{t,u}) \triangleq \left\{ a \in A_u : a \in \arg\max_{b \in A_u} \int_\Omega U_u(\omega, b) B^F_{u,t}(d\omega, h_{t,u}) \right\}.$$

A.2 Feeds construction and auditor-platform interaction

We now switch our focus to formalize the setup for the auditor-platform interaction.

Platform filtering. An important component of our model is related to the question of *how feeds are filtered*? As mentioned before, feeds are chosen by the platform using a black-box filtering algorithm, which is utilized to maximize a certain reward function. The filtering algorithm is fed with an extensive amount of inputs that the platform uses to filter, such as, current available contents, past feeds, users interaction history, users feedback (e.g., users "sentiments" which are certain complex functions of the users beliefs), the users social network topology, and so on. The reward function reflects the platform's objective. For example, it may balance factors like advertising revenue, personalization, user engagement (e.g., the predicted number of clicks), content novelty, acquisition of new information about users, cost of operations, or a combination of these and other factors. We denote the platform's reward function by $\mathrm{Rew}^F_t : X^M \times P_{t-1} \rightarrow R$, where P_{t-1} captures the inputs mentioned above, and accordingly,

$$\mathbf{X}^F_u(t) = \arg\max_{\mathbf{x} \in F^M} \mathrm{Rew}^F_t(\mathbf{X}, P_{u,t-1}), \tag{124}$$

where, again, $P_{u,t-1}$ captures the platform external data used for filtering. For now, we leave both Rew^F_t and P_{t-1} unspecified.

Filtered vs. reference feeds. The discussion in the background section about the regulation boundary and motivation suggests a neat and consistent formulation for the auditor's objective. Following (Wachter and Mittelstadt, 2019; Ghosh, 2019; Cen et al., 2023; Petty, 2000), we define a reference (or, competitive) boundary that is formed based on the users consent, and its location is determined by domain experts. While user *u*'s filtered feed $\mathbf{X}^{F}_{u}(t)$ at time *t* is chosen by the platform in a certain reward-maximizing methodology, On the other hand, the *reference feeds* $\mathbf{X}^{R}_{u}(t)$ could have been hypothetically selected by the platform if it strictly followed the consumer-provider agreement. These reference feeds are specific to each user $u \in [U]$ and time *t.* In this scenario, the platform would construct the feed based solely on the user's interests, without any subjective preferences influencing the content selection. Essentially, the only natural situation where the platform can filter content without introducing any subjective bias into the user's decision-making process and actions is by selecting feasible content that maximizes the user's benefit/reward. This approach ensures that the user's feed reflects their own preferences, which may align with the platform's benefits at times, but not necessarily always. Mathematically, the user's exclusive benefit is quantified by a personal reward function that encompasses only the components measuring the user's benefits. We formulate this objective rigorously, while elucidates the difference between the filtered and reference feeds.

Definition 15 (Construction of reference feeds) *Suppose that the platform's reward objective function can be written as following type*

$$\mathrm{Rew}^{F}_{t}(\mathbf{X}, P_{i,t-1}) \;,\; \mathrm{Rew}_{t,per}(\mathbf{X}, P_{i,t-1}) + \mathrm{Rew}_{t,rev}(\mathbf{X}, P_{i,t-1}) + \mathrm{Rew}_{t,self}(\mathbf{X}, P_{i,t-1}),$$

where $\mathrm{Rew}_{t,per}$ *is the reward gained by those feeds which are personalized to the user,* $\mathrm{Rew}_{t,rev}$ *is the revenue-related reward gained by advertisements, and* $\mathrm{Rew}_{t,self}$ *predicts the reward associated with the information the platform would gain from platform "selfish" aspects (e.g., running a social experiment on the user). Without the loss of generality, assume that the first two types of rewards are consistent with the consumer-provider agreement, but the last one is not. Then, the reference feed could be the one that maximize the contribution of the first two types of rewards, namely,*

$$\mathrm{Rew}^{R}_{t}(\mathbf{X}, P_{i,t-1}) \;,\; \mathrm{Rew}_{t,per}(\mathbf{X}, P_{i,t-1}) + \mathrm{Rew}_{t,rev}(\mathbf{X}, P_{i,t-1}).$$

It should be emphasized here that the specific reference feed construction we described above is just one possible example; our results and algorithms only require that there is some fixed reference feed (per user).

Auditor's generative modeling. The AF mechanism is not known and should not be disclosed to the auditor. Nonetheless, it should be clear that for the auditor to be able to inspect the SMP, something about the feeds generation process must be assumed. In this paper, we assume that from the auditor's point of view, the feeds are generated at random, and we denote the conditional law of the feed at time *t* conditioned on history feeds $h_{t-1,u}$ by $P^{F}_{u,t}(h_{t-1,u})$ ¸ $P(\mathbf{X}^{F}_{u}(t)|H^{F}_{t-1,u} = h_{t-1,u})$, for user *u*. Later on, for the framework to be mathematical tractable, we will place additional assumptions on the family of distributions.

Time dependent counterfactual regulations. Above, we have focused on the "filtered vs. reference" feeds approach. We propose the following as an alternative. Let S be a *regulatory statement* that an inspector (or, perhaps, the platform itself) wish to test. For example, S could be: "*The platform should produce similar feeds, in the course of a given time horizon* T, *for users who are identical except for property P*", where *P* could be ethnicity, sexual orientation, gender, a combination of these factors, etc. Let $\mathrm{U}_P \subset \mathrm{U} \times \mathrm{U}$ be a subset of pairs of users that comply with P. Then, for any pair of users $(i, j) \in \mathrm{U}_P$, the inspector's objective is to determine whether the platform's filtering algorithm cause user i's and user j's beliefs and actions to be significantly different. We formulate this objective rigorously in the next section. We mention here that a similar approach to the above was proposed recently in (Cen and Shah, 2021), assuming a time-independent static model. Our study first focuses on constructing a regulation procedure given the first usable form, filtered vs. reference feeds. However, we will later reveal that a regulation procedure for the second form, counterfactual regulations, could be constructed using two parallel procedures of the first form.

Hypothesis testing. The auditor's goal is to determine whether the platform upholds the consumer-provider agreement, and by doing so, to moderate intense influence on the user's decision-making, which may be caused by observing filtered feed, compared to what would have been the user's decision-making under the reference feed. With the model introduced above, the auditor's task can be formulated as a hypothesis testing problem with the following two hypotheses:

- ***The null hypothesis*** H_0: the auditor (or self-audit) decision is that the platform honors the consumer-provider agreement.

- ***The alternative hypothesis*** H_1: the auditor (or self-audit) decision is to investigate the platform for a possible violation.

Accordingly, relying on a certain from of data, which we will specify in the sequel, the auditor's detection problem is to determine whether H_0 or H_1 is true. We need to specify what kind of "test" is considered. Given a fixed risk $\delta \in (0, 1)$, we expect the auditing procedure to find the true one with probability $1 - \delta$, whichever it is. We call such a procedure δ-correct. We consider the following notion of "frugality", which we name *batch setting* : the auditor specifies in advance the number of samples needed for the test, and announce its decision just after observing the data all at once, and the sample complexity of the test is the smallest sample size of a δ-correct procedure.

Auditor's data. For $t \geq 1$ the auditor observes the filtered and reference feeds $\{\mathbf{X}^{\mathrm{F}}_{\mathbf{i}}(t), \mathbf{X}^{\mathrm{R}}_{\mathbf{i}}(t)\}$, for all (or a subset of) users $u \in \mathrm{U}$, and utilize these to test for regulation violations. There are two ways to access this data without invasions to privacy. First, under self-regulation (currently, almost all platform are entirely self-regulated (Klonick, 2017)), the platform obviously has access to those feeds, and therefore, there are no privacy issues. The second option is to provide anonymized data to the auditor. Indeed, both the users identities and the meaning behind the features should/can be removed since they do not affect regulation enforcement. Note that de-anonymization is not a real concern here because the anonymized datasets will not be publicly shared anyhow. Moreover, since the auditor only requires the numerical features of the feeds, rather their semantic interpre- tation, de-anonymization would require unreasonable significant effort that the auditor is not willing to undertake. Thus, with carefully laid out but reasonable measures, the users data would remain private and anonymous. Finally, notice that in principle the filtered and reference feeds need not necessarily correspond to real users and could represent sufficiently representative sample of hypothetical users.

A.3 Formalizing the auditor's goal

Let $\{\mathbf{X}^{\mathrm{F}}_u(t)\}_{t\geq 1}$ and $\{\mathbf{X}^{\mathrm{R}}_u(t)\}_{t\geq 1}$ denote the sequences of user u's filtered an reference feeds evolved over time, respectively. As discussed above, the users implicitly form beliefs from their feeds. With enough evidence, the users gain confidence, and then take actions. Accordingly, the corresponding user u's beliefs and actions are denoted by $\{\mathsf{B}^{\mathrm{F}}_{u,t}, \mathsf{br}^{\mathrm{F}}_{u,t}\}_{t\geq 1}$ and $\{\mathsf{B}^{\mathrm{R}}_{u,t}, \mathsf{br}^{\mathrm{R}}_{u,t}\}_{t\geq 1}$, implied by the filtered and reference feeds, respectively.

Violation. We now define the meaning of "violation" from the auditor's perspective. Let $\mathsf{T} \in \mathbb{N}$ denote the time horizon, which determines how far into the past the auditor scrutinizes the platform's behavior. Let $\mathsf{d}(\cdot\|\cdot) : \Omega \times \Omega \to \mathbb{R}_{\geq 0}$ be a probability metric between two probability measures defined over Ω. Let $\bar{\mathcal{U}} \subseteq \mathcal{U}$ be a certain subset of users (such representative subset of the entire set of users). Then, define the *total action-variability metric* as follows:

$$\mathsf{V}_{\mathrm{action}} \triangleq \frac{1}{\mathsf{T}\cdot|\bar{\mathcal{U}}|}\sum_{i\in\bar{\mathcal{U}}}\sum_{t=1}^{\mathsf{T}}\max_{h_{t,i}}\mathsf{d}\left(\mathsf{br}^{\mathrm{F}}_{i,t}(h_{t,i})\,\middle\|\,\mathsf{br}^{\mathrm{R}}_{i,t}(h_{t,i})\right). \tag{125}$$

Similarly, define the *total belief-variability metric* as,

$$\mathsf{V}_{\mathrm{belief}} \triangleq \frac{1}{\mathsf{T}\cdot|\bar{\mathcal{U}}|}\sum_{i\in\bar{\mathcal{U}}}\sum_{t=1}^{\mathsf{T}}\max_{h_{t,i}}\mathsf{d}\left(\mathsf{B}^{\mathrm{F}}_{i,t}(h_{t,i})\,\middle\|\,\mathsf{B}^{\mathrm{R}}_{i,t}(h_{t,i})\right). \tag{126}$$

Finally, recall that in Subsection A.1 we also proposed a statistical model for filtering. Accordingly, as we explain below, it is beneficial to define also the *total filtering-variability metric*:

$$\mathsf{V}_{\mathrm{filter}} \triangleq \frac{1}{\mathsf{T}\cdot|\bar{\mathcal{U}}|}\sum_{i\in\bar{\mathcal{U}}}\sum_{t=1}^{\mathsf{T}}\max_{h_{t-1,i}}\mathsf{d}\left(P^{\mathrm{F}}_{i,t}(h_{t-1,i})\,\middle\|\,P^{\mathrm{R}}_{i,t}(h_{t-1,i})\right). \tag{127}$$

It is useful to note that there is an analytical relationship between the above variabilities. Indeed, viewing $\mathsf{br}^{\mathrm{F}}_{i,t}$ as a result of a probabilistic kernel that is applied on the beliefs, and assuming that the metric d satisfies the data processing inequality (Cover and Thomas, 2006), it follows that $\mathsf{V}_{\mathrm{action}} \leq \mathsf{V}_{\mathrm{belief}} \leq \mathsf{V}_{\mathrm{filter}}$. Now, from the auditor's perspective, violation could mean that $\mathsf{V}_{\mathrm{action}} > \varepsilon > 0$, for some $\varepsilon > 0$ which governs the regulation strictness, i.e., higher values of ε indicate greater strictness. Alternatively, violation can also be defined through the belief-variability, namely, $\mathsf{V}_{\mathrm{belief}} > \varepsilon > 0$, for some $\varepsilon > 0$. Accordingly, depending on the auditor's ambition, its testing/decision problem can be formulated as one of the following:

$$\mathcal{H}_0 : \mathsf{V}_{\mathrm{action}} \leq \varepsilon_1 \quad \text{vs.} \quad \mathcal{H}_1 : \mathsf{V}_{\mathrm{action}} \geq \varepsilon_2, \tag{128a}$$

$$\mathcal{H}_0' : \mathsf{V}_{\mathrm{belief}} \leq \varepsilon_1 \quad \text{vs.} \quad \mathcal{H}_1' : \mathsf{V}_{\mathrm{belief}} \geq \varepsilon_2, \tag{128b}$$

$$\mathcal{H}_0'' : \mathsf{V}_{\mathrm{filter}} \leq \varepsilon_1 \quad \text{vs.} \quad \mathcal{H}_1'' : \mathsf{V}_{\mathrm{filter}} \geq \varepsilon_2, \tag{128c}$$

where $\varepsilon_2 > \varepsilon_1 \geq 0$. Devising successful statistical tests which solve (128a) (or, (128b)) with high probability, guarantee that whenever the auditor decision is $\mathcal{H}_0$ (or, $\mathcal{H}_0'$), then the platform honors the consumer-provider agreement, since the beliefs and actions are indistinguishable under the filtered and reference feeds. Conversely, rejecting H_0 (or, H^r_0) with high confidence implies that AF causes significantly different learning outcomes. Note that by the data processing inequality, accepting $H^{rr}{}_0$in (128c) imply immediately that H_0 and $H^r{}_0$ hold as well. Note that the general form of the hypothesis testing problems formulated in (128) reminiscent of the well-studied *tolerant closeness testing* problem (see, e.g., Daskalakis et al. (2018b); Canonne et al. (2022)). In this paper, we focus on the hypothesis test in (128c).

Testing. Solving (128c) is mathematically intractable unless we place further assumptions on the family of distributions that generate the feeds. In this paper, we assume the following quasi-Markov homogeneous model. We divide the time horizon into batches, and assume that in each batch, the platform filtering process is modeled as a large probabilistic state machine. During these batches the platform collect new data to create new successive feeds. From the auditor's point of view, the platform is a rather sequentially-feeds generating system, making a probabilistic relationship of the current feed conditioned on the previous feeds, in time intervals. Under these circumstances, the auditor models problem (4) as a quasi-Markov homogeneous model.

Mathematically, let $T_{total} \in N$ denote the time horizon, which determines how far into the past the auditor scrutinizes the platform's behavior. Assume we have $B \in N$ batches each of length $T \in N$, such that in batch $b \in [B]$ we have a time sampling sequence $b \cdot T < t_{0,b} < t_{1,b} < \cdots < t_{T,b} \leq (b+1) \cdot T$. In each batch, from the auditor's point of view, the piece of content $\mathbf{x}^F_{l,u}(t_{i,b})$, at time $t_{i,b}$, for $\boldsymbol{l} \in [M]$, is drawn from a first-order irreducible Markov chain, namely, $P(\mathbf{x}^F_{l,u}\ (t_{i,b})|\mathbf{x}^F_{l,u}(t_{0,b}), \ldots, \mathbf{x}^F_{l,u}(t_{i-1,b})) = P(\mathbf{x}^F_{l,u}(t_{i,b})|\mathbf{x}^F{}_{l,u}(t_{i-1,b}))$, and $P(\mathbf{x}^F{}_{l,u}(t_{i,b}) = s_2|\mathbf{x}^F{}_{l,u}(t_{i-1,b}) = s_1)$, $Q_{u,b}(s_1, s_2)$, for any two possible states $s_1, s_2 \in X$. We denote the transition probability matrix by in batch $b \in [B]$ by $Q^{\mathbf{F}}_{u,b} = [Q_{u,b}(s_1, s_2)]_{s_1,s_2 \in F}$. We assume further that the M Markov trajectories are i.i.d. Note that over different intervals, indexed by b, the filtering process could be transformed into a new state machine subjected to a different transition probabilities. For example, this transformation may occur over time when new external data incur noticeable changes in the platform's reward. Thus, in the bth batch, the observed feeds are,

$$\underset{\text{Feed 1}}{\text{n}\ \mathbf{x}^F_{l,u}(t_{0,b})\ \text{,}^{M}_{l=1}},\ \underset{\text{Feed 2}}{\text{n}\ \mathbf{x}^F_{l,u}(t_{1,b})\ \text{,}^{M}_{l=1}}, \ldots, \underset{\text{Feed T}}{\text{n}\ \mathbf{x}^F{}_{l,u}(t_{T,b})\ \text{,}^{M}_{l=1}}.$$

The above discussion is relevant to the reference feeds generation process as well; in particular, we denote by $P^{\mathbf{R}}_{u,n}$, $[P_{u,n}(s_1, s_2)]_{i,j \in F}$ the corresponding matrix transition probabilities.

From the auditor point of view, in terms of the reward-based platform filtering, a practical interpretation for the above modeling is as follows. At any interval b, the platform generates some updated Markovian transition-matrix that is subjected to an updated Markov chain by maximizing its reward function, i.e.,

$$Q^{\mathbf{F}}_{u,b+1} = \arg\max_{Q} \sum_t \text{Rew}^F(Q, P_{u,b})$$

$$\text{s.t.} \quad \forall j \in X, \quad \sum_{i \in F} \mathbf{Q}_{i,j} = 1,$$

where $P_{u,b}$ captures the external data and inputs to the platform used for filtering, intended for user u, and was collected during the current time interval $(b \cdot T, (b + 1) \cdot T]$. Similarly, the reference feeds are generated by the same statistical process but by the reference-based rewards objective, i.e.,

$$P^{\mathbf{R}}_{u,b+1} = \arg\max_{P} \sum_{t} \mathrm{Rew}^{R}(P, P_{u,b})$$

$$\text{s.t.} \quad \forall j \in X, \sum_{i \in F} P_{i,j} = 1.$$

References

D. Acemoglu, M. A. Dahleh, I. Lobel, and A. Ozdaglar. Bayesian learning in social networks. *The Review of Economic Studies*, 78(4):1201–1236, 2011.

J. Acharya, C. Daskalakis, and G. Kamath. Optimal testing for properties of distributions. *Advances in Neural Information Processing Systems*, 28, 2015.

J. Anderson and L. Rainie. The future of truth and misinformation online. *Pew Research Center*, 2017.

A. V. Banerjee. A simple model of herd behavior. *The quarterly journal of economics*, 107 (3):797–817, 1992.

T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *Journal of the ACM (JACM)*, 60(1):1–25, 2013.

H. Berghel. Lies, damn lies, and fake news. *Computer*, 50(2):80–85, 2017.

A. Blake. A new study suggests fake news might have won donald trump the 2016 election. 2018.

O. Board. Ensuring respect for free expression, through independent judgment. March 2020.

E. Bozdag. Bias in algorithmic filtering and personalization. *Ethics and information technology*, 15(3):209–227, 2013.

V. C. Brannon. Free speech and the regulation of social media content. *Congressional Research Service*, 27, 2019.

A. Campbell. How data privacy laws can fight fake news. *Just security*, 2019.

C. L. Canonne, A. Jain, G. Kamath, and J. Li. The price of tolerance in distribution testing. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 573–624. PMLR, 02–05 Jul 2022.

R. Caplan. 29 algorithmic filtering. *Mediated Communication*, page 561, 2018.

S. Cen and D. Shah. Regulating algorithmic filtering on social media. In *Advances in Neural Information Processing Systems*, volume 34, pages 6997–7011. Curran Associates, Inc., 2021.

S. H. Cen and D. Shah. Regulating algorithmic filtering on social media. *arXiv preprint: arxiv.org/pdf/2006.09647v3.pdf*, 2020.

S. H. Cen, A. Madry, and D. Shah. A user-driven framework for regulating and auditing social media. *arXiv preprint arXiv:2304.10525*, 2023.

S.-O. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. SIAM, 2014.

S. O. Chan, Q. Ding, and S. H. Li. Learning and testing irreducible markov chains via the *k*-cover time. In V. Feldman, K. Ligett, and S. Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 458–480. PMLR, 16–19 Mar 2021. URL https:
//proceedings.mlr.press/v132/chan21a.html.

M. M. Chau, M. Burgermaster, and L. Mamykina. The use of social media in nutrition interventions for adolescents and young adults—a systematic review. *International journal of medical informatics*, 120:77–91, 2018.

B. Chesney and D. Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecom- munications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

C. Damian, E. Clive, E. Julie, F. Paul, H. Simon, K. Julian, I. C. Lucas, O. Brendan,
P. Rebecca, S. Jo, and W. Giles. Disinformation and 'fake news'. 2019.

C. Daskalakis, N. Dikkala, and N. Gravin. Testing symmetric markov chains from a single trajectory. In *Conference On Learning Theory*, pages 385–409. PMLR, 2018a.

C. Daskalakis, G. Kamath, and J. Wright. Which distribution distances are sublinearly testable? In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2747–2764. SIAM, 2018b.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515, 2017.

M. A. DeVito, D. Gergle, and J. Birnholtz. " algorithms ruin everything" # riptwitter, folk theories, and resistance to algorithmic change in social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3163–3174, 2017.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

D. Ghosh. A new digital social contract is coming for silicon valley. *Harvard Business Review*, 27, 2019.

T. Independent. Amid capitol violence, facebook, youtube remove trump video. January 2021.

M. S. Jahan and M. Oussalah. A systematic review of hate speech automatic detection using natural language processing. *CoRR*, abs/2106.00742, 2021. URL https://arxiv.org/abs/2106.00742.

M. Jane, M. Hagger, J. Foster, S. Ho, and S. Pal. Social media for health promotion and weight management: a critical debate. *BMC public health*, 18(1):1–7, 2018.

K. Klonick. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.*, 131:1598, 2017.

B. Koszegi. Behavioral contract theory. *Journal of Economic Literature*, 52(4):1075–1118, 2014.

S. Kruikemeier, S. C. Boerman, and N. Bol. Breaching the contract? using social contract theory to explain individuals' online behavior to safeguard privacy. *Media Psychology*, 23(2):269–292, 2020. doi: 10.1080/15213269.2019.1598434. URL https://doi.org/10.
1080/15213269.2019.1598434.

J. Kurbalija. *An introduction to internet governance*. Diplo Foundation, 2016.

F. L. Lee. Impact of social media on opinion polarization in varying times. *Communication and the Public*, 1(1):56–71, 2016.

R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. vol- ume 17, pages 179–194, 01 2011.

B. Lewis and A. E. Marwick. Media Manipulation and Disinformation Online. *New York: Data & Society Research Institute*, 2017.

A. Manning. Implicit contract theory. *Current Issues in Labour Economics*, page 63, 1989.

R. Medzini. Enhanced self-regulation: The case of facebook's content governance. *New Media & Society*, page 1461444821989352, 2021.

A. Mitchell, J. Gottfried, M. Barthel, and E. Shearer. The modern news consumer: News attitudes and practices in the digital era. 2016.

S. Mohseni, E. D. Ragan, and X. Hu. Open issues in combating fake news: Interpretability as an opportunity. *arXiv:1711.04024*, 2019.

P. Molavi, A. Tahbaz-Salehi, and A. Jadbabaie. A theory of non-bayesian social learning. *Econometrica*, 86(2):445–490, 2018.

B. News. Twitter suspends 70,000 accounts linked to qanon. January 2021.

J. A. Obar and S. S. Wildman. Social media definition and the governance challenge-an in- troduction to the special issue. *Obar, JA and Wildman, S.(2015). Social media definition and the governance challenge: An introduction to the special issue. Telecommunications policy*, 39(9):745–750, 2015.

E. Pariser. How the new personalized web is changing what we read and how we think. 2011.

J. Paschen. Investigating the emotional appeal of fake news using artificial intelligence and human contributions. *Journal of Product & Brand Management*, 05 2019.

R. D. Petty. Marketing without consent: Consumer choice and costs, privacy, and public policy. *Journal of Public Policy & Marketing*, 19(1):42–53, 2000.

K. Quinn. Why we share: A uses and gratifications approach to privacy regulation in social media use. *Journal of Broadcasting & Electronic Media*, 60(1):61–86, 2016.

M. Z. Rácz and J. Richey. Rumor source detection with multiple observations under adaptive diffusions. *IEEE Transactions on Network Science and Engineering*, 8(1):2–12, 2020.

N. Rivera, T. Sauerwald, and J. Sylvester. Multiple random walks on graphs: mixing few to cover many. *Combinatorics, Probability and Computing*, 32(4):594–637, 2023.

A. Rodr´ıguez, C. Argueta, and Y.-L. Chen. Automatic detection of hate speech on facebook using sentiment and emotion analysis. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 169–174, 2019. doi: 10.1109/ICAIIC.2019.8669073.

K. Sarikakis and L. Winter. Social media users' legal consciousness about privacy. *Social Media+ Society*, 3(1):2056305117695325, 2017.

S. Siersdorfer, S. Chelaru, J. S. Pedro, I. S. Altingovde, and W. Nejdl. Analyzing and mining comments and comment ratings on the social web. *ACM Transactions on the Web (TWEB)*, 8(3):1–39, 2014.

T. Speicher, M. Ali, G. Venkatadri, F. N. Ribeiro, G. Arvanitakis, F. Benevenuto, K. P. Gummadi, P. Loiseau, and A. Mislove. Potential for Discrimination in Online Targeted Advertising. In S. A. Friedler and C. 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. URL http:// proceedings.mlr.press/v81/speicher18a.html.

L. Sweeney. Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue*, 11(3):10–29, 2013.

S. Wachter and B. Mittelstadt. A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Colum. Bus. L. Rev.*, page 494, 2019.

L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

G. Wolfer and A. Kontorovich. Minimax learning of ergodic markov chains. In *Algorithmic Learning Theory*, pages 904–930. PMLR, 2019.

G. Wolfer and A. Kontorovich. Minimax testing of identity to a reference ergodic markov chain. In *International Conference on Artificial Intelligence and Statistics*, pages 191–201. PMLR, 2020.

World Economic Forum. World economic forum global agenda council on the future of software and society. A call for agile governance principles. Technical report, 2016. https://www3.weforum.org/docs/IP/2016/ICT/Agile_Governance_Summary.pdf.