Peer Filtering: Democratic Misinformation Control in Social Networks

Calvin Roth, Ankur Mani, Krishnamurthy Iyer

Department of Industrial and Systems Engineering
University of Minnesota

November 20, 2024

- Misinformation is rampant in online social networks with serious societal implications.

- Misinformation is rampant in online social networks with serious societal implications.
- Governments do not know how to regulate and passes the blame to the platforms.

- Misinformation is rampant in online social networks with serious societal implications.
- Governments do not know how to regulate and passes the blame to the platforms.
  - *"I don't think that Facebook or Internet platforms in general should be arbiters of truth."* - Mark Zuckerberg
    Quoted in [Jackson et al., 2022]

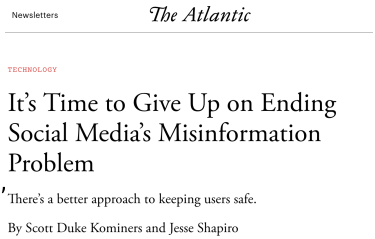
- Misinformation is rampant in online social networks with serious societal implications.
- Governments do not know how to regulate and passes the blame to the platforms.
  - *"I don't think that Facebook or Internet platforms in general should be arbiters of truth."* - Mark Zuckerberg
    Quoted in [Jackson et al., 2022]
- Online social networking platforms do spend significant effort to curb but without much success.

Common Techniques

- Tagging/Removing [Patwa et al., 2021, Pennycook and Rand, 2019, Clayton et al., 2020, Brashier et al., 2021, Mena, 2020, Lyons et al., 2020, Carey et al., 2022, Pennycook et al., 2020a]
- Nudging [Pennycook et al., 2020b, Pennycook et al., 2021, Pennycook and Rand, 2021, Fazio, 2020]
- Debunking [Van Der Linden, 2022, Chan et al., 2017, Nyhan et al., 2014, Schwarz et al., 2016, Bhargava et al., 2023, Ecker et al., 2020, Paynter et al., 2019]
- Pre-bunking [Van der Linden et al., 2017, Pfau and Burgoon, 1988, Van Der Linden, 2022, Niederdeppe et al., 2015, Cook et al., 2017, Banas and Rains, 2010]

Criticism

- Targeting specific users and content is criticized for bias and leads to mistrust [Kominers and Shapiro, 2024b, Kominers and Shapiro, 2024a].


Newsletters
The Atlantic

TECHNOLOGY

It's Time to Give Up on Ending Social Media's Misinformation Problem

There's a better approach to keeping users safe.

By Scott Duke Kominers and Jesse Shapiro

Criticism

- Targeting specific users and content is criticized for bias and leads to mistrust [Kominers and Shapiro, 2024b, Kominers and Shapiro, 2024a].

Newsletters *The Atlantic*

TECHNOLOGY

It's Time to Give Up on Ending Social Media's Misinformation Problem

There's a better approach to keeping users safe.

By Scott Duke Kominers and Jesse Shapiro

What can the platforms do instead?

Research Questions

- Can the social network filter out false content? *Peer Filtering*

Research Questions

- Can the social network filter out false content? *Peer Filtering*
- When is Peer Filtering effective?

Research Questions

- Can the social network filter out false content? *Peer Filtering*
- When is Peer Filtering effective?
- How can the platform enhance the Peer Filtering effect without directly moderating content?

Model

A large social network where users receive content either shared by their peers or from external sources.

New content is arrives into the system and is shown to one user. Content has a *veracity* and an *inclination*:

- Has veracity $\alpha = T$ w.p. p and F w.p. $1 - p$
- Equally likely to be inclined *left* or *right*
- Equally likely to be introduced to each user

Model

A large social network where users receive content either shared by their peers or from external sources.

New content is arrives into the system and is shown to one user.
Content has a *veracity* and an *inclination*:

- Has veracity $\alpha = T$ w.p. p and F w.p. $1 - p$
- Equally likely to be inclined *left* or *right*
- Equally likely to be introduced to each user

Users when exposed to content decide to share/not share the content.

User Model

Users have an inclination and different preferences for alignment and truth.

User Model

Users have an inclination and different preferences for alignment and truth.

	T	F
A	1	-1
M	1	-1

Impartial Truthteller (IT)

	T	F
A	1	-1
M	-1	-1

Partisan Truthteller (PT)

	T	F
A	1	1
M	1	-1

Impartial Fabulist (IF)

	T	F
A	1	1
M	-1	-1

Partisan Fabulist (PF)

Figure: Utility from sharing content for different user types.

$\theta_t :=$ Fraction of users who are type t for $t \in \{IT, IF, PT, PF\}$.

Private Signal

- When the ith user consumes content, k, they will know if it is aligned with their political views.

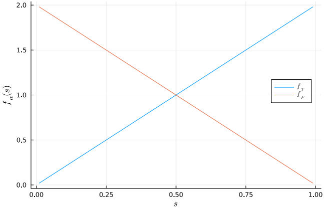
Private Signal

- When the ith user consumes content, k, they will know if it is aligned with their political views.
- This user also receives an independent signal $s_{ik} \in [0, 1]$ regarding its veracity.
- s_{ik} is sampled from f_T or f_F depending on the veracity of the content itself.

Private Signal

- When the ith user consumes content, k, they will know if it is aligned with their political views.
- This user also receives an independent signal $s_{ik} \in [0, 1]$ regarding its veracity.
- s_{ik} is sampled from f_T or f_F depending on the veracity of the content itself.
- We pick

$$f_T(s) = 2s \qquad f_F(s) = 2(1 - s) \qquad s \in [0, 1]$$



Agent Behavior

Agent i's belief about the kth content veracity, where δ is the prior belief, common for all content, and $\hat{\delta}(x)$ is the posterior belief given private signal $s_{ik} \sim f_\alpha$

$$\mathbb{P}(\alpha_k = T | s_{ik} = x) = \frac{\delta x}{\delta x + (1-\delta)(1-x)} := \hat{\delta}(x).$$

Agent Behavior

Agent i's belief about the kth content veracity, where δ is the prior belief, common for all content, and $\hat{\delta}(x)$ is the posterior belief given private signal $s_{ik} \sim f_\alpha$

$$\mathbb{P}(\alpha_k = T | s_{ik} = x) = \frac{\delta x}{\delta x + (1-\delta)(1-x)} := \hat{\delta}(x).$$

Agent utility from sharing

	IT	PT	IF	PF
A	$2\hat{\delta}(x) - 1$	$2\hat{\delta}(x) - 1$	1	1
M	$2\hat{\delta}(x) - 1$	-1	$2\hat{\delta}(x) - 1$	-1

Figure: Expected utility of each user type for sharing an aligned (A) or a misaligned (M) content, as a function of her posterior belief $\hat{\delta}(x)$.

Agent Behavior

Agent sharing decisions

$$\beta^{\alpha}_{a,t} = \begin{cases} 1 & \text{if } a = \mathrm{A} \text{ and } t \in \{\mathrm{IF}, \mathrm{PF}\}; \\ 0 & \text{if } a = \mathrm{M} \text{ and } t \in \{\mathrm{PT}, \mathrm{PF}\}; \\ \overline{\mathrm{F}}_{\alpha}(1-\delta) & \text{otherwise,} \end{cases} \tag{1}$$

Agent Behavior

Agent sharing decisions

$$\beta^{\alpha}_{a,t} = \begin{cases} 1 & \text{if } a = \mathrm{A} \text{ and } t \in \{\mathrm{IF}, \mathrm{PF}\}; \\ 0 & \text{if } a = \mathrm{M} \text{ and } t \in \{\mathrm{PT}, \mathrm{PF}\}; \\ \overline{\mathrm{F}}_{\alpha}(1-\delta) & \text{otherwise,} \end{cases} \tag{1}$$

Fraction of T/F content shared for a given prior δ.

$$\beta^{\alpha}(\delta) := \sum_{a \in \{\mathrm{A},\mathrm{M}\}} \frac{1}{2} \cdot \sum_{t \in \Theta} \theta_t \beta^{\alpha}_{a,t} = \frac{1}{2}\left(\theta_{\mathrm{F}} + (\theta_{\mathrm{T}} + \theta_{\mathrm{I}})\overline{\mathrm{F}}_{\alpha}(1-\delta)\right) \tag{2}$$

Agent Behavior

Agent sharing decisions

$$\beta^{\alpha}_{a,t} = \begin{cases} 1 & \text{if } a = \mathrm{A} \text{ and } t \in \{\mathrm{IF}, \mathrm{PF}\}; \\ 0 & \text{if } a = \mathrm{M} \text{ and } t \in \{\mathrm{PT}, \mathrm{PF}\}; \\ \overline{\mathrm{F}}_{\alpha}(1-\delta) & \text{otherwise,} \end{cases} \tag{1}$$

Fraction of T/F content shared for a given prior δ.

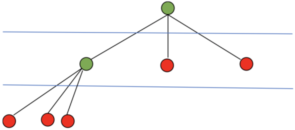
$$\beta^{\alpha}(\delta) := \sum_{a \in \{\mathrm{A},\mathrm{M}\}} \frac{1}{2} \cdot \sum_{t \in \Theta} \theta_t \beta^{\alpha}_{a,t} = \frac{1}{2}\big(\theta_{\mathrm{F}} + (\theta_{\mathrm{T}} + \theta_{\mathrm{I}})\overline{\mathrm{F}}_{\alpha}(1-\delta)\big) \tag{2}$$

Observation

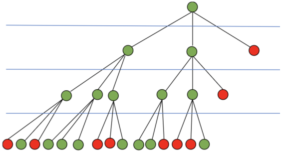
True content is always shared more than False content.

$$\beta^{T} \geq \beta^{F}$$

Spread Process



(a)



(b)

- Content either goes extinct(Bust) or spreads to the whole network(Boom).
- If $\kappa\beta^\alpha < 1$ then content goes bust with certainty and the expected spread is $\frac{1}{1-\kappa\beta^\alpha}$.
- If $\kappa\beta^\alpha > 1$ then the probability of going boom, q^α, is given by

$$\underbrace{1 - q^\alpha}_{\text{Content goes bust}} = \underbrace{1 - \beta^\alpha}_{\text{No first share}} + \underbrace{\beta^\alpha (1 - q^\alpha)^\kappa}_{\text{All descendants go bust}}$$

Proportion of True Content

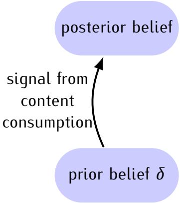
The proportion of $\Phi(\beta^T, \beta^F)$ of true content is

$$\Phi(\beta^T, \beta^F) = \begin{cases} \frac{\frac{p}{1-\kappa\beta^T}}{\frac{p}{1-\kappa\beta^T} + \frac{1-p}{1-\kappa\beta^F}} & \text{If } \kappa\beta^F \leq \kappa\beta^T < 1; \\ 1 & \text{If } \kappa\beta^F < 1 \leq \kappa\beta^T; \\ \frac{pq^T}{pq^T + (1-p)q^F} & \text{If } \kappa\beta^T \geq \kappa\beta^F > 1 \end{cases}$$

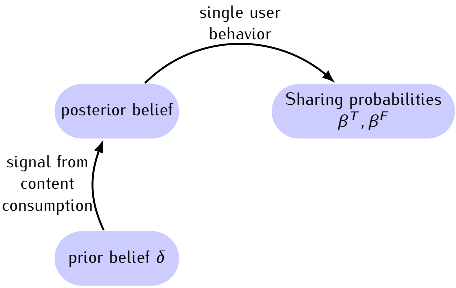
Equilibrium

prior belief δ

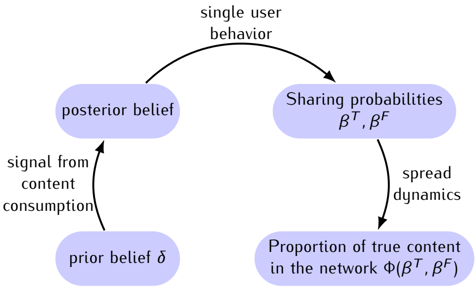
Equilibrium



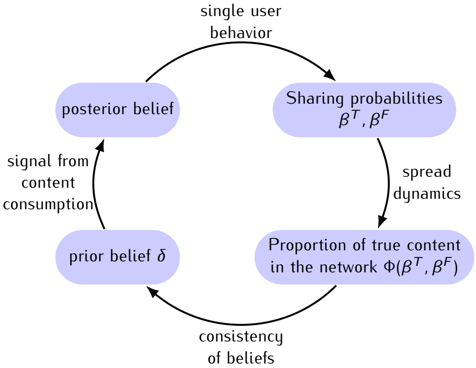
Equilibrium



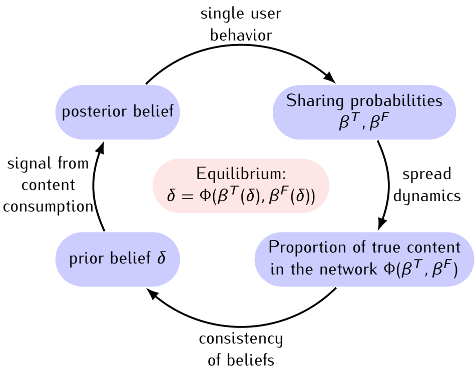
Equilibrium



Equilibrium



Equilibrium



Definition

An equilibrium is a pair of sharing probabilities β^T, β^F that satisfy the equation

$$\delta = \Phi(\beta^T(\delta), \beta^F(\delta))$$

True and false content bloom together

Theorem

In every equilibrium either both true and false content are in a bust equilibrium or both are in a boom equilibrium.

True and false content bloom together

Theorem

In every equilibrium either both true and false content are in a bust equilibrium or both are in a boom equilibrium.

- Since $\beta^F \leq \beta^T$ it seems like we just need to find a κ that makes $\kappa\beta^F < 1 \leq \kappa\beta^T$

True and false content bloom together

Theorem

In every equilibrium either both true and false content are in a bust equilibrium or both are in a boom equilibrium.

- Since $\beta^F \leq \beta^T$ it seems like we just need to find a κ that makes $\kappa\beta^F < 1 \leq \kappa\beta^T$
- But if only true news goes viral then users are more trusting $\rightarrow$ false news can spread.

Existence of Equilibria

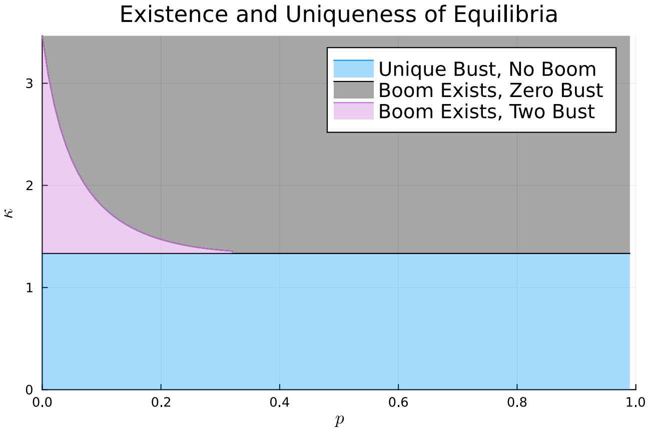


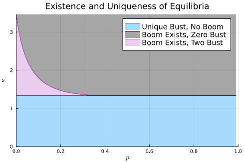
Figure: Visual Representation of how p and κ effect the equilibria. Here $\theta_t = \frac{1}{4}$ for all types.

Theorem: Existence and Uniqueness

Theorem

There exists a boom equilibrium if and only if $\kappa > \frac{2}{1+\theta_l}$. *On the other hand, there exists a bust equilibrium if any of the following conditions are met:*

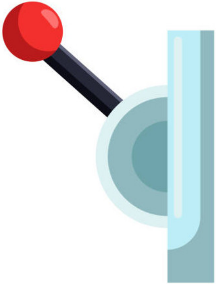
- $\kappa < \frac{2}{1+\theta_l}$ *(a unique bust equilibrium exists);*
- $\kappa = \frac{2}{1+\theta_l}$ *and* $p < \frac{1}{3}$ *(a unique bust equilibrium exists);*
- *If* $\kappa \in \left(\frac{2}{1+\theta_l}, \kappa(p) \right)$ *and* $p < \frac{1}{3}$ *(two bust equilibria exist);*
- *If* $\kappa = \kappa(p)$ *and* $p < \frac{1}{3}$ *(a unique bust equilibrium exists),*



What levers does the platform have?

What levers does the platform have?

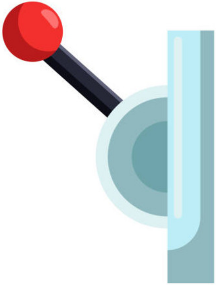
The platform can change κ.



What levers does the platform have?

The platform can change κ.
Accuracy metrics:

- Specificity: q^T
- Sensitivity: $1 - q^F$



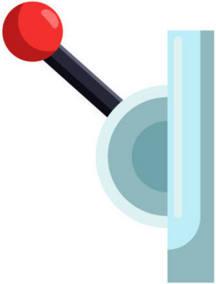
What levers does the platform have?

The platform can change κ.
Accuracy metrics:

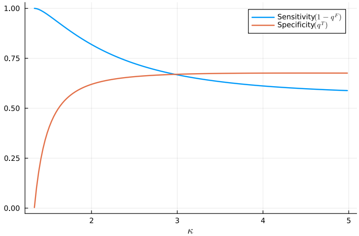
- Specificity: q^T
- Sensitivity: $1 - q^F$

Usage metrics:

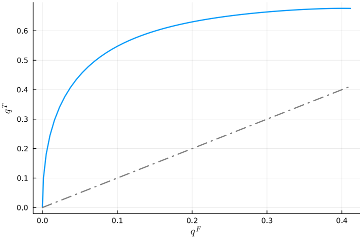
- Engagement: $p\beta^T + (1-p)\beta^F$
  - Probability the average user shares content.
- Volume: $pq^T + (1-p)q^F$
  - The amount of viral content.



Results

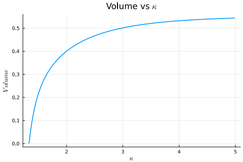


(a) Sensitivity(q^T) and specificity($1 - q^F$) vs κ



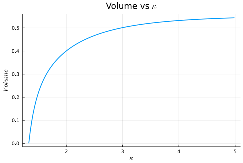
(b) Sensitivity(q^T) vs specificity(q^F).

Usage

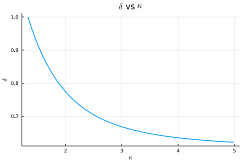


(a)
Volume($pq^T + (1-p)q^F$). This is the amount of content that goes viral

Usage

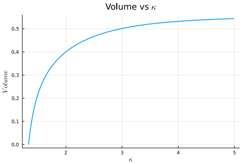


(a) Volume$(pq^T + (1-p)q^F)$. This is the amount of content that goes viral

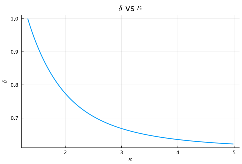


(b) Average proportion of true content in the network

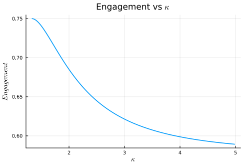
Usage



(a) Volume($pq^T + (1-p)q^F$). This is the amount of content that goes viral



(b) Average proportion of true content in the network



(c) Engagement($p\beta^T + (1-p)\beta^F$). This is the probability any particular article is shared.

References

Banas, J. A. and Rains, S. A. (2010).
A meta-analysis of research on inoculation theory.
Communication monographs, 77(3):281–311.

Bhargava, P., MacDonald, K., Newton, C., Lin, H., and Pennycook, G. (2023).
How effective are tiktok misinformation debunking videos?
Harvard Kennedy School Misinformation Review.

Brashier, N. M., Pennycook, G., Berinsky, A. J., and Rand, D. G. (2021).
Timing matters when correcting fake news.
Proceedings of the National Academy of Sciences, 118(5):e2020043118.

Carey, J. M., Guess, A. M., Loewen, P. J., Merkley, E., Nyhan, B., Phillips, J. B., and Reifler, J. (2022).
The ephemeral effects of fact-checks on covid-19 misperceptions in the united states, great britain and canada.