Mathematical Framework for Online Social Media Auditing

Wasim Huleihel, Yehonathan Refael

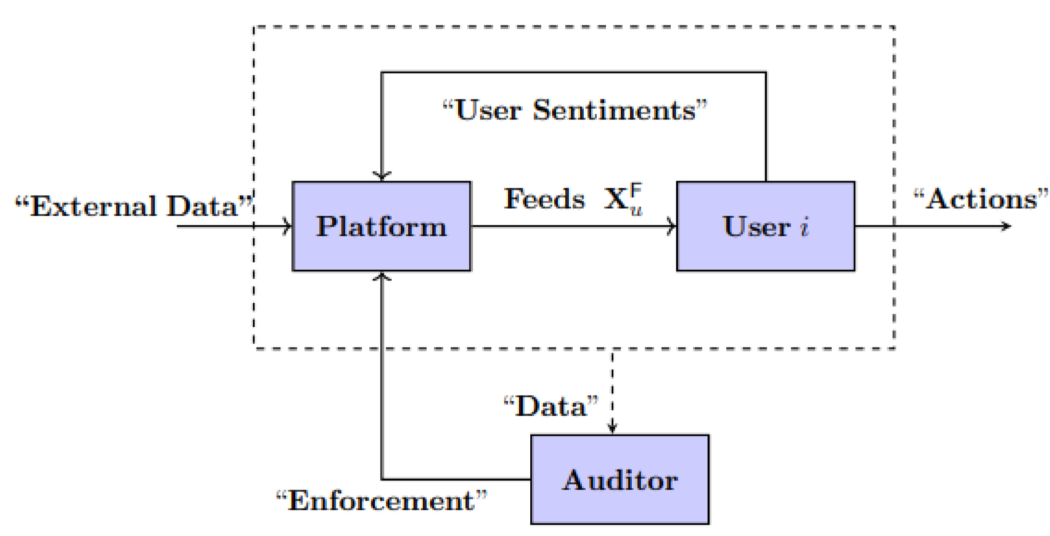
Department of Electrical Engineering, Engineering Faculty, Tel Aviv University

Introduction

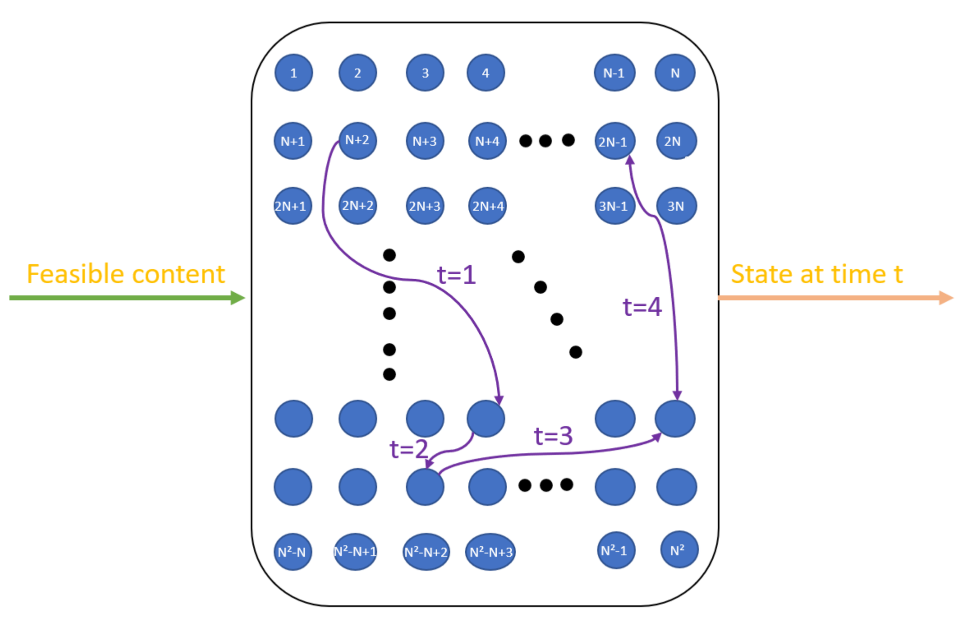
Objective: is to moderate any intense influence on the user's decision-making, which may be caused by observing filtered possible contents, compared to what would have been the user's decision-making under randomizing from possible contents.

The filtered feed shown to user $u \in [\mathsf{U}]$ at time $t \in \mathbb{N}$ by $\mathbf{X}_u^{\mathsf{F}}(t)$, and assume that it consists of $\mathsf{M} \in \mathbb{N}$ pieces of contents, namely, $\mathbf{X}_u^{\mathsf{F}}(t) = \{\boldsymbol{x}_{1,u}^{\mathsf{F}}(t), \ldots, \boldsymbol{x}_{\mathsf{M},u}^{\mathsf{F}}(t)\}$, where $\boldsymbol{x}_{j,u}^{\mathsf{F}}(t) \in \mathcal{X}$ denotes a piece of content, for $1 \leq j \leq \mathsf{M}$. Similarly, *reference feeds* $\mathbf{X}_u^{\mathsf{R}}(t)$ is the one that could have hypothetically selected by the platform if it strictly followed the consumer-provider agreement.

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



We define $\mathbb{P}(\boldsymbol{x}_{\ell,u}^{\mathsf{F}}(t_{i,b})|\boldsymbol{x}_{\ell,u}^{\mathsf{F}}(t_{0,b}), \ldots, \boldsymbol{x}_{\ell,u}^{\mathsf{F}}(t_{i-1,b})) = \mathbb{P}(\boldsymbol{x}_{\ell,u}^{\mathsf{F}}(t_{i,b})|\boldsymbol{x}_{\ell,u}^{\mathsf{F}}(t_{i-1,b}))$, and $\mathbb{P}(\boldsymbol{x}_{\ell,u}^{\mathsf{F}}(t_{i,b}) = s_2|\boldsymbol{x}_{\ell,u}^{\mathsf{F}}(t_{i-1,b}) = s_1) \triangleq Q_{u,b}(s_1, s_2)$, for any two possible states $s_1, s_2 \in \mathcal{X}$. Similarily, for the reference feed we define $\mathsf{P}_{u,b}^{\boldsymbol{R}} \triangleq [P_{u,b}(s_1, s_2)]_{i,j \in \mathcal{X}}$.

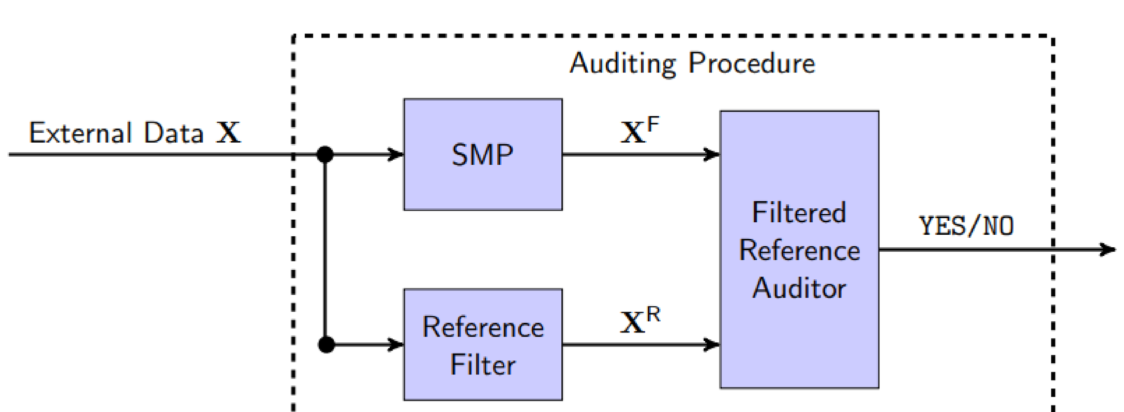


The *total filtering-variability metric* as,

$$\mathbb{V}_{\mathsf{filter}} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \left\|\mathsf{P}_{u,b}^{\boldsymbol{R}} - \mathsf{Q}_{u,b}^{\boldsymbol{F}}\right\|_{\infty},$$

where $\mathbf{P}_{u,b}^{\boldsymbol{R}}(i) \triangleq [P_{u,b}(i,j)]_{j \in \mathcal{X}}$, $\mathbf{Q}_{u,b}^{\boldsymbol{F}}(i) \triangleq [Q_{u,b}(i,j)]_{j \in \mathcal{X}}$, and $\mathcal{U} = [\mathsf{U}]$.

Auditor's goal



Violation. we define a violation event as the case where $\mathbb{V}_{\mathsf{filter}}$ is "unusually large". Specifically, the audit's decision task is formulated as the following hypothesis testing problem,

$$\mathcal{H}_0 : \mathbb{V}_{\mathsf{filter}} \leq \varepsilon_1 \quad \text{vs.} \quad \mathcal{H}_1 : \mathbb{V}_{\mathsf{filter}} \geq \varepsilon_2,$$

where $\varepsilon_2 > \varepsilon_1 \geq 0$ govern the auditing strictness.
Devising successful statistical tests which solve the above test with high probability, guarantees that whenever the auditor decision is $\mathcal{H}_0$, then the platform honors the consumer-provider agreement, since the beliefs and actions are indistinguishable under the filtered and reference feeds.

Auditing formulation

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

$$\tau_{\mathsf{cov}}^{(k)}(\ell; \mathscr{M}) \triangleq \inf\left\{t \geq 0 : \forall i \in [n], \sum_{j=1}^{\ell} \mathcal{N}_i^{\mathsf{Z}_j}(t) \geq k\right\}.$$

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

$$t_{\mathsf{cov}}^{(k)}(\ell; \mathscr{M}) \triangleq \max_{\mathbf{v} \in [n]^{\ell}} \mathbb{E}\left[\tau_{\mathsf{cov}}^{(k)}(\ell; \mathscr{M}) \mid \mathsf{Z}_{1,1} = v_1, \mathsf{Z}_{2,1} = v_2, \ldots, \mathsf{Z}_{\ell,1} = v_{\ell}\right],$$

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

Auditing algorithm

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

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

Input: U, n, m, ε_1, δ, and samples $\mathcal{S}_P$ and $\mathcal{S}_Q$ from $\{(P_u, Q_u)\}_{u \in [\mathsf{U}]}$.
Set $\tau \longleftarrow c \min\left(\frac{m^{3/2}\varepsilon_2}{n^{\frac{1}{2}}}, \frac{\mathsf{U}m^2\varepsilon_2^2}{n}\right)$
Compute G in (13).
If $G < \tau$, then **Return** `YES`
Else $G \geq \tau$, then **Return** `NO`

Auditor testing problem 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 $\left[\left(\mathbf{X}_u^{\mathsf{F}}(t_1), \mathbf{X}_u^{\mathsf{R}}(t_1)\right), \ldots, \left(\mathbf{X}_u^{\mathsf{F}}(t_{\mathsf{T}}), \mathbf{X}_u^{\mathsf{R}}(t_{\mathsf{T}})\right)\right]$ drawn from an *unknown* corresponding pair of Markov chains $\left(\mathsf{Q}_u^{\boldsymbol{F}}, \mathsf{P}_u^{\boldsymbol{R}}\right)$, for each user $u \in \mathsf{U}$, an $(\varepsilon_1, \varepsilon_2, \delta)$-sum of pairs tolerant closeness testing algorithm outputs `YES` if $\mathbb{V}_{\mathsf{filter}} \leq \varepsilon_1$ and '`NO` if $\mathbb{V}_{\mathsf{filter}} \geq \varepsilon_2$, with probability at least $1 - \delta$.

Algorithm 2: Filtered vs. reference auditing procedure

Input: T, $n \triangleq |\mathcal{X}|$, $\varepsilon_1, \varepsilon_1, \delta, \bar{m}$, and feeds $\{\mathbf{X}_u^{\mathsf{R}}(t), \mathbf{X}_u^{\mathsf{F}}(t)\}_{t=1}^{\mathsf{T}}$, for $u \in [\mathsf{U}]$.
Output: `YES` if $\mathbb{V}_{\mathsf{filter}} \leq \varepsilon_1$ / `NO` if $\mathbb{V}_{\mathsf{filter}} \geq \varepsilon_2$.
For $i \leftarrow 1, 2 \ldots\ldots, n$
Set $\mathcal{S}^{\mathsf{R}} \leftarrow \emptyset$ and $\mathcal{S}^{\mathsf{F}} \leftarrow \emptyset$
For every user $u \leftarrow 1, 2 \ldots\ldots, \mathsf{U}$
If $\sum_{j=1}^{\mathsf{M}} \mathcal{N}_i^{\boldsymbol{x}_{j,u}^{\mathsf{R}}} < \bar{m}$ **or** $\sum_{j=1}^{\mathsf{M}} \mathcal{N}_i^{\boldsymbol{x}_{j,u}^{\mathsf{F}}} < \bar{m}$
Return `NO`
Calculate $\mathcal{S}_u^{\mathsf{R}} \leftarrow \cup_{j=1}^{\mathsf{M}} \psi_{\bar{m}}^{(i)}\left(\{\boldsymbol{x}_{j,u}^{\mathsf{R}}(t)\}_{t=1}^{\mathsf{T}}\right)$ and $\mathcal{S}_u^{\mathsf{F}} \leftarrow \cup_{j=1}^{\mathsf{M}} \psi_{\bar{m}}^{(i)}\left(\{\boldsymbol{x}_{j,u}^{\mathsf{F}}(t)\}_{t=1}^{\mathsf{T}}\right)$
Do $\mathcal{S}^{\mathsf{R}} \leftarrow \mathcal{S}^{\mathsf{R}} \cup \mathcal{S}_u^{\mathsf{R}}$ and $\mathcal{S}^{\mathsf{F}} \leftarrow \mathcal{S}^{\mathsf{F}} \cup \mathcal{S}_u^{\mathsf{F}}$
If $\textsc{IIDTester}(\mathcal{S}^{\mathsf{R}}, \mathcal{S}^{\mathsf{F}}, \delta, \varepsilon_1, \varepsilon_2, \bar{m}, n) =$ `NO`
Return `NO`
Return `YES`

The mapping $\psi_k^{(i)}(\mathsf{Z}_1^q)$ is define as follows: we look at the first k visits to state i (i.e., at times $t = t_1, \ldots, t_k$ with $\mathsf{Z}_t = i$) and write down the corresponding transitions in Z_1^q, i.e., Z_{t+1}.

Sample complexity

Theorem 4 (Sample complexity of the sum closeness testing). *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*

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

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

Theorem 5 (Auditing 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,*

$$\mathsf{T} = \mathcal{O}\left(\max_{u \in [\mathsf{U}]} \max_{\mathsf{W} \in \{\mathsf{Q}_u^{\mathsf{F}}, \mathsf{P}_u^{\mathsf{R}}\}} t_{\mathsf{cov}}^{\bar{m}}(\mathsf{M}; \mathsf{W}) \log\frac{\mathsf{U}}{\delta}\right),$$

samples per user.

Counterfactual regulation

Let $\mathcal{S}$ be a *regulatory statement* that an inspector (or, perhaps, the platform itself) wish to test. For example, $\mathcal{S}$ 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* $\mathscr{P}$", where $\mathscr{P}$ could be ethnicity, sexual orientation, gender, a combination of these factors, etc. Let $\mathcal{U}_{\mathscr{P}} \subset [\mathsf{U}] \times [\mathsf{U}]$ be a subset of pairs of users that comply with $\mathscr{P}$. Then, for any pair of users $(i,j) \in \mathcal{U}_{\mathscr{P}}$, 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.

Definition 7 (Counterfactual total variability). *Let $\mathcal{U}_{\mathscr{P}} \subset [\mathsf{U}] \times [\mathsf{U}]$ be a subset of pairs of users that comply with $\mathscr{P}$. Then, for any pair of users $(i,j) \in \mathcal{U}_{\mathscr{P}}$, the total variability in algorithmic filtering behavior for counterfactual users is given by*

$$\begin{aligned}\bar{\mathbb{V}}_{\mathsf{cu}}(\mathcal{S}, \mathcal{U}_{\mathscr{P}}) &\triangleq \frac{1}{|\mathcal{U}_{\mathscr{P}}|} \sum_{(i,j) \in \mathcal{U}_{\mathscr{P}}} \max_{\ell \in \mathcal{X}} \mathsf{d}_{\mathsf{TV}}\left(Q_i(\ell, \cdot), Q_j(\ell, \cdot)\right) \\ &= \frac{1}{|\mathcal{U}_{\mathscr{P}}|} \sum_{(i,j) \in \mathcal{U}_{\mathscr{P}}} \max_{\ell \in \mathcal{X}} \|\mathbf{Q}_i(\ell) - \mathbf{Q}_j(\ell)\|_1 \\ &= \frac{1}{|\mathcal{U}_{\mathscr{P}}|} \sum_{(i,j) \in \mathcal{U}_{\mathscr{P}}} \|\mathsf{Q}_i^{\boldsymbol{F}} - \mathsf{Q}_j^{\boldsymbol{F}}\|_{\infty}.\end{aligned}$$

The investigator's task to test for violations in the following sense:

$$\mathcal{H}_0^{\mathcal{S}} : \bar{\mathbb{V}}_{\mathsf{cu}}(\mathcal{S}, \mathcal{U}_{\mathscr{P}}) \leq \varepsilon_1 \quad \text{vs.} \quad \mathcal{H}_1^{\mathcal{S}} : \bar{\mathbb{V}}_{\mathsf{cu}}(\mathcal{S}, \mathcal{U}_{\mathscr{P}}) \geq \varepsilon_2.$$

Conclusions

The study presents an auditing method that tests for unexpected deviations in the user's decision-making process over a predefined time horizon. These deviations could be due to selective content filtering by the platform. We developed metrics for effectiveness and implementability methods with sample complexity guarantees.