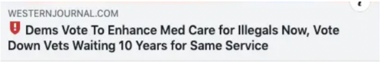
Misinformation: Strategic Sharing, Homophily, and Endogenous Echo Chambers

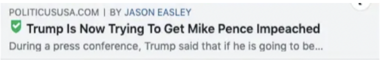
Daron Acemoglu Asu Ozdaglar James Siderius

Economics Theory Lunch
February 23, 2021

Motivation

- In 2016, 14% of Americans said they use social media as their primary source of news (Allcott and Gentzkow (2017)) with over 70% of Americans getting at least *some* of their news from social media (Levy (2020)).





- Falsehood diffused significantly farther, faster, deeper, and more broadly than truth (Vosoughi et al (2018)). Widespread concerns about the diffusion and propagation of misinformation.
- Exacerbated by "filter bubble" algorithms of social media platforms (Levy (2020)): platform shows users what they think users will engage with most based on their beliefs.
- Such misinformation may have had political and social effects (Allcott and Gentzkow (2017)).

This Paper

- A model of online content sharing, where the content may contain *misinformation*.

- *Key Decision*: as a user of the platform, when to share and when to inspect ("fact-check") information for truthfulness.

- Key findings:
  - Effects of political polarization and homophily of the social network.
  - Characterize and clarify why the platform's incentives may propagate misinformation.
  - Possible policy remedies taking into account potential backfiring of interventions.

Model: Preliminaries

- Underlying state of the world $\theta \in \{L, R\}$, corresponding to whether the left-wing or right-wing candidate is more qualified.

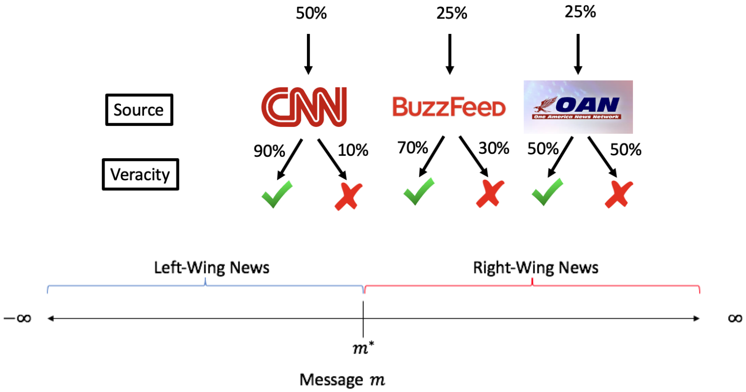
- There are N agents in the population, and each agent i has a heterogenous prior $b_i \in (0,1)$ that the state is $\theta = R$ which is drawn according to a continuous distribution $H_i(\cdot)$ at $t = 0$ with lower support $\underline{b}_i$ and upper support $\bar{b}_i$.

- Agents are arranged in a stochastic social network defined by a matrix $\mathbf{P}$ of link probabilities with p_{ij} being the probability that agent i has a link to agent j.

- We denote by $\mathcal{N}_i$ as the set of agents attached to agent i with an outgoing link (i.e., agent i's neighborhood).

Model: News Generation

- News story generated at $t = 0$ with a three-dimensional type (s, ν, m).



- When the news is truthful, message is drawn from density $f(\cdot|\theta)$ satisfying MLRP. When it contains misinformation, drawn as if state is actually $\neg\theta$.
- Neither s nor ν is known to the agents. The ex-ante (before looking at the message m) probability the article is truthful is q.

Model: Agents' Actions

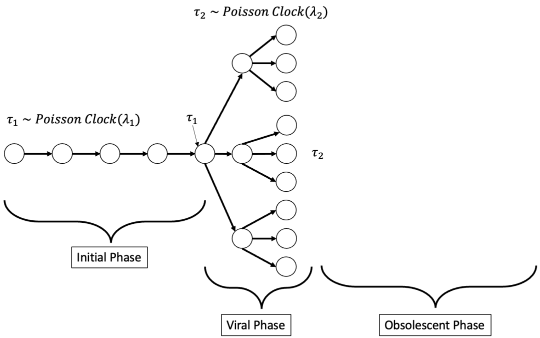
- Time is discrete $t = 1, 2, \ldots$. The article starts at agent i^* chosen uniformly at random at $t = 1$.

- An agent i who receives the article reads the message m and then chooses an action $a_i \in \{\mathcal{S}, \mathcal{I}, \mathcal{K}\}$:
  - $\mathcal{S}$: immediately share the article.
  - $\mathcal{I}$: first inspect the article for veracity before sharing it (i.e., "fact-check").
  - $\mathcal{K}$: immediately kill the article by not sharing it with others.

- *Type-I error*: kill a truthful article because the agent strongly disagrees with it, even though it contains accurate news (e.g., right-extremist kills a left-wing article).

- *Type-II error*: share an article with misinformation because it happens to confirm your own beliefs.

Model: Phases of the Article

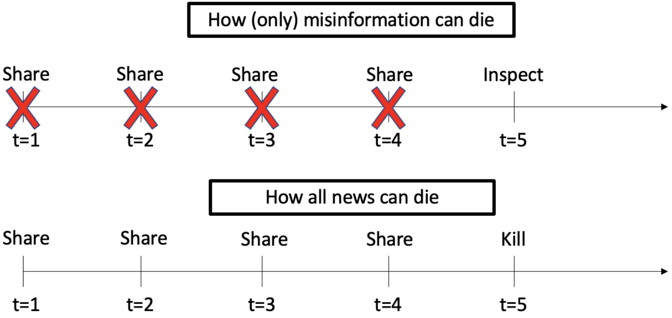


- *Initial Phase*
  - Agent i shares: article is passed onto **exactly one agent** (chosen uniformly at random from $\mathcal{N}_i$).
  - Agent i kills: the article moves directly to the obsolescent phase.
- *Viral Phase*
  - Agent i shares: article is passed onto γ agents (chosen uniformly at random from $\mathcal{N}_i$).
- *Obsolescent Phase*
  - Article becomes obsolete and is inspected by an outside source.

Model: Payoffs

- The game "ends" when either: (i) the obsolescent phase ends or (ii) the article is inspected and found to contain misinformation.
- $\mathcal{K}$: Normalize payoff to 0.
- $\mathcal{S}$
  - Let t_i be the time in which agent i receives the article.
  - Share utility given by $S_i \equiv \kappa \sum_{\tau=1}^{\infty} \beta^{\tau-1} S_{i,\tau}$ where β is the discount factor, κ is the marginal share utility, and $S_{i,\tau}$ is the number of (indirect) shares occurring τ periods later resulting from i's share.
  - If article is inspected at time t and found to contain misinformation, agent i faces a social punishment from sharing misinformation $C\beta^{t-t_i-1}$; for instance, reputational concerns.
- $\mathcal{I}$
  - Inspection is costly; agents pay a cost $K > 0$ to inspect.
  - Receive a benefit $\delta > 0$ from "exposing" a viral article that contains misinformation; get 0 benefit from exposing an initial phase article.
  - If article is truthful, receive the same payoff as playing $\mathcal{S}$ after paying K.
- Let $v_{initial}$ and v_{viral} be the share payoff when it is common knowledge the article is truthful (exogenous). Assume parameter values satisfy $v_{initial} < K < \min\{v_{viral}, \delta\}$.

Model: Payoff Illustration

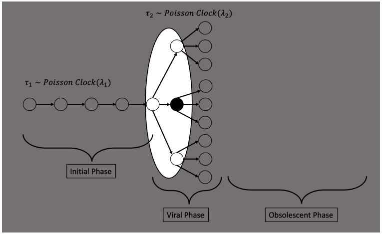


Red X denotes agents who are punished

Model: Information Structure

- Agents do not have knowledge of the social network, the prior sharing process, or calendar time.

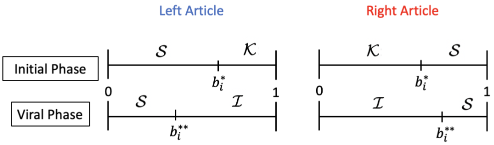
- If agent i receives the article from agent j, she observes how many *other* agents received the article from agent j as well.
  - While agents do not know calendar time, they are aware which phase the article is in.



- Solution Concept: Sequential equilibria.

Equilibrium Characterization: Cutoff Form

Recall that b_i is the prior (or "ideology") of agent i about $\theta = R$. Define a cutoff strategy as:



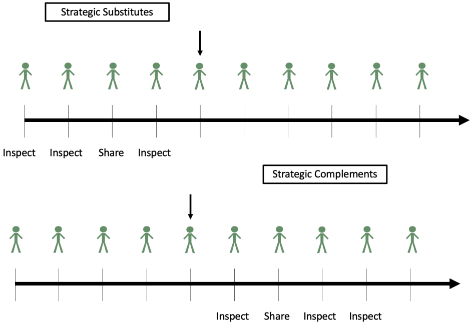
Theorem

There exists a cutoff-strategy equilibrium and all equilibria are in cutoff strategies.

Proof Sketch: WLOG we assume that $m > m^*$ for the remainder of this talk.

- Easy to show inspecting is dominated in initial phase and killing dominated in viral phase.
- The posterior belief π_i that the article is truthful given m is increasing in b_i.
- The payoff from sharing over inspecting and killing is increasing in π_i.
- *Existence*: Define map $\phi : [0,1]^{2N} \to [0,1]^{2N}$ from cutoff space to best-response cutoff space. Apply Brouwer's fixed point theorem for existence.

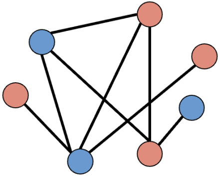
Equilibrium Characterization: Strategic Complements in the Viral Phase



- Strategic substitutes with past agents: more inspections increase my belief the article is truthful conditional on it coming to me; no need to inspect.
- Strategic complements with future agents: more inspections means sharing misinformation is more dangerous; should be cautious and inspect first.
- In some basic simulations, vast majority (over 98%) of all-share equilibria (where $(b_1^*, b_1^{**}, \ldots, b_N^*, b_N^{**}) = \mathbf{0}$) satisfy net strategic complements property.

Single-Island Model: Symmetric Equilibria

Assume the network is an Erdos-Renyi network with link probability $p_{ij} = p \in (0,1)$ and the distribution of priors is the same for every agent, i.e., $H_i = H$.



Proposition

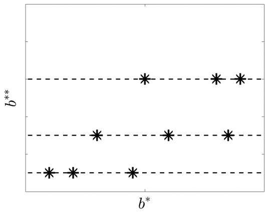
As $N \to \infty$, only symmetric equilibria survive; that is, $b_i^* = b^*$ and $b_i^{**} = b^{**}$ for all agents i.

Intuition. When the population is large (and connections are uniform), the payoff from agent i's action a_i is the same as the payoff from agent j's action a_j. Both agents must employ identical cutoff strategies in equilibrium.

Single-Island Model: Lattice Structure

Theorem

The equilibrium set of cutoffs (b^*, b^{**}) form a lattice structure according to the natural order.



Proof Sketch

- *Viral phase* $\implies$ the cutoff b^* does not matter in a sequential equilibrium for the best-response of agents. Solve for the set of b^{**} that can be supported in the viral phase (independent of what happens in the initial phase).
- Strategic complements in the *initial phase*: killing simultaneously increases the chance of punishment and decreases the share utility.
- Set of b^* also monotone in b^{**}: more inspections in the viral phase decrease incentives of initial phase agents to share.

Single-Island Model: Extremism

Proposition

Consider an extremal equilibrium (b^*, b^{**}) that satisfies the strategic complements property with message $m > m^*$.

- (a) *If no left-wing agents ever share, then if $m' > m$ (the message becomes more extreme), there are more shares in both phases.*
- (b) *If some left-wing agents share in both phases, then if $m' > m$ (the message becomes more extreme), there are fewer shares in both phases.*

- Levy (2020): Engagement with counter-attitudinal news can reduce strong attitudes about politically-congruent, extremist content.
- Left-wing agents in the population act as a firebreak in the spread of extremist right-wing news that contains misinformation.
- When views on social media are homogenous, extremism fuels aggressive sharing without fact-checking.

Single-Island Model: Polarization

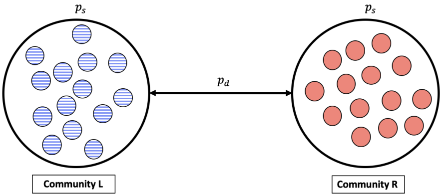
Proposition

Suppose there is an extremal equilibrium (b^*, b^{**}) that satisfies the strategic complements property with prior distribution H.

1. *If H' is more polarized than H and no left-wing agents ever share, then there are more shares in both phases.*
2. *If H' is more polarized than H and some left-wing agents share in both phases, then there are fewer shares in both phases.*

- When inspections are high, more polarization hurts the spread of misinformation because it encourages right-wing agents to stop inspecting knowing much of society is extreme right.
- With healthy skepticism from left-wing agents, more polarization increases scrutiny and promotes accountability for news sharing.
- Polarization in networks with uniform connections (no homophily) can help reduce the spread of viral misinformation.

Multiple Island Networks: Preliminaries



- Partition agents into k blocks of size $N_1, N_2, \ldots, N_k$, called *islands*.
- Let $\ell_i \in \{1, \ldots, k\}$ denote the island that agent i is in.
- Link probabilities are given by:

$$p_{ij} = \begin{cases} p_s, & \text{if } \ell_i = \ell_j \\ p_d, & \text{if } \ell_i \neq \ell_j \end{cases}$$

  where $p_s > p_d$.
- This is known as the *homophily structure* of the network. Special case is the *segregated islands* model whereby $p_s > 0$ but $p_d = 0$.
- Also assume that island ℓ has prior distribution H_ℓ and there exists a chain $H_1 \succeq_{FOSD} H_2 \succeq_{FOSD} \cdots \succeq_{FOSD} H_k$.

Multiple Island Networks: Virality

Proposition

In the stochastic-block model, as $N \to \infty$, all equilibria are in symmetric island-dependent cutoff strategies. In other words, in every equilibrium, there exists $\{(b_\ell^*, b_\ell^{**})\}_{\ell=1}^k$ such that $b_i^* = b_{\ell_i}^*$ and $b_i^{**} = b_{\ell_i}^{**}$ for all agents i.

- However, cutoffs do not necessarily satisfy any lattice order: greater inspections on island 1 can lead to greater or fewer inspections on island 2, depending on whether strategic complements or substitutes dominates.

Definition

Let $T_1(i^*), T_2(i^*)$ be the (random) times at which the game ends under equilibria σ_1, σ_2, respectively, conditional on agent i^* being seeded. Call $S_{T(i^*)}$ the total amount of sharing that occurs before stopping time $T(i^*)$. We say that σ_1 is *more viral* than σ_2 if $\max_{i^*} \mathbb{E}[S_{T_1(i^*)}] > \max_{i^*} \mathbb{E}[S_{T_2(i^*)}]$.

- Social media platforms often target the initial agent who sees the article. "Virality" captures the total expected shares *conditional* on a good initial recommendation.

Multiple Island Networks: Homophily

Lemma

Suppose island ℓ has (net) strategic complements and has all-share on its island only. Then an increase in homophily preserves the all-share equilibrium on this island.

Theorem

Suppose some island with an all-share equilibrium has net strategic complements. Then there exists $\underline{p}$ such that if $p_s / p_d \geq \underline{p}$, misinformation always becomes (weakly) more viral following an increase in homophily.

Intuition:

- If seed agent i^* is on this island, need to consider the likelihood of the article to "jump" to a different island (which may or may not have more inspections following the increase in homophily).
- Show that when p_s / p_d is big enough to start, virality always (weakly) increases with more homophily even if inspections on all other islands go from 0 to 1 (and we keep all-share on island ℓ).

Two Islands: Extremism and Polarization

- H_R and H_L have distinct support, i.e., H_R has support on $[\underline{b}_R, \bar{b}_R]$ and H_L has support on $[\underline{b}_L, \bar{b}_L]$, with $\bar{b}_L < 1/2 < \underline{b}_R$.

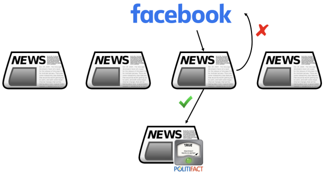
Theorem

There exists $\underline{p}$ such that if $p_s / p_d > \underline{p}$, either an increase in the extremity of the message or an increase in polarization (weakly) increases the virality of misinformation for the most viral equilibrium.

- *Intuition*: Extreme message on an extremist island will spread like wildfire (no inspections) and with significant homophily is unlikely to jump to a more scrutinizing island.
- Uniform connections: Extreme messages and polarization do not allow misinformation to spread very far.
- Extreme homophily: Extreme messages and polarization fuel the flames among pro-attitudinal agents.
- E.g., More inclined to share "All Lives Matter" if supporters of "Black Lives Matter" are unlikely to see it.

Platform: Design Problem

- The platform wants to maximize engagement (i.e., shares) on the platform and is indifferent to the veracity of the content.
- Let there be k communities with disjoint prior distributions $\underline{b}_1 < \bar{b}_1 < \underline{b}_2 < \cdots < \underline{b}_k < \bar{b}_k$. Communities are *ideologically symmetric*[1] and there is a least one fully left and one fully right-wing community.
- At $t = 0$ the platform makes the following choices:



- We assume the cost of inspection is minimal for the platform relative to the payoff (e.g., ad revenue) they receive from shares on the platform. Only do not fact-check if indifferent.
- Finally, Facebook can choose the network $\mathbf{P}$ by using any recommendation algorithm it would like.

[1] In the sense that $\underline{b}_\ell = 1 - \bar{b}_{k-\ell+1}$ and $\bar{b}_\ell = 1 - \underline{b}_{k-\ell+1}$ holds for all ℓ.

Platform: Filter Bubble Algorithm

Definition

Let $\mathcal{L}(m) \equiv f(m|\theta = R)/f(m|\theta = L)$. We say message m is *more extreme* than message m' if $\max\{1/\mathcal{L}(m), \mathcal{L}(m)\} \geq \max\{1/\mathcal{L}(m'), \mathcal{L}(m')\}$.

- Extremity does not differentiate between left or right-wing news. Only depends on the likelihood of the message coming from one side or the other.

Theorem

There exists η such that:

a. *If $\max_m \max\{1/\mathcal{L}(m), \mathcal{L}(m)\} < \eta$, there exists a sequence $\{a_1, a_2, \ldots, a_n\}$ such that the platform chooses articles in this sequential order until one can be verified, "tags" it as truthful, and then adopts any network model;*
b. *If $\max_m \max\{1/\mathcal{L}(m), \mathcal{L}(m)\} > \eta$, the platform chooses the most extreme article, does not inspect it, and adopts the segregated islands connection model.*

- Platform inclined to pick extreme articles and recommend them to extremist communities.
- Platform's optimal recommendation algorithm gives rise to an endogenous echo chamber where misinformation goes entirely unchecked.

Planner's Problem: Provenance

- Assume WLOG we are in the single-island model with lattice structure.
- Does revealing the source of the news cut down on the spread of misinformation?
  - *Effective inspection cost*: By providing the source, one reduces the "effective" inspection cost K of the agent:

Proposition

If the effective inspection cost K decreases, then there is more inspecting in both the most and least sharing equilibria.

- One can reveal the provenance s of the news source. Two types of sources: reputable (probability ϕ) and sketchy (probability $1 - \phi$).

Proposition

There exists $\bar{\phi} < 1$ such that:

1. *If $\phi > \bar{\phi}$, a policy that reveals the source of the news reduces the virality of misinformation in both the most and least sharing equilibria;*
2. *If $\phi < 1 - \bar{\phi}$, a policy that reveals the source of the news increases the virality of misinformation in both the most and least sharing equilibria.*

Planner's Problem: Backfire Example

Example

Suppose the reputable news has ex-ante probability $q_r = 0.9$ whereas the sketchy news has probability $q_s = 0.5$, and both are equally likely. When the news source is not revealed, the probability the article is truthful is $q = 0.7$. There is a unique equilibrium for all three instances:

1. *Revelation, reputable*: An all-share equilibrium is the unique equilibrium for reputable sources because it is unlikely the article is fake (and is a waste of resources to verify it).
2. *Revelation, sketchy*: An all-inspect equilibrium is the unique equilibrium for sketchy sources because it is quite likely the article is fake (and sharing is potentially very costly if the article is revealed as so).
3. *No revelation*: Because the article *may* (with 50% probability) be coming from a sketchy source, over 90% of the population inspects the article before sharing, just to be safe.

However, on average, revealing the article's provenance is much worse: inspections drop from 90% to 50% on average, and leads to a 5% likelihood of a fake article not being inspected as opposed to only a 3% likelihood when the source is kept hidden. The population is much too trusting of reputable articles.

Planner's Problem: Censorship

1. At time $t = 0$, the planner observes the article's message m but not its source s or veracity ν.
2. The planner can either choose to *censor* the message or *allow* the message. In the former case, the article is killed and does not propagate on the platform. In the latter case, the article is introduced to a seed agent at $t = 1$ as usual.
3. Planner's objective: Censor articles that, for the optimal recommendation algorithm of the platform, will not be fact-checked ever by the users.

Proposition

There exists a threshold η such that if the message m satisfies $\{\mathcal{L}(m), 1/\mathcal{L}(m)\} > \eta$, the article is censored; otherwise, it is allowed.

- Key takeaway: aligned interests in fact-checking between the platform and the users.
  - Moderate articles are fact-checked by the platform before recommendation, but would be inspected anyway by the users of the platform.
  - Extreme articles that ought to be fact-checked by the platform go unchecked by the platform *and* users.
  - To correct the latter, need to censor these extreme articles.

Conclusion

- First known model of strategic content sharing that shows how echo chambers exacerbate the spread of misinformation.
- Polarization can act as a deterrent in a well-integrated society but fuels extremism and single-mindedness in the presence of echo chambers.
- Social media platform who wishes to maximize engagement can capitalize on this effect: share extremist articles with extremist communities that may or may not contain accurate information.
- Policies that demonstrate the provenance of the article or censor extreme articles can often be effective at combatting misinformation.
- Model can be used to understand the efficacy of other policies on the control of misinformation in social networks.