Challenges For Causal Inference On Digital Platforms

Moritz Hardt
Max Planck Institute for Intelligent Systems

Causal stories about algorithmic moderation on social platforms

Filter bubbles

Echo chambers

Political polarization

Radicalization

Amplification

Misinformation

All urgent, fiercely debated problems

Limited theoretical and empirical understanding

Two projects about causal inference on social platforms

Algorithmic amplification of politics on Twitter (2022)
Ferenc Huszár, Sofia Ira Ktena, Conor O’Brien, Luca Belli, Andrew Schlaikjer, MH

Causal inference struggles with agency on online platforms (2022)
Smitha Milli, Luca Belli, MH

Based on two years consulting at Twitter (2019--2021)

Twitter’s Home timeline

What you see when you log on

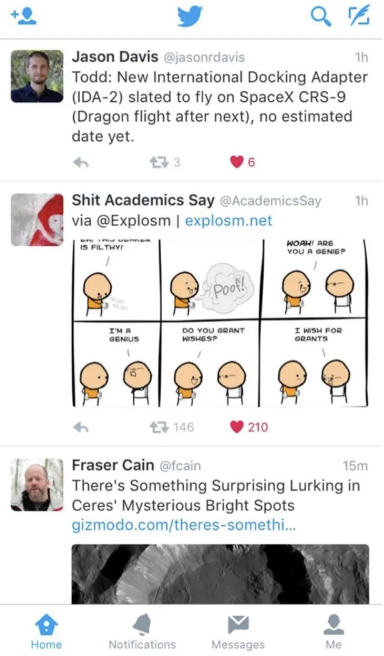
Personalized algorithmic ranking since 2016

Machine learning model trained on various data

Before: Reverse-chronological ordering (and some filtering)

Intense public debate about the effects of algorithmic ranking

Especially in the political context



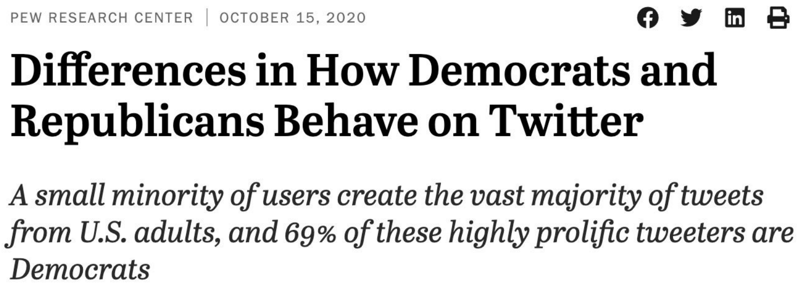
Jason Davis @jasonrdavis 1h
Todd: New International Docking Adapter (IDA-2) slated to fly on SpaceX CRS-9 (Dragon flight after next), no estimated date yet.
3 6

Shit Academics Say @AcademicsSay 1h
via @Explosm | explosm.net
146 210

Fraser Cain @fcain 15m
There's Something Surprising Lurking in Ceres' Mysterious Bright Spots
gizmodo.com/theres-somethi...

Home timeline ca 2016

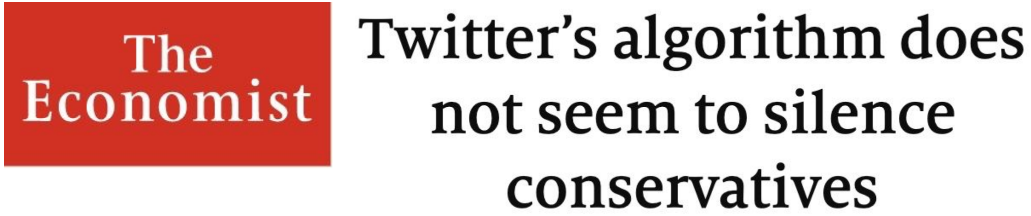
Who is benefitting from the algorithmic timeline?



PEW RESEARCH CENTER | OCTOBER 15, 2020

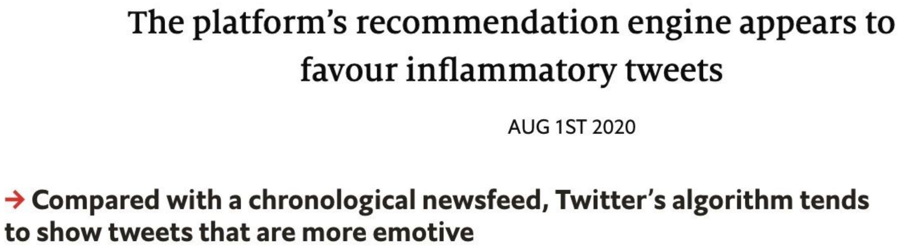
Differences in How Democrats and Republicans Behave on Twitter

A small minority of users create the vast majority of tweets from U.S. adults, and 69% of these highly prolific tweeters are Democrats



The Economist

Twitter's algorithm does not seem to silence conservatives



The platform's recommendation engine appears to favour inflammatory tweets

AUG 1ST 2020

→ **Compared with a chronological newsfeed, Twitter's algorithm tends to show tweets that are more emotive**

False Accusation:

The Unfounded Claim that Social Media Companies Censor Conservatives

PAUL M. BARRETT AND J. GRANT SIMS

Question: Does algorithmic personalization cause an advantage along established political lines?

Experimental setup starting in 2016

Control group: Randomly chosen 1% of all global users assigned reverse-chronological timeline

Treatment group: Randomly chosen 4% of all global users assigned new algorithmically ranked control group

Primarily used for product tweaks over the years

Idiosyncrasies of the experimental setup

Network effects (“violation of SUTVA”)

Control group mostly sees content written by non-control users

Treatment changes over time (updates to algorithm)

Control changes over time (safety filters etc)

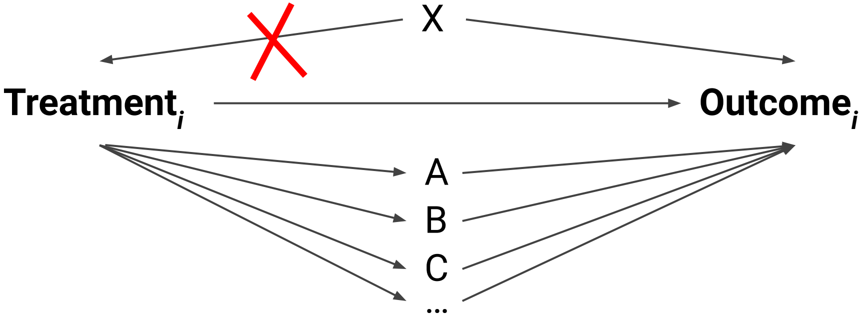
Twitter used experimental setup for tweaking the platform

Hodge-podge causal effects

Randomization breaks confounding

But: *All mediators at play simultaneously* (hodge-podge causal effect)

- Network effects active
- How well different actors strategically respond to algorithmic timeline
- Twitter’s own optimizations based on experimental setup



Different mechanisms have different political, moral, and sociological meaning.

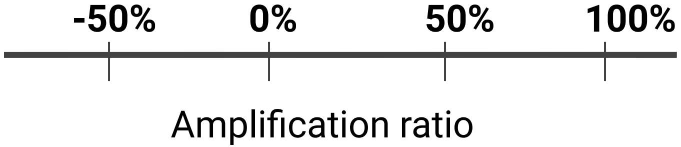
Defining and measuring algorithmic amplification

Amplification ratio of a set *T* of tweets in a set *U* of users:

> number of treatment users in *U* who encountered a tweet in *T* *divided by*
> number of control users in *U* who encountered a tweet in *T*

Example: *U* is all German Twitter users, *T* is all tweets by politicians of the CDU in from April 1, 2020 to August 15, 2020.

Normalize amplification ratio so that **0% is equal proportion**, i.e.,
random user from *U* in treatment is just as likely to see a tweet in *T*
as a random user from *U* in control.



Scope of algorithmic audit

Millions of Tweets from individual politicians

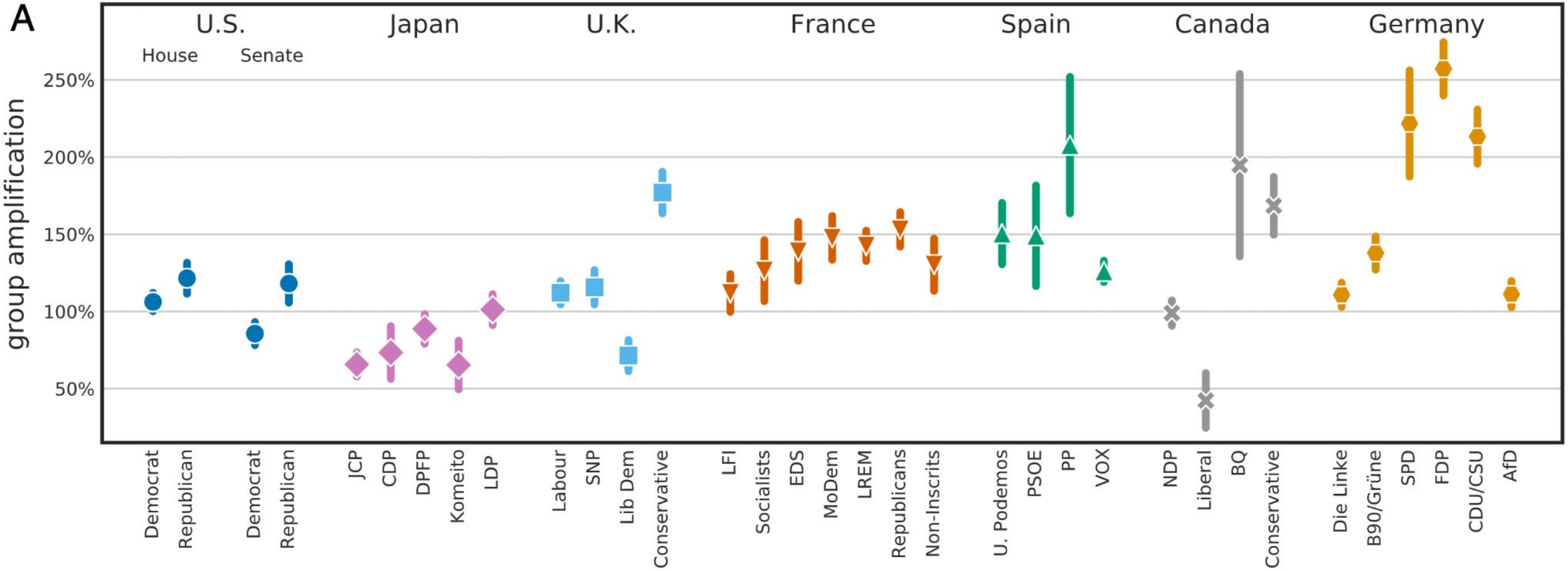
Fine-grained analysis of the major political parties in seven countries

Canada, France, Germany, Japan, Spain, U.K., U.S.

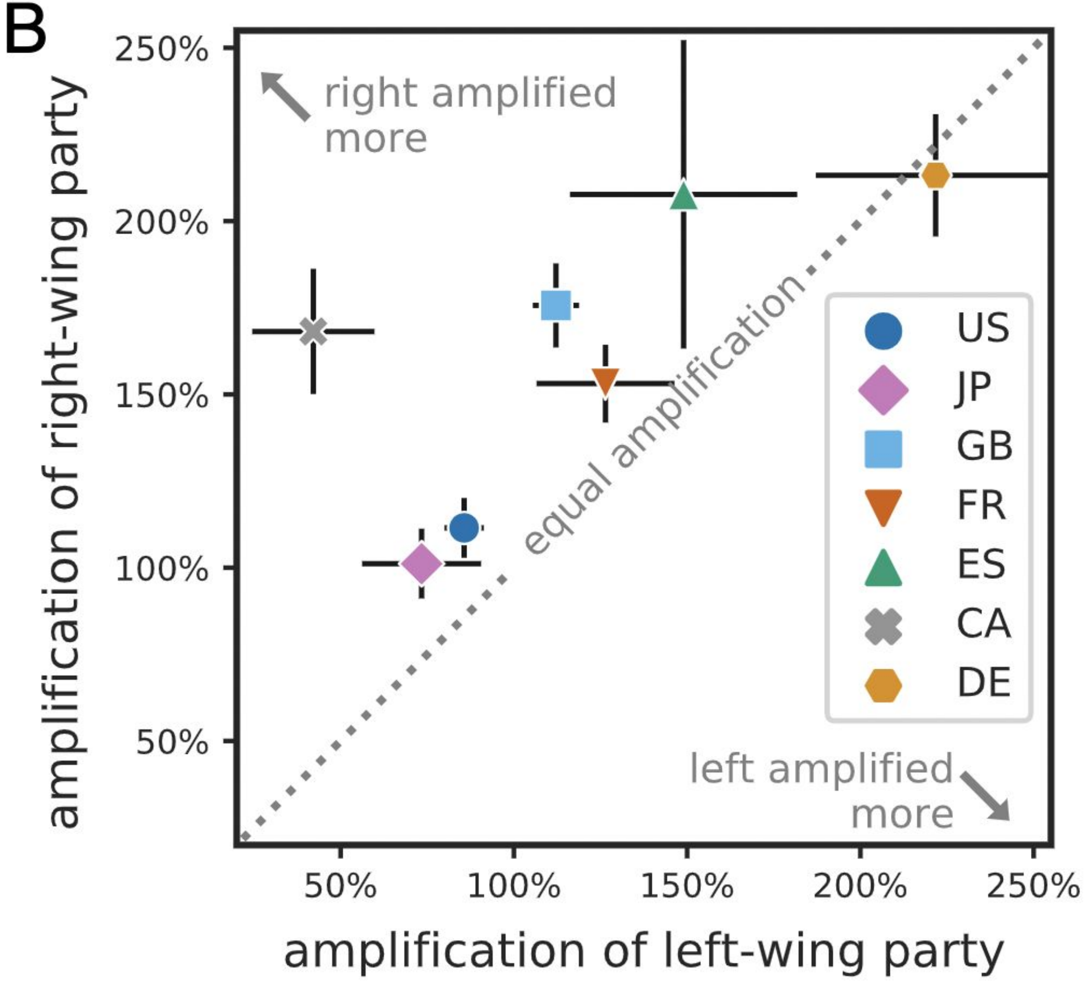
6.2 million news articles shared in the United States

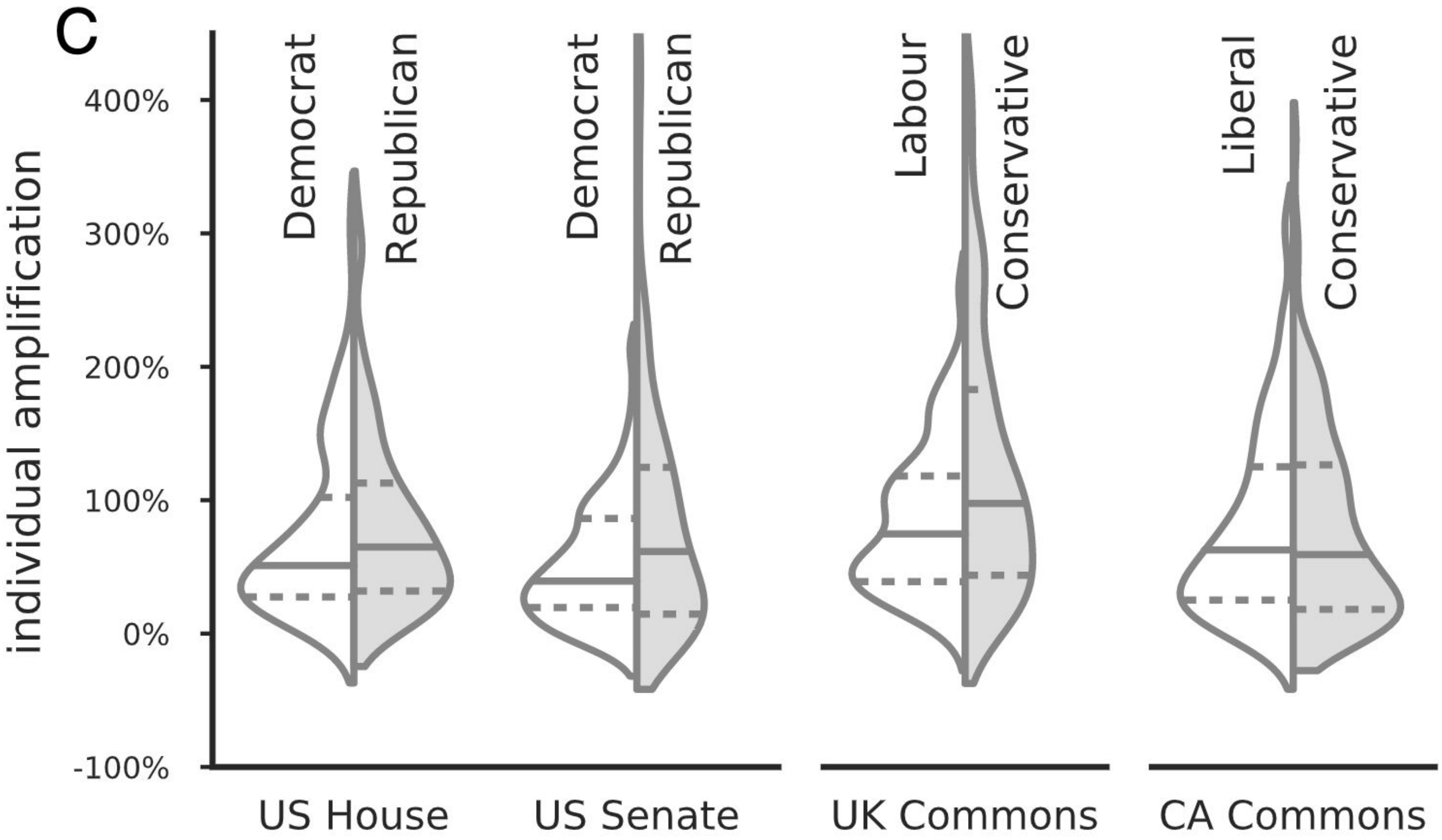
Tweets from April through August 2020

Algorithmic amplification for each party

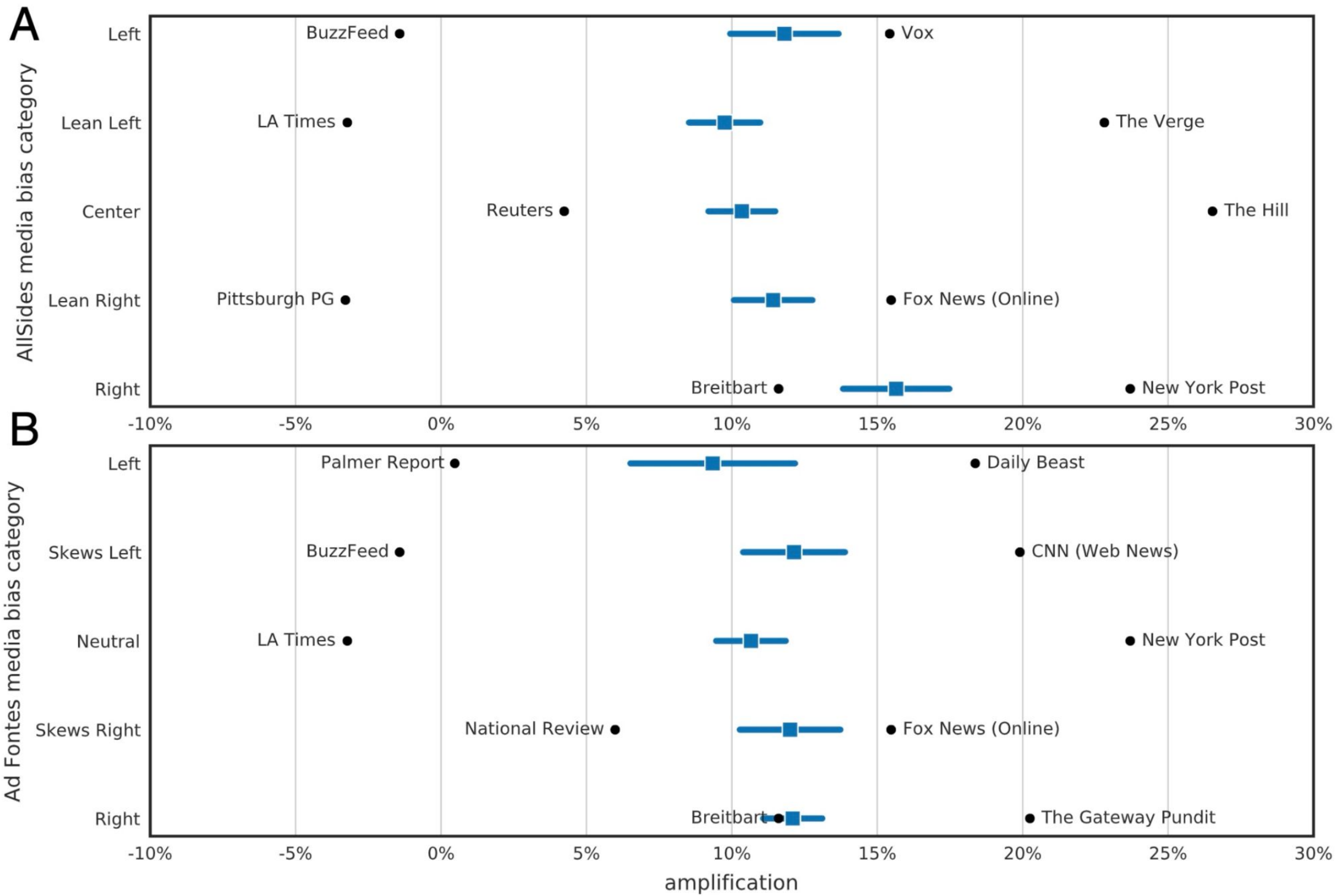


Right versus left of the spectrum





C
individual amplification
400%
300%
200%
100%
0%
-100%
Democrat
Republican
Democrat
Republican
Labour
Conservative
Liberal
Conservative
US House
US Senate
UK Commons
CA Commons


A
AllSides media bias category
Left
Lean Left
Center
Lean Right
Right
BuzzFeed
Vox
LA Times
The Verge
Reuters
The Hill
Pittsburgh PG
Fox News (Online)
Breitbart
New York Post
-10%
-5%
0%
5%
10%
15%
20%
25%
30%
B
Ad Fontes media bias category
Left
Skews Left
Neutral
Skews Right
Right
Palmer Report
Daily Beast
BuzzFeed
CNN (Web News)
LA Times
New York Post
National Review
Fox News (Online)
Breitbart
The Gateway Pundit
-10%
-5%
0%
5%
10%
15%
20%
25%
30%
amplification

Discussion

- Across seven countries, right-wing parties benefit as much, and often more, from algorithmic personalization than left-wing parties
- US media outlets with a right-leaning bias are amplified marginally more
- Among individual politicians party membership is not strongly associated with amplification
- Study does not pin down mechanism(s) behind the effect
  - Growing evidence that different parties utilize Twitter differently, e.g., Parmelee, Bichard (2011), Freelon, Marwich, Kreiss (2020)
- Focus is on relative differences among parties and politicians, not the question whether we'd be better off with chronological timeline for everyone.

Agency and control on social platforms

User choice and controls

A common response to concerns with algorithmic moderation:
“Let’s give users more control over what they see and how they see it.”

Twitter offers numerous user controls, including:

- Personalized push notifications
- Personalized email notifications
- Personalized algorithmic timeline
- Quality filter

All opt-in/treatment by default, but users can opt-out

Understanding causal effects of user agency

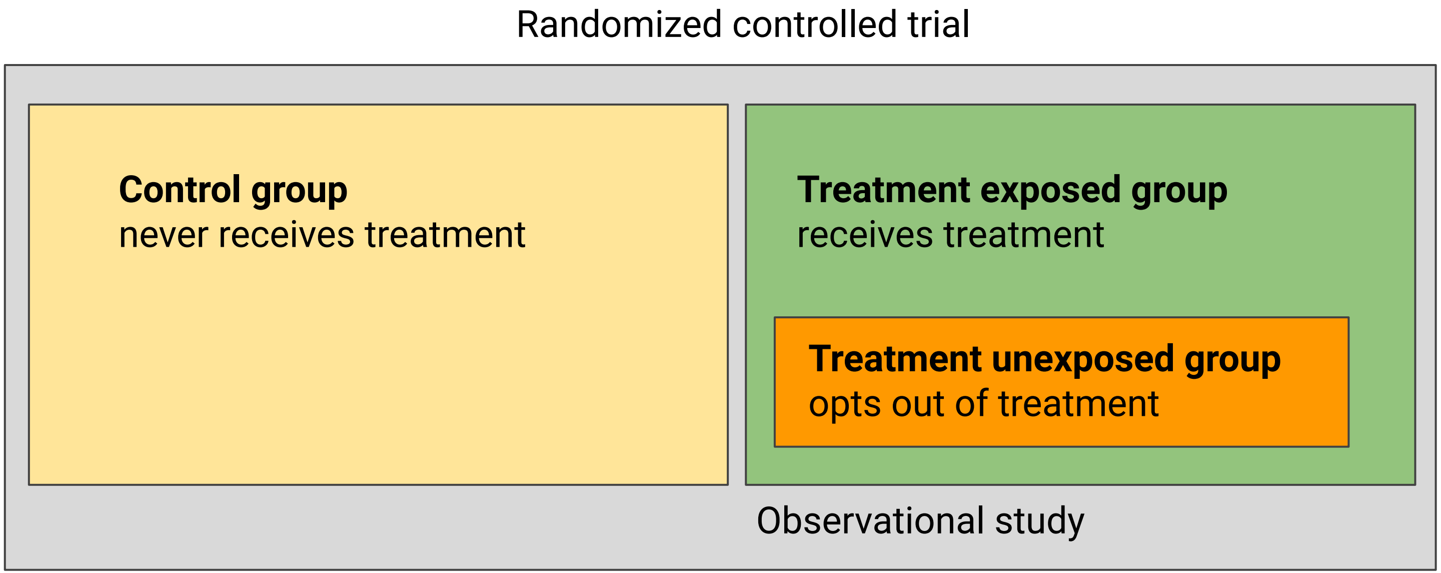
Some users opting out gives us data about both treatment and control.

1. Can we estimate the causal effect of opt-in from observational data?
2. Do randomized experiments (A/B tests) anticipate the effect that opt-out has on those who choose to do so?

Positive answer to (1) would allow us to avoid costly, and possibly unethical, randomized experiments

Positive answer to (2) would allow us to anticipate effects of offering user control

Experimental setup: Within-study comparison



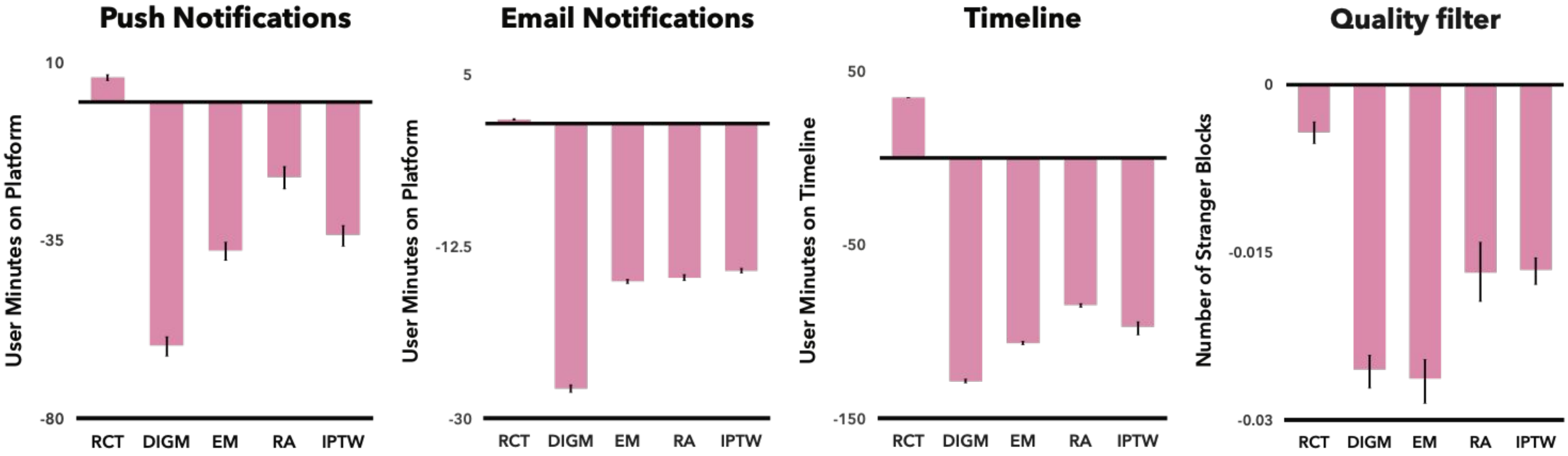
Similar to setup in Gordon, Zettelmeyer, Bhargava, Chapsky (2019) study on failure of observational methods in the context of Facebook ads

Scope of study

Four large-scale within-study comparison of experimental and observational causal inference on the Twitter platform

Four user settings: Push notifications, email notifications, algorithmic timeline, quality filter

Four standard observational causal methods

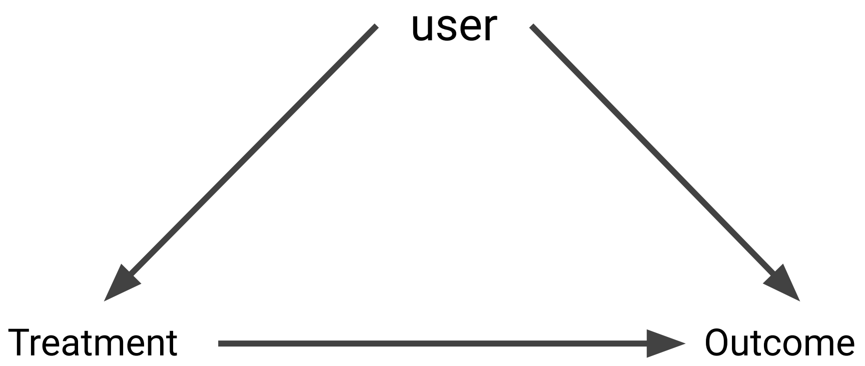


RCT = Randomized controlled trial
DIGM = difference in group means
EM = exact matching
RA = regression adjustment
IPTW = inverse probability of treatment weighting

Persistent confounding

Adjusted for 14 available variables that could be confounding behavior

```
“Power user”
“Privacy sensitive”
“Restricted account”
```



What drives user choice is poorly described by observable user features.

Caveat: Impossible to rule out that there could be an observational design that works.

Catch 22

Platforms enable user controls, because human behavior is complex and hard to predict from observable features.

This difficulty of predicting user agency makes it hard to deconfound treatment in an observational study.

Example: Propensity score Pr(Treatment | user observables) asks us to predict user agency from observable features about the user.

Conclusions and challenges

Causal effects of algorithms in social systems

Why is it so hard to understand the causal effects of algorithms?

Methodologically, not just micro, also macro:

- RCTs surface valuable empirical understanding
- RCTs alone tell us what the dynamics are that bring macroscopic changes

Microfoundations for algorithmic decisions:

- How do individuals respond to algorithmic decisions?
- We currently lack adequate microfoundations for algorithmic decisions, cf., Mendler-Dünner, Jagadeesan, *H* (2021)

Broader directions

More theoretical/conceptual work should provide definitions that clarify hypothesized causal mechanisms

More empirical work should attempt to test and establish causal relationships

What causal questions do we want to answer?

What experiments do we want platforms to conduct?

Thank you.