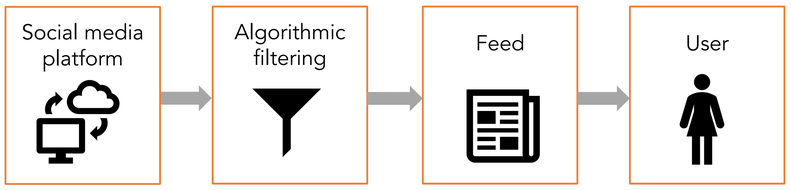
Regulating algorithmic filtering on social media

Sarah H. Cen (shcen@mit.edu), Aleksander Mądry and Devavrat Shah
Massachusetts Institute of Technology

Algorithmic filtering (AF)



1. How should one translate a regulation → auditing procedure?

Main contribution: an **audit** to check platform's compliance.
Strong statistical **guarantees** on how well the audit enforces the regulation.

2. How does the audit affect the platform & its users?

Find that there is not necessarily a performance-regulation **trade-off**.
Show **content diversity** aligns interests of the regulator & platform.

Calls to regulate

There are increasing calls to regulate.

Example: That advertisements not be based on user's sexual orientation.
Example: That information on public health (e.g., COVID-19) do not reflect political affiliation.

However, translating a **regulation into an auditing procedure** is challenging.

Obstacles to regulations

- Current approaches tend to be reactive (respond to issues as they arise).
- Regulations can impose performance cost (bad for user and platform).
- Others require removal of content (free speech issue).
- Some audits require access to users' personal data (data privacy issue).

Contributions

Main contribution: Auditing procedure such that ...

Given a regulation in counterfactual form,
an auditor can test the platform's compliance.

Advantages:

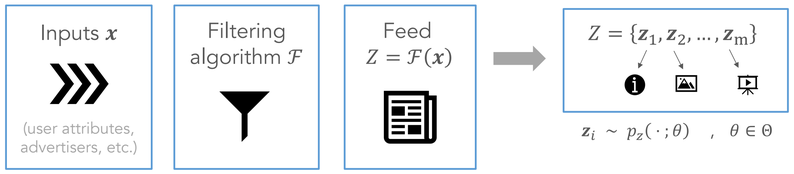
Black-box access to $\mathcal{F}$	Does not need user data	Modular	Intuitive tunable parameter	No content removal

We study the audit from three stakeholder perspectives:

Auditor	Platform	User
Provide guarantee on how well procedure enforces regulation.	Show not necessarily a trade- off btw regulation & performance.	Find audit incentivizes platform to add some content diversity.

Problem statement

The platform selects the content shown to its users by ...



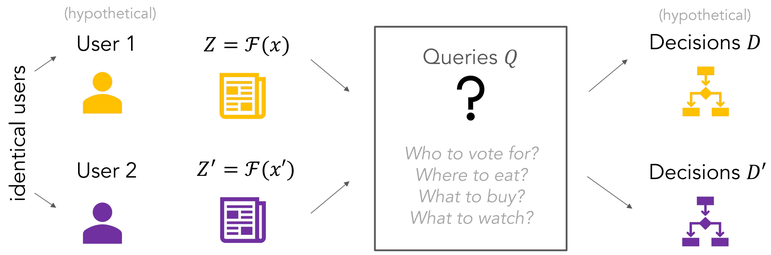
Auditor's task: Given a *counterfactual* regulation and black-box access to $\mathcal{F}$, check if the platform is compliant.

What is a counterfactual regulation?
"Algorithm $\mathcal{F}$ must behave similarly under x and x' for $(x, x') \in S$."

Ex 1: Weather forecasts should be consistent in the same location.
=
"The weather forecasts filtered by $\mathcal{F}$ should be *similar* for all users who are in the same geographical location."

Ex 2: Targeted ads cannot use sexual orientation as an input.
=
"The ads shown by $\mathcal{F}$ should be *similar* for two users who are identical except for sexual orientation."

Decision robustness



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

can formalize as hypothesis test

Auditing procedure

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

Inputs: $\mathcal{F}$ | x | x' | Θ | ϵ | $\mathcal{L}^+$ is MVUE

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

Black-box access to $\mathcal{F}$
No user data
Modular
Intuitive parameter
No content removal

Main results

1. 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**.

If $\mathcal{F}$ fails the audit, can be $(1-\epsilon)$-confident $\mathcal{F}$ is not decision-robust as $m \to \infty$.

2. Insight on MVUE.

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

To audit w/o access to users or their decisions (which may be unethical to get), use the MVUE. It gives an "upper bound" on the sensitivity of users to content.

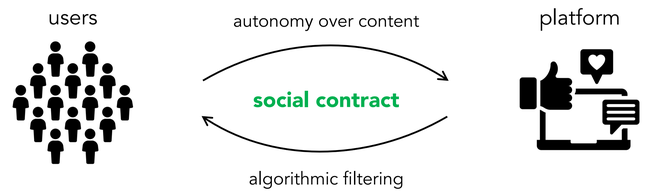
3. Conditions under which there is no performance-regulation trade-off.

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 available content is expressive enough, there is no regulation-performance trade-off.

There are conditions under which the platform does not sacrifice performance.

Content diversity can lower the cost of regulation: The lower the diversity of Z and Z', the more easily an auditor can distinguish between $\mathcal{F}(x)$ and $\mathcal{F}(x')$.

Example implementation: Social contract



Baseline content: For a given user, collection of content generated by the user's friends, users that she follows, pages that she subscribes to, and so on.

Baseline feed: Feed generated by drawing items UAR from baseline content.

Auditor is given two feeds: the baseline feed $\mathcal{B}$ and filtered feed $\mathcal{Z}$.
Auditor does not know *a priori* which feed is the baseline feed.

Running the audit ensures decision-robustness of $\mathcal{Z}$ w.r.t. $\mathcal{B}$
→ The content in the filtered feed is similar to the content to which the user has given consent.