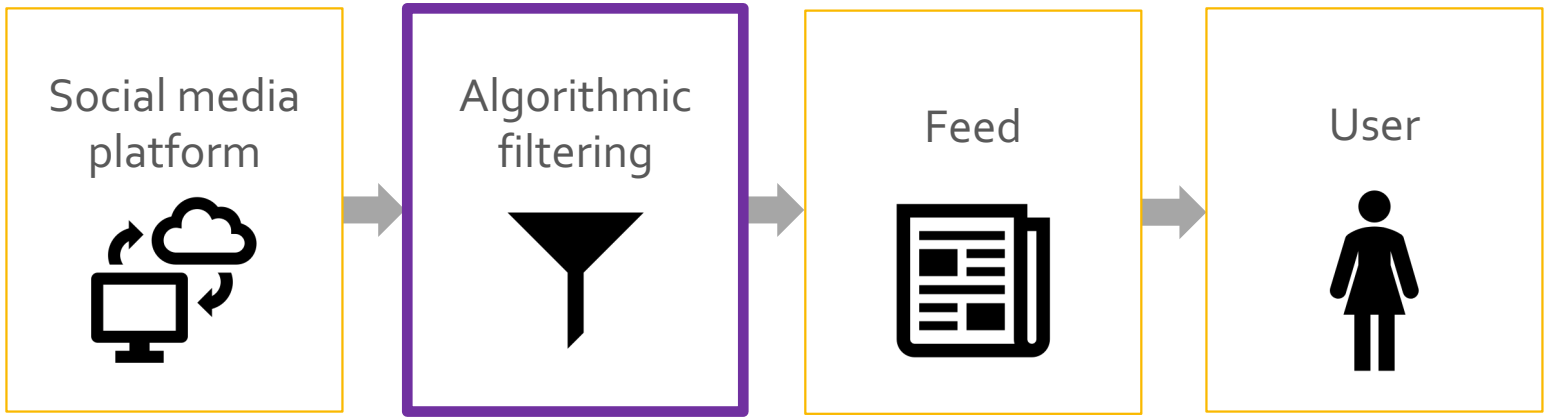
Regulating algorithmic filtering on social media

Sarah H. Cen and Devavrat Shah | MIT EECS

Spotlight at NeurIPS 2021

What is this talk about?



How to audit?

How should one translate a regulation → auditing procedure?

Provide an **auditing procedure** to check platform's compliance.

Strong statistical **guarantees** how well it enforces the regulation.

How does the audit affect the platform & its users?

Not necessarily a performance-regulation **trade-off**.

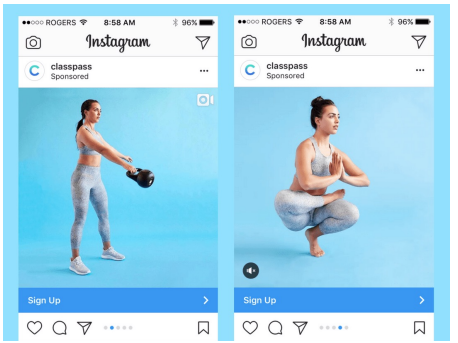
Content diversity aligns interests of the regulator & platform.

Social media platforms *influence* through information

Social media platforms provide information …

source: iphonelife.com

source: gizbot.com



source: later.com

Algorithmic filtering

Filtered feed

$Z = \{z_1, \ldots, z_m\}$

user

Calls to regulate

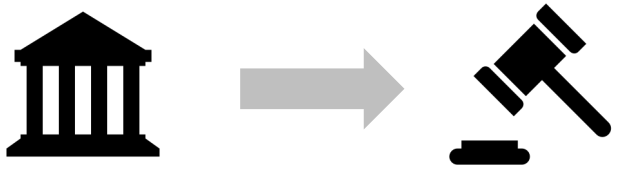
Increasing calls to regulate

Ex 1: Ads not be based on user's sexual orientation.

Ex 2: Info on public health (e.g., COVID-19) not reflect political affiliation.

Ex 3: Not sway voting preferences beyond serving as a social network.

Translating from **regulation → auditing procedure** is difficult.



- **Reactive**: narrow & delayed.
- **Performance cost**: hurts users & platform.
- **Censorship**: removal of content.
- **Privacy**: of user's personal data.
- **Trade secrets**: access to algorithms is limited.
- …

Main takeaways

Can we translate a regulatory guideline → auditing procedure?

Main contribution: Auditing procedure

Given a regulation in CF form, an auditor can test whether the platform is in compliance.

- Black-box access
- Without users' personal data
- Modular
- Intuitive tunable parameter
- No content removal

Auditor	Platform	User
Guarantee on how well procedure enforces regulation.	Not necessarily a trade-off btw regulation & performance.	Audit incentivizes platform to ensure content diversity.

Setup

Problem setup

The platform selects the content shown to its users by …

Inputs $\boldsymbol{x}$

(user attributes, advertisers, etc.)

Filtering algorithm $\mathcal{F}$

Feed $Z = \mathcal{F}(\boldsymbol{x})$

Run $\mathcal{F}$ on $\{\boldsymbol{x}_j\}$ and observe $\{Z_j\}$.

Auditor's task

Given a regulation and black-box access to $\mathcal{F}$, check if platform is compliant.

$Z = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m\}$

$\boldsymbol{z}_i \sim p_z(\cdot\,; \theta)$, $\theta \in \Theta$

Form of regulation

Counterfactual regulation

hypothetical!

"Algorithm $\mathcal{F}$ must behave similarly under $\boldsymbol{x}$ and $\boldsymbol{x}'$ for all $(\boldsymbol{x}, \boldsymbol{x}') \in S$."

Articles containing medical advice on COVID-19 must be robust to user's political affiliation.

=

"The articles shown by $\mathcal{F}$ that contain medical advice on COVID-19 should be **similar** whether a user is left-leaning or right-leaning.

It is not the platform's job to sway voting preferences beyond serving as a social network.

=

"Posts about political candidates that are injected by $\mathcal{F}$ should be **similar** to the content a user would see from its social network without any algorithmic filtering."

What is an appropriate notion of "similarity" ?

Algorithmic filtering affects how users make decisions

Observation

Algorithmic filtering is powerful and sometimes harmful because information influences decisions.

Examples. The content that $\mathcal{F}$ filters affects …

- How the user votes
- Whether they get vaccinated
- Where they eat
- What items they purchase

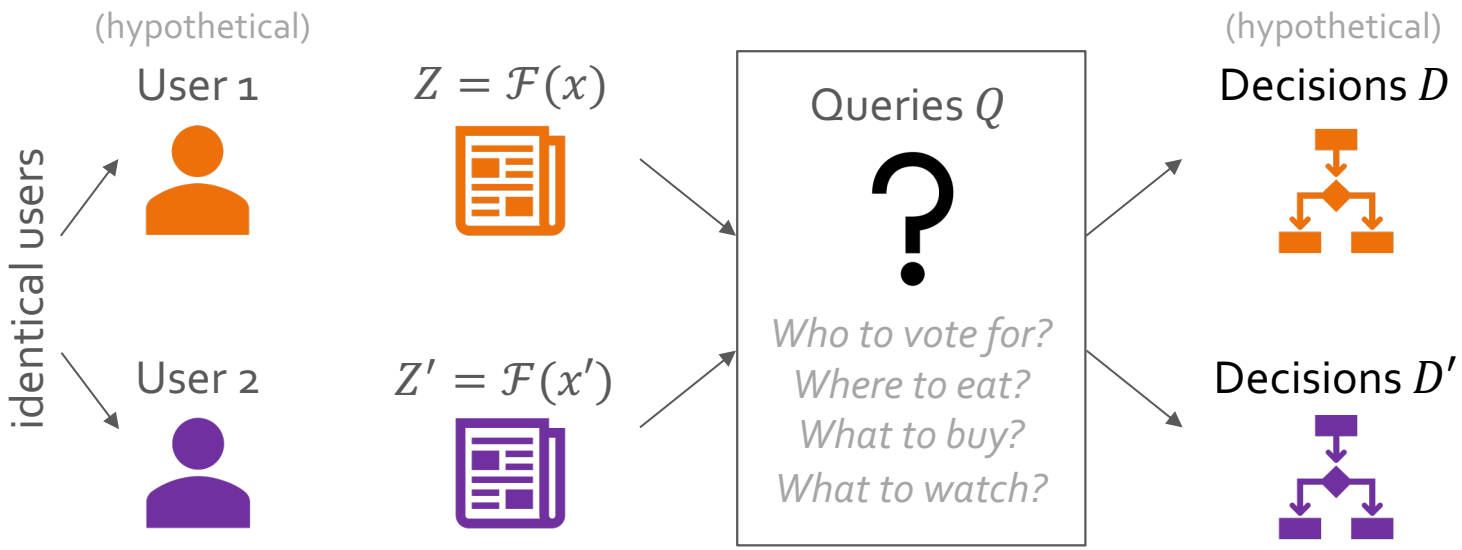
Implication on "similarity".

When we enforce similarity between $\mathcal{F}(\boldsymbol{x})$ and $\mathcal{F}(\boldsymbol{x}')$,
it should be w.r.t. the outcome of interest: **the users' decisions**.

Decision robustness

Counterfactual regulation

"Algorithm $\mathcal{F}$ must behave similarly under $\boldsymbol{x}$ and $\boldsymbol{x}'$ for all $(\boldsymbol{x}, \boldsymbol{x}') \in S$."



$\mathcal{F}$ is *decision-robust* to $(\boldsymbol{x}, \boldsymbol{x}')$ if and only if, for any Q, one cannot confidently determine that $\boldsymbol{x} \neq \boldsymbol{x}'$ from D and D'.

↳ can formalize as hypothesis test

Auditing procedure

Auditing procedure

Counterfactual regulation

"Algorithm $\mathcal{F}$ must behave similarly under $\boldsymbol{x}$ and $\boldsymbol{x}'$ for all $(\boldsymbol{x}, \boldsymbol{x}') \in S$."

Auditing procedure

Inputs: $\mathcal{F}$ $\boldsymbol{x}$ $\boldsymbol{x}'$ Θ ϵ

$\tilde{\boldsymbol{\theta}} \leftarrow \mathcal{L}^{+}(\mathcal{F}(\mathbf{x}))$;
$\tilde{\boldsymbol{\theta}}' \leftarrow \mathcal{L}^{+}(\mathcal{F}(\mathbf{x}'))$;
if $(\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}')^{\top} I(\tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}') \geq \frac{2}{m}\chi_r^2(1-\epsilon)$ **then**
  Does not pass the test for $(\mathbf{x}, \mathbf{x}')$;
end
Passes the test for $(\mathbf{x}, \mathbf{x}')$;

Minimum-variance unbiased estimator (MVUE)

Advantages

Advantages

- Only needs black-box access to $\mathcal{F}$.
- Does not require access to users or their personal data.
- Modular.
- Intuitive tunable parameter.
- No content removal.

(Can be combined with other methods!)

What does the audit do?

Main result

Guarantee on how well the audit enforces the regulation.

Theorem (informal). If the filtering algorithm $\mathcal{F}$ passes the audit, then $\mathcal{F}$ is guaranteed to be approximately asymptotically decision-robust.

Alternative statement

If $\mathcal{F}$ does not pass the audit, then the auditor is $(1 - \epsilon)$-confident that $\mathcal{F}$ is not decision-robust as $m \to \infty$.

Takeaways

- The audit enforces strong similarity between $\mathcal{F}(\boldsymbol{x})$ and $\mathcal{F}(\boldsymbol{x}')$.
- ϵ is the allowable false positive rate: increasing ϵ increases strictness.

Why the MVUE?

Insight on MVUE.

Proposition (informal). Faced with a decision between a finite number of options, the decision of the hypothetical user whose belief after viewing content Z is given by the MVUE is more sensitive to Z than any other user.

Takeaway

To audit without access to users or their decisions, use the MVUE.

The user whose decisions are **most sensitive to the content that they see** is the hypothetical user given by the MVUE.

MVUE allows us to reason about how content affects users *without access to users' decisions* → expensive or unethical to obtain.

Trade-off between regulation & performance

Conditions under which there is no trade-off.

Theorem (informal). When the platform's performance is independent of elements in $\boldsymbol{\theta}$ and those elements have sufficient leverage over the Fisher information, then as long as the feed is finite and the available content is expressive enough, then there is no regulation-performance trade-off.

Takeaway

There are conditions under which the platform passes the audit without sacrificing performance.

Content diversity can reduce the cost of regulation

The lower the content diversity of Z and Z', the more easily an auditor can distinguish between how $\mathcal{F}$ behaves under $\boldsymbol{x}$ and $\boldsymbol{x}'$.

Review

Can we translate a regulatory guideline → auditing procedure?

Main contribution: Auditing procedure

Given a regulation in CF form, an auditor can test whether the platform is in compliance.

Black-box access

Without users' personal data

Modular

Intuitive tunable parameter

No content removal

Auditor	Platform	User
Guarantee on how well procedure enforces regulation.	Not necessarily a trade-off btw regulation & performance.	Audit incentivizes platform to ensure content diversity.

Thank you!

shcen@mit.edu