

Social Media Suspensions and Shadow Banning: Political Bias or Genuine Disinformation Control?

Tauhid Zaman

tauhid.zaman@yale.edu



Censorship on Social Media

Trump sues Twitter, Google and Facebook alleging 'censorship'

© 7 July 2021



Former US president Donald Trump has filed a lawsuit against tech giants Google, Twitter and Facebook, claiming that he is the victim of censorship.

Censorship on Social Media

- Two ways to censor content on social media

Censorship on Social Media

- Two ways to censor content on social media

- 1. Suspension** – remove a user's account from the platform

Censorship on Social Media

- Two ways to censor content on social media

1. Suspension – remove a user's account from the platform

2. Shadow banning – quietly limit the reach of a user's content

Suspension on Social Media

- Suspensions are used for the most dangerous accounts

Suspension on Social Media

- Suspensions are used for the most dangerous accounts

- Terrorists inciting violence



Suspension on Social Media

- Suspensions are used for the most dangerous accounts

- Terrorists inciting violence



- Bots distorting online conversations



Suspension on Social Media

- Suspensions are used for the most dangerous accounts

- Terrorists inciting violence



- Bots distorting online conversations



- Users engaging in hate speech or spreading disinformation



Shadow Banning on Social Media

← Post

Donald J. Trump 
@realDonaldTrump

DONALDJTRUMP.COM

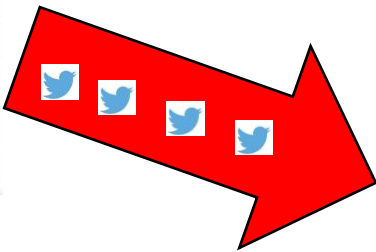
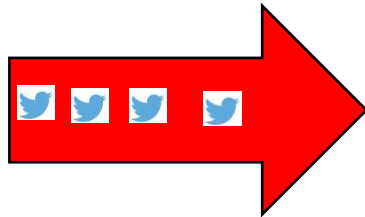
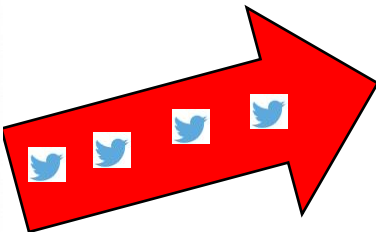
MUG SHOT — AUGUST 24, 2023



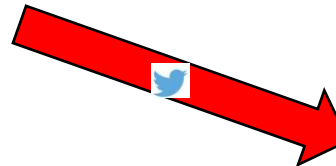
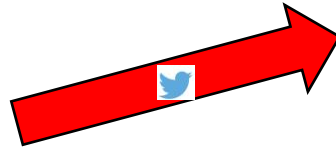
ELECTION INTERFERENCE
NEVER SURRENDER!
DONALDJTRUMP.COM

7:08 am · 25 Aug 2023 · 2.4M Views

37.1K Reposts 14.8K Quotes 104K Likes 3,299 Bookmarks



Shadow Banning on Social Media



Shadow Banning on Social Media



Donald J. Trump ✓

@realDonaldTrump

Following



Twitter “SHADOW BANNING” prominent Republicans. Not good. We will look into this discriminatory and illegal practice at once! Many complaints.

6:46 AM - 26 Jul 2018

1,569 Retweets 4,222 Likes



1.2K 1.6K 4.2K



Political Bias in Social Media Censorship

Political Bias in Social Media Censorship

- Question 1: Is there political bias in Twitter **suspensions**?

Political Bias in Social Media Censorship

- Question 1: Is there political bias in Twitter **suspensions**?
- Answer: Maybe

Political Bias in Social Media Censorship

- Question 1: Is there political bias in Twitter **suspensions**?
- Answer: Maybe

- Question 2: Can a social media platform impose politically biased **shadow banning** without appearing to do so?

Political Bias in Social Media Censorship

- Question 1: Is there political bias in Twitter **suspensions**?
- Answer: Maybe

- Question 2: Can a social media platform impose politically biased **shadow banning** without appearing to do so?
- Answer: Yes



Political Bias in Twitter Suspensions

Data

- We tracked 9,000 Twitter users who used political hashtags during the 2020 U.S. presidential election on October 6, 2020

#VoteBidenHarris2020



4,500 Twitter users

#Trump2020

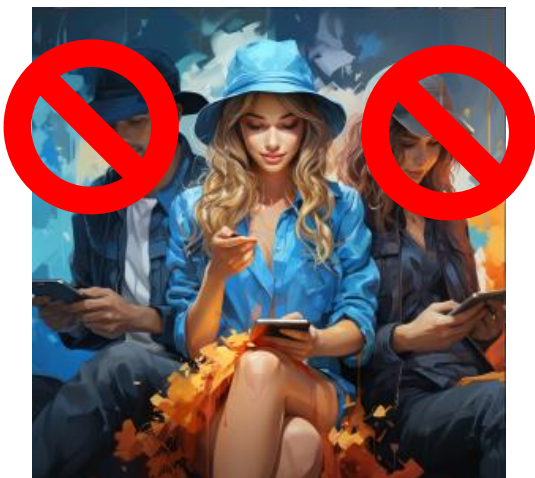


4,500 Twitter users

Suspension Data

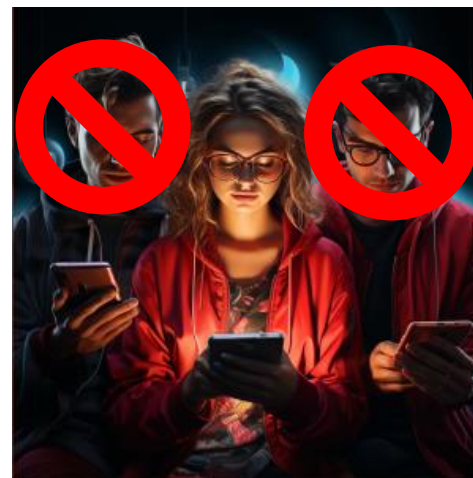
- After 9 months, we checked who was suspended on Twitter

#VoteBidenHarris2020



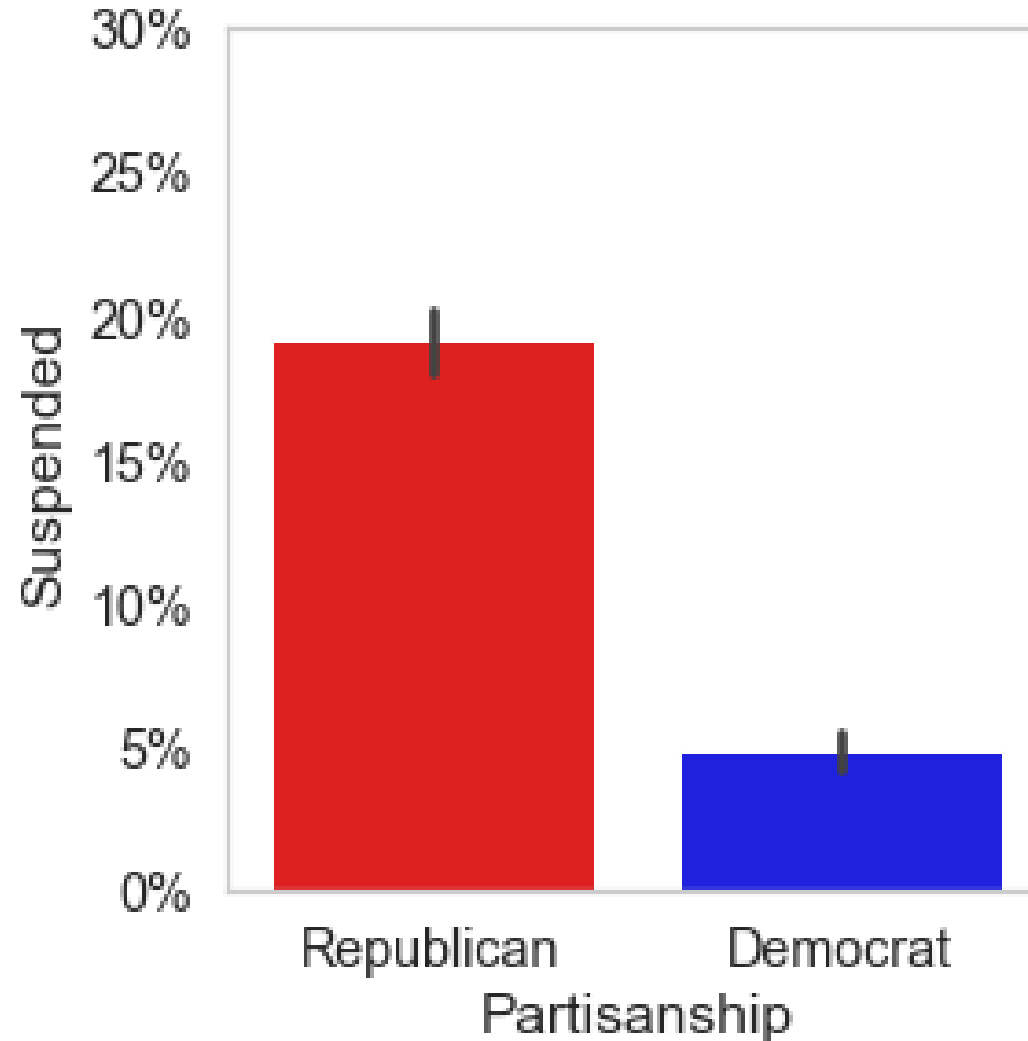
4,500 Twitter users

#Trump2020



4,500 Twitter users

Political Bias in Suspensions



Political Bias in Suspensions?

- The asymmetry in suspensions may be due to **political bias by** Twitter

Political Bias in Suspensions?

- The asymmetry in suspensions may be due to **political bias by** Twitter
- The asymmetry in suspensions could be the result of a **non-political suspension policy**

Political Bias in Suspensions?

- The asymmetry in suspensions may be due to **political bias by** Twitter
- The asymmetry in suspensions could be the result of a **non-political suspension policy**
- Example: mitigate the spread of **disinformation**

Media Quality Scores

- Scores exist that rate the quality of news sites
 - Low score means the news is more likely to be misinformation or fake news
- **Professional fact-checker** trustworthiness ratings ¹
- **Politically-balanced layperson crowd-sourced** trustworthiness ²

1. Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. *Tweeting from left to right: Is online political communication more than an echo chamber?* *Psychological science* 26(10):1531–1542.
2. Pennycook, G., and Rand, D. G. 2019. *Fighting misinformation on social media using crowdsourced judgments of news source quality.* *Proceedings of the National Academy of Sciences* 116(7):2521–2526.

Media Quality Scores

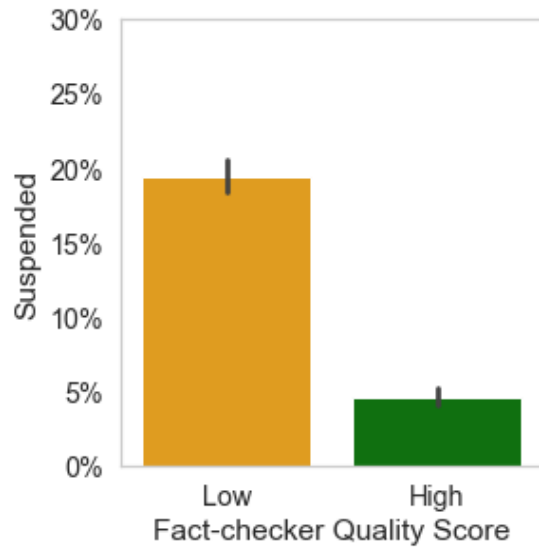
Mainstream			Hyper-partisan			Fake news		
Domain	Politically balanced layperson rating	Fact-checker rating	Domain	Politically balanced layperson rating	Fact-checker rating	Domain	Politically balanced layperson rating	Fact-checker rating
abcnews.go.com	0.45	0.56	activepost.com	0.2	0	americannews.com	0.22	0
aol.com/news	0.35	0.41	antiwar.com	0.18	0	angrypatriotmovement.com	0.18	0
bbc.co.uk	0.38	0.81	blacklistednews.com	0.18	0	bb4sp.com	0.18	0
bostonglobe.com	0.33	0.75	breitbart.com	0.22	0.16	beforeitsnews.com	0.19	0
cbsnews.com	0.48	0.66	commondreams.org	0.18	0.03	channel24news.com	0.25	0.06
chicagotribune.com	0.38	0.53	conservativetribune.com	0.24	0.03	clashdaily.com	0.18	0
cnn.com	0.47	0.84	crooksandliars.com	0.18	0.13	conservativedaily.com	0.23	0
dailymail.co.uk	0.3	0.44	dailycaller.com	0.21	0.13	dailybuzzlive.com	0.24	0
foxnews.com	0.45	0.44	dailykos.com	0.2	0.16	downtrend.com	0.19	0
huffingtonpost.com	0.41	0.47	dailysignal.com	0.2	0	freedomdaily.com	0.2	0.03
latimes.com	0.33	0.75	dailywire.com	0.25	0.16	newsbreakshere.com	0.19	0
msnbc.com	0.44	0.66	ijr.com	0.19	0.09	notallowedto.com	0.17	0
news.yahoo.com	0.4	0.59	infowars.com	0.21	0.03	now8news.com	0.2	0
nydailynews.com	0.33	0.34	newsmax.com	0.23	0.13	onepoliticalplaza.com	0.19	0
nypost.com	0.38	0.38	patriotpost.us	0.21	0	react365.com	0.17	0
nytimes.com	0.45	0.91	rawstory.com	0.19	0.09	realnewsrightnow.com	0.21	0
sfchronicle.com	0.26	0.59	redstate.com	0.2	0.06	socialeverything.com	0.18	0
usatoday.com	0.45	0.66	thedailysheep.com	0.18	0.09	thenewyorkevening.com	0.24	0
washingtonpost.com	0.45	0.91	thepoliticalinsider.com	0.22	0.03	whatdoesitmean.com	0.19	0
wsj.com	0.34	0.72	westernjournal.com	0.22	0.06	yournewswire.com	0.19	0.06

Media Quality Scores

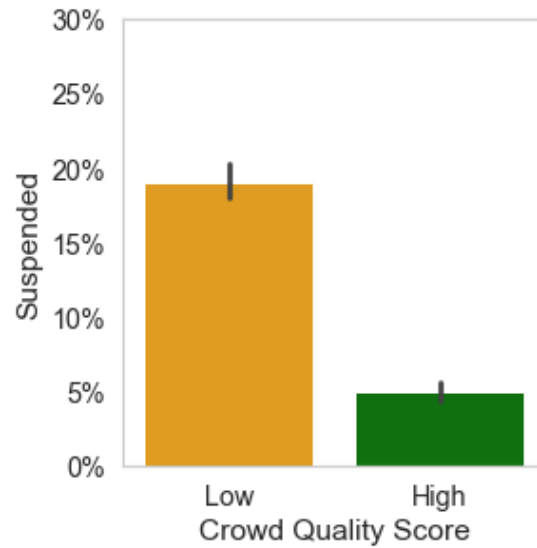
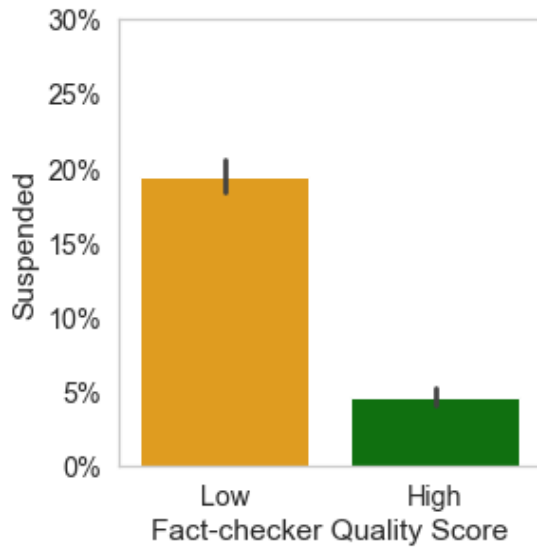
Mainstream			Hyper-partisan			Fake news		
Domain	Politically balanced layperson rating	Fact-checker rating	Domain	Politically balanced layperson rating	Fact-checker rating	Domain	Politically balanced layperson rating	Fact-checker rating
abcnews.go.com	0.45	0.56	activepost.com	0.2	0	americannews.com	0.22	0
aol.com/news	0.35	0.41	antiwar.com	0.18	0	angrypatriotmovement.com	0.18	0
bbc.co.uk	0.38	0.81	blacklistednews.com	0.18	0	bb4sp.com	0.18	0
bostonglobe.com	0.33	0.75	breitbart.com	0.22	0.16	beforeitsnews.com	0.19	0
cbsnews.com	0.48	0.66	commondreams.org	0.18	0.03	channel24news.com	0.25	0.06
chicagotribune.com	0.38	0.53	conservativetribune.com	0.24	0.03	clashdaily.com	0.18	0
cnn.com	0.47	0.84	crooksandliars.com	0.18	0.13	conservativedaily.com	0.23	0
dailymail.co.uk	0.3	0.44	dailycaller.com	0.21	0.13	dailybuzzlive.com	0.24	0
foxnews.com	0.45	0.44	dailykos.com	0.2	0.16	downtrend.com	0.19	0
huffingtonpost.com	0.41	0.47	dailysignal.com	0.2	0	freedomdaily.com	0.2	0.03
latimes.com	0.33	0.75	dailywire.com	0.25	0.16	newsbreakshere.com	0.19	0
msnbc.com	0.44	0.66	ijr.com	0.19	0.09	notallowedto.com	0.17	0
news.yahoo.com	0.4	0.59	infowars.com	0.21	0.03	now8news.com	0.2	0
nydailynews.com	0.33	0.34	newsmax.com	0.23	0.13	onepoliticalplaza.com	0.19	0
nypost.com	0.38	0.38	patriotpost.us	0.21	0	react365.com	0.17	0
nytimes.com	0.45	0.91	rawstory.com	0.19	0.09	realnewsrightnow.com	0.21	0
sfchronicle.com	0.26	0.59	redstate.com	0.2	0.06	socialeverything.com	0.18	0
usatoday.com	0.45	0.66	thedailysheep.com	0.18	0.09	thenewyorkevening.com	0.24	0
washingtonpost.com	0.45	0.91	thepoliticalinsider.com	0.22	0.03	whatdoesitmean.com	0.19	0
wsj.com	0.34	0.72	westernjournal.com	0.22	0.06	yournewswire.com	0.19	0.06

Media Quality and Suspensions

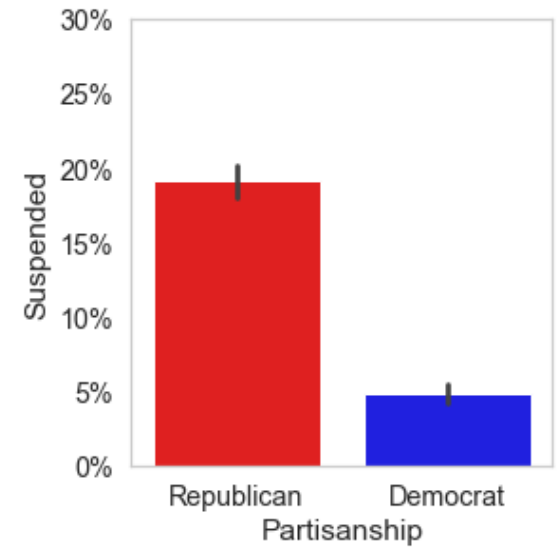
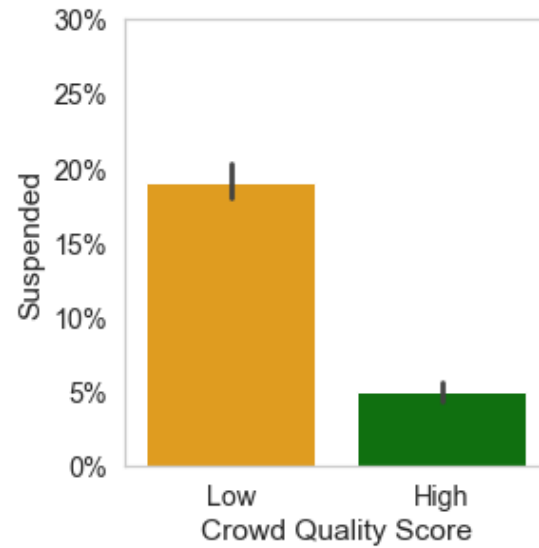
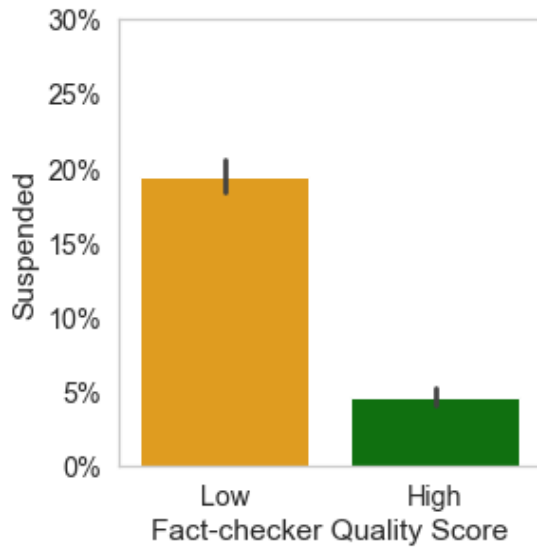
Media Quality and Suspensions



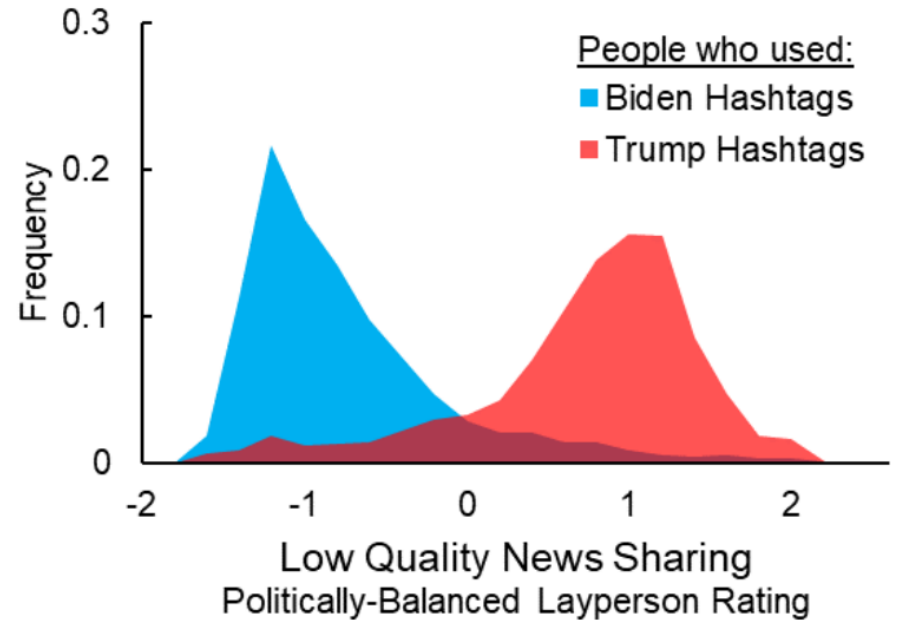
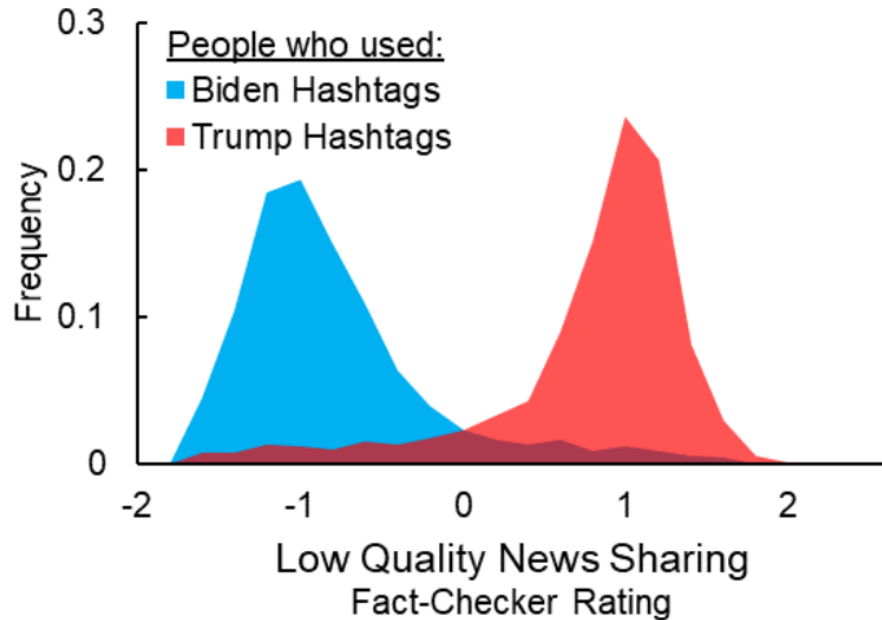
Media Quality and Suspensions



Media Quality and Suspensions



Media Quality and Political Sentiment



Political Bias in Suspensions?

- Political bias exists in suspensions
- The bias can be explained by good-faith efforts to **mitigate the spread of low quality news**
- It is still possible the bias is due to **political partisanship by Twitter**

New Management



Elon Musk  
@elonmusk

Subscribe ...

Twitter obv has a strong left wing bias

9:17 AM · May 9, 2022

 4.6K

 10K

 69K

 229



New Management





Shadow Banning

Shadow Banning on Twitter

ELON MUSK · Published May 12, 2023 4:05pm EDT

Musk says new Twitter CEO will not shadow ban users: 'That will not be the case'

Elon Musk has hired Linda Yaccarino to run Twitter

By [Julia Musto](#) | [FOXBusiness](#) |



U.S. Stock Market Quotes

Quote Lookup



Shadow Banning on Twitter

ELON MUSK · Published May 12, 2023 4:05pm EDT

Musk says new Twitter CEO will not shadow ban users: 'That will not be the case'

Elon Musk has hired Linda Yaccarino to run Twitter

By [Julia Musto](#) | [FOXBusiness](#) |



U.S. Stock Market Quotes

Quote Lookup



Frederic Legrand - COMEO/Shutterstock (Licensed)

Elon Musk railed against shadow bans—now he's using them on his critics

Musk is a self-appointed free speech crusader. Sometimes.



Steven Monacelli

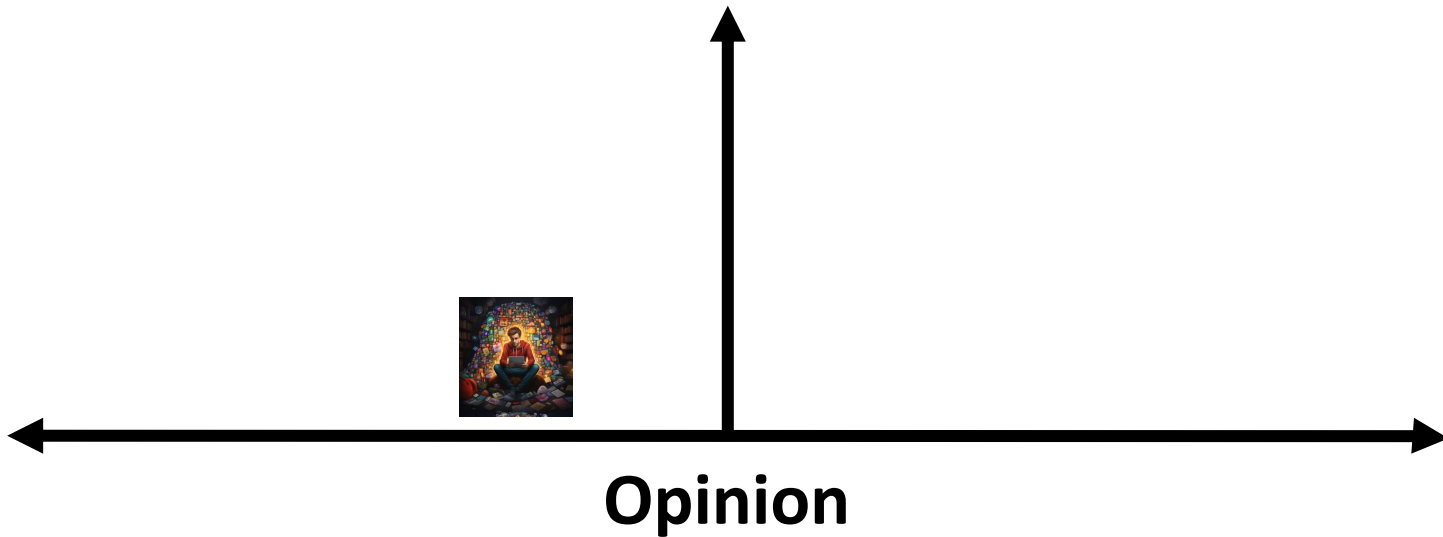
Tech

Posted on Nov 13, 2023

Updated on Nov 13, 2023, 2:24 pm CST

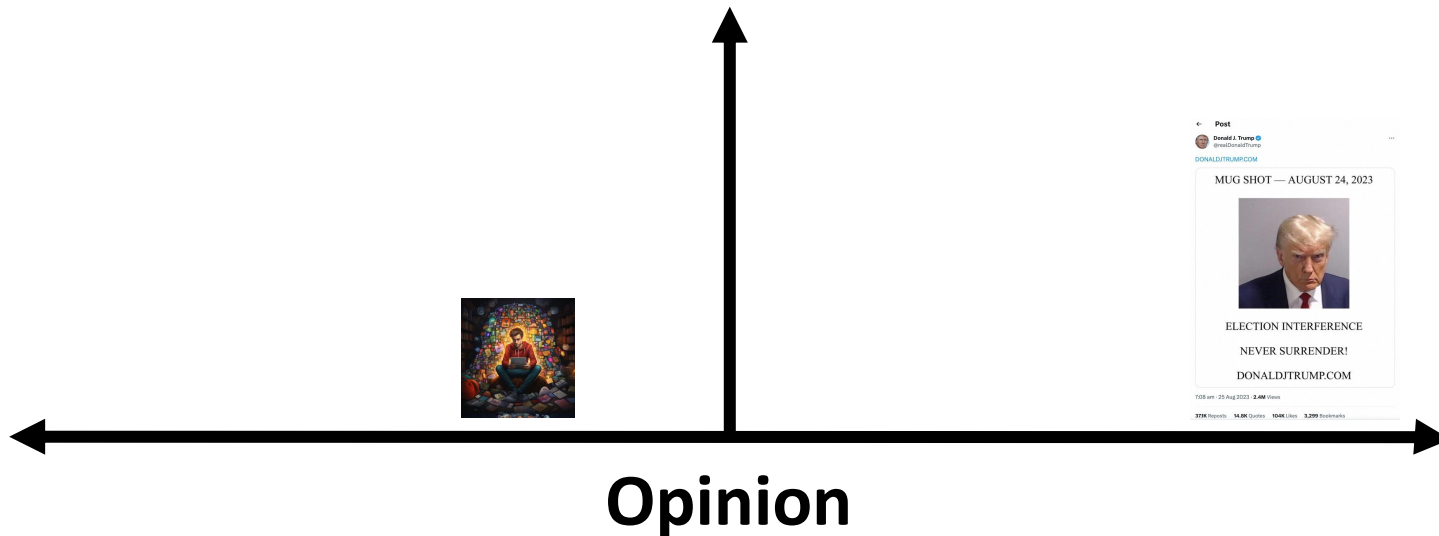
Persuasion

- People have opinions



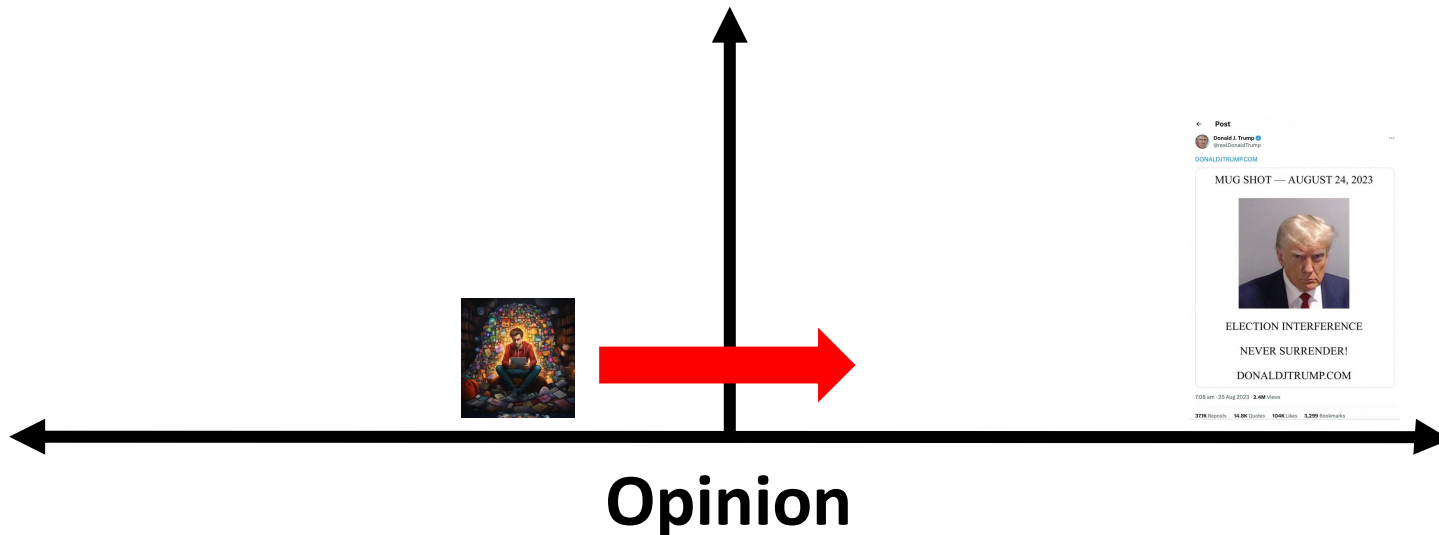
Persuasion

- People have opinions
- Tweets have opinions



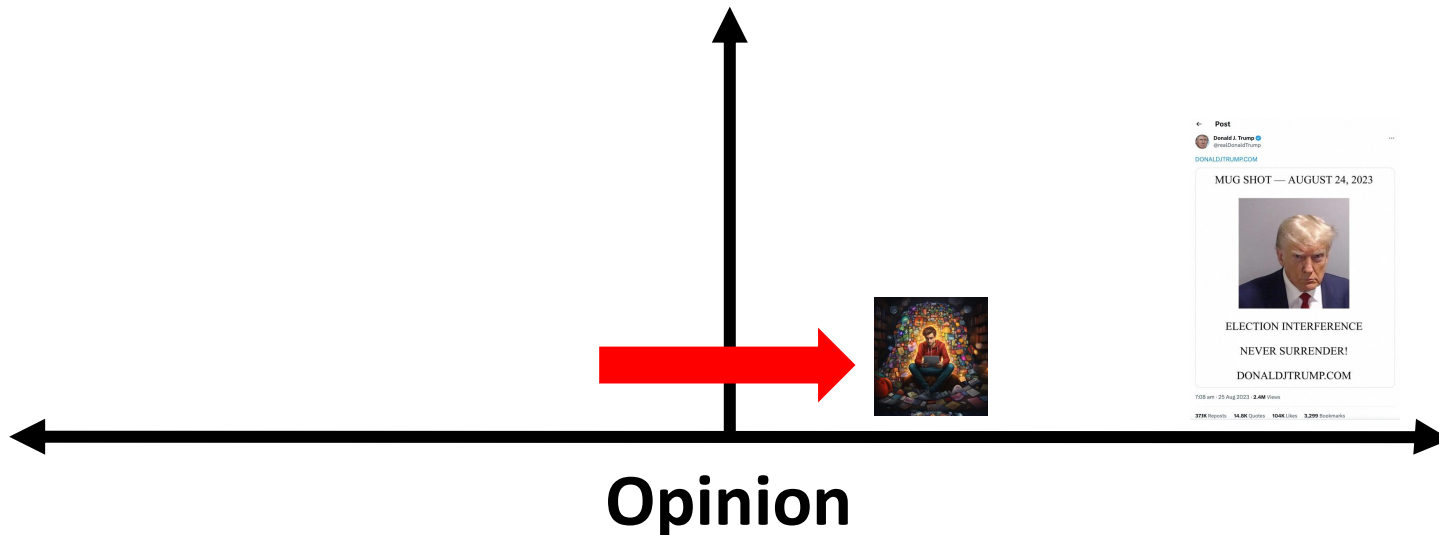
Persuasion

- People have opinions
- Tweets have opinions
- **Persuasion** -tweets move people's opinions towards their own

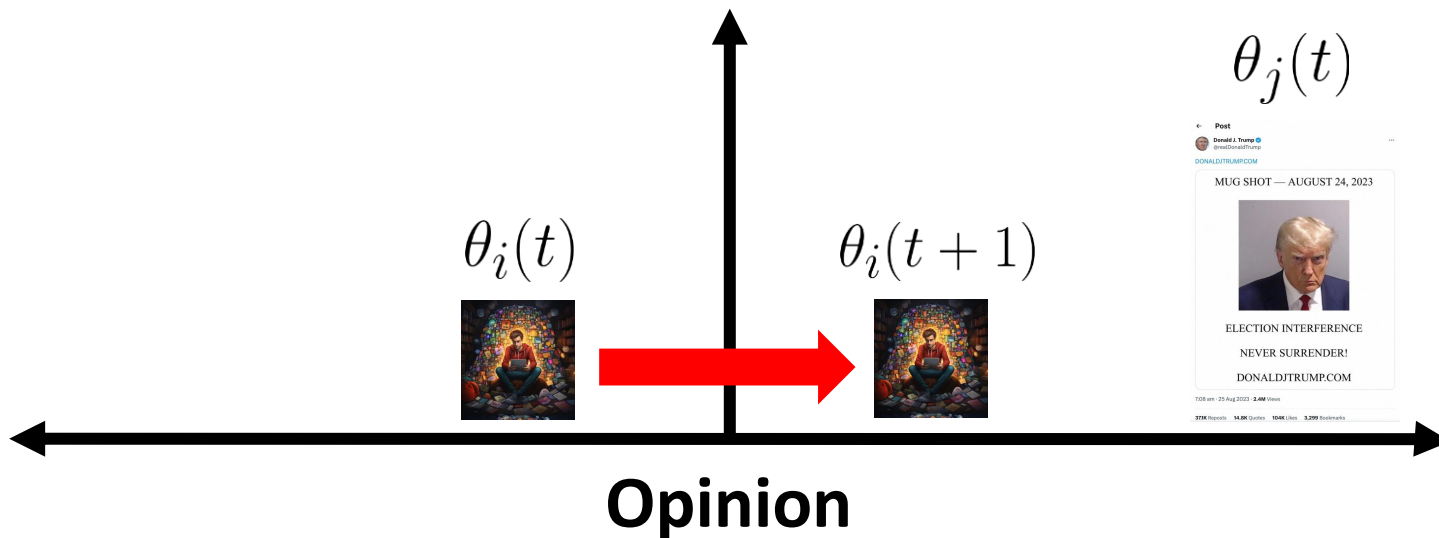


Persuasion

- People have opinions
- Tweets have opinions
- **Persuasion** -tweets move people's opinions towards their own

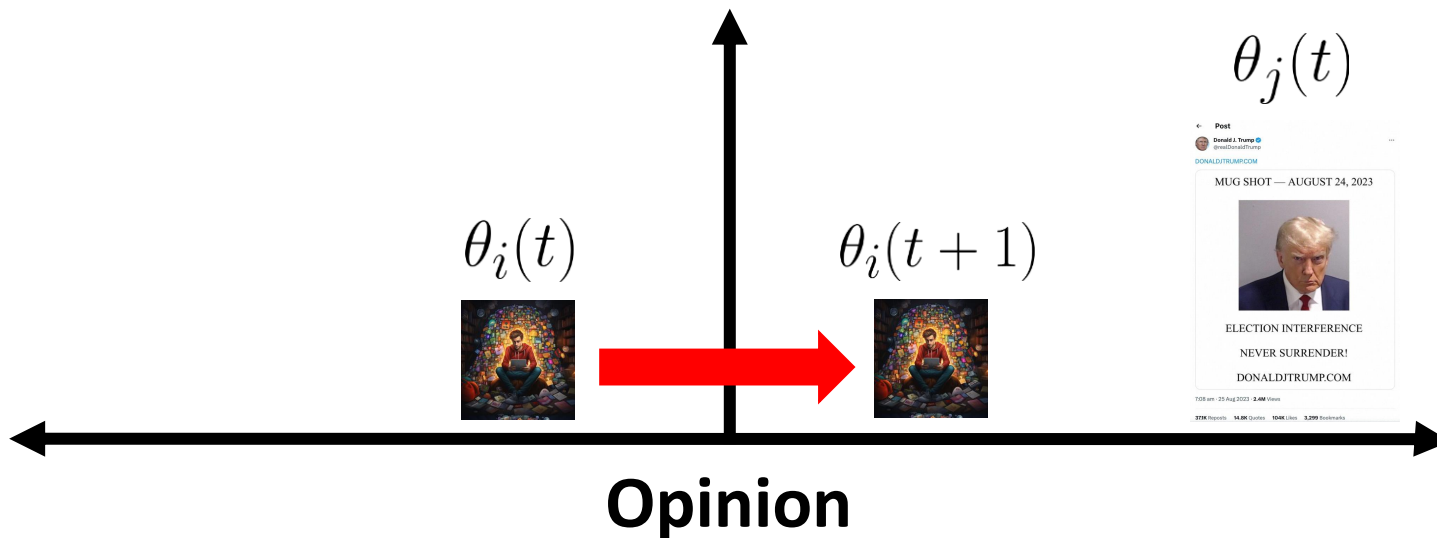


DeGroot Model



DeGroot Model

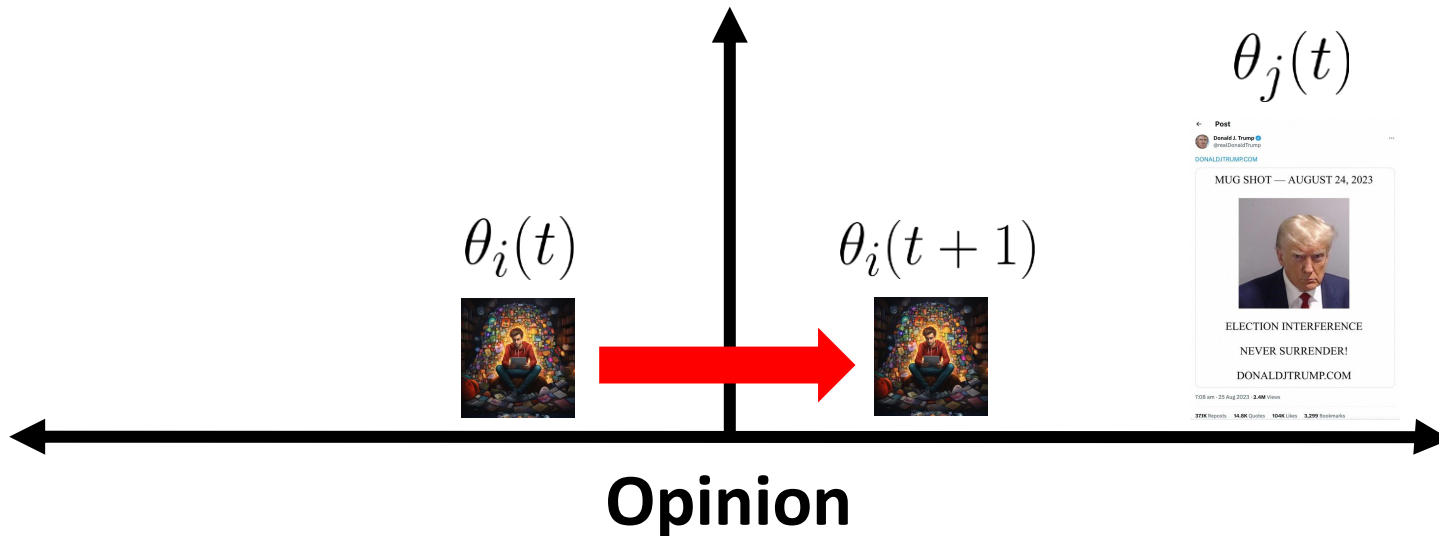
- DeGroot model describes how opinions move



DeGroot Model

- DeGroot model describes how opinions move

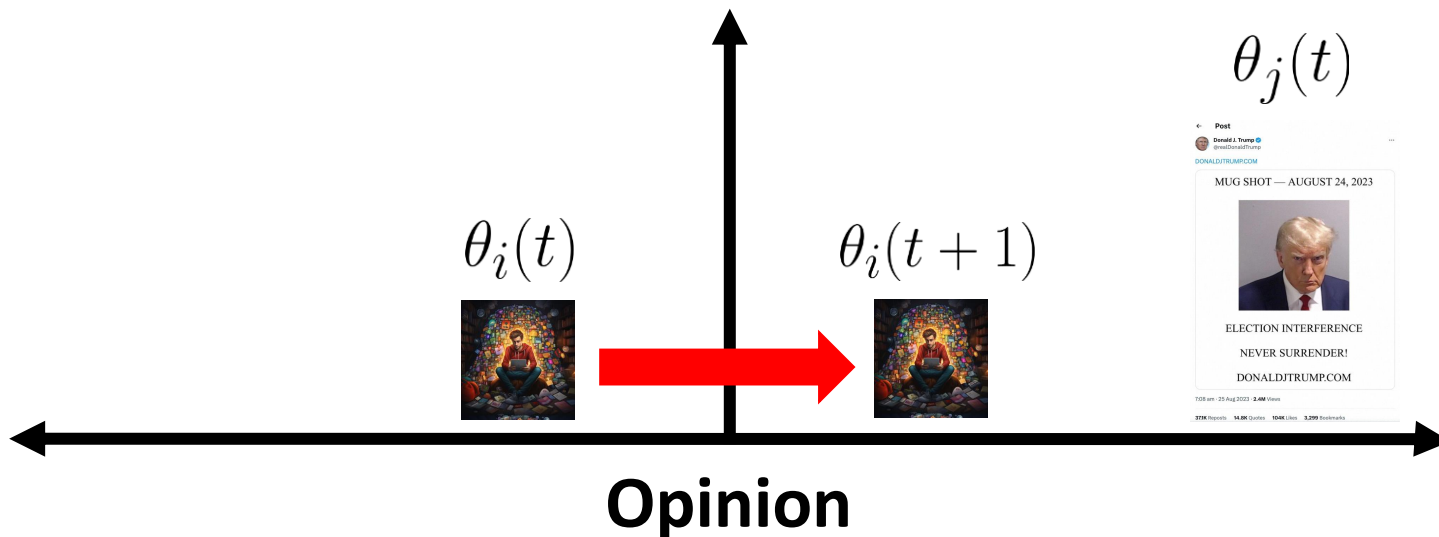
$$\theta_i(t + 1) = \theta_i(t) + \omega(\theta_j(t) - \theta_i(t))$$



DeGroot Model in Large Networks

- For large networks, we can make time continuous because tweets happen so fast

$$\frac{d\theta_i}{dt} = \sum_j \lambda_{ji} \omega(\theta_j - \theta_i)$$

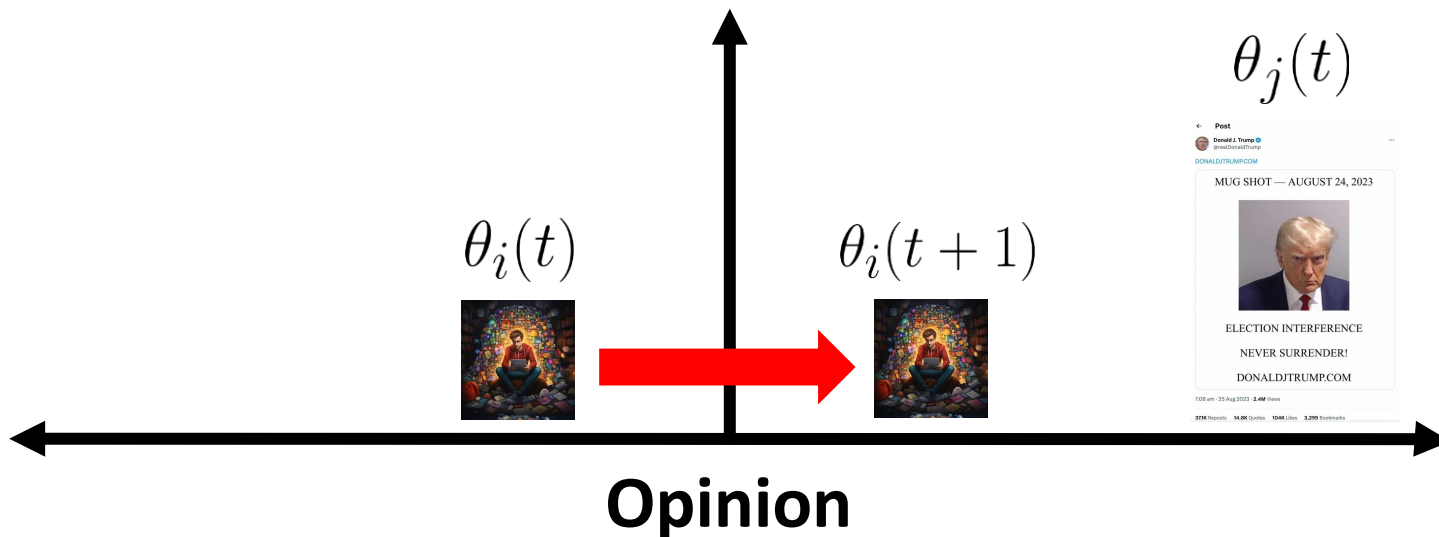


DeGroot Model in Large Networks

- For large networks, we can make time continuous because tweets happen so fast

$$\frac{d\theta_i}{dt} = \sum_j \lambda_{ji} \omega(\theta_j - \theta_i)$$

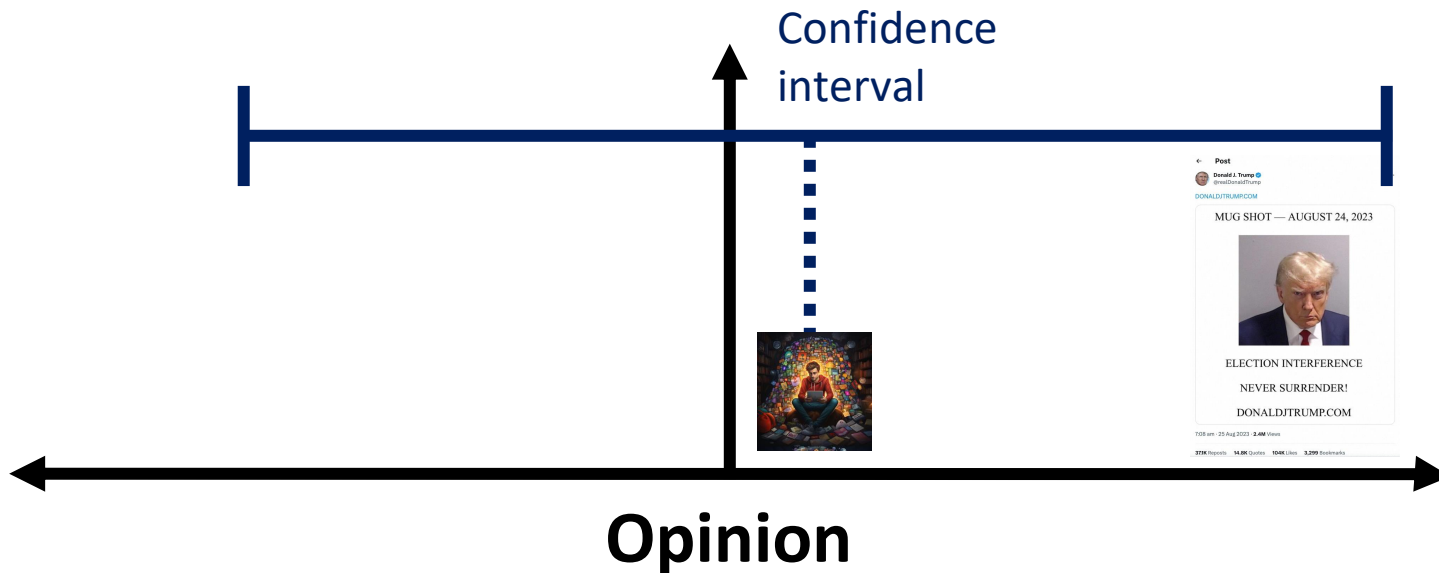
Tweet rate



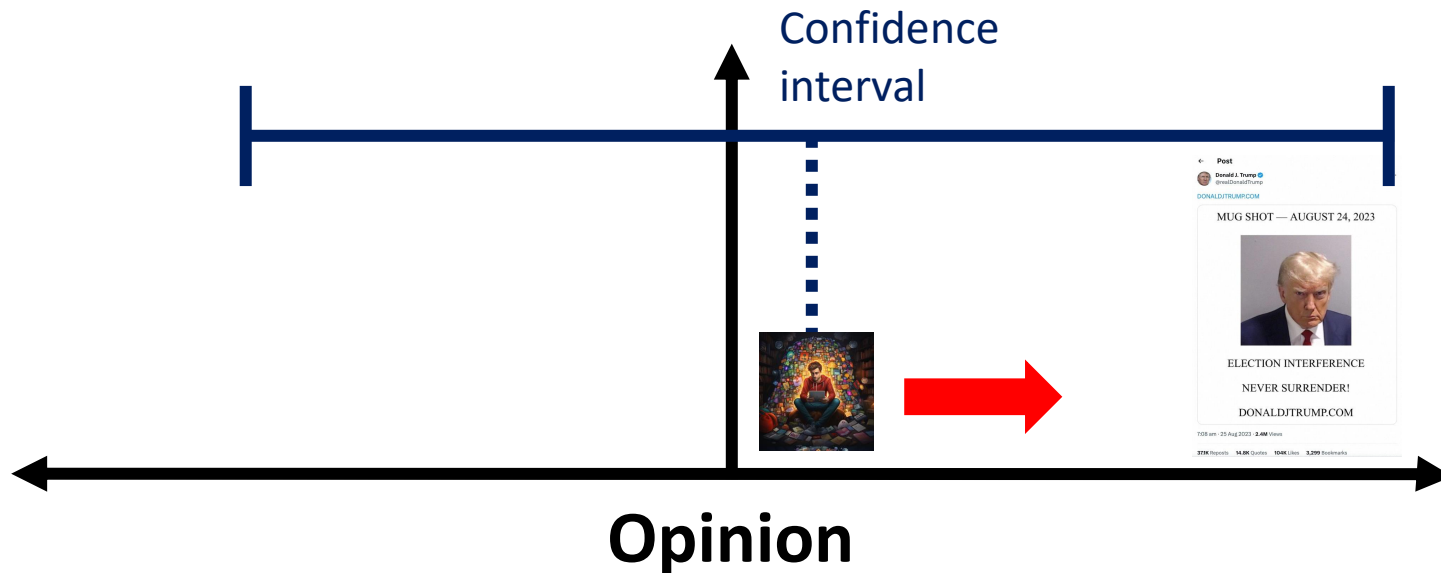
Bounded Confidence Model

- DeGroot model results in consensus – everyone has the same opinion eventually
- In reality there is polarization in social networks
- **Bounded confidence model** – if opinion of tweet is too far away, then there is no persuasion

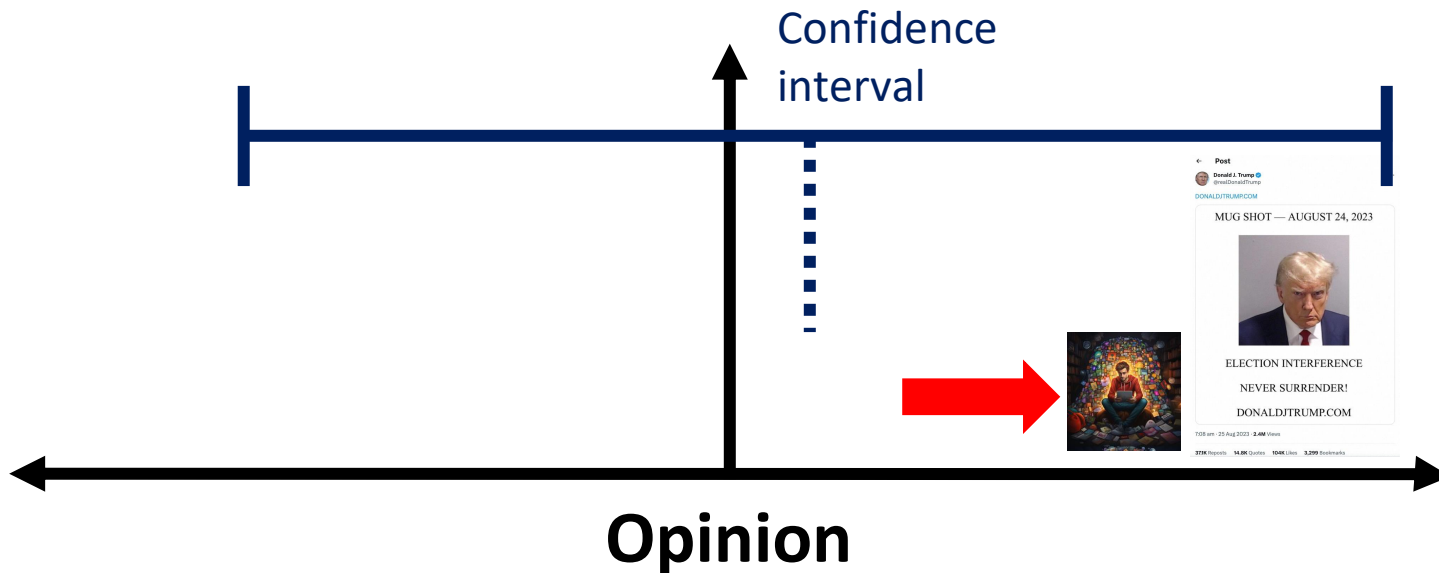
Bounded Confidence Model



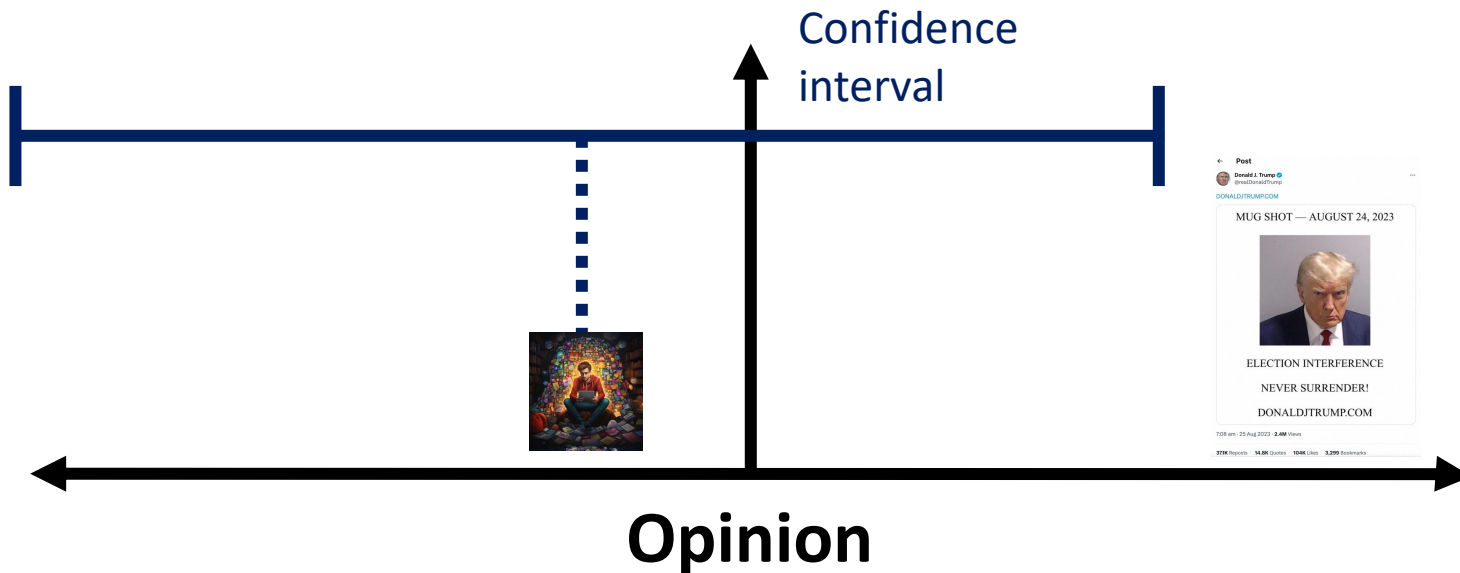
Bounded Confidence Model



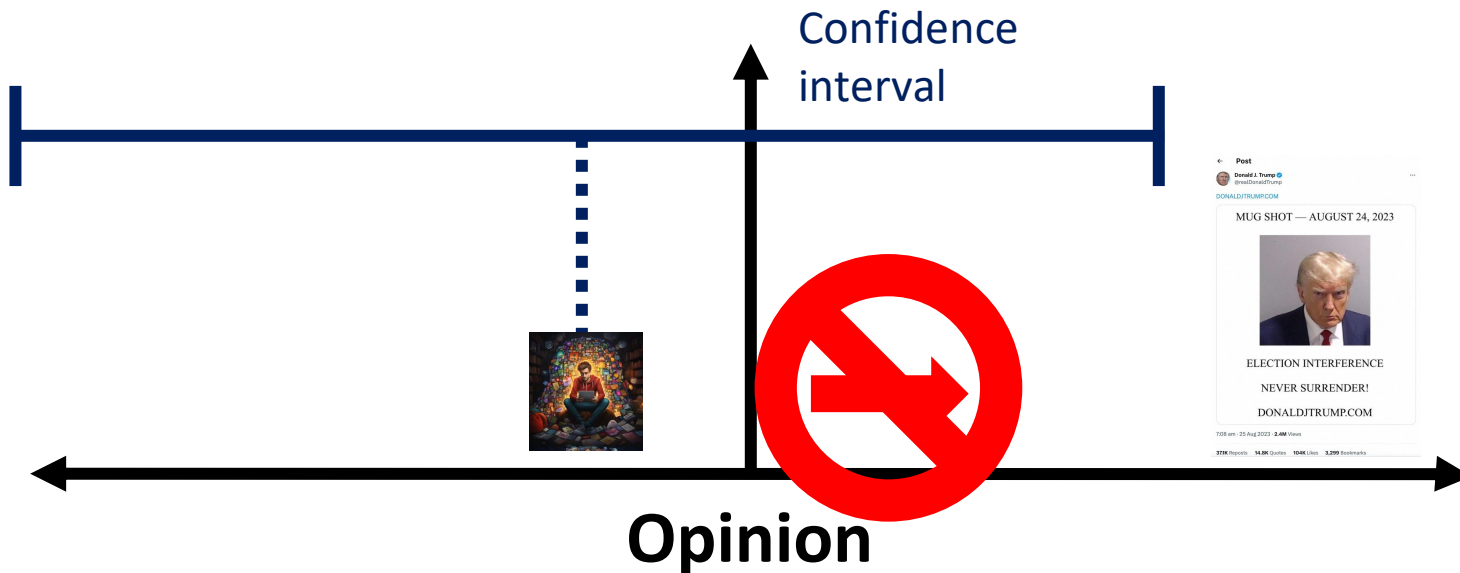
Bounded Confidence Model



Bounded Confidence Model



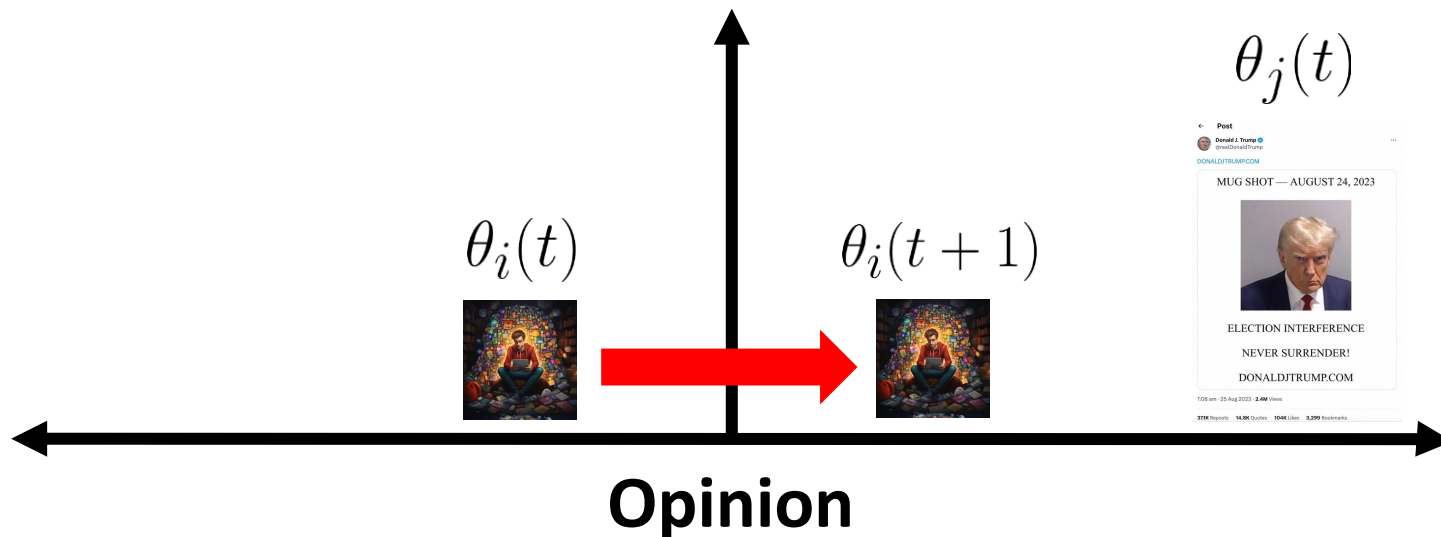
Bounded Confidence Model



Bounded Confidence Model in Large Networks

- For large networks, we can make time continuous because tweets happen so fast

$$\frac{d\theta_i}{dt} = \sum_j \lambda_{ji} f(\theta_j - \theta_i)$$

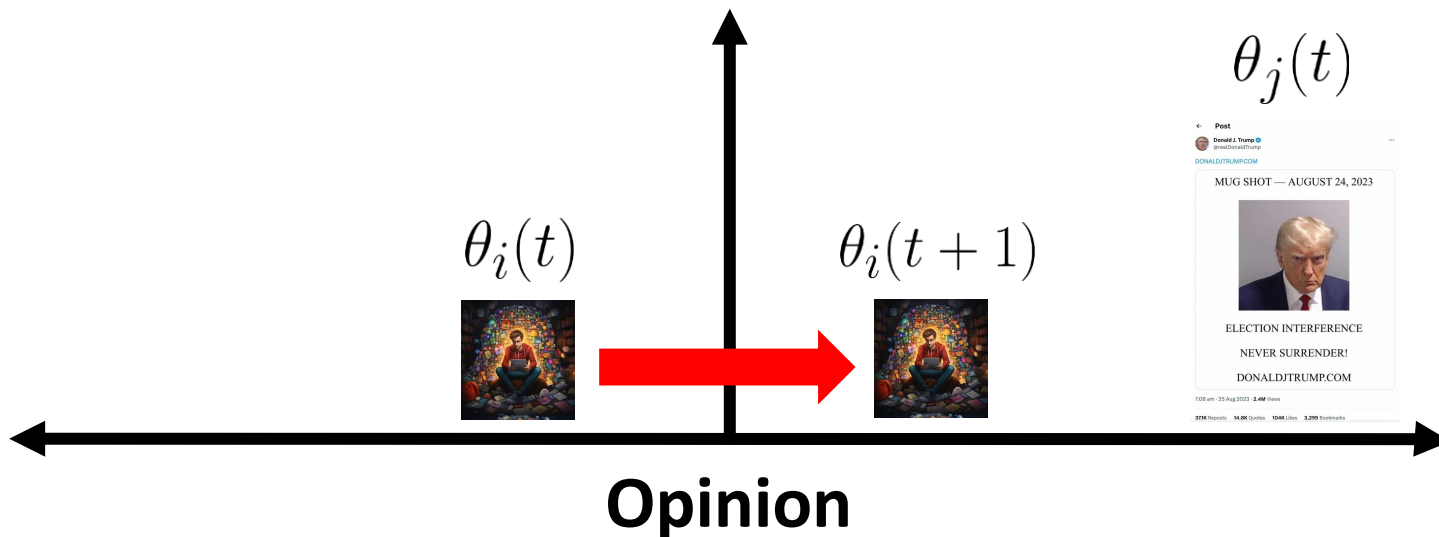


Bounded Confidence Model in Large Networks

- For large networks, we can make time continuous because tweets happen so fast

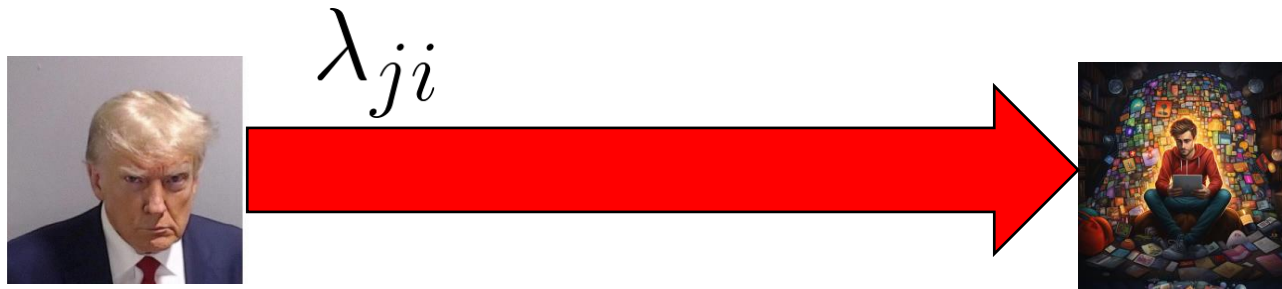
$$\frac{d\theta_i}{dt} = \sum_j \lambda_{ji} f(\theta_j - \theta_i)$$

Opinion shift function



Shadow Banning Control

- Shadow banning reduces the tweet rate on an edge

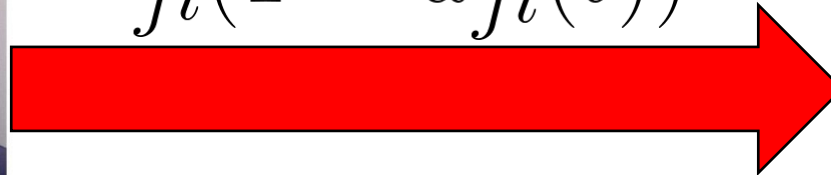


Shadow Banning Control

- Shadow banning reduces the tweet rate on an edge

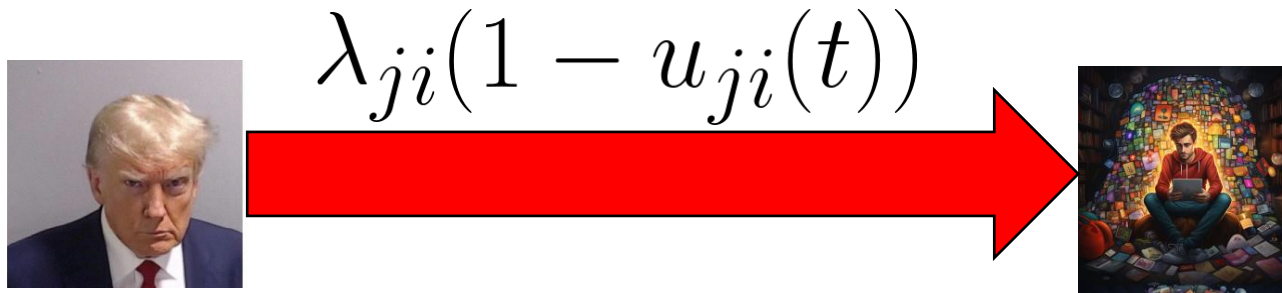


$$\lambda_{ji}(1 - u_{ji}(t))$$



Shadow Banning Control

- Shadow banning reduces the tweet rate on an edge



$u_{ji}(t)$ = shadow banning strength between i and j

Opinion Objectives

- Social media platform wants to maximize some **objective function** of the opinions in the network
- **Opinion mean** – make everyone support a political extreme
- **Opinion variance** – create consensus or polarization

Maximizing Opinion Objective

- Let's call the opinion objective $U(\theta)$
- We will maximize the rate at which the objective increases with respect to the shadow ban strength

$$\max_{u(t)} \frac{dU}{dt}$$

Maximizing Opinion Objective

- Time derivative of objective is linear in the shadow band controls

$$\frac{dU}{dt} = \sum_{i,j} \frac{\partial U}{\partial \theta_i} \lambda_{ji} (1 - u_{ji}(t)) f(\theta_j - \theta_i)$$

Maximizing Opinion Objective

- Time derivative of objective is linear in the shadow band controls

$$\frac{dU}{dt} = \sum_{i,j} \frac{\partial U}{\partial \theta_i} \lambda_{ji} (1 - u_{ji}(t)) f(\theta_j - \theta_i)$$

- We can solve for the shadow band controls for **large networks** very easily (solve a linear program)

Maximizing Opinion Objective

- Time derivative of objective is linear in the shadow band controls

$$\frac{dU}{dt} = \sum_{i,j} \frac{\partial U}{\partial \theta_i} \lambda_{ji} (1 - u_{ji}(t)) f(\theta_j - \theta_i)$$

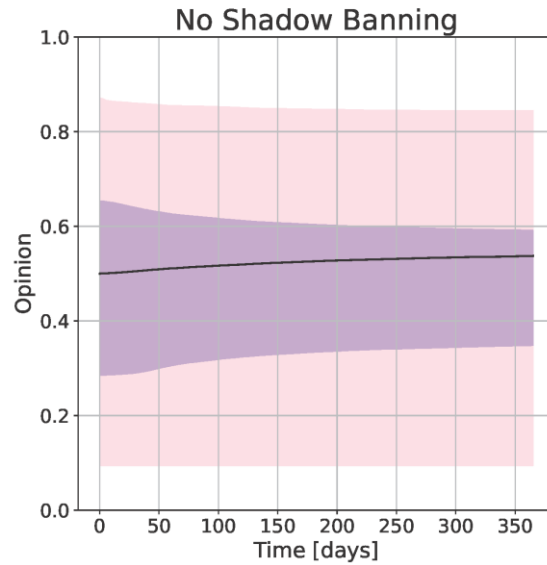
- We can solve for the shadow ban controls for **large networks** very easily (solve a linear program)
- We can even limit the maximum shadow banning strength on the edges

Real Twitter Network

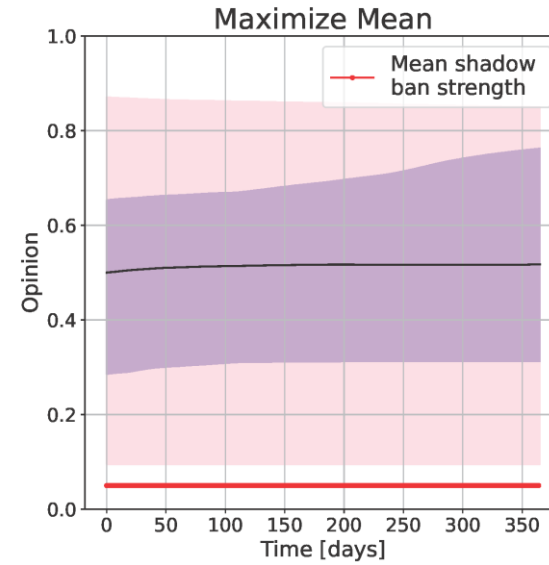
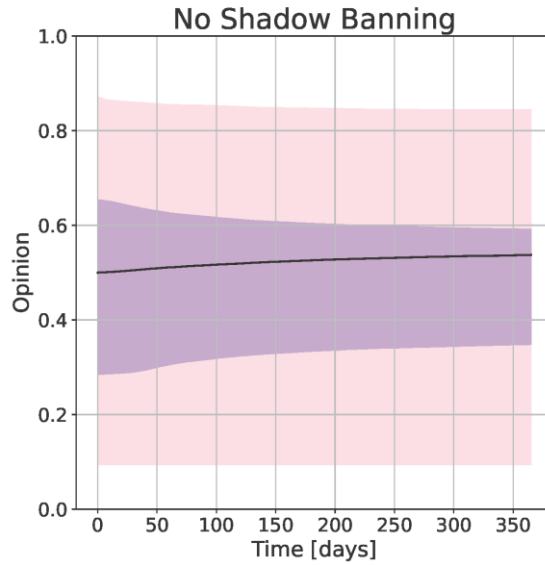
- Twitter follower network of users tweeting about the 2016 US presidential election¹
- Network has 30,000 users and 800,000 edges
 - Tweet rates: λ_{ji}
 - Network structure: $(j, i) \in E$
 - Opinions (measured with a neural network): $\theta_i(0)$

1. N. Guenon des Mesnards. et.al., *Operations Research* (2021)

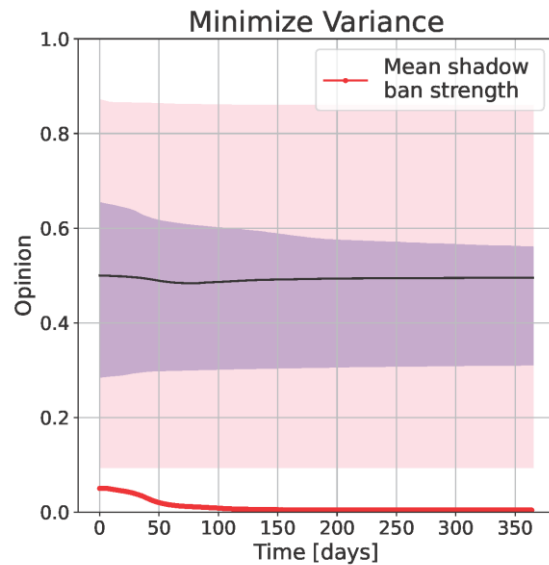
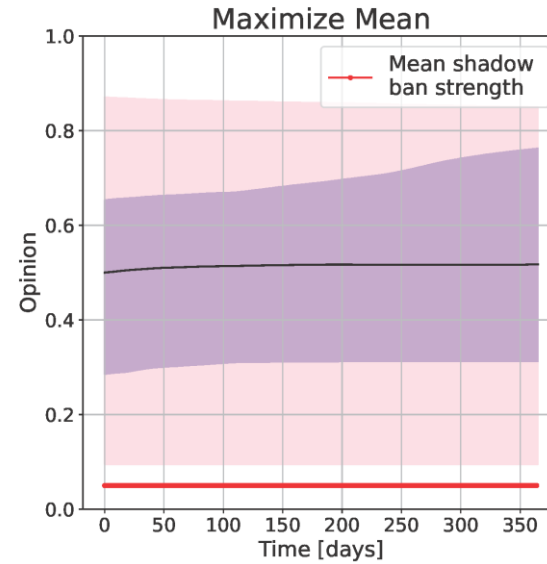
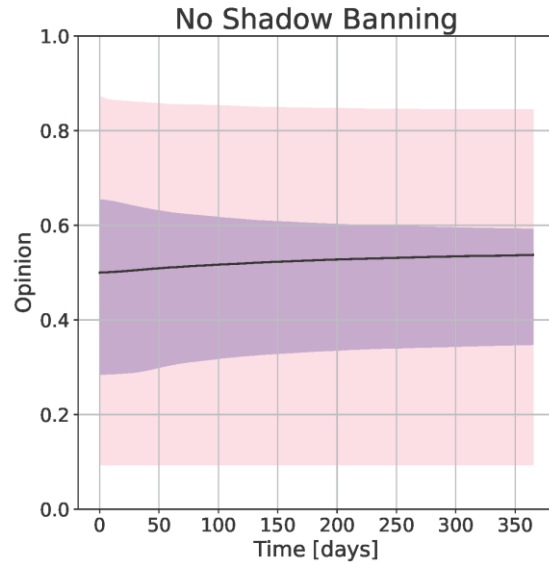
Real Twitter Network



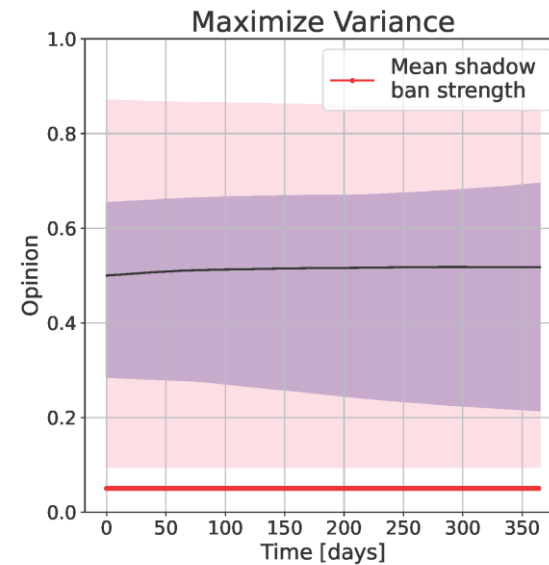
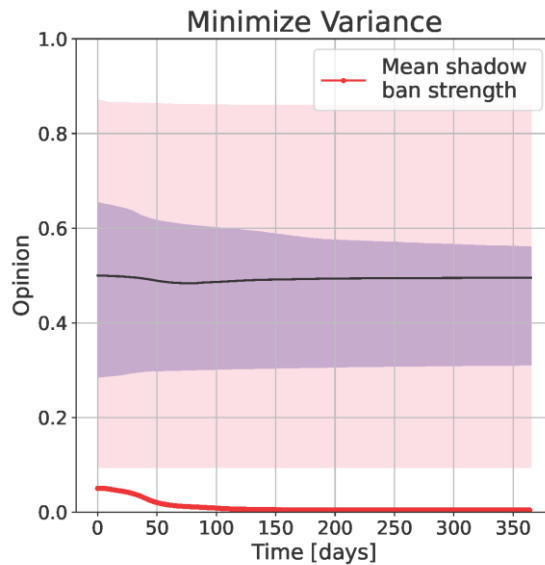
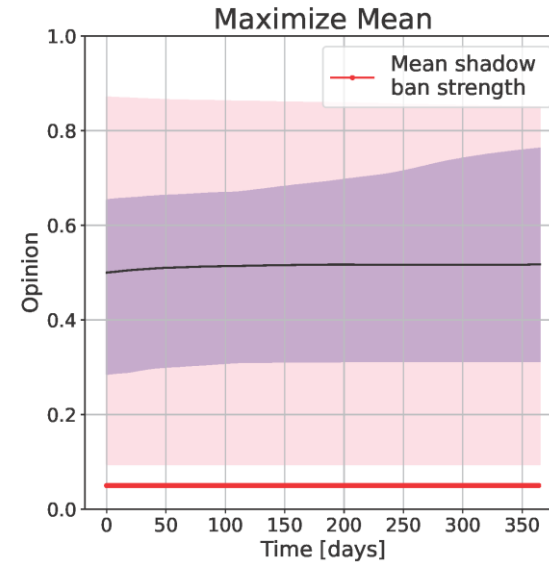
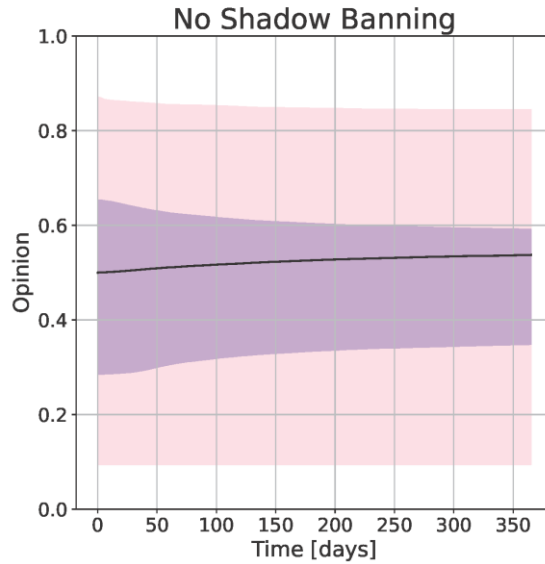
Real Twitter Network



Real Twitter Network



Real Twitter Network



Biased Shadow Banning on Twitter



[Frederic Legrand - COMEO/Shutterstock](#) (Licensed)

Elon Musk railed against shadow bans—now he's using them on his critics

Musk is a self-appointed free speech crusader. Sometimes.



Steven Monacelli

Tech

Posted on Nov 13, 2023 Updated on Nov 13, 2023, 2:24 pm CST

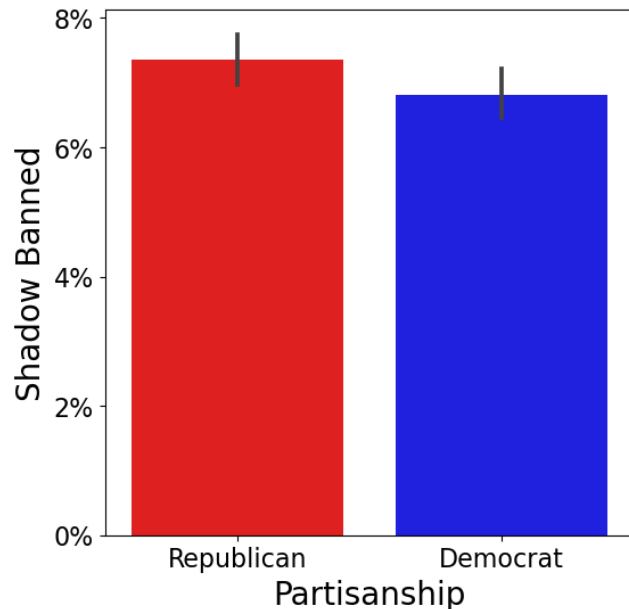
Biased Shadow Banning Looks
Unbiased

Biased Shadow Banning Looks Unbiased

- Consider the mean objective – this favors republicans, so we expect democrats to be shadow banned

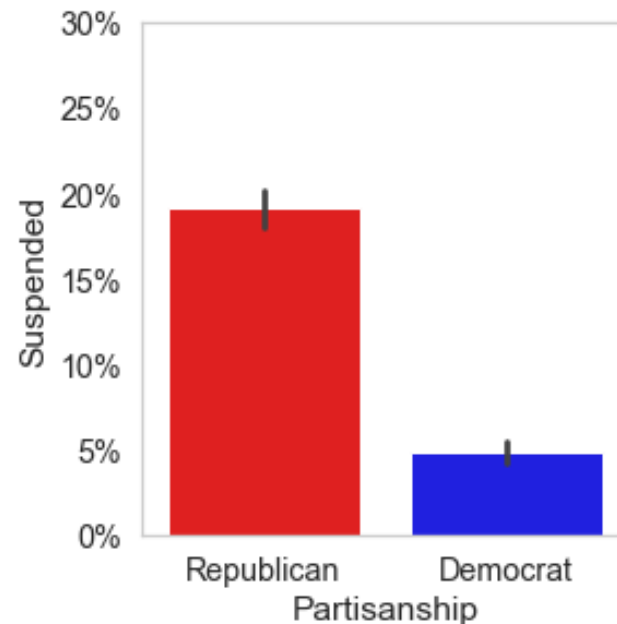
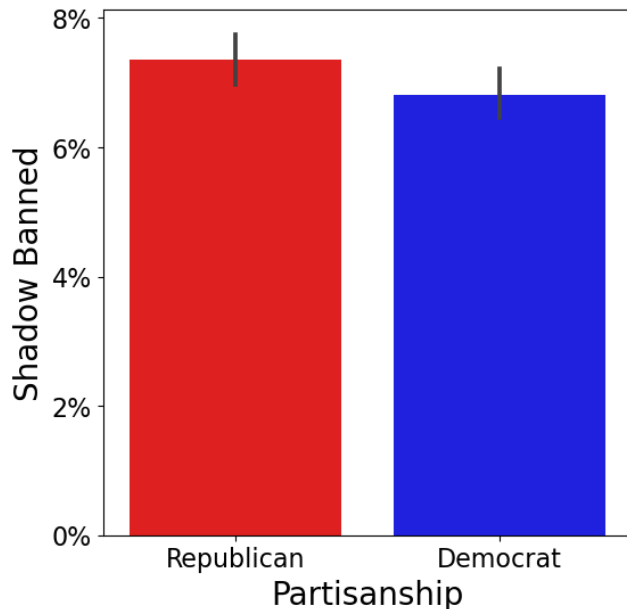
Biased Shadow Banning Looks Unbiased

- Consider the mean objective – this favors republicans, so we expect democrats to be shadow banned
- Surprisingly, the policy shadow bans both parties nearly equally



Biased Shadow Banning Looks Unbiased

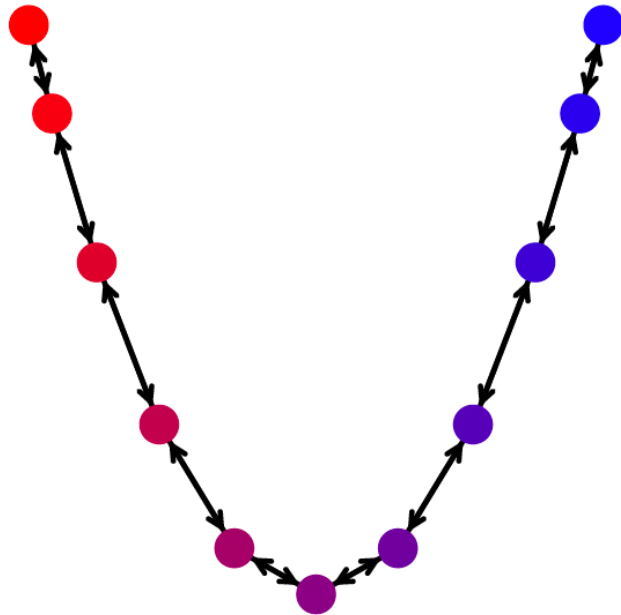
- Consider the mean objective – this favors republicans, so we expect democrats to be shadow banned
- Surprisingly, the policy shadow bans both parties nearly equally
- In comparison, actual suspensions were biased against republicans by 4X



Why Biased Shadow Banning Looks Unbiased

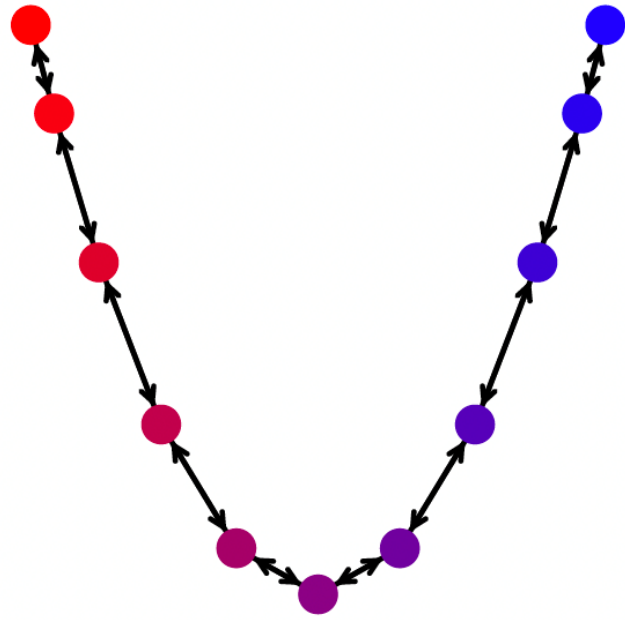
Why Biased Shadow Banning Looks Unbiased

No Shadow Banning

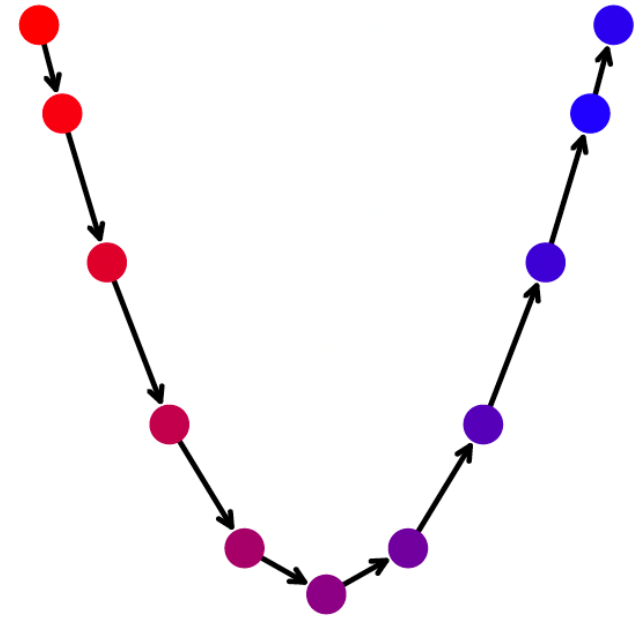


Why Biased Shadow Banning Looks Unbiased

No Shadow Banning

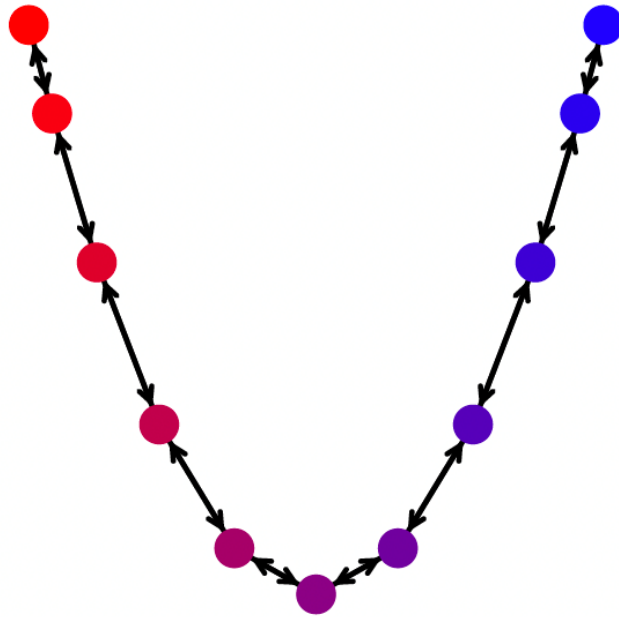


Maximize Mean

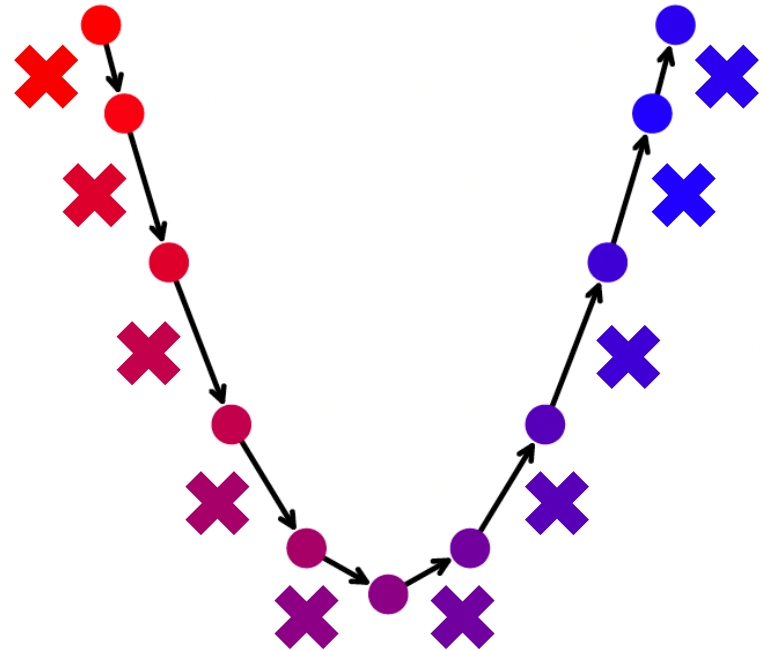


Why Biased Shadow Banning Looks Unbiased

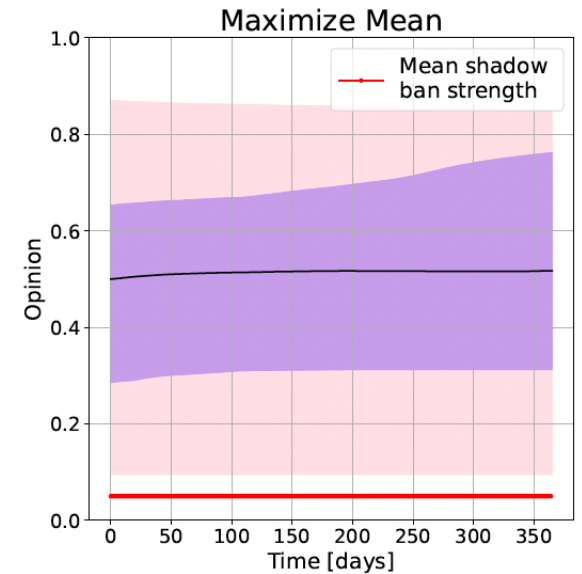
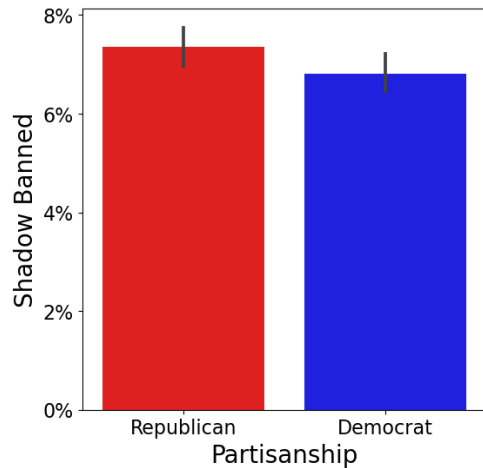
No Shadow Banning



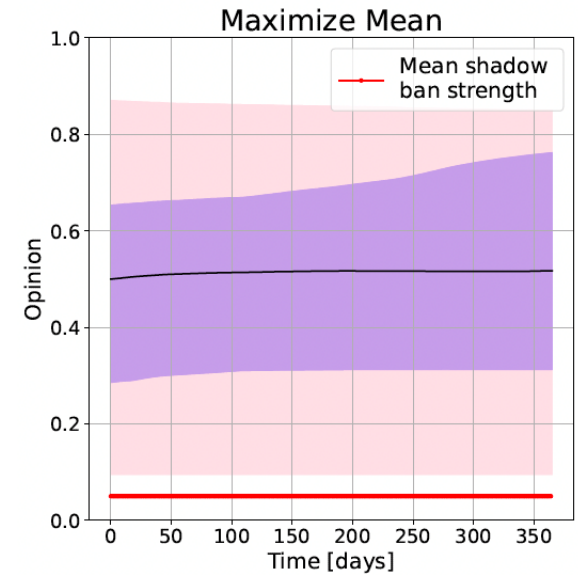
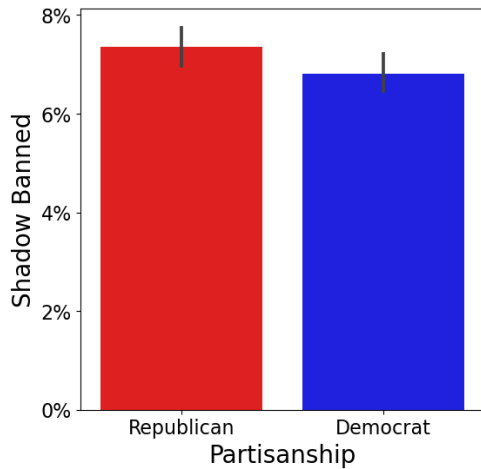
Maximize Mean



Biased Shadow Banning Looks Unbiased



Biased Shadow Banning Looks Unbiased



Conclusion

- Censorship by social media platforms is a serious issue with huge implications for societies and democracies
- Politically biased suspensions may be due to good faith attempts to mitigate disinformation
- Politically biased shadow banning may not look politically biased at all, until its too late