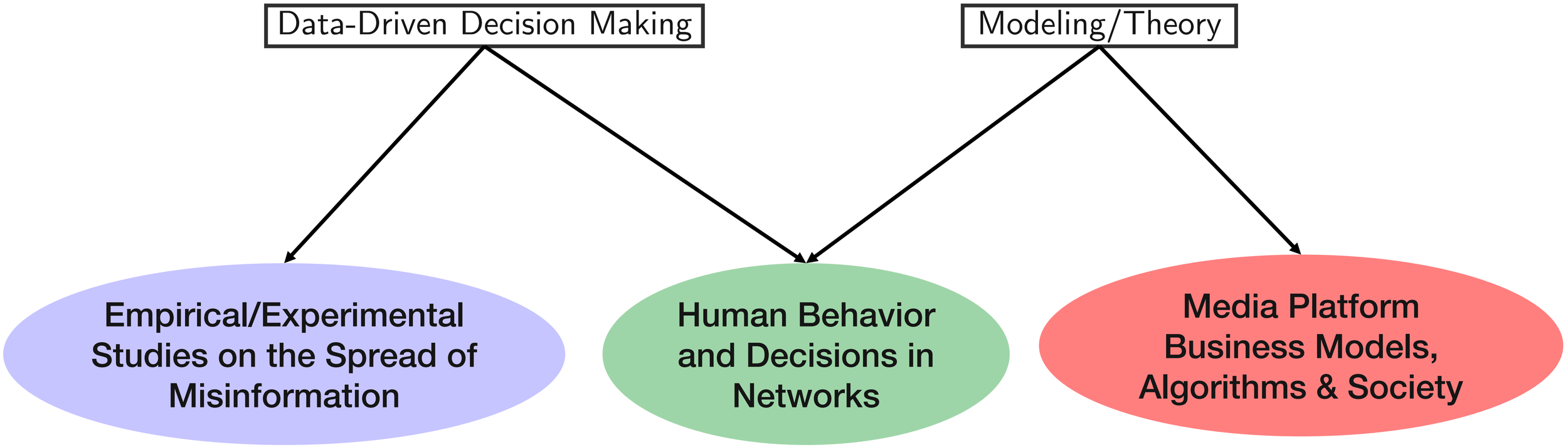
A Model of Online Misinformation

James Siderius (Dartmouth College – Tuck School of Business)
Linköping University – ELLIIT Network Dynamics and Control Focus Period
September 26, 2023

Broader Research Agenda

- At the interface of **computer science**, **operations research**, and **economics**.



Empirical/Experimental Misinformation

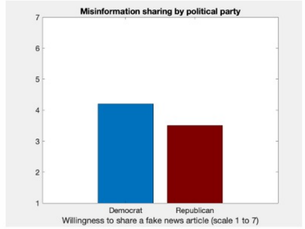
Fighting Fire with Fire: An Experiment on Misinformation Sharing Incentives
Daron Acemoglu, Adam Berinsky, Asu Ozdaglar, David Rand, and James Siderius

Informational Interventions

Which of the following best describes your political preference?

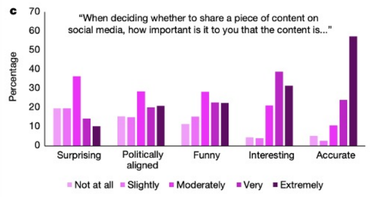
Strongly Democratic	Democratic	Lean Democratic	Lean Republican	Republican	Strongly Republican
○	◉	○	○	○	○

Below is a figure from a second study that classifies the political groups that spread the most misinformation. Despite common misperceptions, the study found that **Democrat voters spread the most misinformation on social media**.



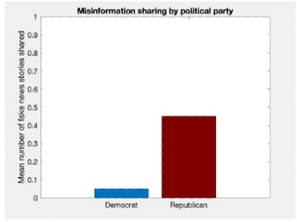
In-Group Treatment

Below is a figure from the first study that asks survey participants about the reasons for sharing social media content. Most participants responded that it is **very important or extremely important that a piece of content be interesting before sharing it**.



Control Treatment

Below is a figure from a second study that classifies the political groups that spread the most misinformation. Despite common misperceptions, the study found that **Republican voters spread the most misinformation on social media**.



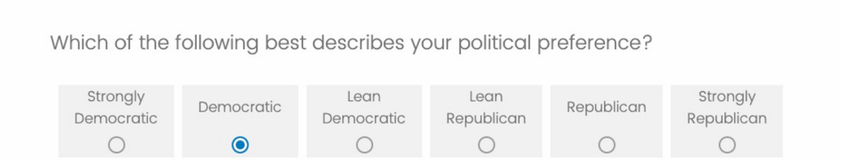
Out-Group Treatment

Empirical/Experimental Misinformation

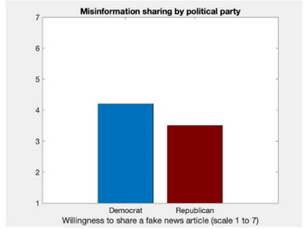
Fighting Fire with Fire: An Experiment on Misinformation Sharing Incentives
Daron Acemoglu, Adam Berinsky, Asu Ozdaglar, David Rand, and James Siderius

Generative AI, Algorithmic Ranking, and Social Media Engagement
Daniel Huttenlocher, Asu Ozdaglar, Charles Lyu, James Siderius, others @ MIT Media Lab

Informational Interventions

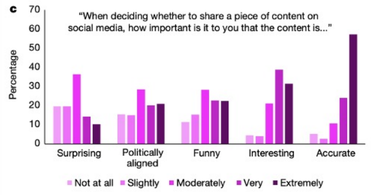


Below is a figure from a second study that classifies the political groups that spread the most misinformation. Despite common misperceptions, the study found that **Democrat voters spread the most misinformation on social media**.



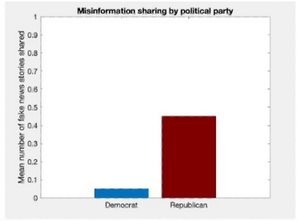
In-Group Treatment

Below is a figure from the first study that asks survey participants about the reasons for sharing social media content. Most participants responded that it is **very important or extremely important that a piece of content be interesting before sharing it**.

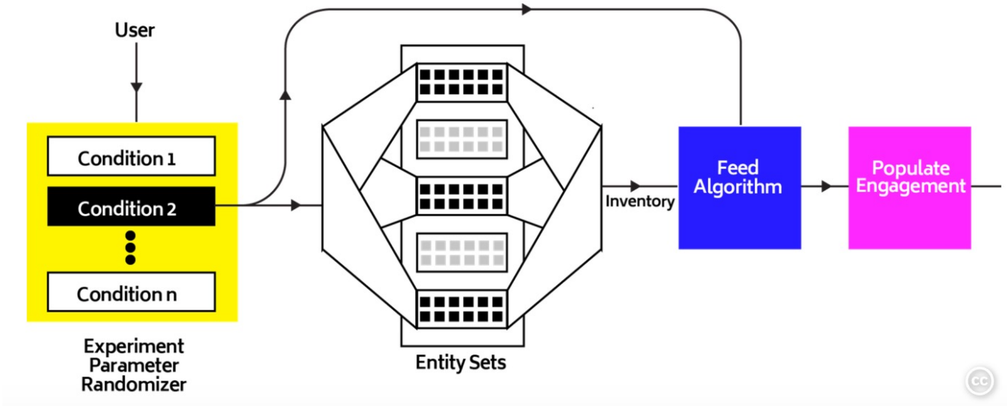


Control Treatment

Below is a figure from a second study that classifies the political groups that spread the most misinformation. Despite common misperceptions, the study found that **Republican voters spread the most misinformation on social media**.



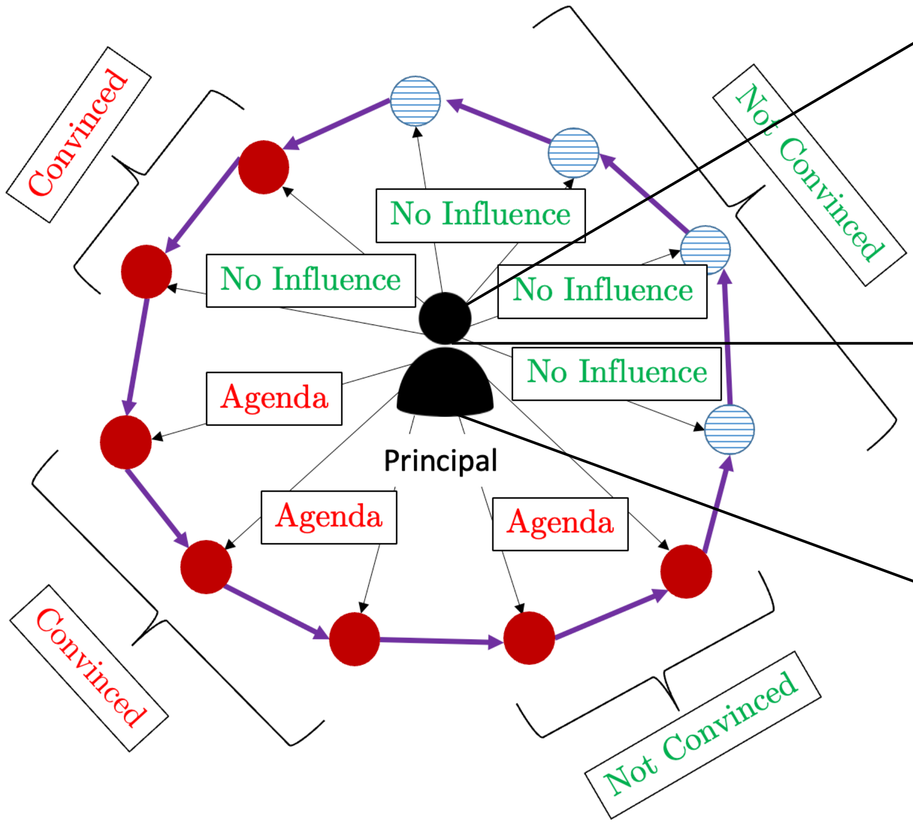
Out-Group Treatment



Credit: Ziv Epstein and MIT Media Lab

Human Behavior and Decisions in Networks

When is Society Susceptible to Manipulation?
Mohamed Mostagir, Asu Ozdaglar, and **James Siderius**



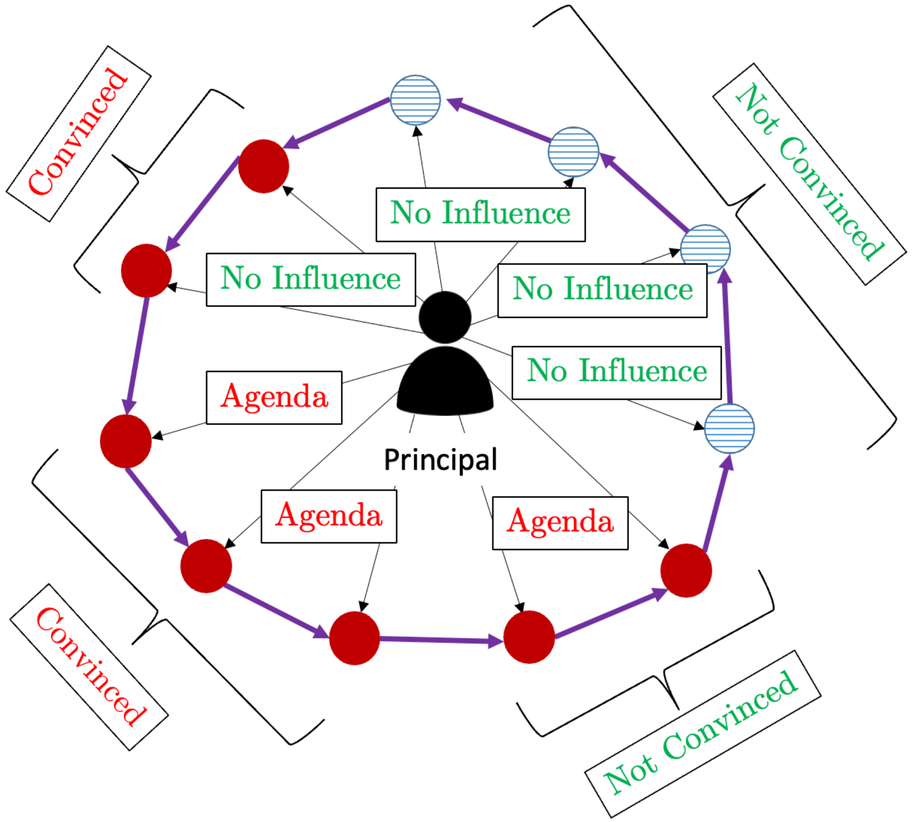
US Government

Disinformation/
Propaganda

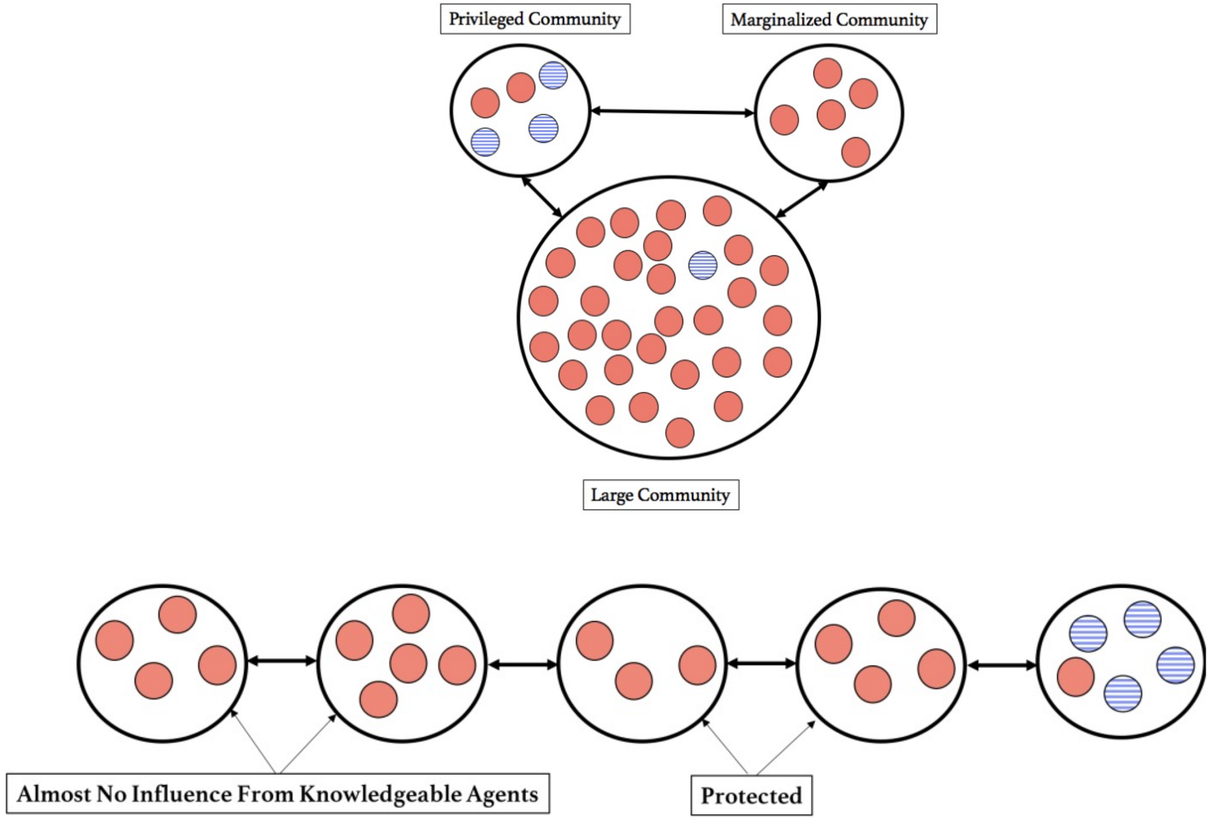
Social Media
Influencer

Human Behavior and Decisions in Networks

When is Society Susceptible to Manipulation?
Mohamed Mostagir, Asu Ozdaglar, and **James Siderius**

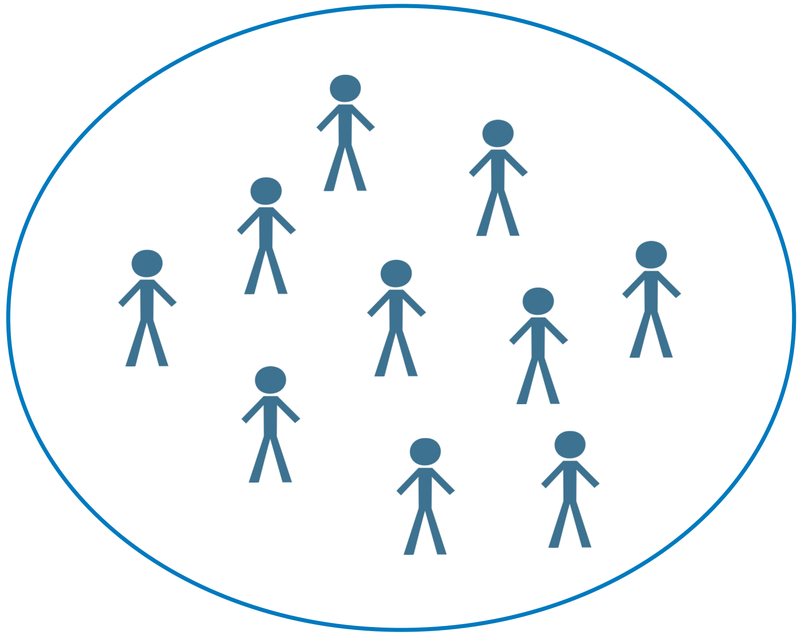


Social Inequality and the Spread of Misinformation
Mohamed Mostagir and **James Siderius**



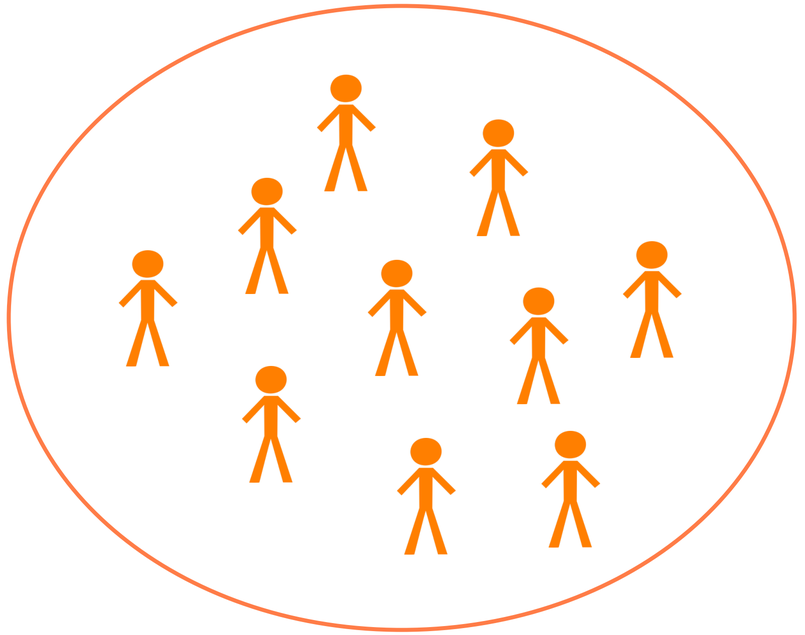
Human Behavior and Decisions in Networks

Learning in a Post-Truth World
Mohamed Mostagir and **James Siderius**



Bayesian Agents Perform Full Bayesian Inference

Naïve and Bayesian Learning under Misinformation Policies
Mohamed Mostagir and **James Siderius**



DeGroot Agents Perform Linear Updating with their Neighbors' Beliefs

Human Behavior and Decisions in Networks

Learning in a Post-Truth World
Mohamed Mostagir and **James Siderius**

Naïve and Bayesian Learning under Misinformation Policies
Mohamed Mostagir and **James Siderius**

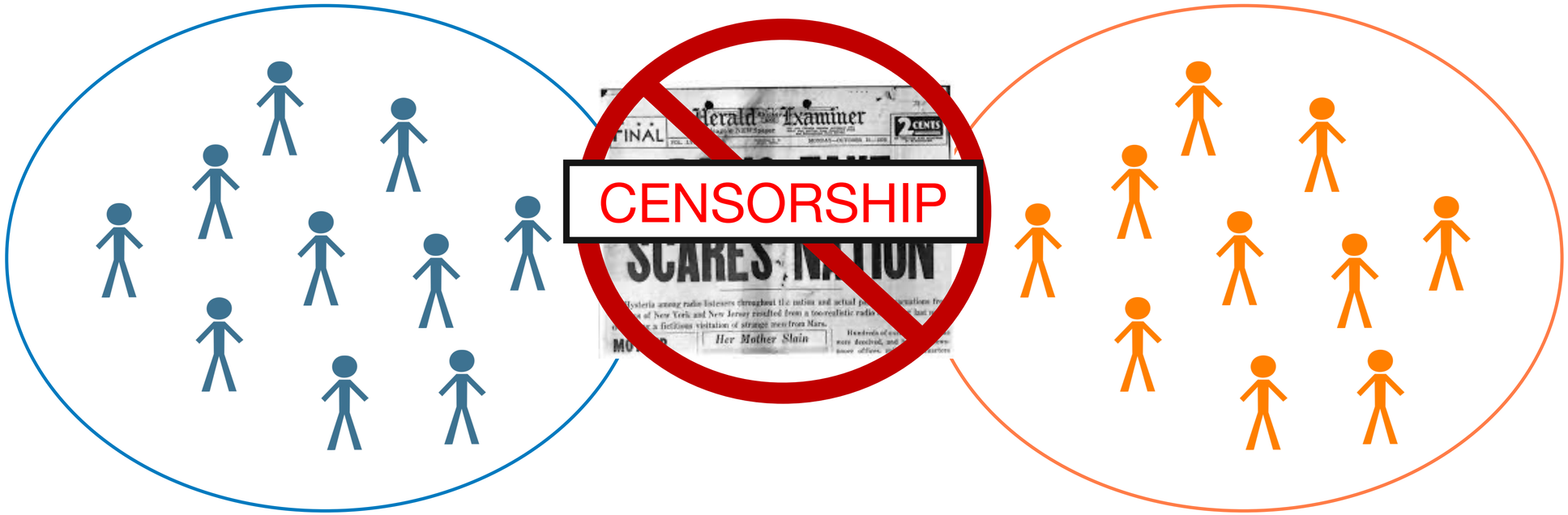
Bayesian Agents Perform Full Bayesian Inference

DeGroot Agents Perform Linear Updating with their Neighbors' Beliefs

Human Behavior and Decisions in Networks

Learning in a Post-Truth World
Mohamed Mostagir and **James Siderius**

Naïve and Bayesian Learning under Misinformation Policies
Mohamed Mostagir and **James Siderius**



Bayesian Agents Perform Full Bayesian Inference

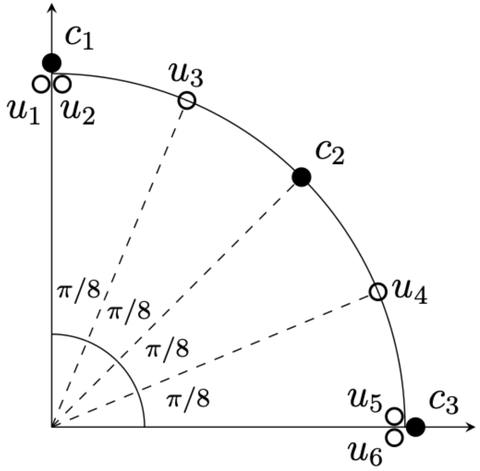
DeGroot Agents Perform Linear Updating with their Neighbors' Beliefs

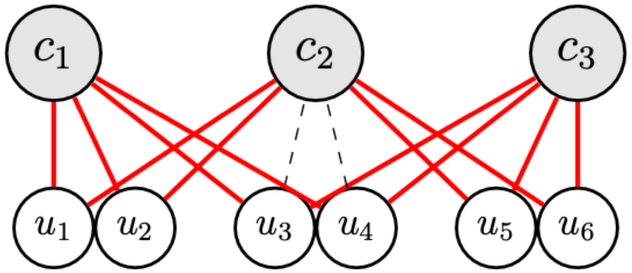
Media Platforms, Tech & Society

Two-Sided Media Matching: Hardness, Algorithms, and Social Impact
Daniel Huttenlocher, Hannah Li, Charles Lyu, Asu Ozdaglar, and **James Siderius**

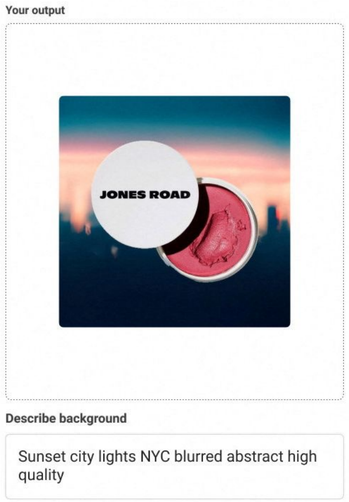
Welfare Implications of Online Media Business Models
Daron Acemoglu, Daniel Huttenlocher, Asu Ozdaglar, and **James Siderius**

When Should Platforms Break Echo Chambers?
Mohamed Mostagir and **James Siderius**

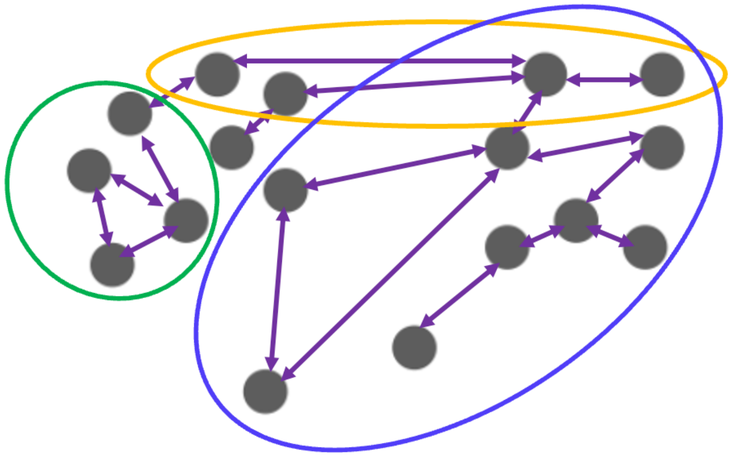


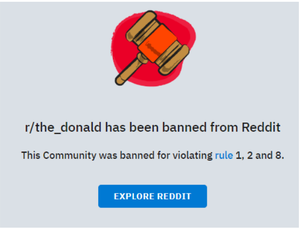


Elon Musk says Twitter, now X, could charge all users subscription fees



Structural Network Interventions





A Model of Online Misinformation

The Review of Economic Studies

Daron Acemoglu
MIT Economics

Asuman Ozdaglar
MIT EECS

James Siderius
Dartmouth / MIT

What We Do

- **Model** of diffusion of an article on a social media network

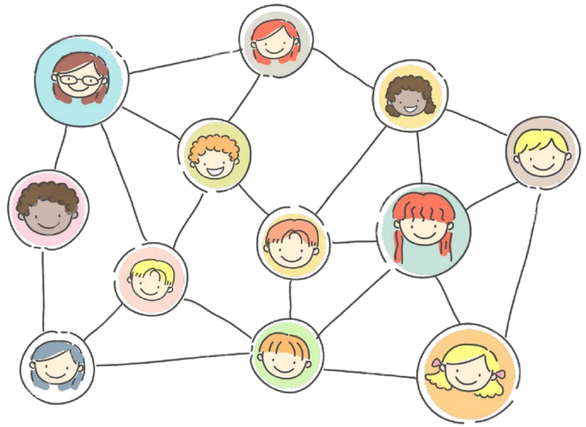
What We Do

- **Model** of diffusion of an article on a social media network
  - Game-theoretic model of user sharing decisions ("Bayesian framework")

What We Do

- **Model** of diffusion of an article on a social media network
  - Game-theoretic model of user sharing decisions ("Bayesian framework")
  - How does the social media **sharing network** affect total diffusion?

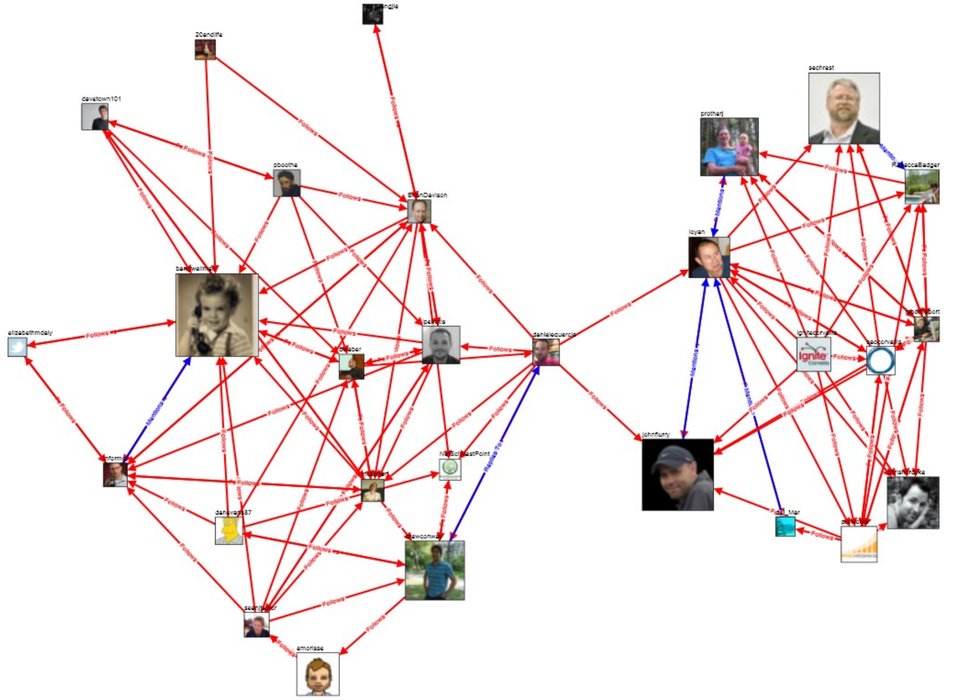
Undirected Network



What We Do

- **Model** of diffusion of an article on a social media network
  - Game-theoretic model of user sharing decisions ("Bayesian framework")
  - How does the social media **sharing network** affect total diffusion?

Directed Network

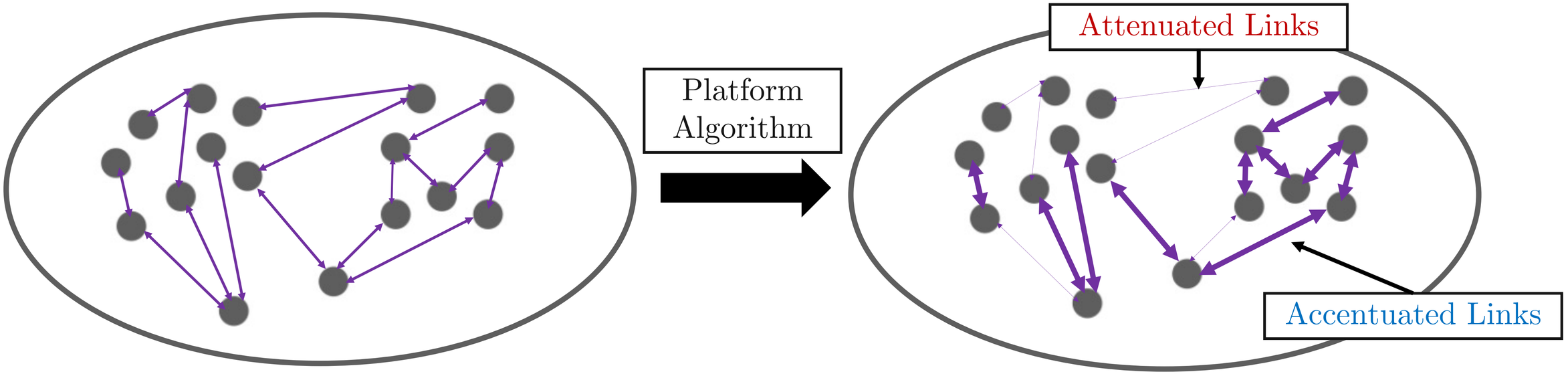


What We Do

- **Model** of diffusion of an article on a social media network
  - Game-theoretic model of user sharing decisions ("Bayesian framework")
  - How does the social media **sharing network** affect total diffusion?
- Platform incentives and **algorithms** that boost content

What We Do

- **Model** of diffusion of an article on a social media network
  - Game-theoretic model of user sharing decisions ("Bayesian framework")
  - How does the social media **sharing network** affect total diffusion?
- Platform incentives and **algorithms** that boost content
  - If the platform can "shape" the sharing network, how should it do so?



What We Do

- **Model** of diffusion of an article on a social media network
  - Game-theoretic model of user sharing decisions ("Bayesian framework")
  - How does the social media **sharing network** affect total diffusion?
- Platform incentives and **algorithms** that boost content
  - If the platform can "shape" the sharing network, how should it do so?
  - What are the **societal impacts** of these algorithms?

What We Do

- **Model** of diffusion of an article on a social media network
  - Game-theoretic model of user sharing decisions ("Bayesian framework")
  - How does the social media **sharing network** affect total diffusion?
- Platform incentives and **algorithms** that boost content
  - If the platform can "shape" the sharing network, how should it do so?
  - What are the **societal impacts** of these algorithms?
- Regulatory solutions

What We Do

- **Model** of diffusion of an article on a social media network
  - Game-theoretic model of user sharing decisions ("Bayesian framework")
  - How does the social media **sharing network** affect total diffusion?
- Platform incentives and **algorithms** that boost content
  - If the platform can "shape" the sharing network, how should it do so?
  - What are the **societal impacts** of these algorithms?
- Regulatory solutions
  - Effective design to mitigate the spread of harmful content

How do the social media sharing network and the attributes of the content impact its diffusion?

How do the social media sharing network and the attributes of the content impact its diffusion?

Platform's AI: Maximize total shares (proxy for user engagement) in equilibrium.

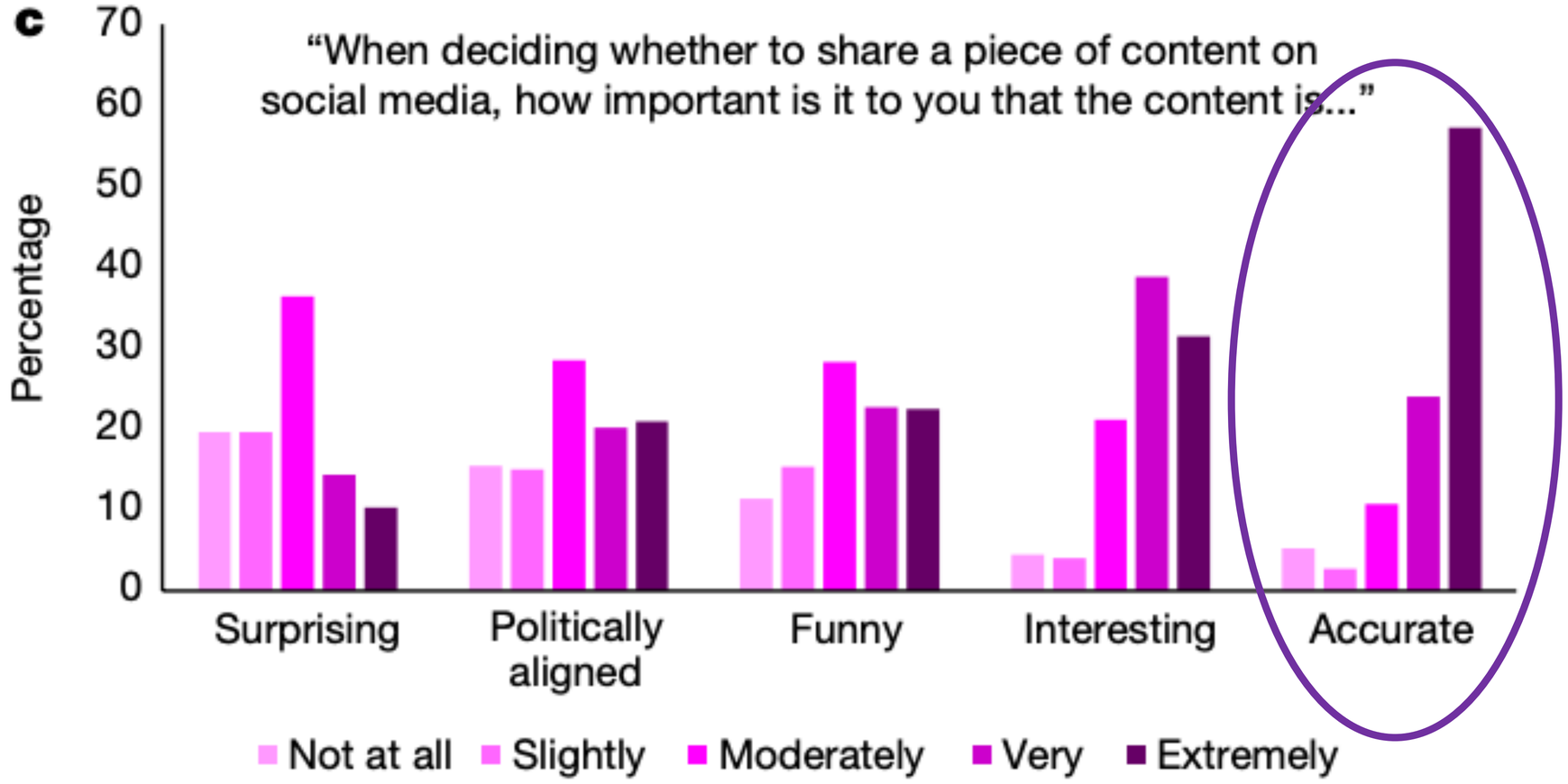
How do the social media sharing network and the attributes of the content impact its diffusion?

Platform's AI: Maximize total shares (proxy for user engagement) in equilibrium.

Societal Objective: Minimize divergence of beliefs from the truth (*ex ante* unknown).

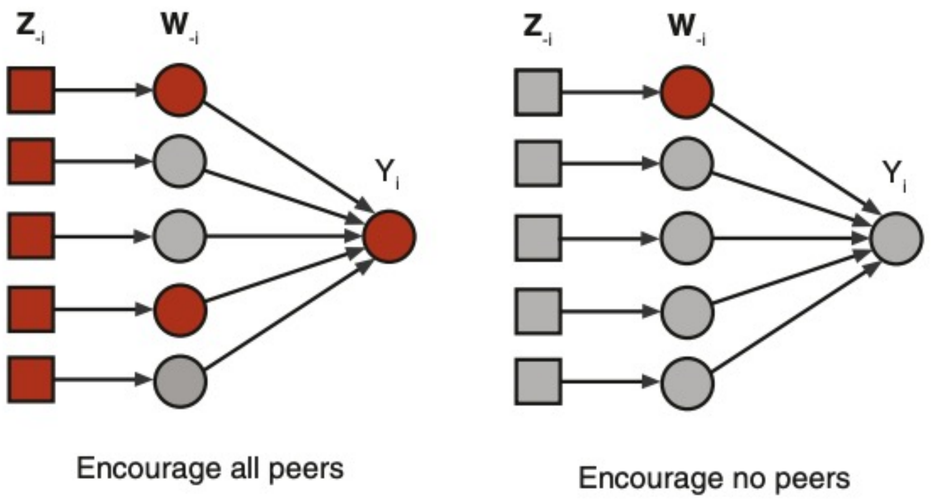
Empirical Facts: Why Users Share

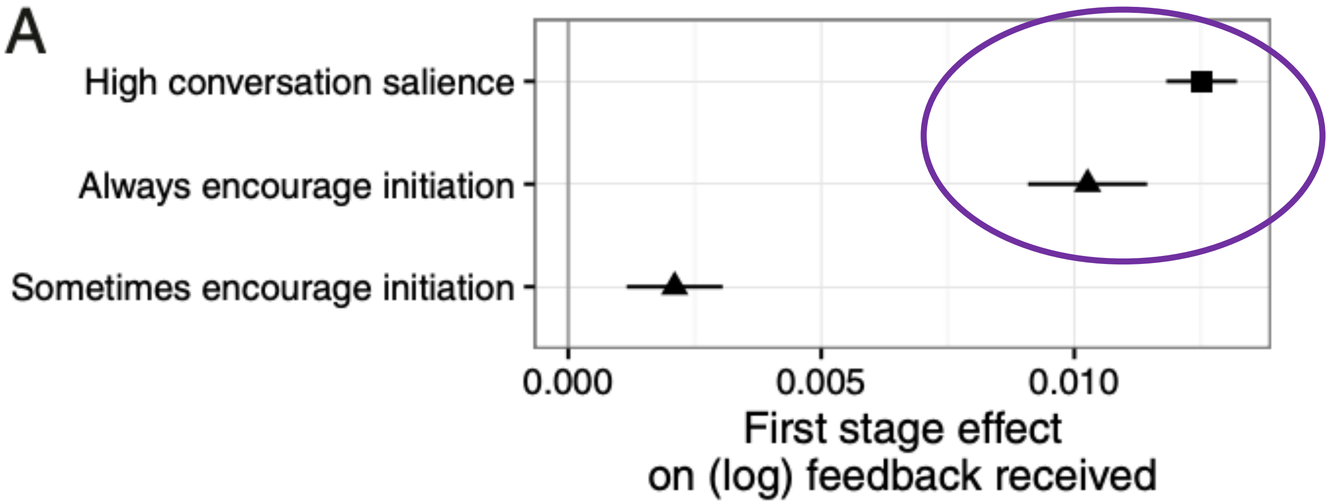
- Users want to share content they believe to be truthful and not contain misinformation (Pennycook et al (2021)).



Empirical Facts: Why Users Share

- Users want to share content they believe to be truthful and not contain misinformation (Pennycook et al (2021)).
- Users derive value from positive peer encouragement on social media (Eckles et al (2016); Duffy et al (2020)), aka network effects.





Empirical Facts: Why Users Share

- Users want to share content they believe to be truthful and not contain misinformation (Pennycook et al (2021)).
- Users derive value from positive peer encouragement on social media (Eckles et al (2016); Duffy et al (2020)), aka network effects.
- Users suffer reputational costs for getting called out for sharing misinformation (Altay et al (2020)).

Empirical Facts: Why Users Share

- Users want to share content they believe to be truthful and not contain misinformation (Pennycook et al (2021)).
- Users derive value from positive peer encouragement on social media (Eckles et al (2016); Duffy et al (2020)), aka network effects.
- Users suffer reputational costs for getting called out for sharing misinformation (Altay et al (2020)).
- Users often engage in criticisms of available content and inform others of misinformation they share on social media (Kim et al (2020) during 2018 midterm elections).

A Single Article

- Study the diffusion of **a single article** that exogenously arrives.

A Single Article

- Study the diffusion of **a single article** that exogenously arrives.
- Underlying state of the world $\theta \in \{L, R\}$ that is unknown.

A Single Article

- Study the diffusion of **a single article** that exogenously arrives.
- Underlying state of the world $\theta \in \{L, R\}$ that is unknown.
- Article has three properties

A Single Article

- Study the diffusion of **a single article** that exogenously arrives.
- Underlying state of the world $\theta \in \{L, R\}$ that is unknown.
- Article has three properties
  - Message: A binary message that either argues for L or R. [Observed]

A Single Article

- Study the diffusion of **a single article** that exogenously arrives.
- Underlying state of the world $\theta \in \{L, R\}$ that is unknown.
- Article has three properties
  - **Message**: A binary message that either argues for L or R. [Observed]
  - **Veracity**: Binary indicator of whether the article contains misinformation or not. [Unobserved]

A Single Article

- Study the diffusion of **a single article** that exogenously arrives.
- Underlying state of the world $\theta \in \{L, R\}$ that is unknown.
- Article has three properties
  - **Message**: A binary message that either argues for L or R. [Observed]
  - **Veracity**: Binary indicator of whether the article contains misinformation or not. [Unobserved]
  - **Reliability**: A score between 0 and 1 indicating the probability of containing misinformation unconditional on the message. [Observed]

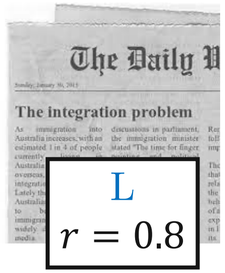
A Single Article

- Study the diffusion of **a single article** that exogenously arrives.
- Underlying state of the world $\theta \in \{L, R\}$ that is unknown.
- Article has three properties
  - **Message**: A binary message that either argues for L or R. [Observed]
  - **Veracity**: Binary indicator of whether the article contains misinformation or not. [Unobserved]
  - **Reliability**: A score between 0 and 1 indicating the probability of containing misinformation unconditional on the message. [Observed]
- Assumption: Truthful articles more often argue for θ; misinformation articles (weakly) more often argue for the opposite of θ.

Social Media Network

- Consists of users with heterogenous priors ("biases") b_i about $\theta = R$.

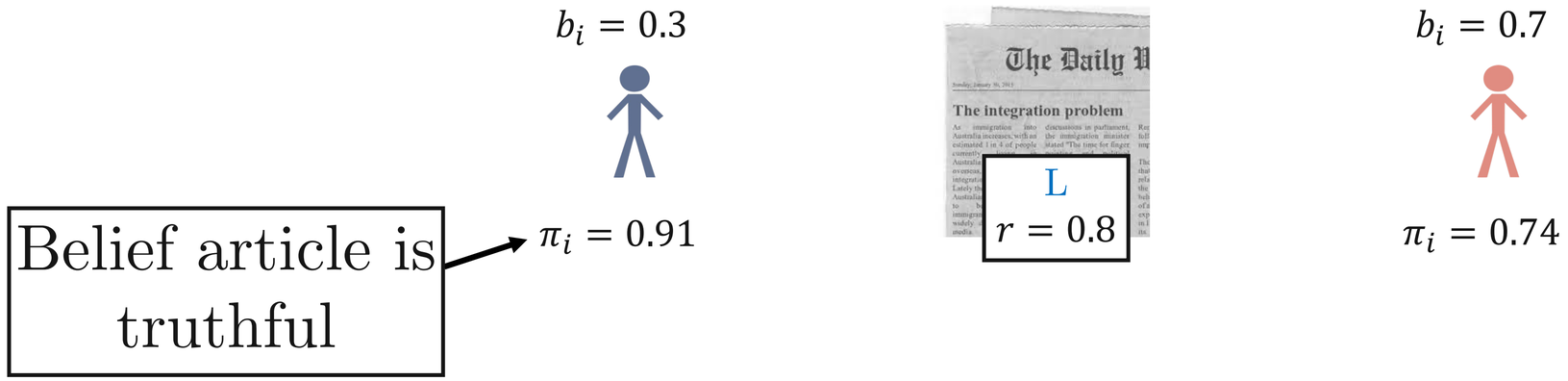
$b_i = 0.3$



$b_i = 0.7$

Social Media Network

- Consists of users with heterogenous priors ("biases") b_i about $\theta = R$.



- π_i can be computed straightforwardly by applying Bayes' rule:

$$\pi_i = \frac{\big(pb_i + (1-p)(1-b_i)\big)r}{\big(pb_i + (1-p)(1-b_i)\big)r + \big(qb_i + (1-q)(1-b_i)\big)(1-r)}$$

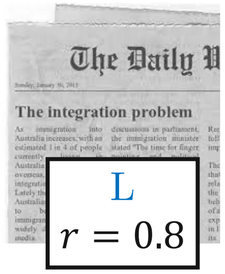
where $p = P(\theta|v = T) > 1/2$ and $q = P(\theta|v = M) \leq 1/2$.

Social Media Network

- Consists of users with heterogenous priors ("biases") b_i about $\theta = R$.

$b_i = 0.3$

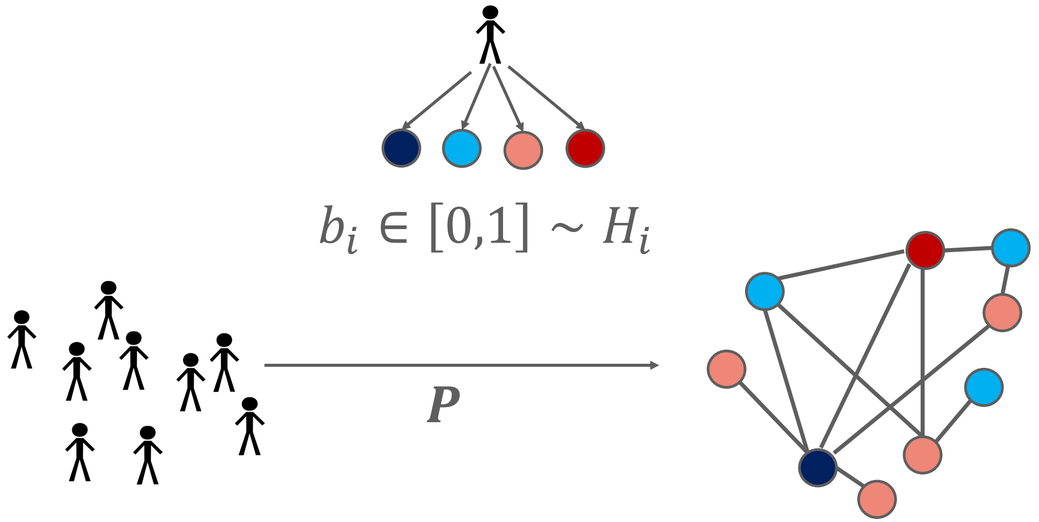
$\pi_i = 0.91$



$b_i = 0.7$

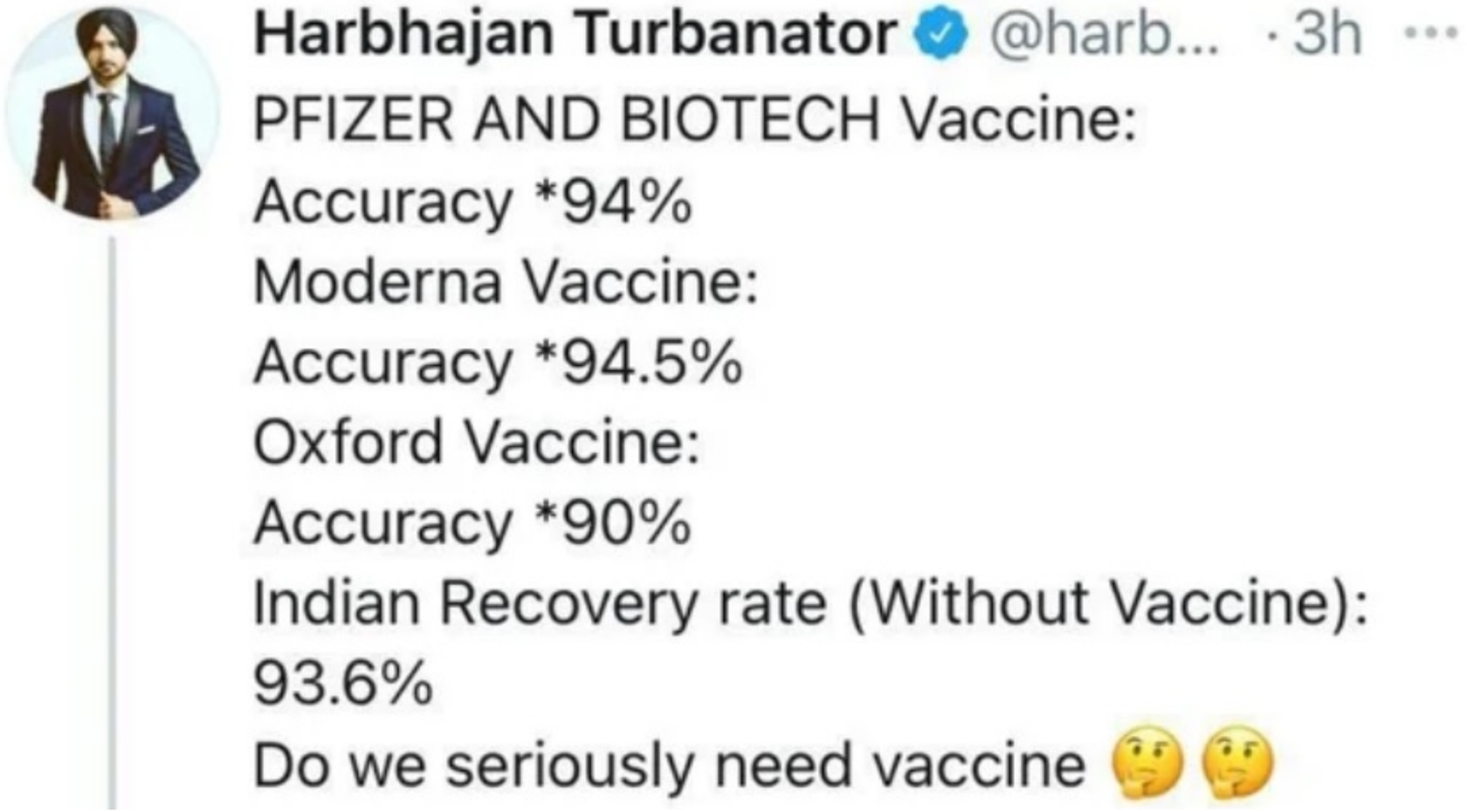
$\pi_i = 0.74$

- Agents arranged in a (stochastic) "sharing" network (link matrix P).



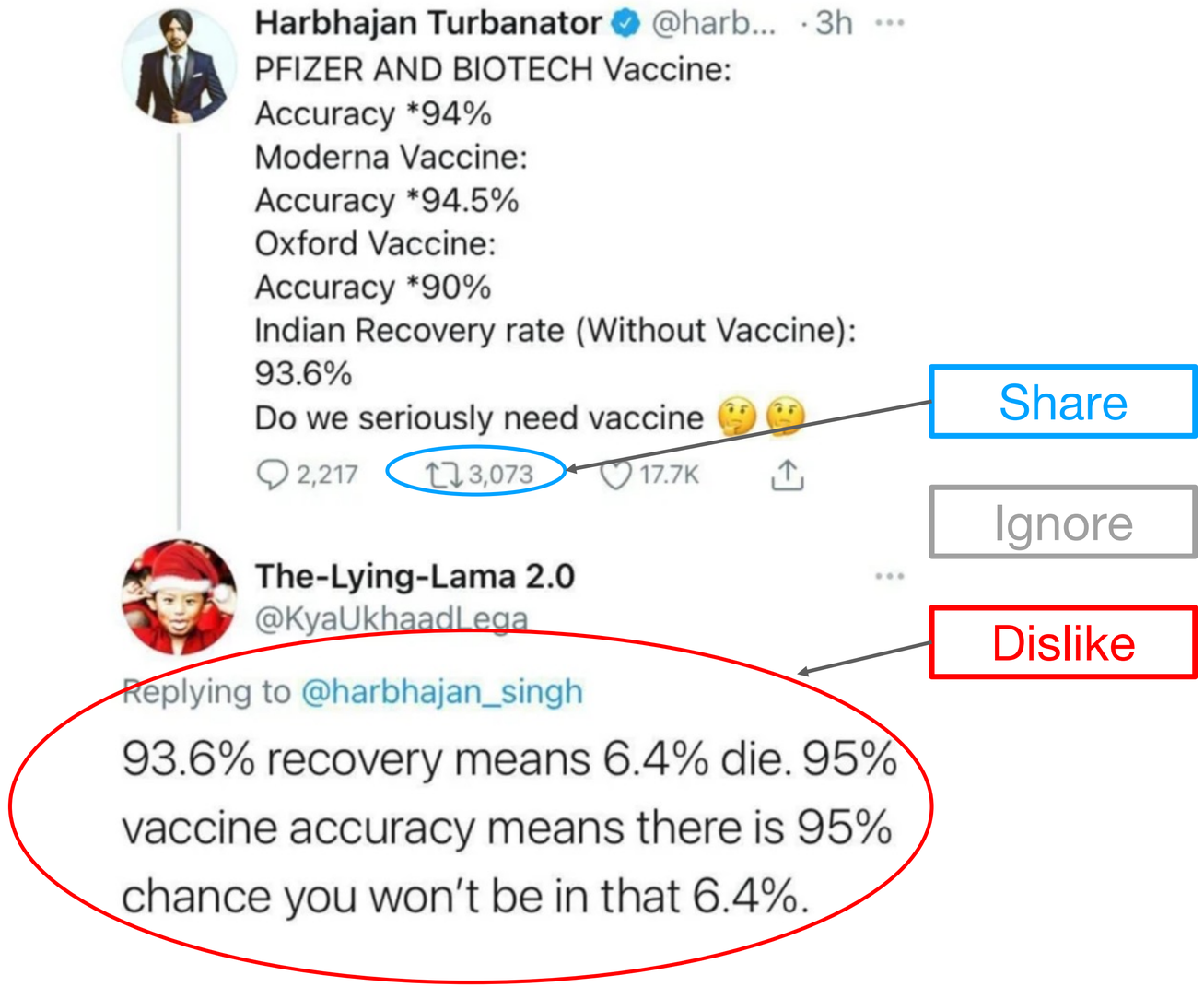
Strategic User Behavior

- Users can Share, Ignore, or Dislike (call out) an article.



Strategic User Behavior

- Users can Share, Ignore, or Dislike (call out) an article.



Strategic User Behavior

- Users can **Share**, **Ignore**, or **Dislike** (call out) an article.
- Payoff to **Ignore** action is normalized at 0.

Strategic User Behavior

- Users can **Share**, **Ignore**, or **Dislike** (call out) an article.
- Payoff to **Ignore** action is normalized at 0.
- Payoff to **Dislike** action depends only on π_i (and is decreasing in π_i).

$$U_i = \tilde{u}(1 - \pi_i) - \tilde{c}$$

Strategic User Behavior

- Users can **Share**, **Ignore**, or **Dislike** (call out) an article.
- Payoff to **Ignore** action is normalized at 0.
- Payoff to **Dislike** action depends only on π_i (and is decreasing in π_i).
- **Share** action causes article to *spread to all of agent i's neighbors*.

Strategic User Behavior

- Users can **Share**, **Ignore**, or **Dislike** (call out) an article.
- Payoff to **Ignore** action is normalized at 0.
- Payoff to **Dislike** action depends only on π_i (and is decreasing in π_i).
- **Share** action causes article to *spread to all of agent i's neighbors*.
- Payoff to **Share** action has two components:

Strategic User Behavior

- Users can **Share**, **Ignore**, or **Dislike** (call out) an article.
- Payoff to **Ignore** action is normalized at 0.
- Payoff to **Dislike** action depends only on π_i (and is decreasing in π_i).
- **Share** action causes article to *spread to all of agent i's neighbors*.
- Payoff to **Share** action has two components:
  - Network-independent component that is increasing in π_i.

$$U_i^{(1)} = u\pi_i - c(1 - \pi_i)$$

Strategic User Behavior

- Users can **Share**, **Ignore**, or **Dislike** (call out) an article.
- Payoff to **Ignore** action is normalized at 0.
- Payoff to **Dislike** action depends only on π_i (and is decreasing in π_i).
- **Share** action causes article to *spread to all of agent i's neighbors*.
- Payoff to **Share** action has two components:
  - Network-independent component that is increasing in π_i.
  - Network-dependent component that is increasing in S_i (number of re-shares from peers) and decreasing in D_i (number of dislikes from peers).

$$U_i^{(2)} = \kappa S_i - dD_i$$

Can be extended to more general supermodular functional forms

Strategic User Behavior

- Users can **Share**, **Ignore**, or **Dislike** (call out) an article.
- Payoff to **Ignore** action is normalized at 0.
- Payoff to **Dislike** action depends only on π_i (and is decreasing in π_i).
- **Share** action causes article to *spread to all of agent i's neighbors*.
- Payoff to **Share** action has two components:
  - Network-independent component that is increasing in π_i.
  - Network-dependent component that is increasing in S_i (number of re-shares from peers) and decreasing in D_i (number of dislikes from peers).

$$U_i = U_i^{(1)} + U_i^{(2)}$$

Diffusion Process

- Article diffuses depending on **Share** actions taken in equilibrium.

Diffusion Process

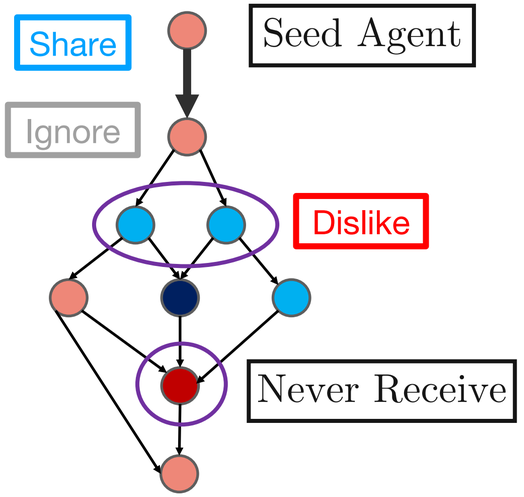
- Article diffuses depending on **Share** actions taken in equilibrium.
  - Markovian process: Make the same decision regardless of the history of the article's spread. (Solution concept: Bayes-Nash equilibrium.)

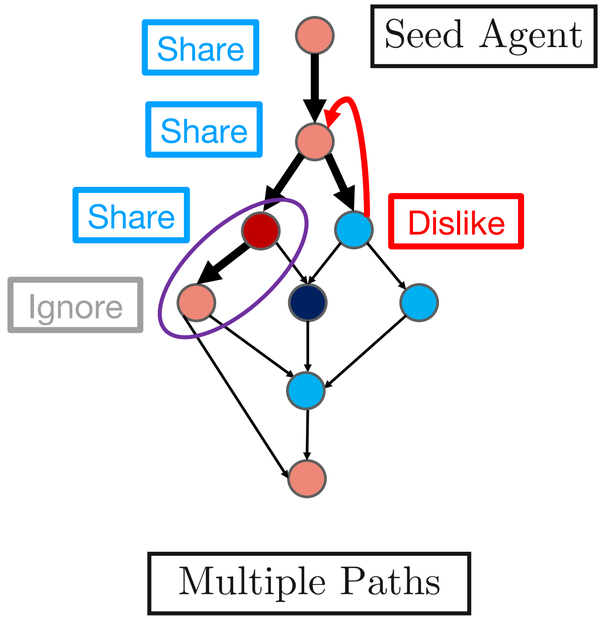
Diffusion Process

- Article diffuses depending on **Share** actions taken in equilibrium.
  - Markovian process: Make the same decision regardless of the history of the article's spread. (Solution concept: Bayes-Nash equilibrium.)
  - Interdependent process: Cannot **Share** if one never receives the article.

Diffusion Process

- Article diffuses depending on **Share** actions taken in equilibrium.
  - Markovian process: Make the same decision regardless of the history of the article's spread. (Solution concept: Bayes-Nash equilibrium.)
  - Interdependent process: Cannot **Share** if one never receives the article.





Characterizing the Equilibria

- To understand diffusion must first understand the **equilibrium strategies** of the agents in the network.

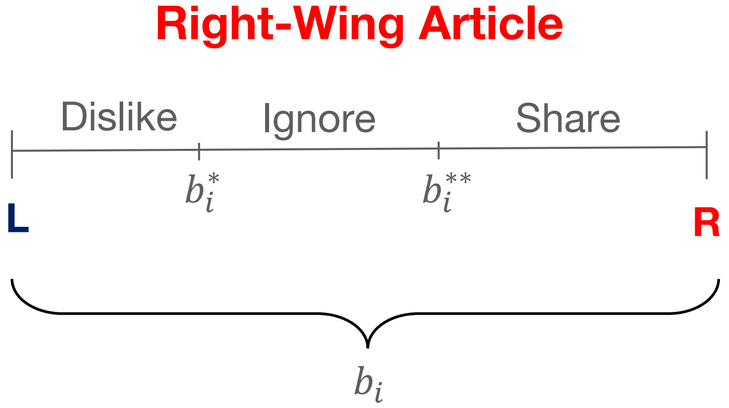
Characterizing the Equilibria

- To understand diffusion must first understand the **equilibrium strategies** of the agents in the network.

- Fix message $m = R$ without loss of generality.

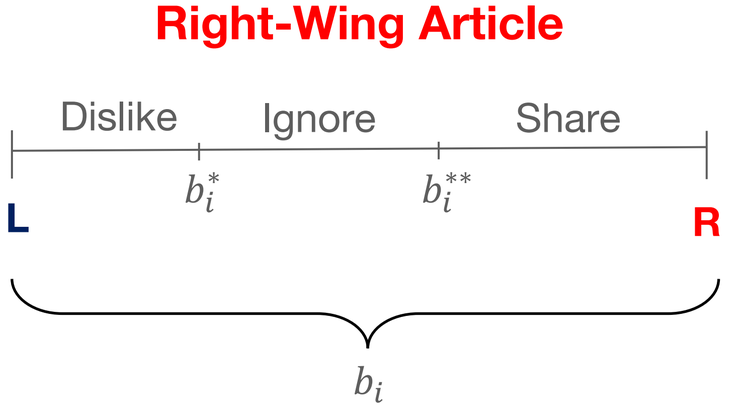
Characterizing the Equilibria

- To understand diffusion must first understand the **equilibrium strategies** of the agents in the network.
- Fix message $m = R$ without loss of generality.
- Definition: A cutoff strategy is one where for every agent i, there exist cutoffs $0 \leq b_i^* \leq b_i^{**} \leq 1$ such that
  - If $b_i < b_i^*$, the agent plays **Dislike**;
  - If $b_i^* < b_i < b_i^{**}$, the agent plays **Ignore**;
  - If $b_i > b_i^{**}$, the agent plays **Share**.



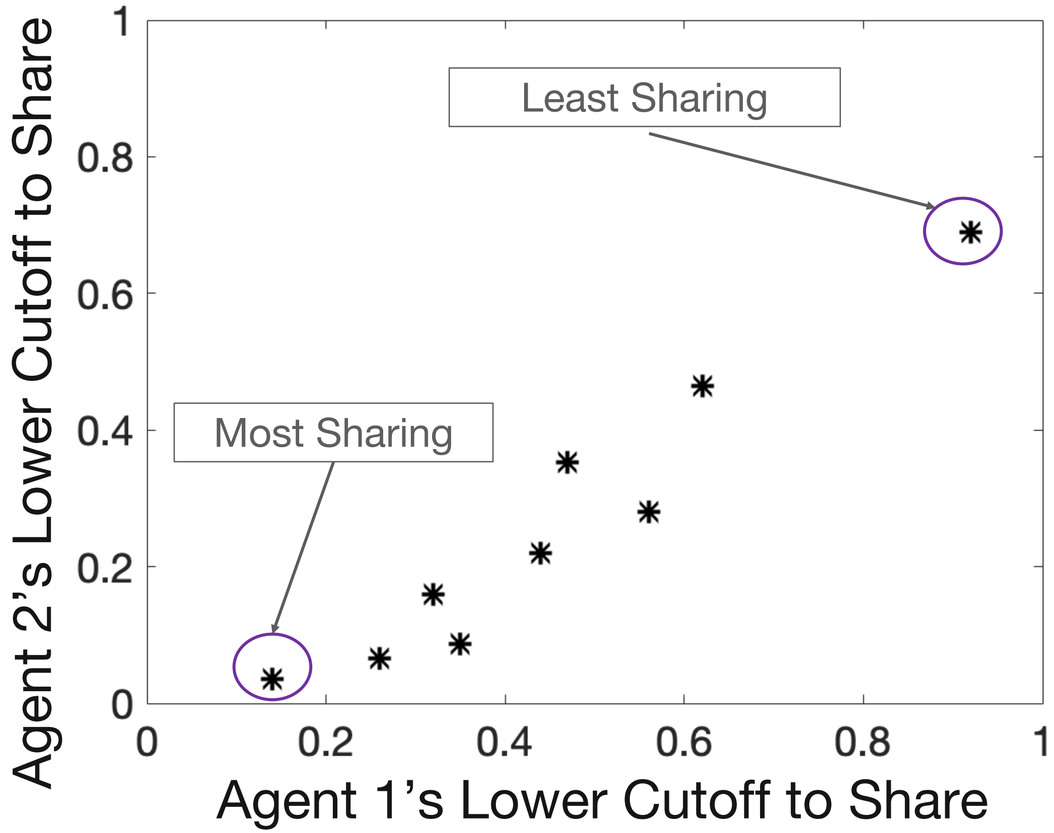
Characterizing the Equilibria

- To understand diffusion must first understand the **equilibrium strategies** of the agents in the network.
- Fix message $m = R$ without loss of generality.
- Definition: A cutoff strategy is one where for every agent i, there exist cutoffs $0 \leq b_i^* \leq b_i^{**} \leq 1$ such that
  - If $b_i < b_i^*$, the agent plays **Dislike**;
  - If $b_i^* < b_i < b_i^{**}$, the agent plays **Ignore**;
  - If $b_i > b_i^{**}$, the agent plays **Share**.



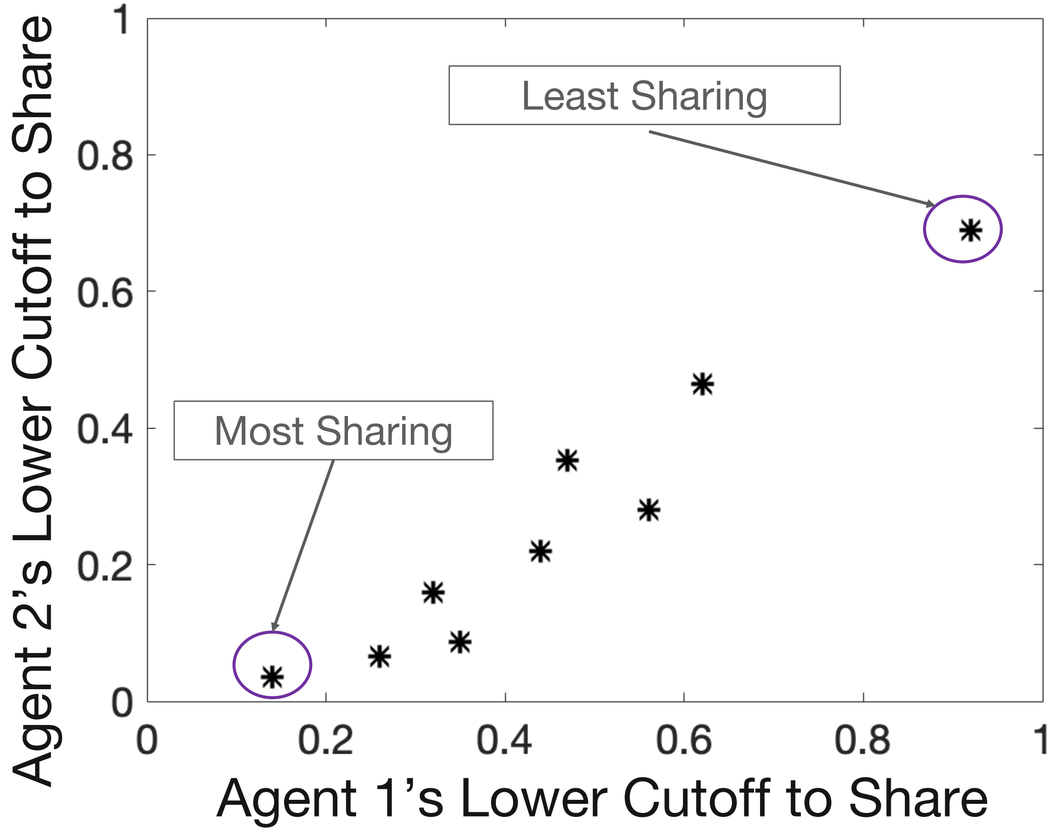
- **Theorem 1**: All equilibria are in cutoff strategies, there exists at least one equilibrium, and there is a most-sharing and a least-sharing equilibrium.

Lattice Structure of Equilibria



- **Theorem 1**: All equilibria are in cutoff strategies, there exists at least one equilibrium, and there is a most-sharing and a least-sharing equilibrium.

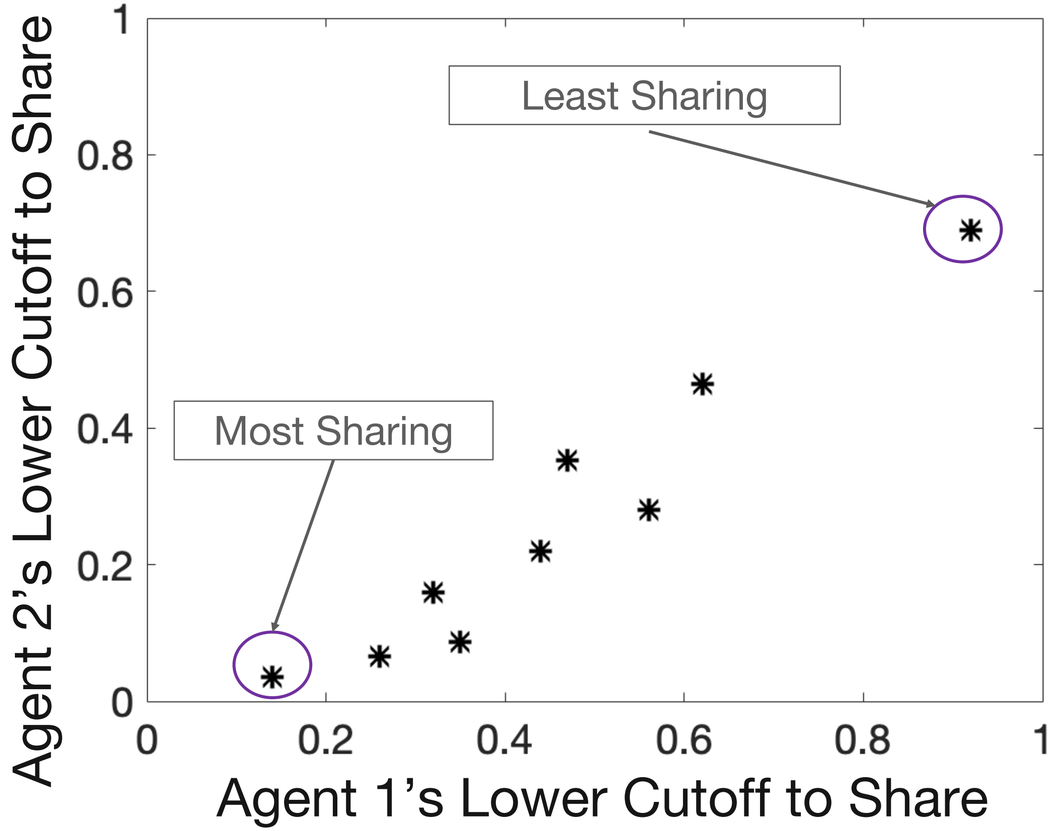
Lattice Structure of Equilibria



- Supermodular game.
- Strategic complementarity in sharing actions.

- **Theorem 1**: All equilibria are in cutoff strategies, there exists at least one equilibrium, and there is a most-sharing and a least-sharing equilibrium.

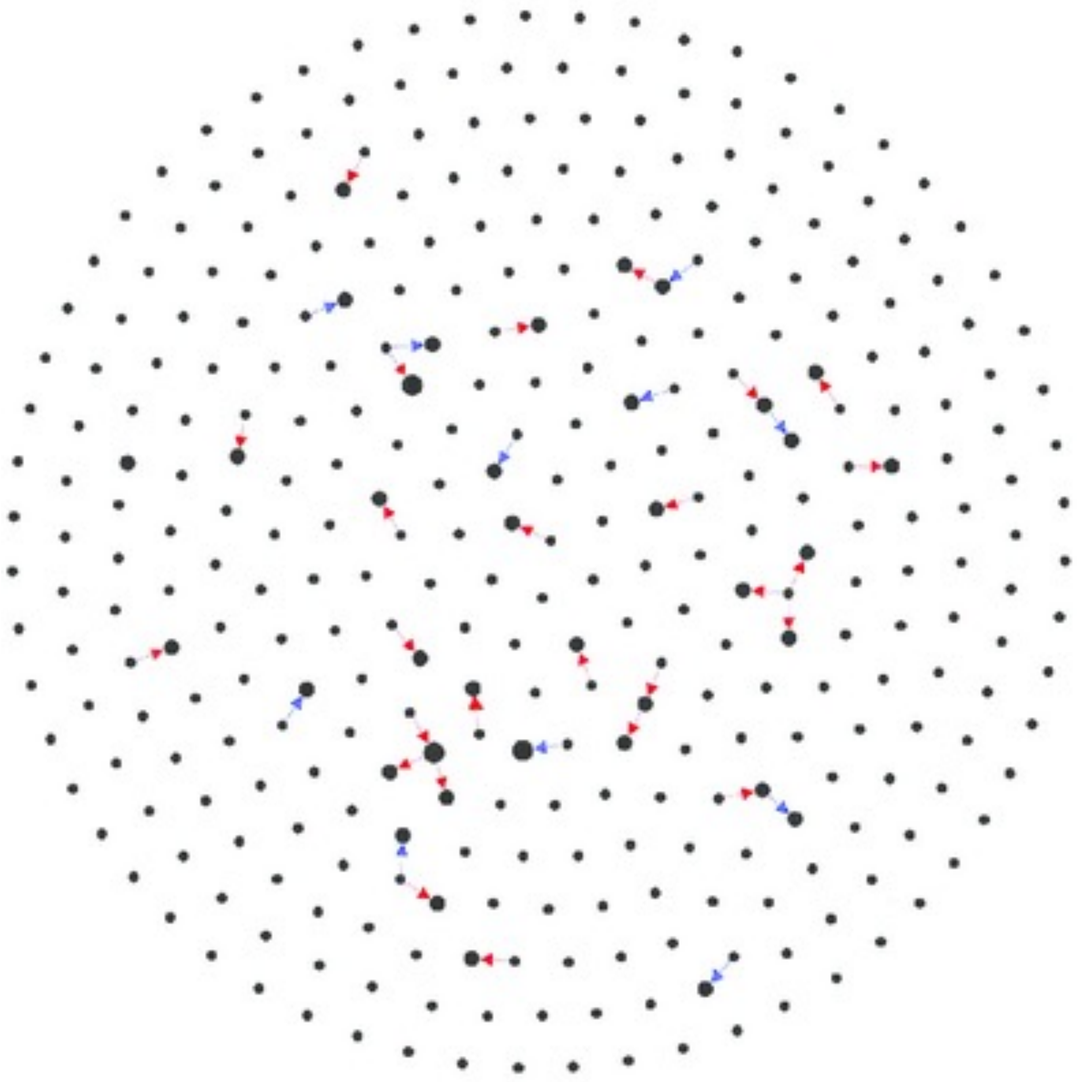
Lattice Structure of Equilibria



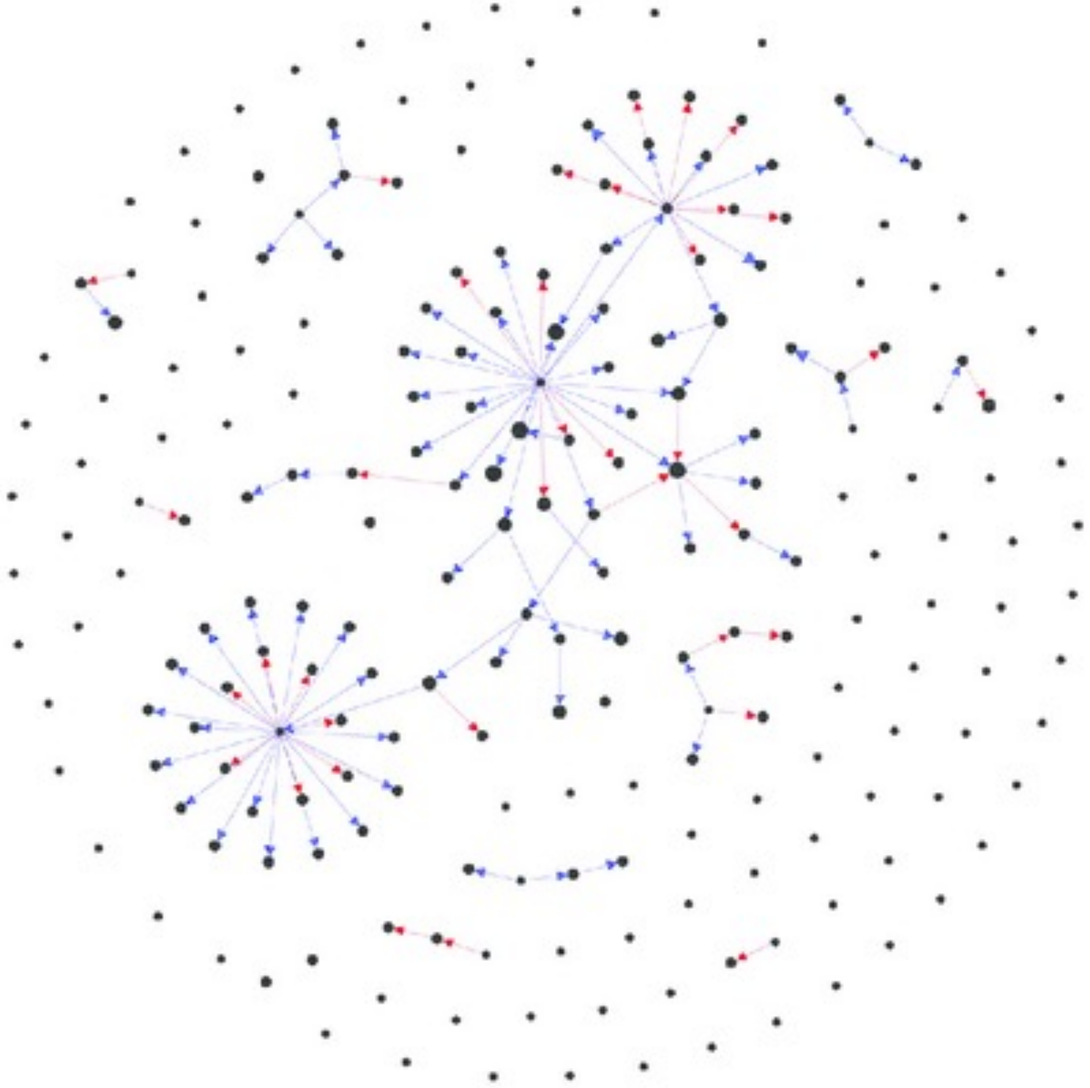
- Supermodular game.
- Strategic complementarity in sharing actions.
- Concentrate on **most sharing**.
  - Well-behaved comparative statics for extremal equilibria.
  - Most concerning for the spread of misinformation.

- Theorem 1: All equilibria are in cutoff strategies, there exists at least one equilibrium, and there is a most-sharing and a least-sharing equilibrium.

Diffusion Process ➔ User Engagement



Low Diffusion / Engagement



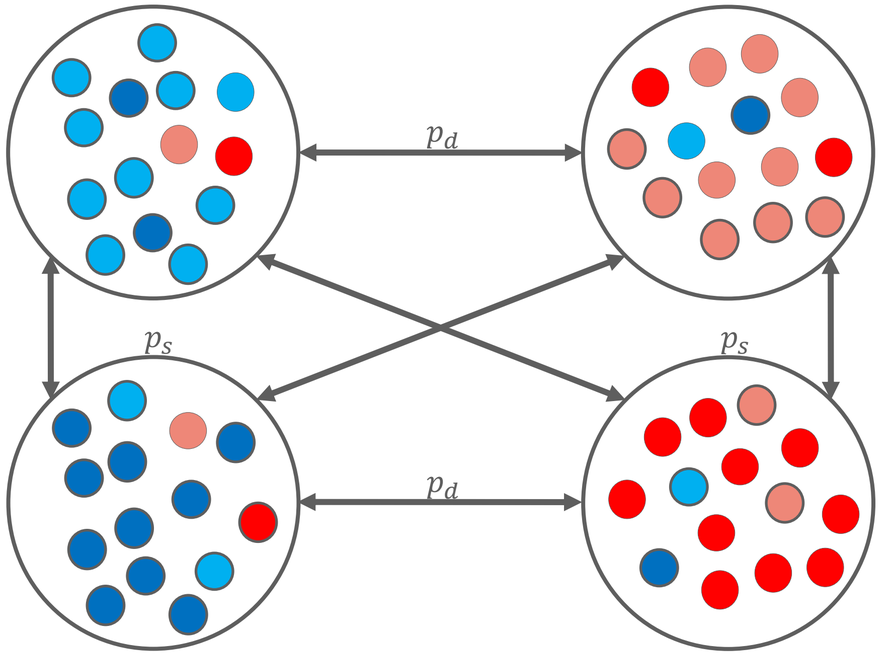
High Diffusion / Engagement

User Engagement in Island Networks

- Focus on the special case of **island network** topologies (with k islands).
  - Users are more likely to be connected with those of similar beliefs.

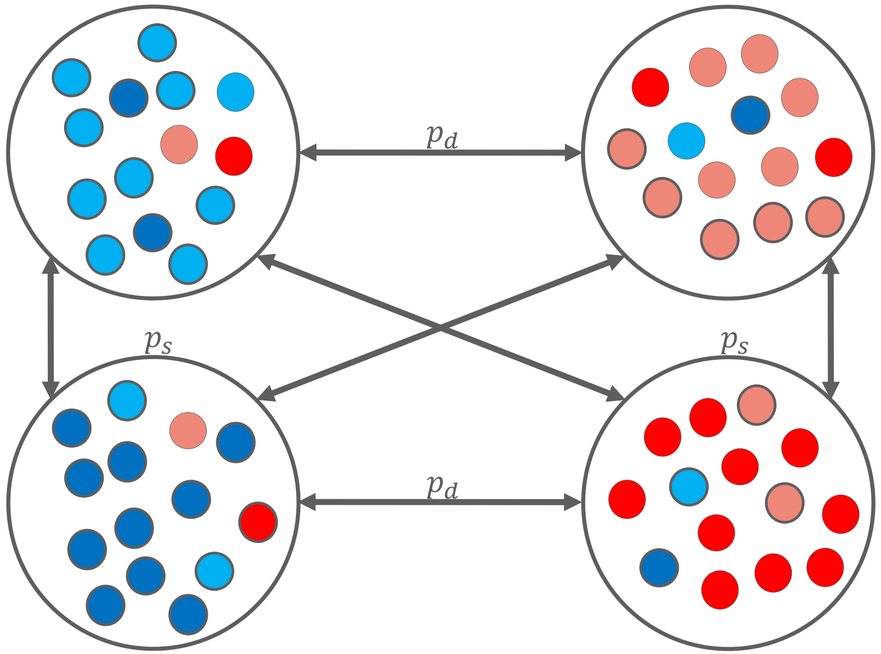
User Engagement in Island Networks

- Focus on the special case of **island network** topologies (with k islands).
  - Users are more likely to be connected with those of similar beliefs.



User Engagement in Island Networks

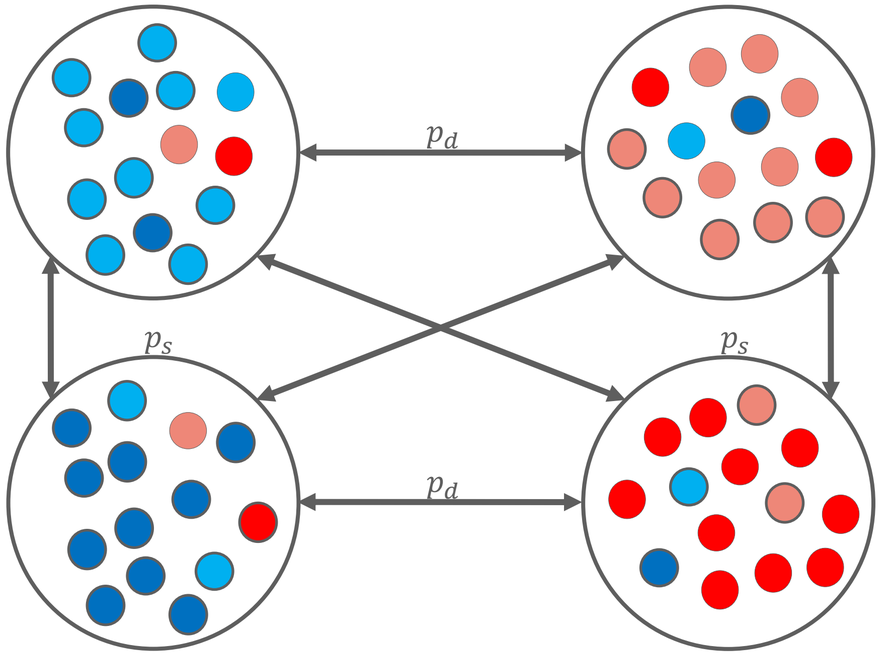
- Focus on the special case of **island network** topologies (with k islands).
  - Users are more likely to be connected with those of similar beliefs.



- Belief distributions satisfy $H_1 \succcurlyeq H_2 \succcurlyeq \cdots \succcurlyeq H_k$ in the FOSD sense.

User Engagement in Island Networks

- Focus on the special case of **island network** topologies (with k islands).
  - Users are more likely to be connected with those of similar beliefs.



- Belief distributions satisfy $H_1 \succcurlyeq H_2 \succcurlyeq \cdots \succcurlyeq H_k$ in the FOSD sense.
- The degree of "homophily" is measured by p_s and p_d.

How does homophily affect the diffusion of content likely to contain misinformation?

Discipline Effect: Low Reliability

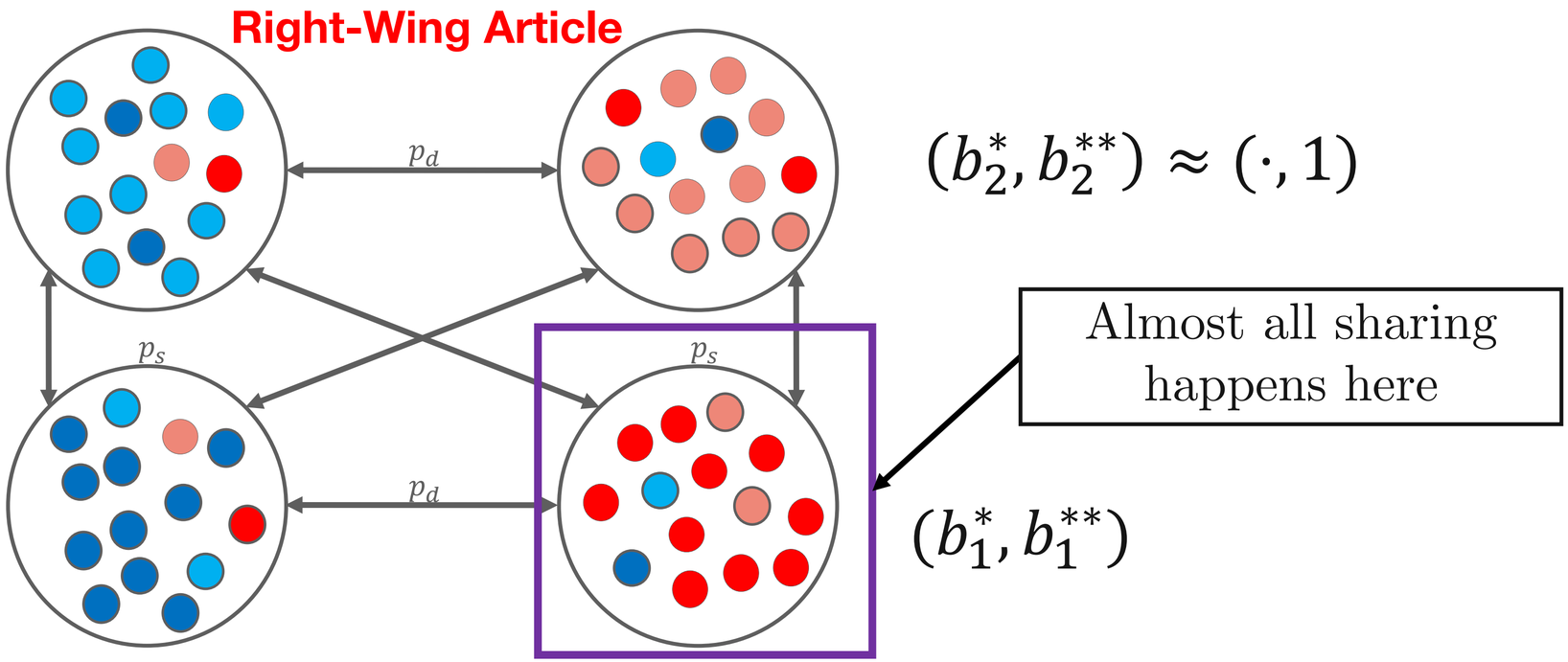
- When r is small, share payoff (from truth) is low, can bound the cutoffs on less extreme islands.

Discipline Effect: Low Reliability

- When r is small, share payoff (from truth) is low, can bound the cutoffs on less extreme islands.

$(b_3^*, b_3^{**}) \approx (\cdot, 1)$

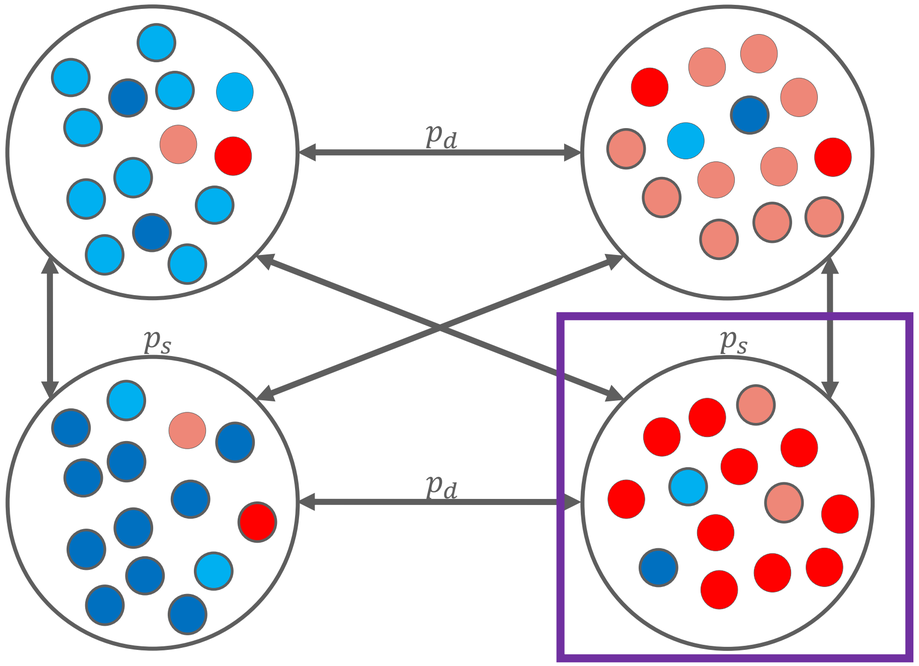
$(b_4^*, b_4^{**}) \approx (\cdot, 1)$



Discipline Effect: Low Reliability

- When r is small, share payoff (from truth) is low, can bound the cutoffs on less extreme islands.

- Topkis's theorem (Monotone Comparative Statics): Equilibrium cutoffs decrease on island 1 with an increase in homophily.



More sharing

$$(b_1^*, b_1^{**}) \geqslant (b_1^{*\prime}, b_1^{**\prime})$$

Discipline Effect: Low Reliability

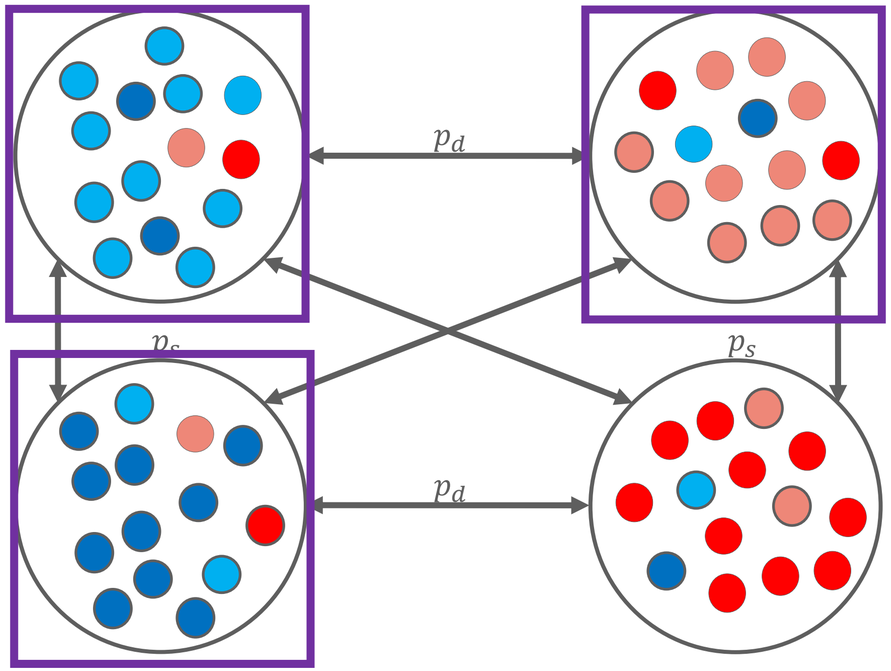
- When r is small, share payoff (from truth) is low, can bound the cutoffs on less extreme islands.
- Topkis's theorem (**Monotone Comparative Statics**): Equilibrium cutoffs increase on other islands with a decrease in homophily.

Less sharing

$(b_3^*, b_3^{**}) \preccurlyeq (b_3^{*\prime}, b_3^{**\prime})$

Less sharing

$(b_4^*, b_4^{**}) \preccurlyeq (b_4^{*\prime}, b_4^{**\prime})$



Less sharing

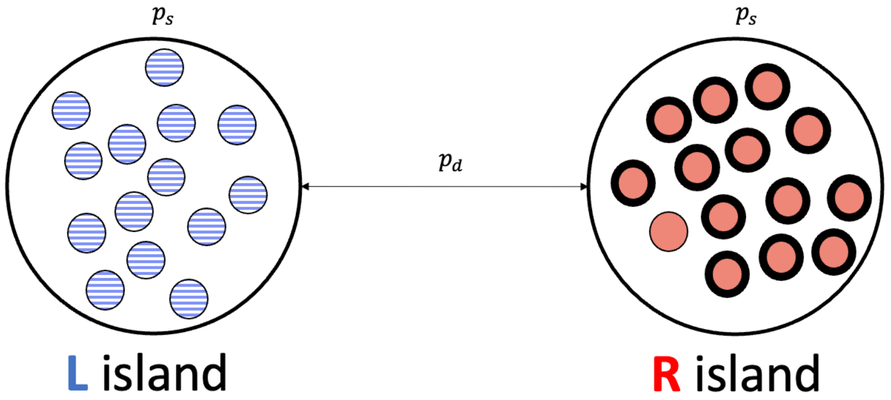
$(b_2^*, b_2^{**}) \preccurlyeq (b_2^{*\prime}, b_2^{**\prime})$

More sharing

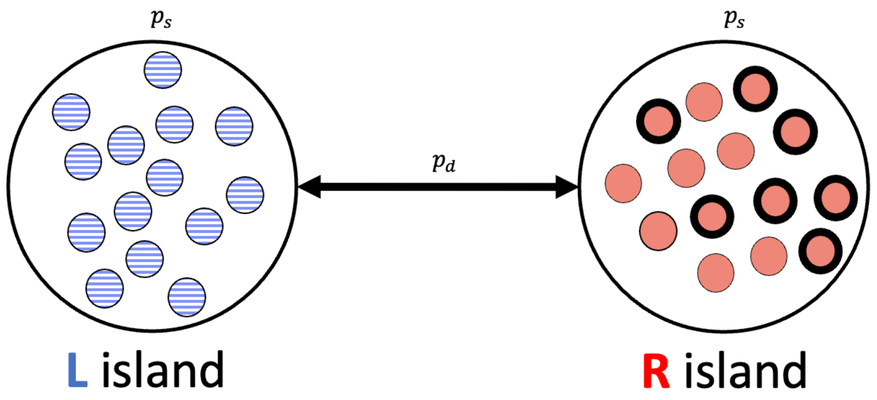
$(b_1^*, b_1^{**}) \succcurlyeq (b_1^{*\prime}, b_1^{**\prime})$

The Discipline Effect

- Consider just two islands for simplicity.



High Homophily
(Low p_d / High p_s)

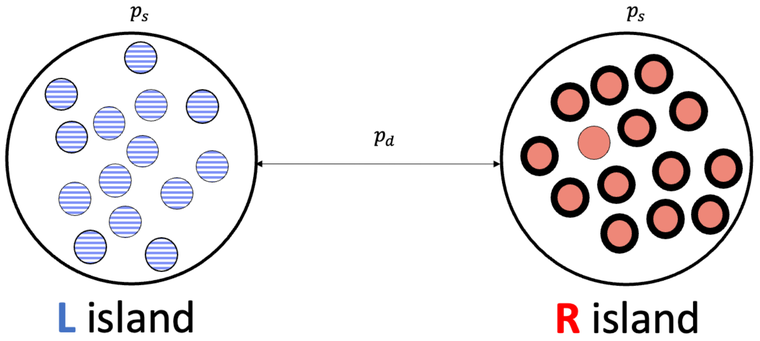


Low Homophily
(High p_d / Low p_s)

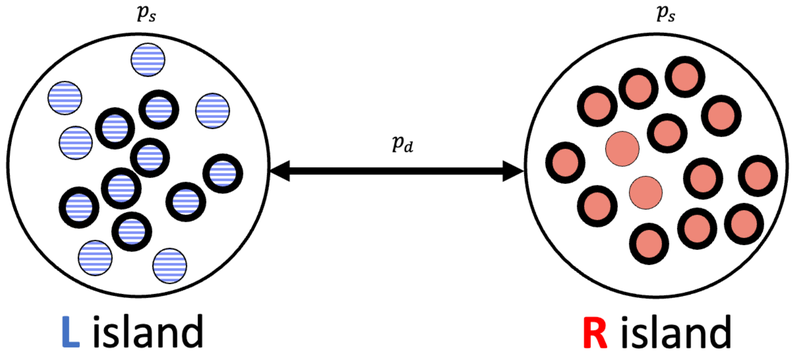
- Discipline drops (and sharing increases) when homophily increases.
  - Neighbors look more like you and will have similar assessments of truth.
  - Less cautious about how the article you share might perceived.

The Circulation Effect

- Once again, consider just two islands for simplicity.



High Homophily
(Low p_d / High p_s)



Low Homophily
(High p_d / Low p_s)

- Circulation increases (and sharing increases) when homophily decreases.
  - Few connections to outside groups – article is less likely to break out.
  - Diffusion process may be confined to small subset of users.

Impact of Misinformation

- Theorem 2: There exist $0 < r_1 < r_2 < 1$ such that:

Impact of Misinformation

- Theorem 2: There exist $0 < r_1 < r_2 < 1$ such that:
  - If $r < r_1$, diffusion increases when homophily increases;

Discipline Effect > Circulation Effect

Impact of Misinformation

- **Theorem 2**: There exist $0 < r_1 < r_2 < 1$ such that:
  - If $r < r_1$, diffusion increases when homophily increases;

Discipline Effect $>$ Circulation Effect

  - If $r > r_2$, diffusion increases when homophily decreases.

Discipline Effect $<$ Circulation Effect

Impact of Misinformation

- **Theorem 2**: There exist $0 < r_1 < r_2 < 1$ such that:
  - If $r < r_1$, diffusion increases when homophily increases;

Discipline Effect $>$ Circulation Effect

  - If $r > r_2$, diffusion increases when homophily decreases.

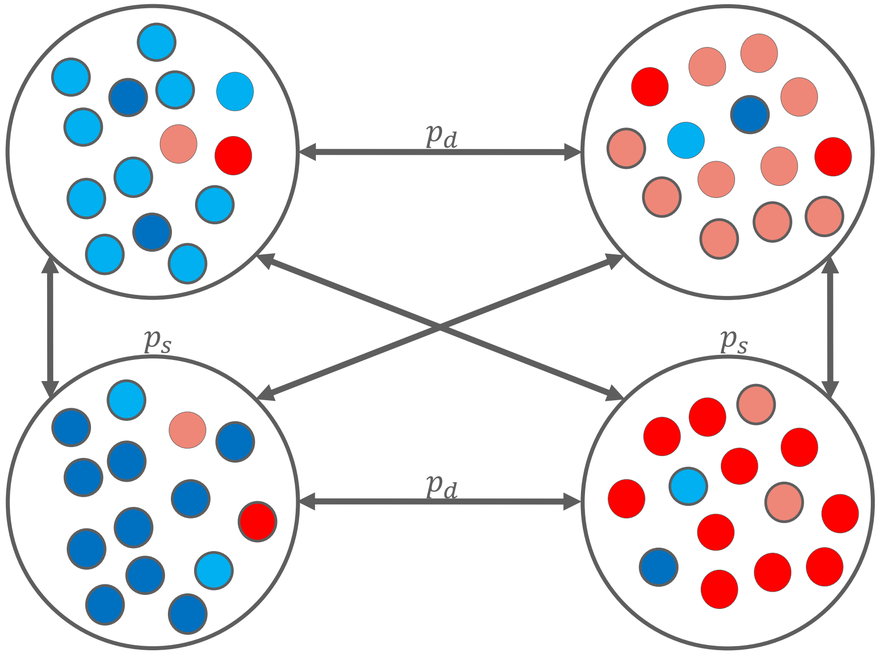
Discipline Effect $<$ Circulation Effect

  - **Higher homophily** in the network increases the spread of the article when it is **likely to contain misinformation**.

How should the platform shape the sharing network to maximize user engagement?

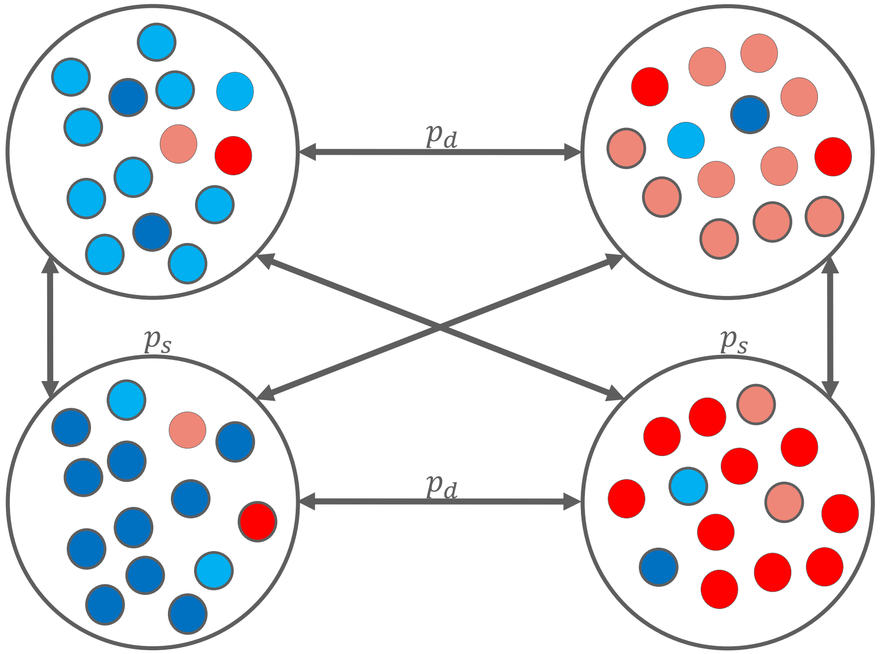
Platform's Problem

- Initially start from some underlying social network with many islands.



Platform's Problem

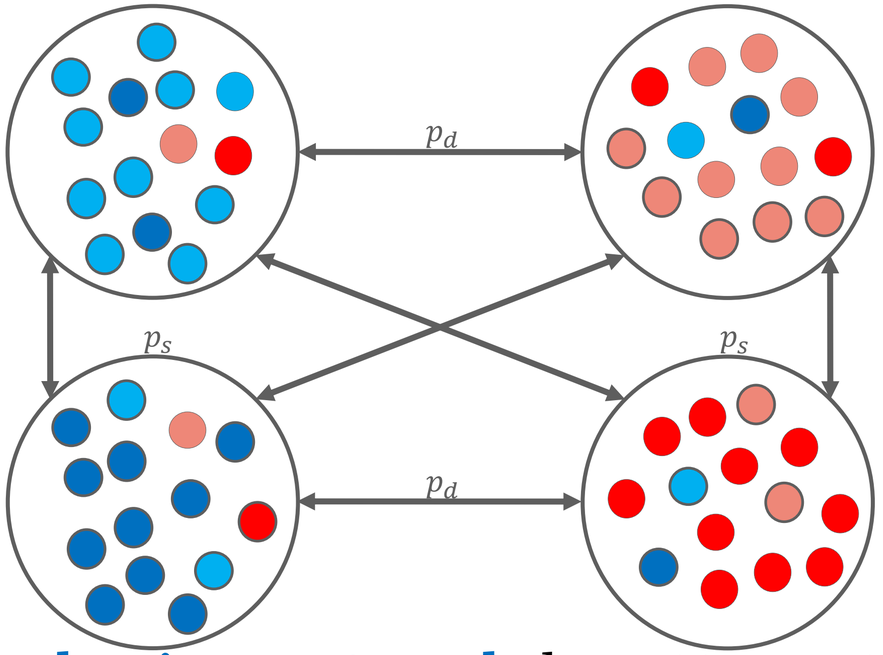
- Initially start from some underlying social network with many islands.



- Platform shapes the sharing network by attenuating or accentuating links in the network (e.g., through boosting or targeted recommendations).

Platform's Problem

- Initially start from some underlying social network with many islands.



- Platform shapes the sharing network by attenuating or accentuating links in the network (e.g., through boosting or targeted recommendations).

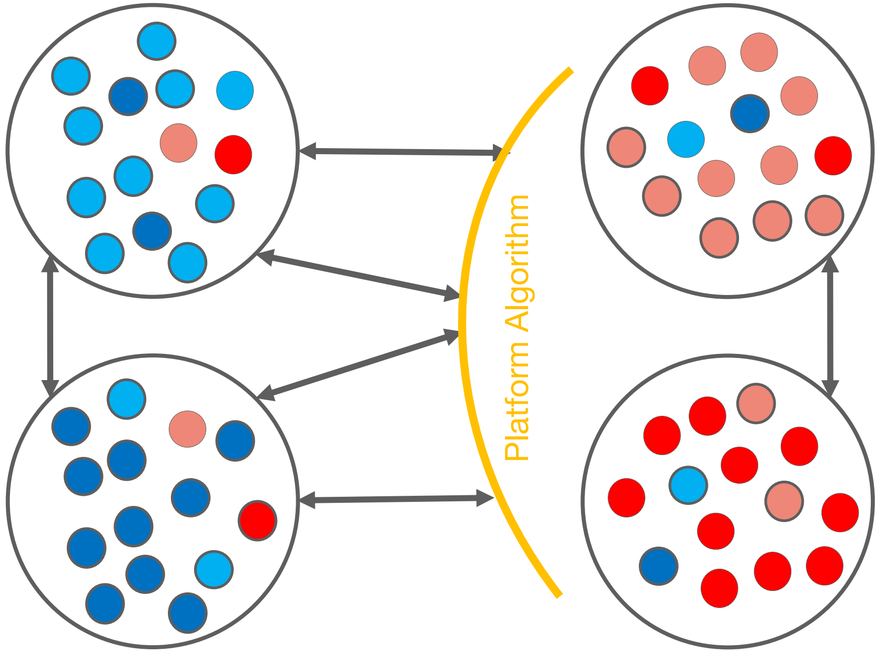
- Platform also selects the seed agent to maximize diffusion (proxy for profit).

Platform's Profit-Maximizing Solution

- Profit-maximizing (PM) sharing network also takes the form of an island structure.

Platform's Profit-Maximizing Solution

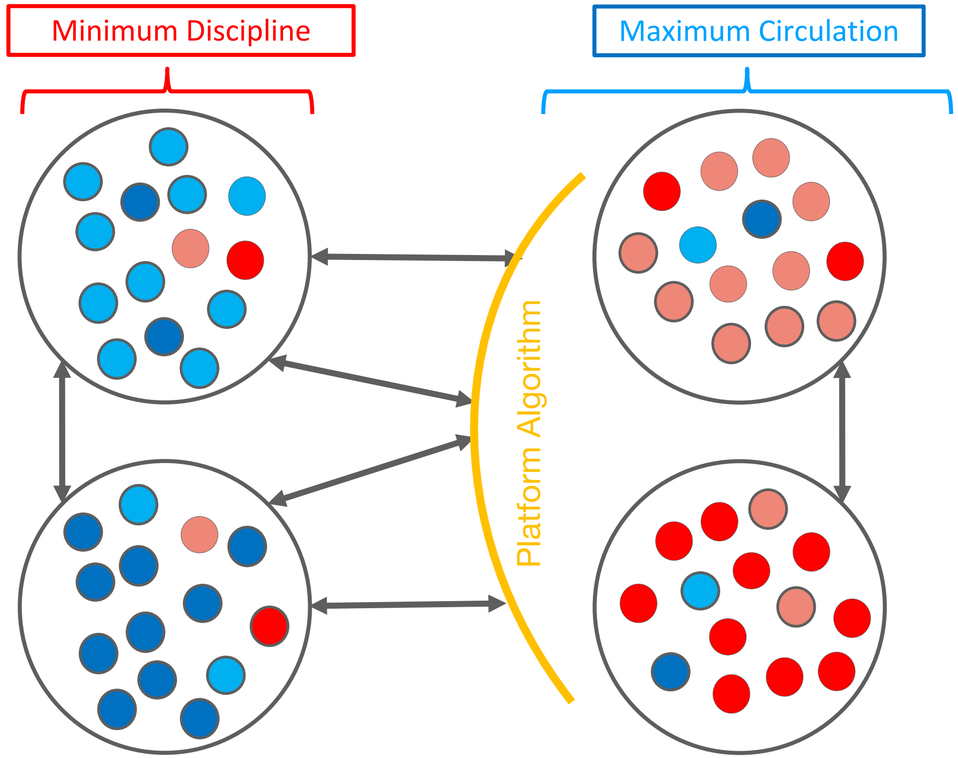
- Profit-maximizing (PM) sharing network also takes the form of an **island structure**.



- **Theorem 3**: There exists a reliability threshold $r^* \in (0,1)$ such that:
  - If $r > r^*$, the PM sharing network has maximal connectivity;
  - If $r < r^*$, the PM sharing network has maximal homophily.

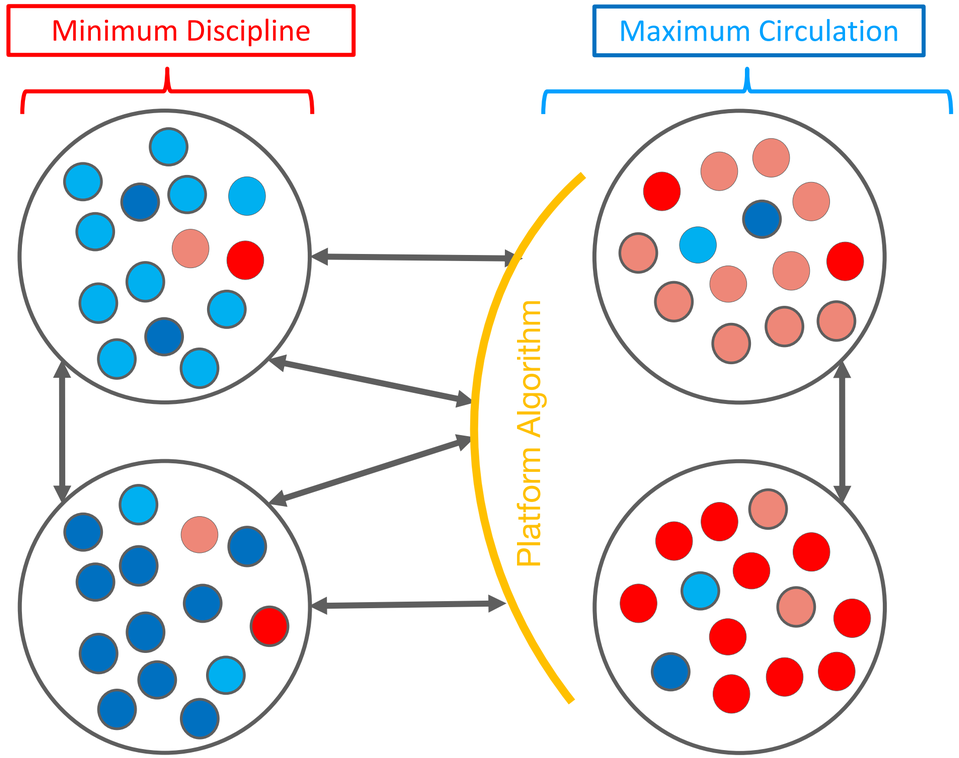
Intuition

- Balance between discipline and circulation effects:



Intuition

- Balance between discipline and circulation effects:



- Algorithmically-induced echo chamber (“**filter bubble**”) created by the platform to maximize diffusion precisely when content tends to be low reliability.

Impact of the Result

1. Global characterization of the profit-maximizing sharing network for the platform.

Impact of the Result

1. Global characterization of the profit-maximizing sharing network for the platform.

2. Intuitive interpretation in terms of empirically-documented filter bubble algorithms (Levy (2021)).

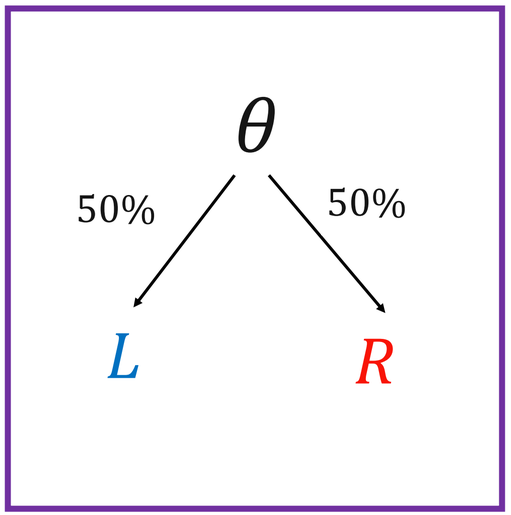
Impact of the Result

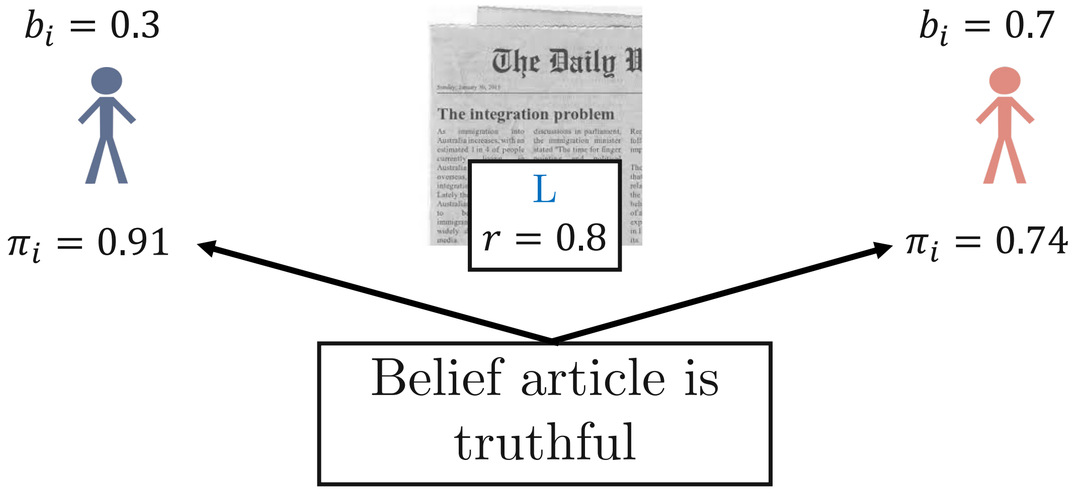
1. Global characterization of the profit-maximizing sharing network for the platform.

2. Intuitive interpretation in terms of empirically-documented filter bubble algorithms (Levy (2021)).

3. Computational simulations confirm similar (but less sharp) algorithms for coarser initial social network topologies (i.e., with fewer initial islands).

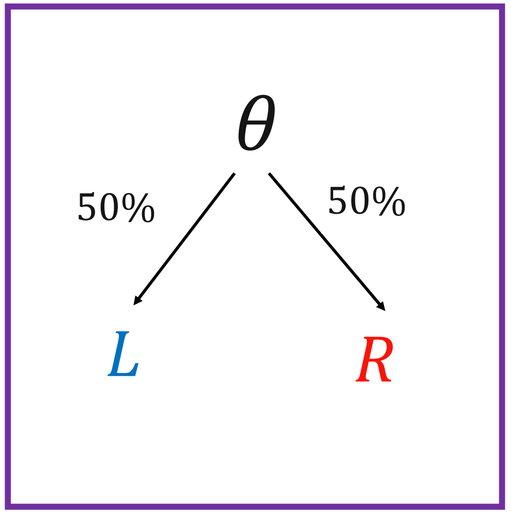
How should a regulator implement policies to counteract the spread of misinformation?

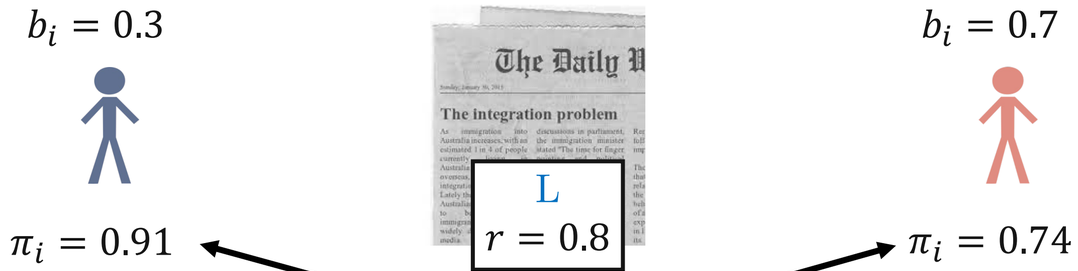




$\hat{b}_i = 0.23$ ← Belief $\theta = R$ → $\hat{b}_i = 0.68$

How should a regulator implement policies to counteract the spread of misinformation?





Belief article is truthful

$\hat{b}_i = 0.23$ ← Belief $\theta = R$ → $\hat{b}_i = 0.68$

There exists $r_{Reg} \in (0,1)$ such that if $r > r_{Reg}$ (resp. $r < r_{Reg}$) , higher (resp. lower) content diffusion leads to greater welfare.

Potential Policies

- **Content moderation**: A regulator removes a fraction of misinformation.
- **Provenance / Accuracy Nudging**: Equip users themselves with the tools to fact-check and verify content.
- **Performance Targets**: Make platforms responsible for self-monitoring by setting necessary misinformation "targets".
- **Network-based (AI) Regulations**: Regulate the algorithms that lead to problematic social media sharing networks.

Potential Policies

- **Content moderation**: A regulator removes a fraction of misinformation.
- **Provenance / Accuracy Nudging**: Equip users themselves with the tools to fact-check and verify content.
- **Performance Targets**: Make platforms responsible for self-monitoring by setting necessary misinformation "targets".
- **Network-based (AI) Regulations**: Regulate the algorithms that lead to problematic social media sharing networks.
- All can work if designed well, but all can "backfire" if not.

Potential Policies

- **Content moderation**: A regulator removes a fraction of misinformation.
- **Provenance / Accuracy Nudging**: Equip users themselves with the tools to fact-check and verify content.
- **Performance Targets**: Make platforms responsible for self-monitoring by setting necessary misinformation "targets".
- **Network-based (AI) Regulations**: Regulate the algorithms that lead to problematic social media sharing networks.
- All can work if designed well, but all can "backfire" if not.
- Different advantages/disadvantages of each (see paper).

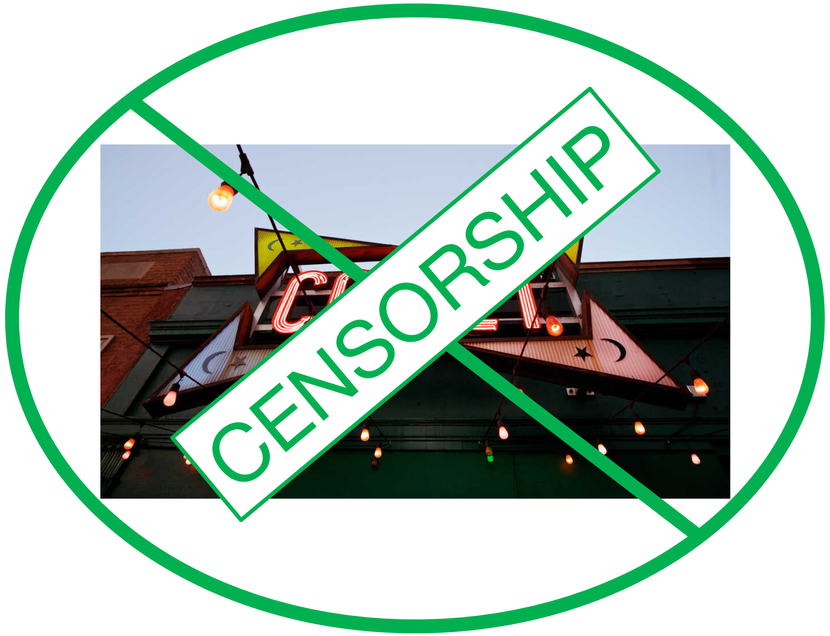
An Example of Backfire

Censorship / Content Moderation

(remove some misinformation)

Content Moderation

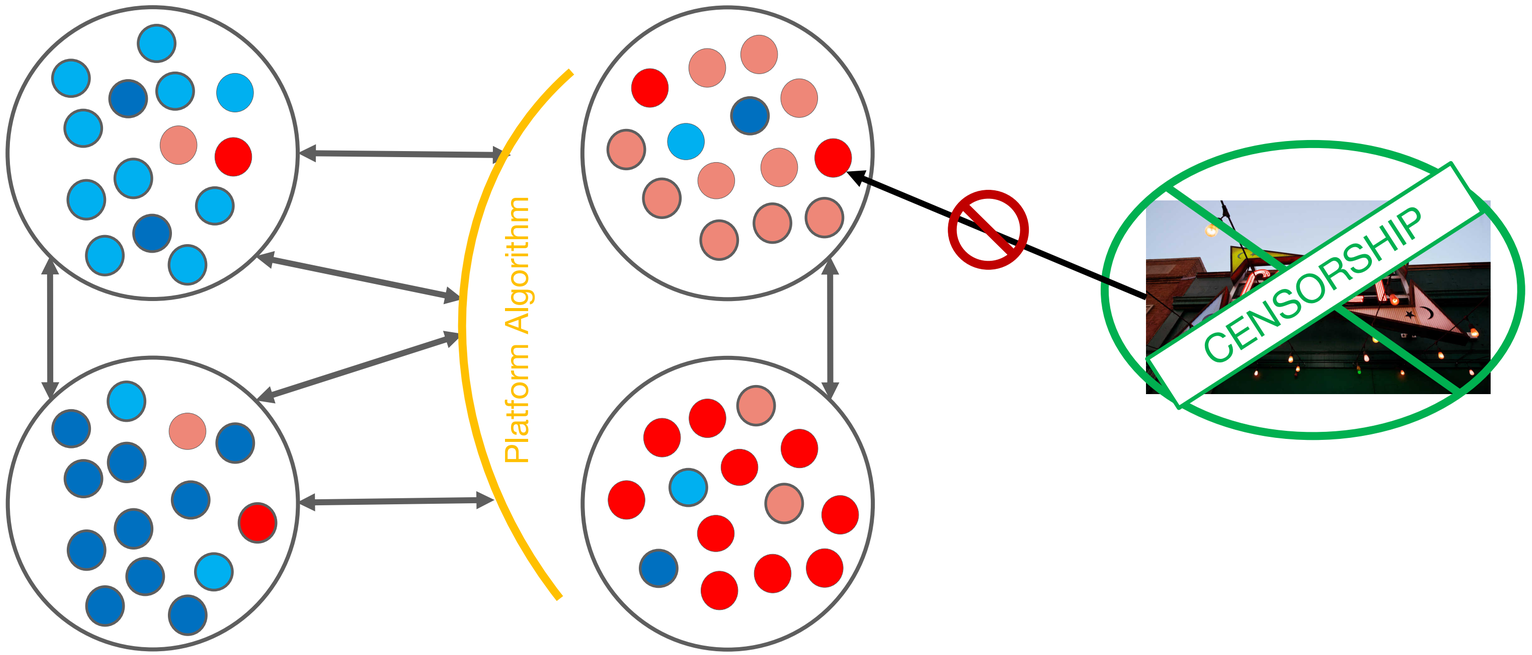
Content Moderation



Detect 1/3 of the misinformation immediately and remove it.

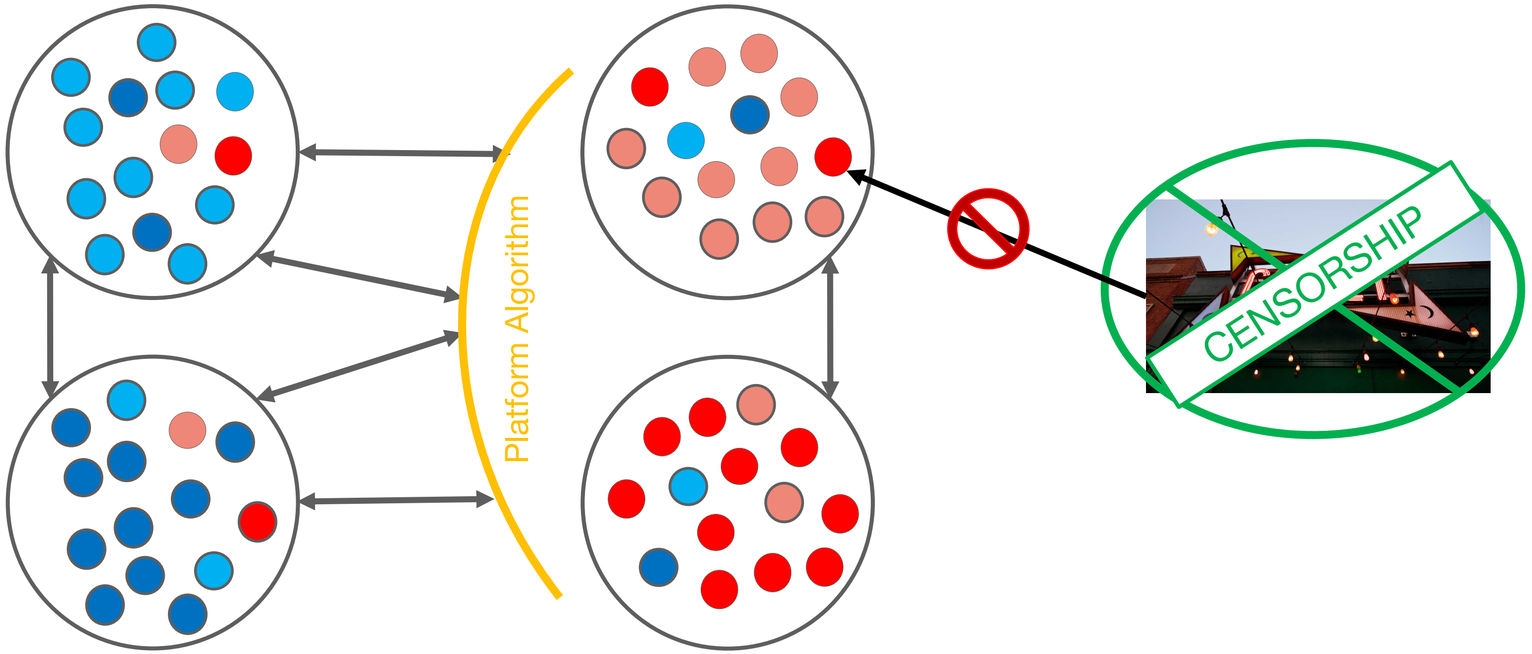
Removed Article

- Recommended article profit-maximizing sharing network:



Removed Article

- Recommended article profit-maximizing sharing network:



- Content moderation policy removes the article from circulation, reduces sharing (and diffusion) of misinformation.

Undetected Article

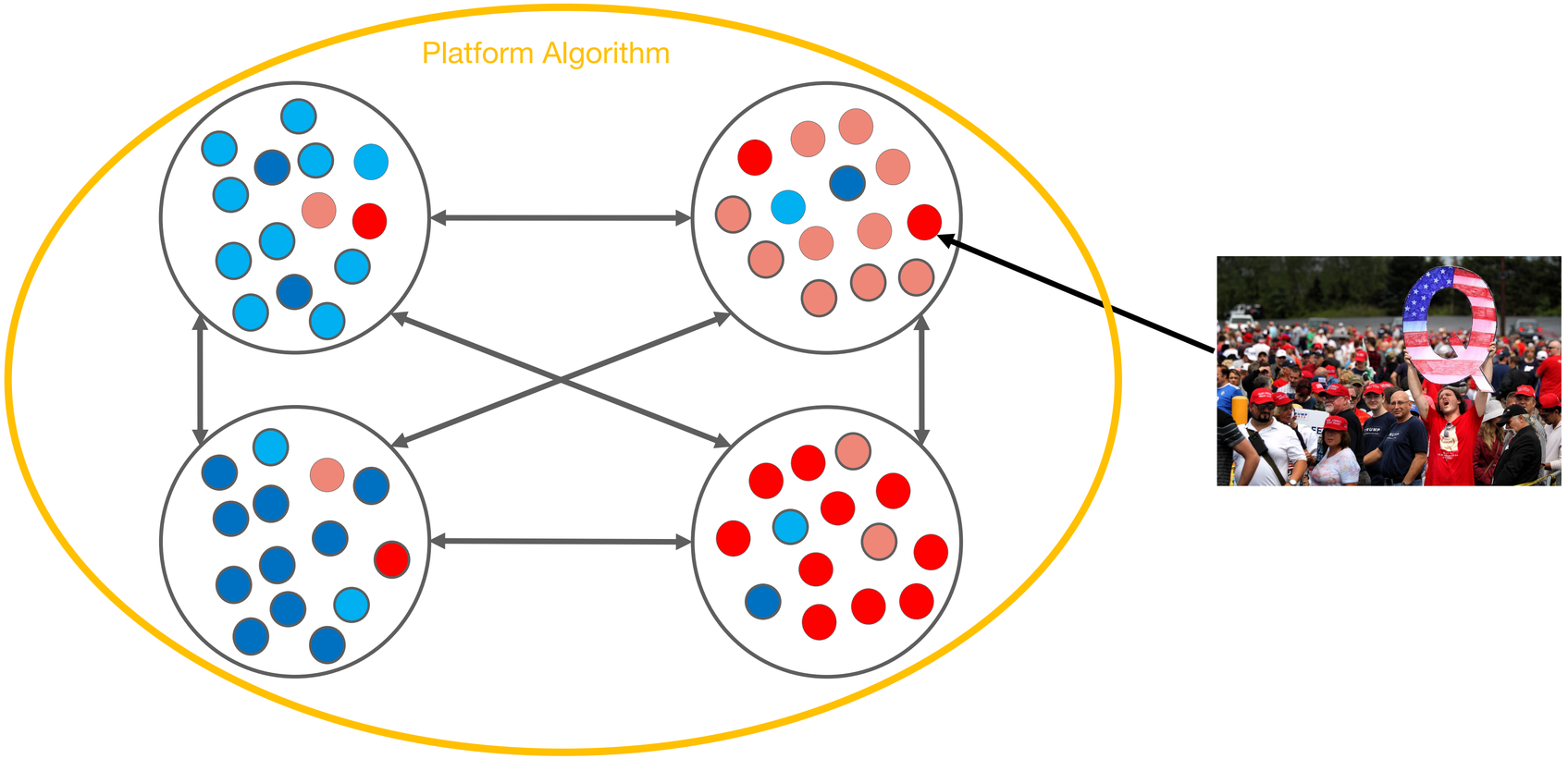
- If the article is not detected, generates an implied truth effect.

Undetected Article

- If the article is not detected, generates an implied truth effect.
  - Platform algorithm adapts as well.

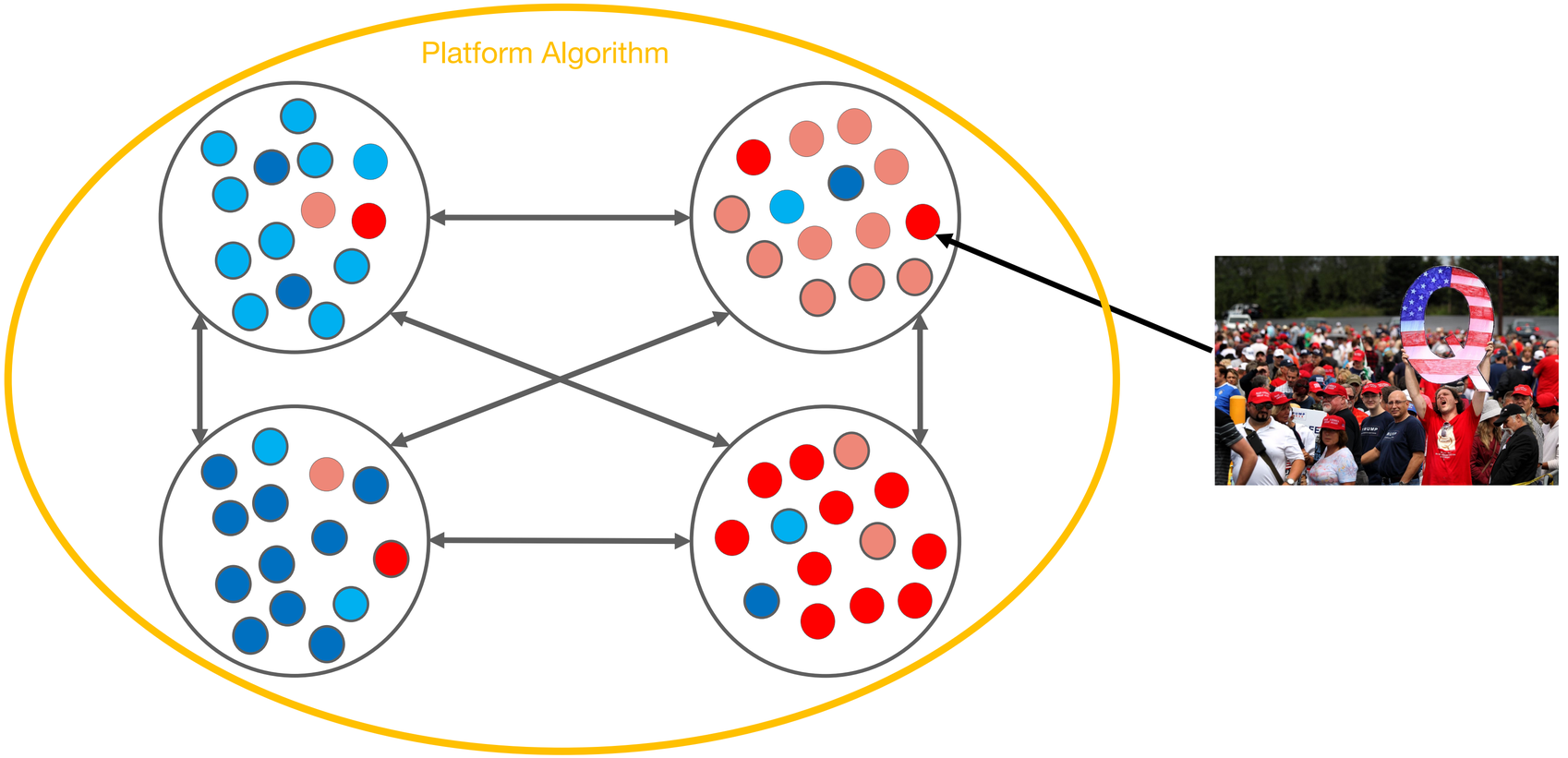
Undetected Article

- If the article is not detected, generates an implied truth effect.
  - Platform algorithm adapts as well.



Undetected Article

- If the article is not detected, generates an implied truth effect.
  - Platform algorithm adapts as well.



- Article may spread **at a rate greater than 3/2** the original rate!

Conclusion

- **Strategic model** of user sharing behavior and diffusion of an article online.

Conclusion

- **Strategic model** of user sharing behavior and diffusion of an article online.
  - Network homophily aids an article's spread when it is more likely to contain misinformation (and hurts the spread of more reliable content).

Conclusion

- **Strategic model** of user sharing behavior and diffusion of an article online.
  - Network homophily aids an article's spread when it is more likely to contain misinformation (and hurts the spread of more reliable content).

- **Platform algorithms** leverage this fact to increase engagement and diffusion.

Conclusion

- **Strategic model** of user sharing behavior and diffusion of an article online.
  - Network homophily aids an article's spread when it is more likely to contain misinformation (and hurts the spread of more reliable content).

- **Platform algorithms** leverage this fact to increase engagement and diffusion.
  - Generate artificial echo chambers ("filter bubbles") for low-reliability content. Platform algorithms play smaller role for more reliable content.

Conclusion

- **Strategic model** of user sharing behavior and diffusion of an article online.
  - Network homophily aids an article's spread when it is more likely to contain misinformation (and hurts the spread of more reliable content).
- **Platform algorithms** leverage this fact to increase engagement and diffusion.
  - Generate artificial echo chambers ("filter bubbles") for low-reliability content. Platform algorithms play smaller role for more reliable content.
- Regulatory policy can be effective, but if not carefully calibrated, can lead to even worse societal outcomes.

THANK YOU!