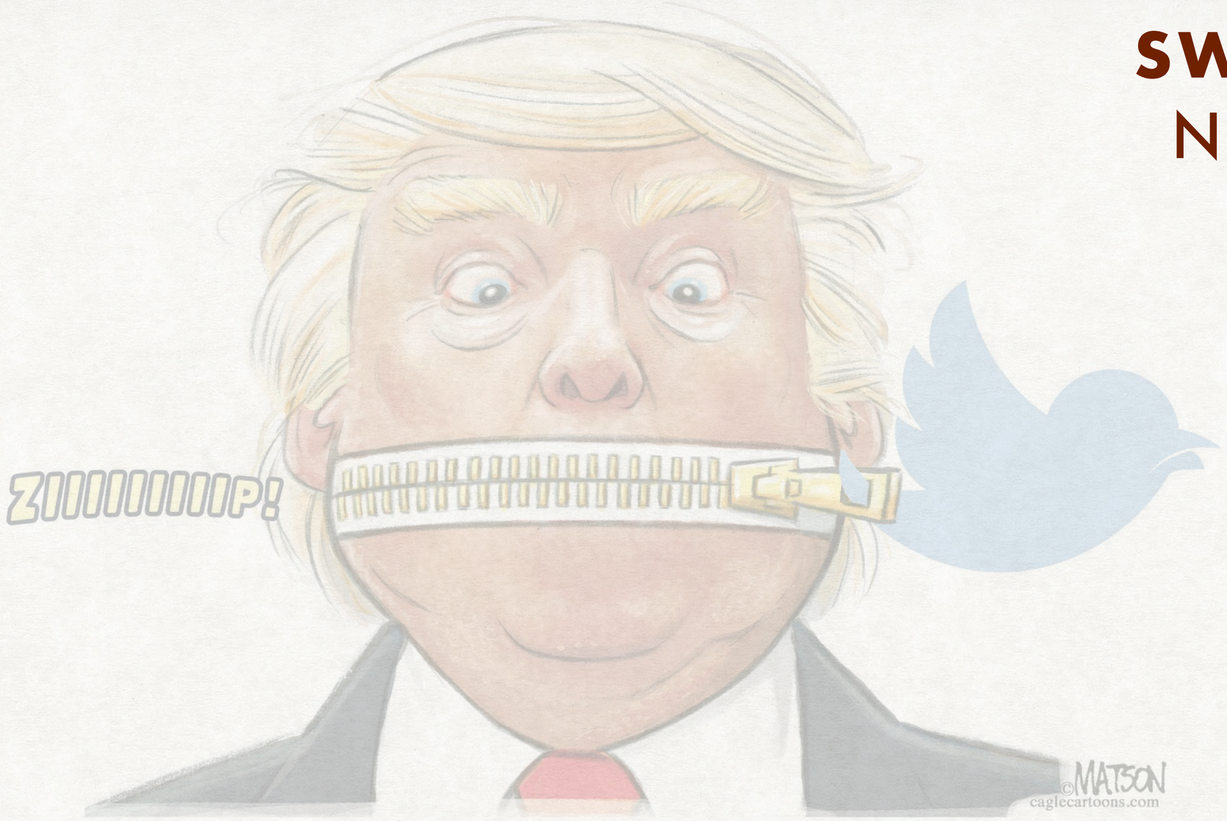
SWAPNEEL MEHTA
NYU CDS, CSMAP



ESTIMATING THE CAUSAL EFFECT OF TWITTER'S INTERVENTIONS ON ENGAGEMENT WITH TRUMP'S TWEETS

Image courtesy R.J. Matson

ABOUT ME

mehtaver.se

2019 - 23 Ph.D. @ NYU Data Science, Center for Social Media + Politics

Social Networks, Causal Inference, Probabilistic Models

2022 Ph.D. ML Engineering Intern at Twitter

Civic Integrity, Misinformation

2020 - 21 Data Science Research Intern at Adobe

Trending Hashtag Recommendation for Videos

2018 - 19 ML + Physics at the European Org. for Nuclear Research

VIRAL DISINFORMATION...

Conspiracy Theories About Facebook Outage Spread Even Without Facebook

Some people believe the hourslong outage may be linked to a supposed data breach that is, most likely, actually a scam.

Fact Check › Fake News

Nope Francis

Reports that His Holiness has endorsed Republican presidential candidate Donald Trump originated with a fake news web site.

Dan Evon
Updated: Jul 24, 2016

SHARE 69.7K

Nicki Minaj
@NICKIMINAJ

My cousin in Trinidad won't get the vaccine cuz his friend got it & became impotent. His testicles became swollen. His friend was weeks away from getting married, now the girl called off the wedding. So just pray on it & make sure you're comfortable with ur decision, not bullied

5:44 PM · Sep 13, 2021 · Twitter for iPhone

26K Retweets **94.1K** Quote Tweets **151.9K** Likes

GAMING THE ELECTION —

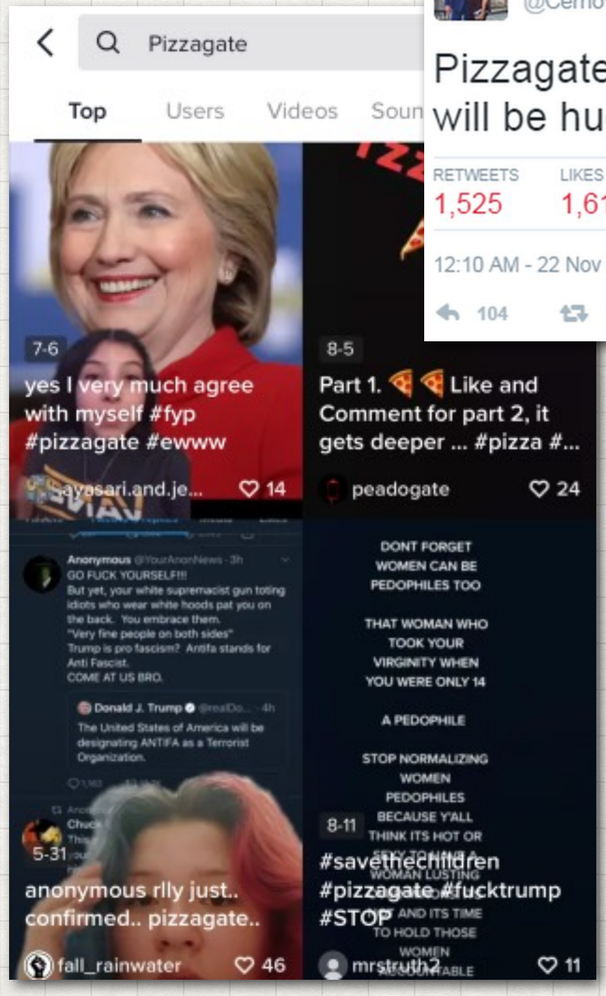
"Hacker X"—the American who built a pro-Trump fake news empire—unmasks himself

He was hired to build a fake news op but now wants to put things right.

AX SHARMA - 10/14/2021, 8:00 AM

Nicki Minaj, Ars Technica, Vox, Vice, SciAm, Jimmy Kimmel Live

...HAS REAL-WORLD CONSEQUENCES!



Man Dead From Taking Chloroquine Product After Trump Touts Drug For Coronavirus

Tara Haelle Senior Contributor
Healthcare
I offer straight talk on science, medicine, health and vaccines.

One man told CNN that in a pharmacy near his home on the Lagos mainland, he witnessed the price rise by more than 400% in a matter of minutes.

CNN, Forbes, TechCrunch, Mike Cernovich, NPR

PLATFORMS ARE TRYING INTERVENTIONS...

Technology

Twitter is sweeping out fake accounts like never before, putting user growth at risk

Twitter suspended more than 70 million accounts in May and June, and the pace has continued in July

Unprecedented Facebook URLs Dataset now Available for Academic Research through Social Science One

February 13, 2020

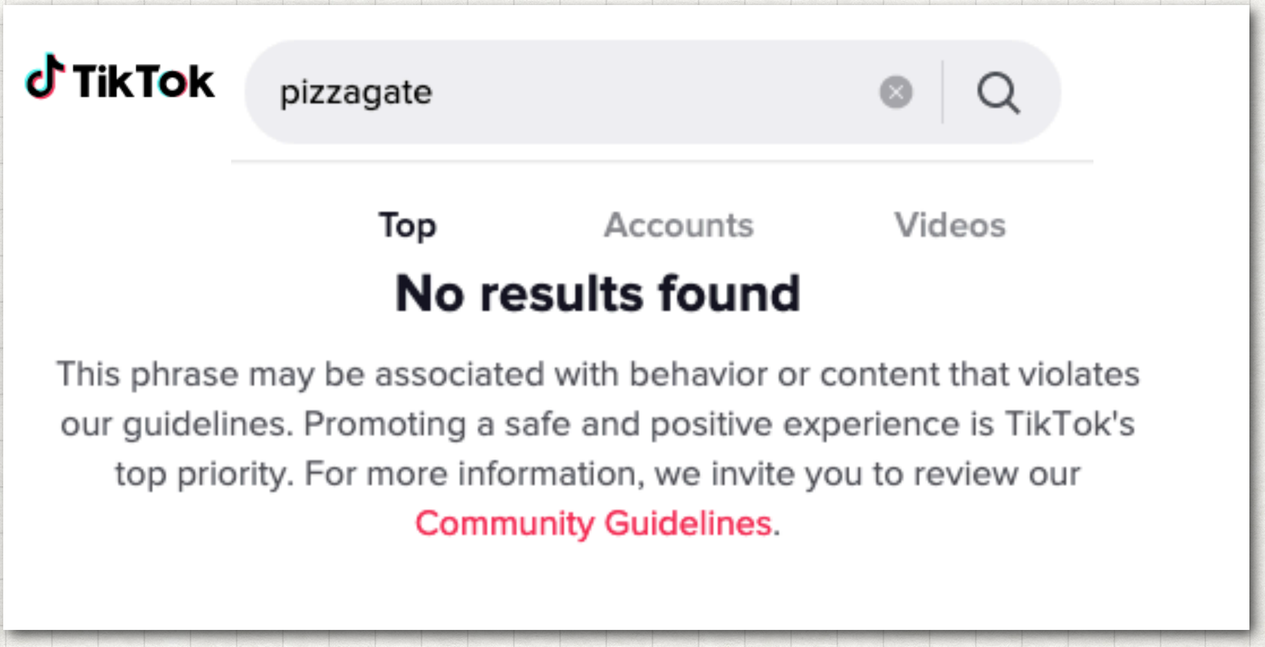
Gary King and Nathaniel Persily

Meta

New Facebook and Instagram Research Initiative to Look at US 2020 Presidential Election

August 31, 2020

By Nick Clegg, VP of Global Affairs and Communications; Chaya Nayak, Head of Facebook's Open Research and Transparency Team



Social Science One + FB, 2020 Election Research, Washington Post

...WITH LITTLE SUCCESS

PEER REVIEWED

Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform

nytimes.com

A Genocide Incited on Facebook, With Posts From Myanmar's Military

Paul Mozur

In India, Facebook Struggles to Combat Misinformation and Hate Speech

Sheera Frenkel, Davey Alba

On TikTok, audio gives new virality to misinformation

The Institute for Strategic Dialogue analyzed 124 TikTok videos featuring vaccine misinformation that garnered more than 20 million views and 2 million likes, comments and shares.

NYT, MIT, NBC News, Sanderson+, (2021)

HOW TO ADDRESS THIS?

HOW TO ADDRESS THIS?

DEBUG POLICY INTERVENTIONS!

WHICH POLICY INTERVENTIONS?

WHICH POLICY INTERVENTIONS?

- Tweets that violate terms of service are ‘intervened upon’ by the platform in various ways

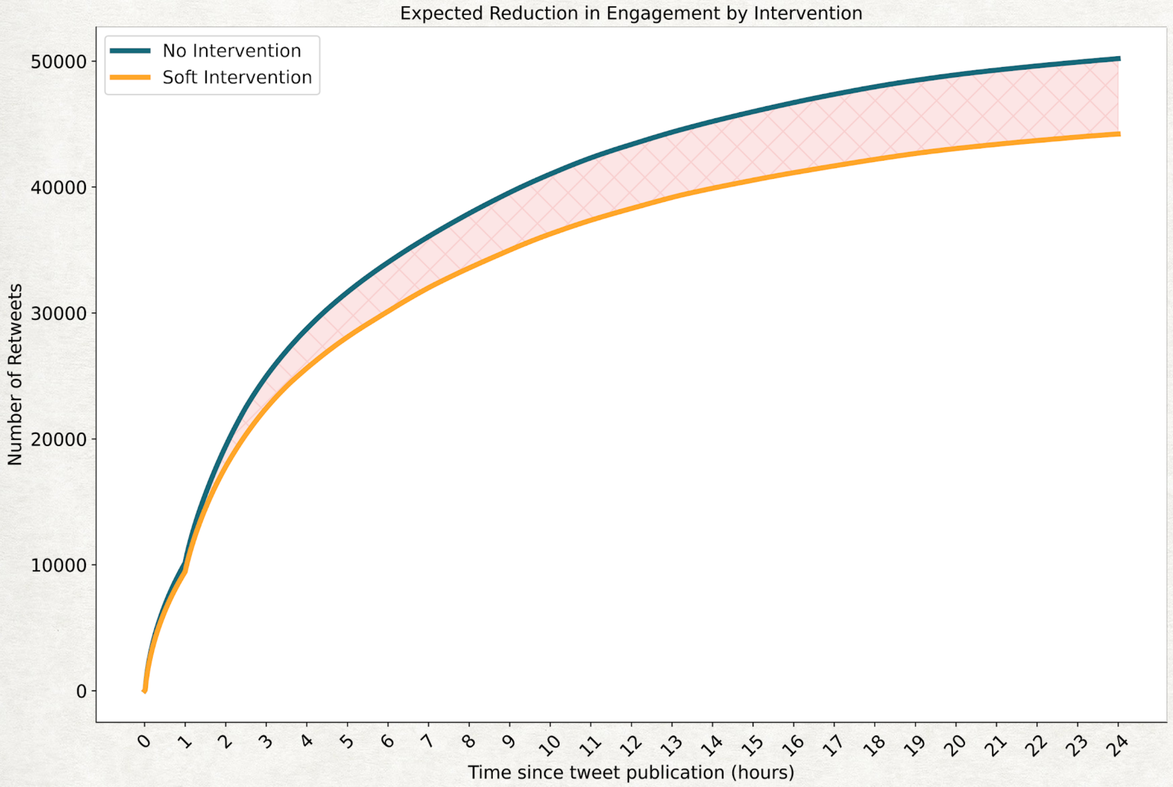
WHICH POLICY INTERVENTIONS?

- Tweets that violate terms of service are ‘intervened upon’ by the platform in various ways
- Two types of interventions: **warning labels** and **removal**

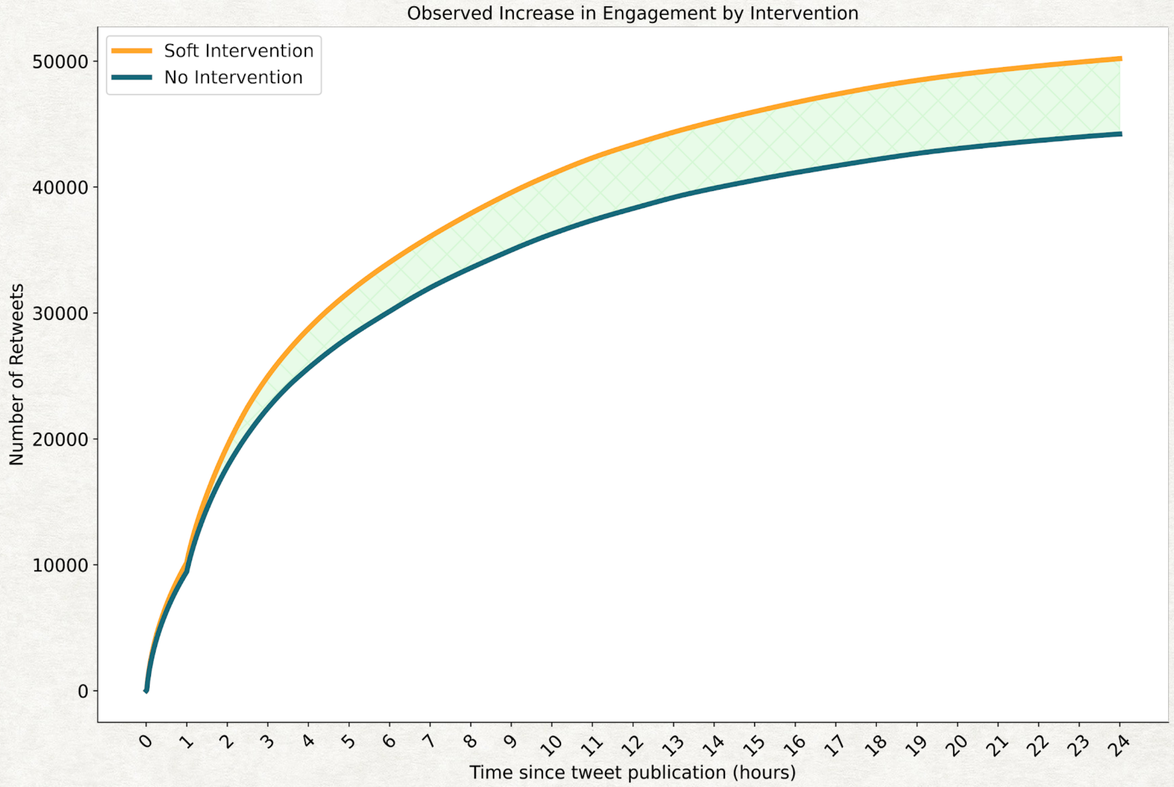
WHICH POLICY INTERVENTIONS?

- Tweets that violate terms of service are ‘intervened upon’ by the platform in various ways
- Two types of interventions: **warning labels** and **removal**
- Not much literature available on cross-platform causal effects of such policies despite their widespread deployment across politics and public health

WHAT IS THE IMPACT OF INTERVENTIONS?



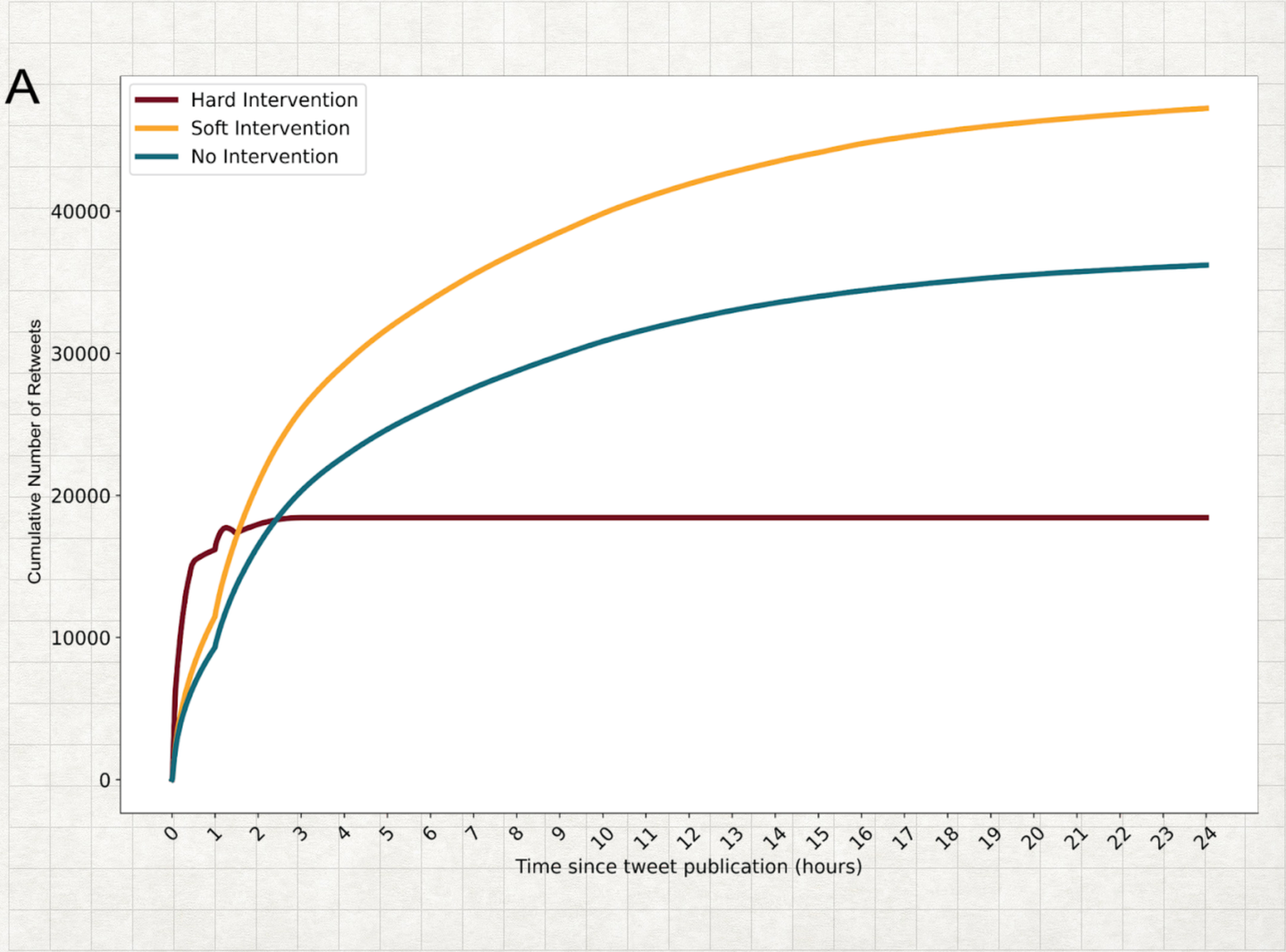
Expectation: Reduced Engagement



Reality: Increased Engagement

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

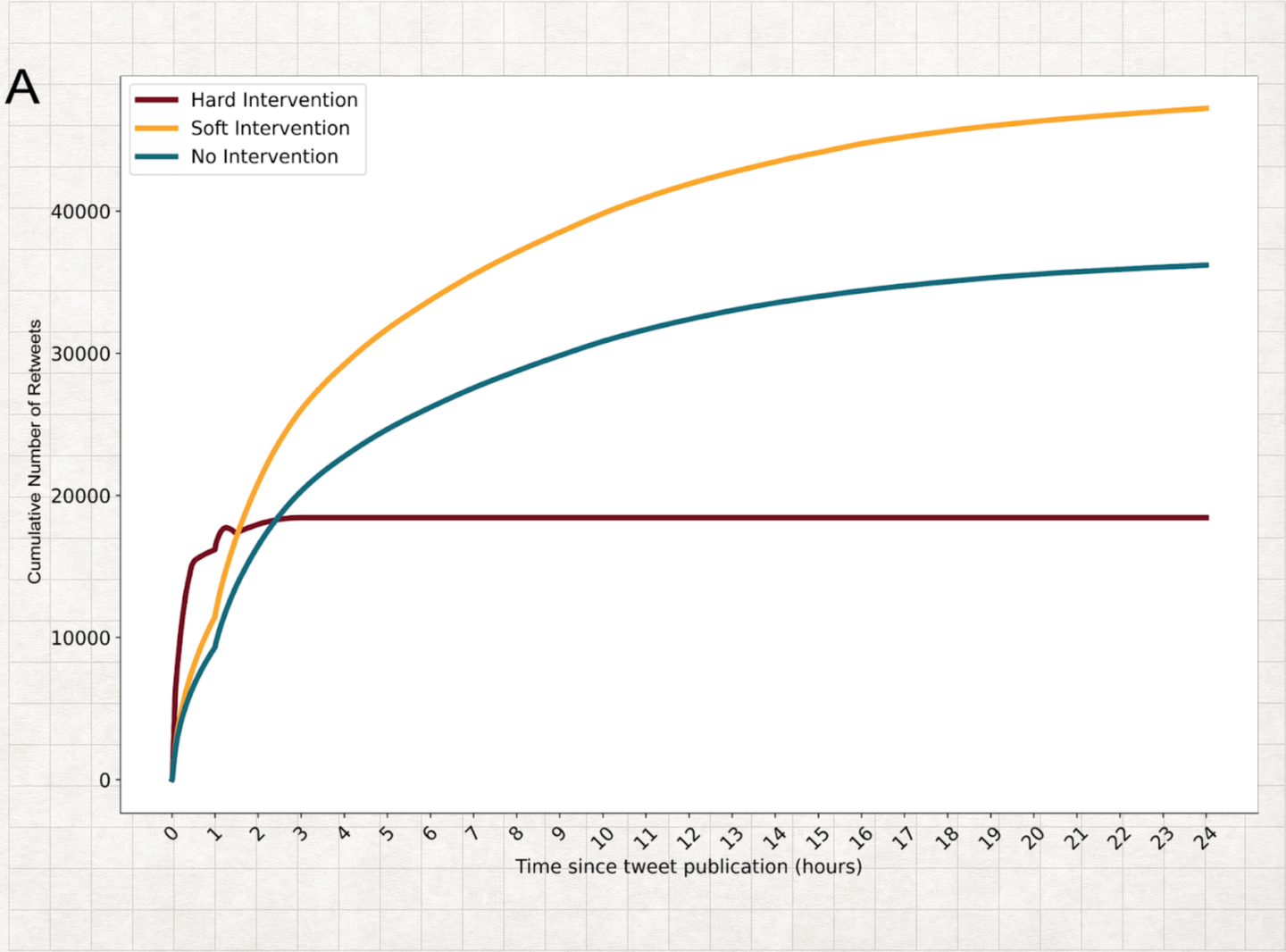
CAUSAL EFFECTS



HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

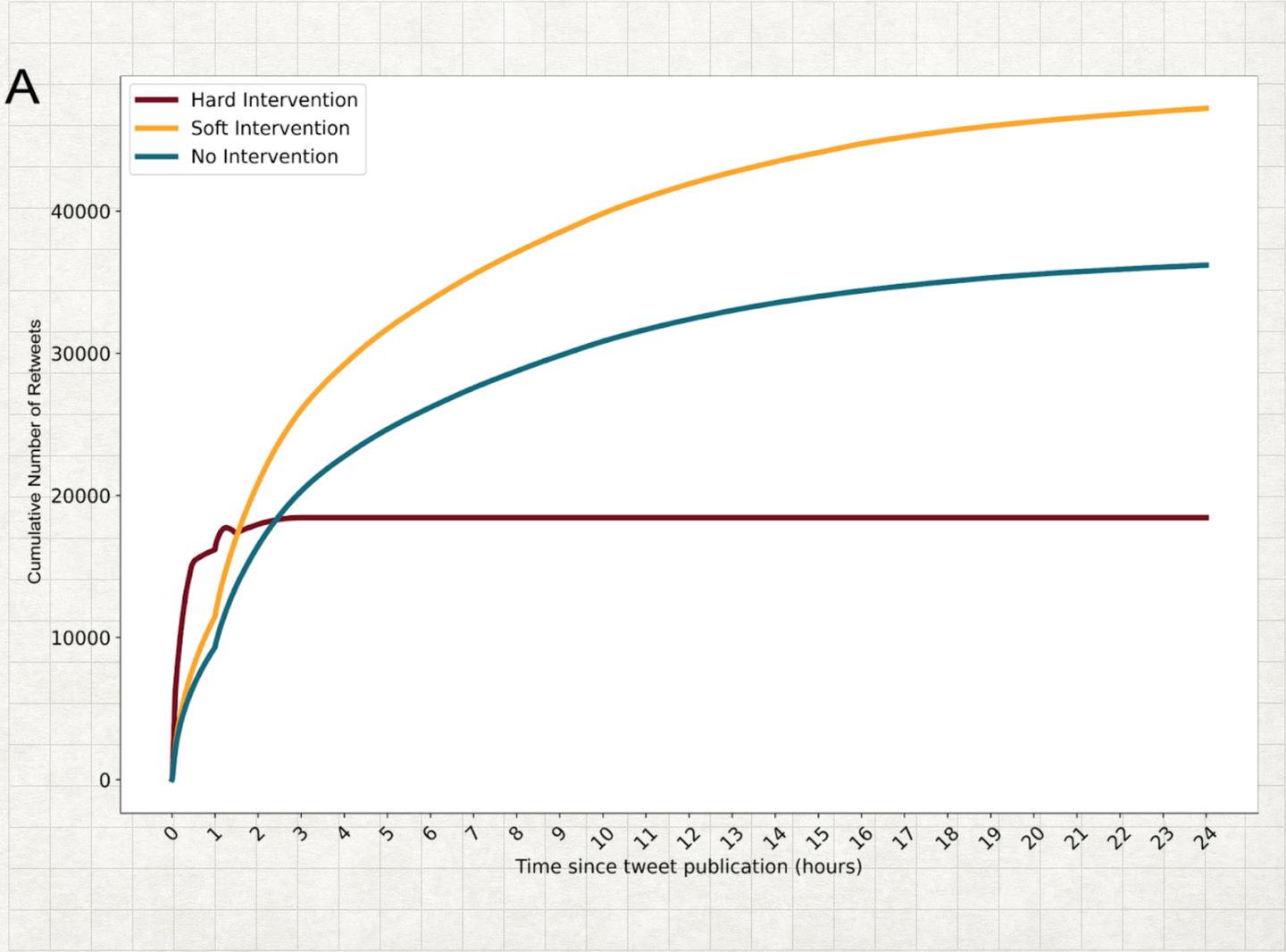
- Naive estimates by Sanderson et. al (2021), indicate strong Streisand effect!



HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

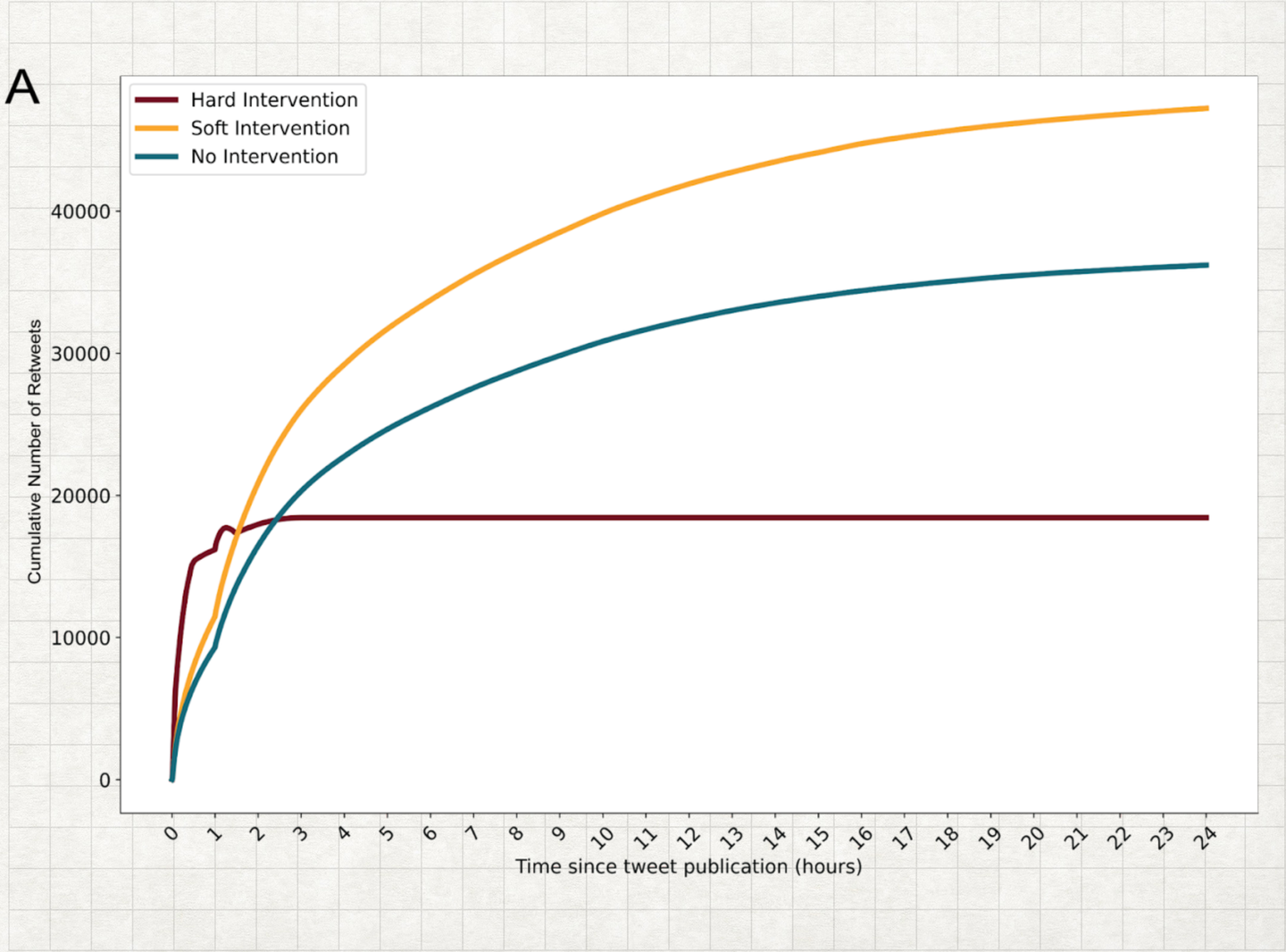
- Naive estimates by Sanderson et. al (2021), indicate strong Streisand effect!
- **Wait does this mean interventions are bad?!**



HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- Naive estimates by Sanderson et. al (2021), indicate strong Streisand effect!
- **Wait does this mean interventions are bad?!**
- What tweets are we really comparing here and what are their features like?



HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

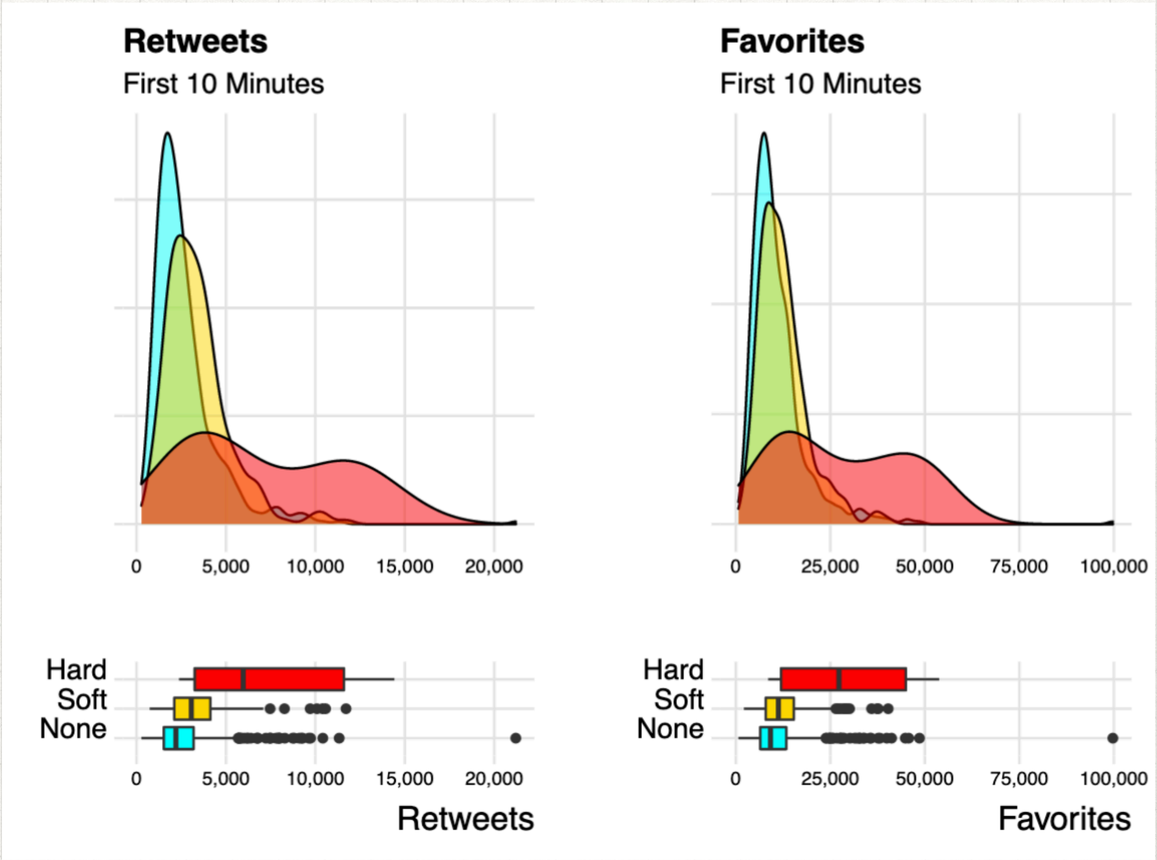
CAUSAL EFFECTS

- The intervened and non-intervened tweets are very different -> biased estimate

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

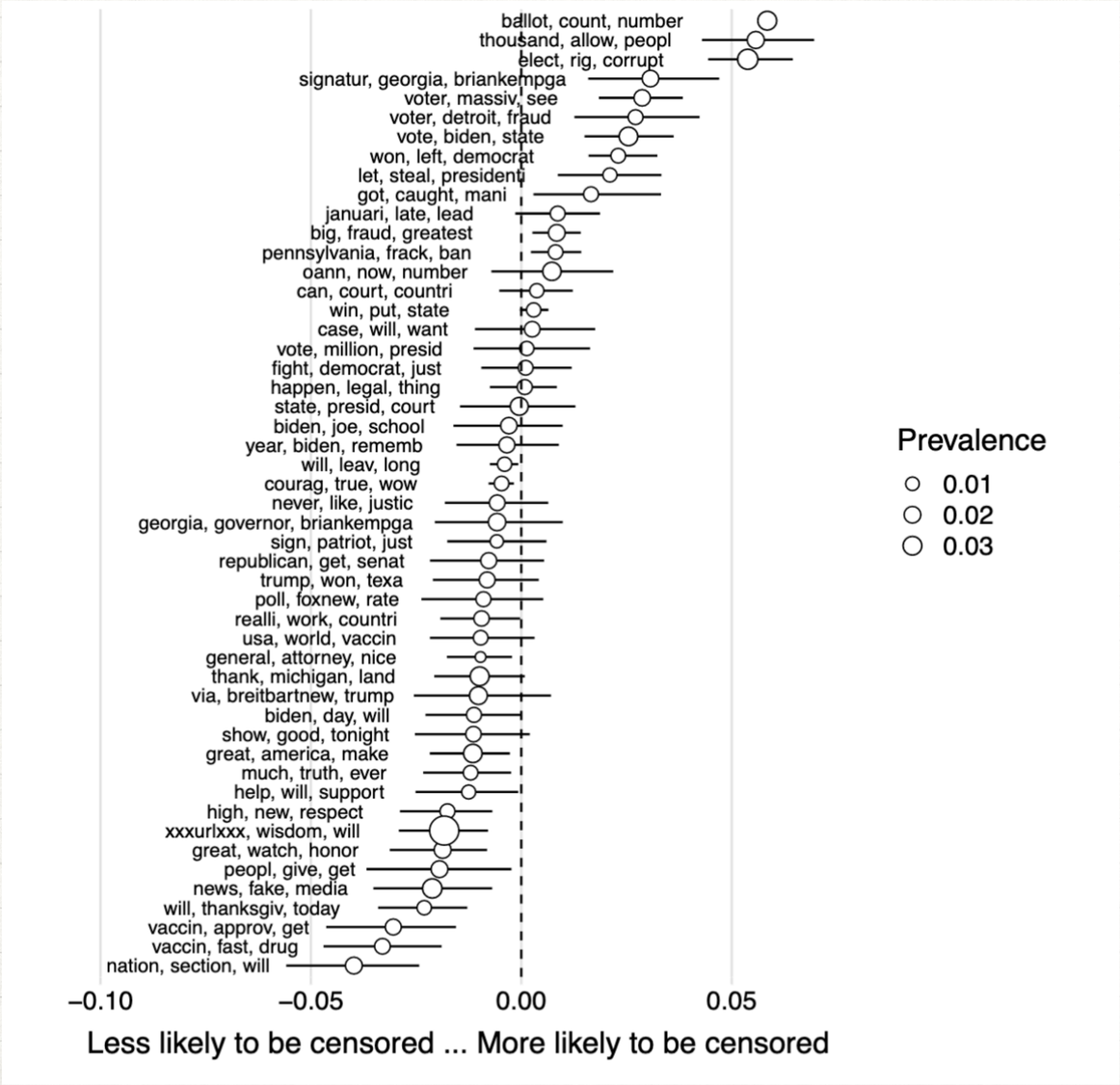
- The intervened and non-intervened tweets are very different -> biased estimate



HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

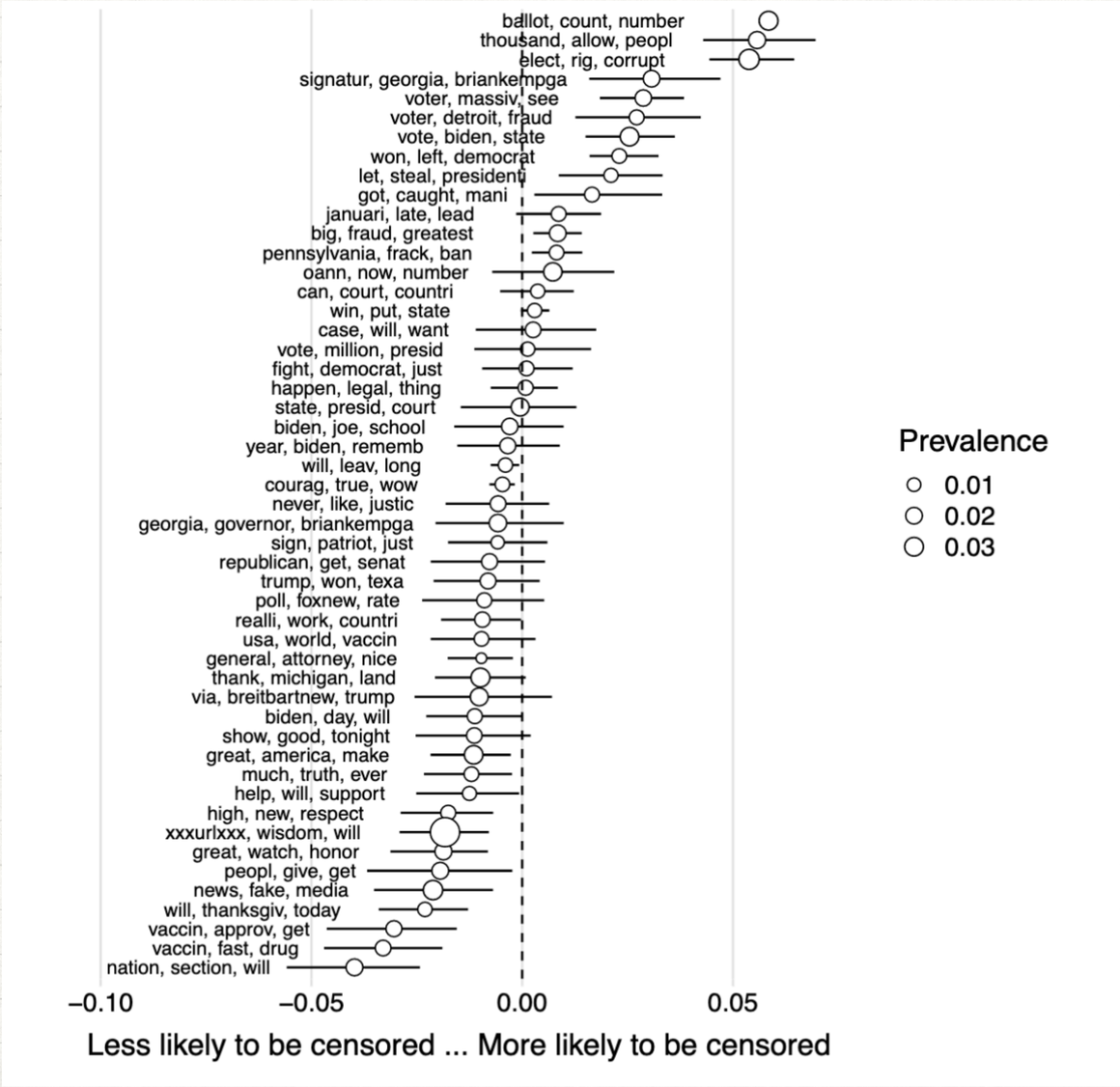
- The intervened and non-intervened tweets are very different -> biased estimate



HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

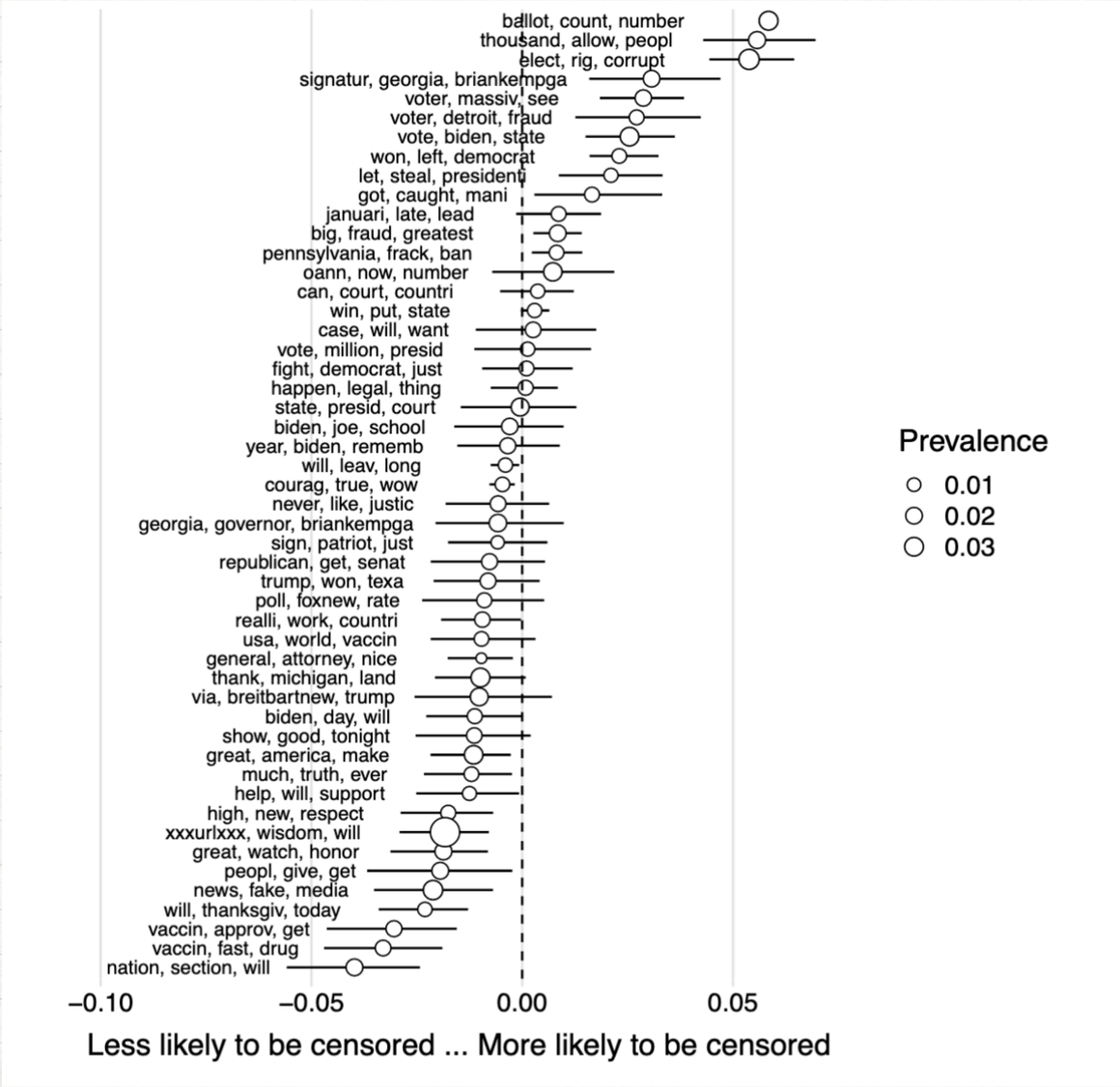
- The intervened and non-intervened tweets are very different -> biased estimate
- Causal Inference 101: Matching helps reduce biased estimates



HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

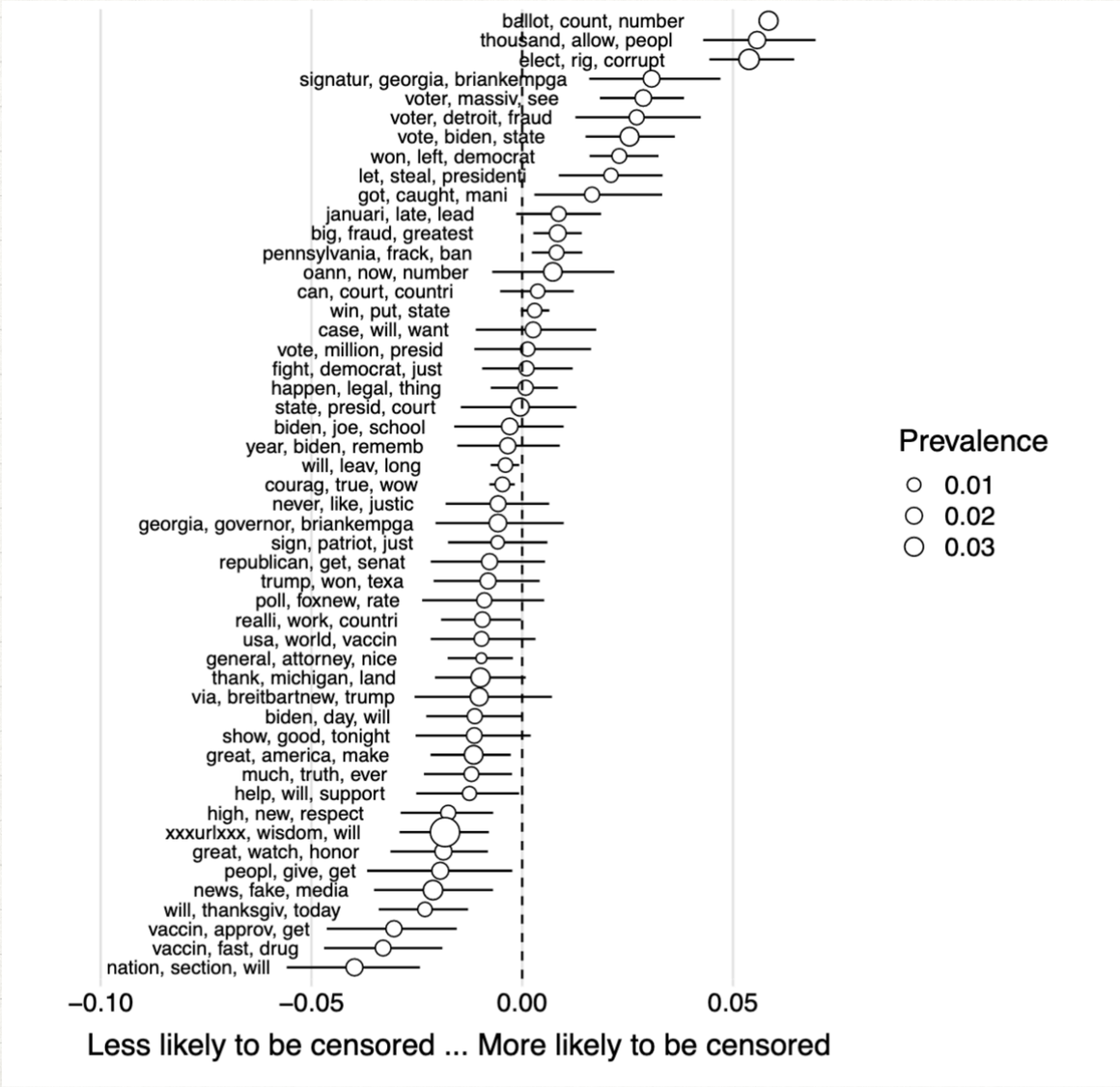
- The intervened and non-intervened tweets are very different -> biased estimate
- Causal Inference 101: Matching helps reduce biased estimates
- Let's Match Tweets!



HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- The intervened and non-intervened tweets are very different -> biased estimate
- Causal Inference 101: Matching helps reduce biased estimates
- Let's Match Tweets!
- Cool new matching technique by Hazlett and Xu (2019) called tjbal



HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- Robustness check over assumptions of earliest to latest time they could have intervened!

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- Robustness check over assumptions of earliest to latest time they could have intervened!
- T0 = earliest possible intervention

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- Robustness check over assumptions of earliest to latest time they could have intervened!
- T0 = earliest possible intervention
- T1 = latest possible intervention

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

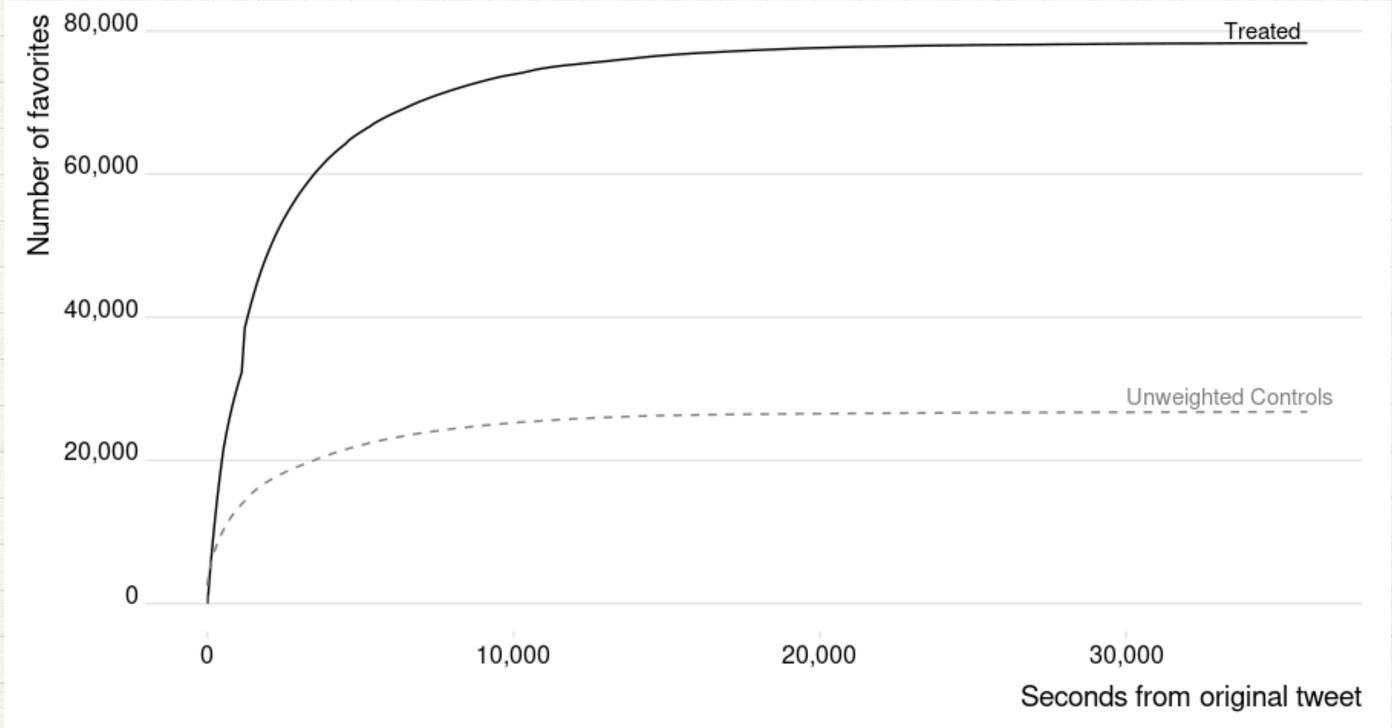
CAUSAL EFFECTS

- Robustness check over assumptions of earliest to latest time they could have intervened!
- T0 = earliest possible intervention
- T1 = latest possible intervention
- Kernel balanced estimates are most robust to changes

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

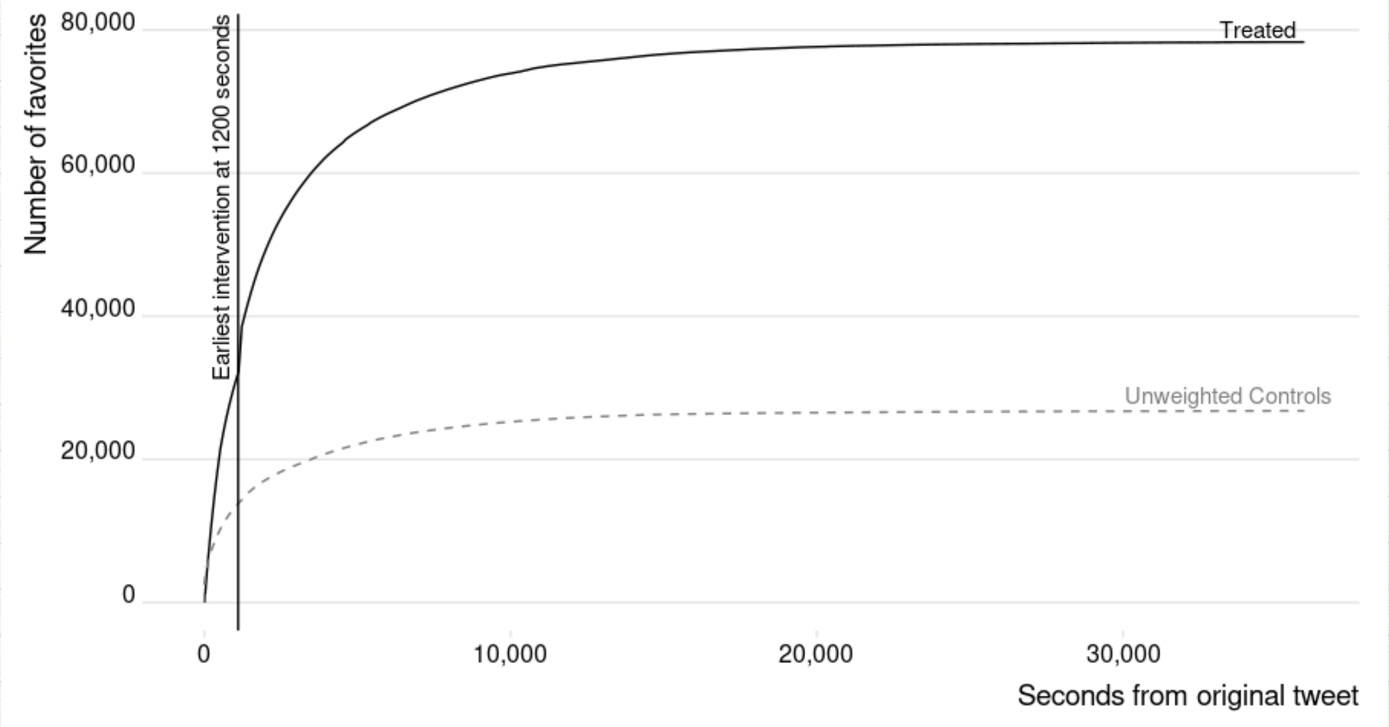
- Robustness check over assumptions of earliest to latest time they could have intervened!
- T0 = earliest possible intervention
- T1 = latest possible intervention
- Kernel balanced estimates are most robust to changes



HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

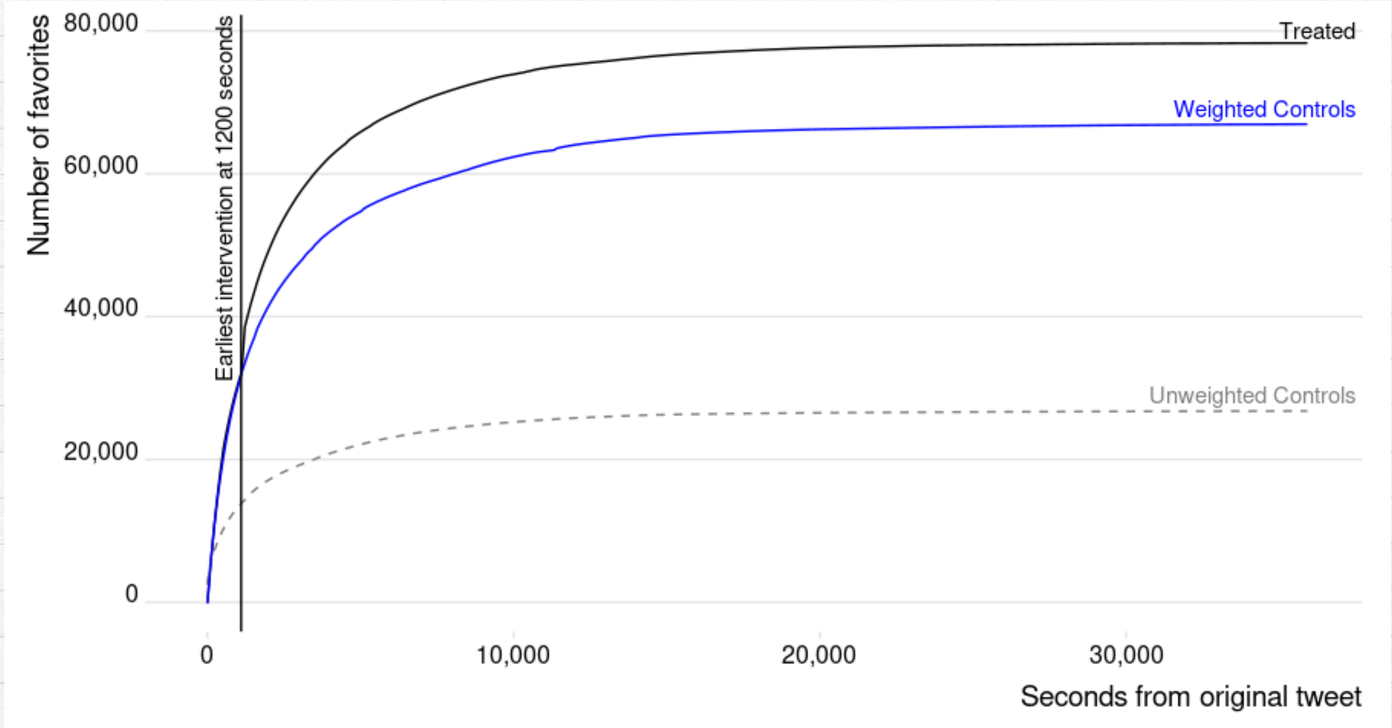
- Robustness check over assumptions of earliest to latest time they could have intervened!
- T0 = earliest possible intervention
- T1 = latest possible intervention
- Kernel balanced estimates are most robust to changes



HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

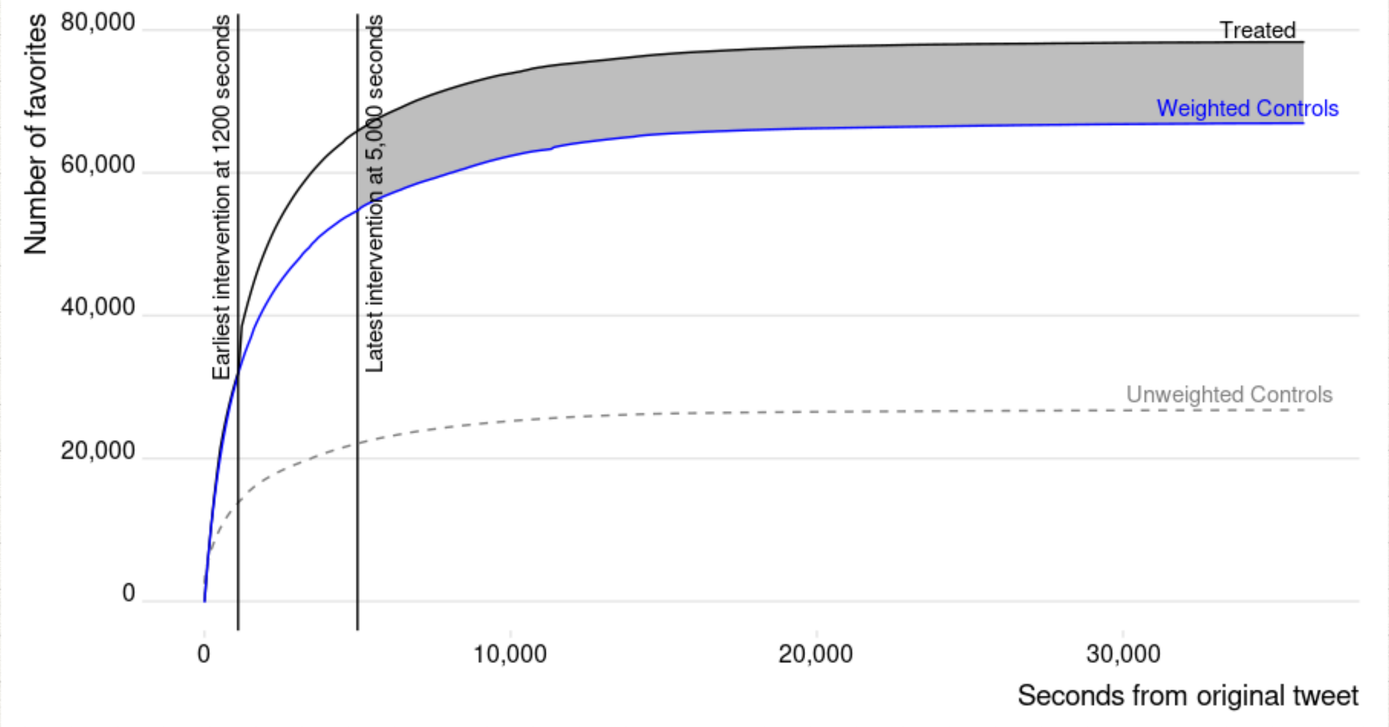
- Robustness check over assumptions of earliest to latest time they could have intervened!
- T0 = earliest possible intervention
- T1 = latest possible intervention
- Kernel balanced estimates are most robust to changes



HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- Robustness check over assumptions of earliest to latest time they could have intervened!
- T0 = earliest possible intervention
- T1 = latest possible intervention
- Kernel balanced estimates are most robust to changes



HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

$$ATT = \mathbb{E}[Y_{it}^{1} - Y_{it}^{0} \mid G_i = 1]$$

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

$$ATT = \mathbb{E}[Y_{it}^1 - Y_{it}^0 \mid G_i = 1]$$

Requires Linearity in Prior Outcomes (LPO)

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

$$ATT = \mathbb{E}[Y_{it}^1 - Y_{it}^0 \mid G_i = 1]$$

Requires Linearity in Prior Outcomes (LPO)

$$\mathbb{E}[Y_{it}^0 \mid \mathbf{Y}_{i,pre}] = (1\ \mathbf{Y}_{i,pre})^T\ \theta_t$$

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- Trajectory Balancing for matching the tweets using two methods:

$$ATT = \mathbb{E}[Y_{it}^1 - Y_{it}^0 \mid G_i = 1]$$

Requires Linearity in Prior Outcomes (LPO)

$$\mathbb{E}[Y_{it}^0 \mid \mathbf{Y}_{i,pre}] = (1\ \mathbf{Y}_{i,pre})^T\ \theta_t$$

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- Trajectory Balancing for matching the tweets using two methods:
  - **Mean Balancing**

$$ATT = \mathbb{E}[Y_{it}^1 - Y_{it}^0 \mid G_i = 1]$$

Requires Linearity in Prior Outcomes (LPO)

$$\mathbb{E}[Y_{it}^0 \mid \mathbf{Y}_{i,pre}] = (1\ \mathbf{Y}_{i,pre})^T\ \theta_t$$

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- Trajectory Balancing for matching the tweets using two methods:
  - **Mean Balancing**

$$ATT = \mathbb{E}[Y_{it}^1 - Y_{it}^0 \mid G_i = 1]$$

Requires Linearity in Prior Outcomes (LPO)

$$\mathbb{E}[Y_{it}^0 \mid \mathbf{Y}_{i,pre}] = (1\ \mathbf{Y}_{i,pre})^T\ \theta_t$$

$$\frac{1}{N_{tr}} \sum_{G_i=1} \mathbf{Y}_{i,pre} = \sum_{G_i=0} w_i \mathbf{Y}_{i,pre}$$

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- Trajectory Balancing for matching the tweets using two methods:
  - **Mean Balancing**
  - **Kernel Balancing**

$$ATT = \mathbb{E}[Y_{it}^1 - Y_{it}^0 \mid G_i = 1]$$

Requires Linearity in Prior Outcomes (LPO)

$$\mathbb{E}[Y_{it}^0 \mid \mathbf{Y}_{i,pre}] = (1\ \mathbf{Y}_{i,pre})^T\ \theta_t$$

$$\frac{1}{N_{tr}} \sum_{G_i=1} \mathbf{Y}_{i,pre} = \sum_{G_i=0} w_i \mathbf{Y}_{i,pre}$$

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- Trajectory Balancing for matching the tweets using two methods:
  - **Mean Balancing**
  - **Kernel Balancing**

$$ATT = \mathbb{E}[Y_{it}^1 - Y_{it}^0 \mid G_i = 1]$$

Requires Linearity in Prior Outcomes (LPO)

$$\mathbb{E}[Y_{it}^0 \mid \mathbf{Y}_{i,pre}] = (1\ \mathbf{Y}_{i,pre})^T\ \theta_t$$

$$\frac{1}{N_{tr}} \sum_{G_i=1} \mathbf{Y}_{i,pre} = \sum_{G_i=0} w_i \mathbf{Y}_{i,pre}$$

Relaxes Linearity in Prior Outcomes (LPO)

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- Trajectory Balancing for matching the tweets using two methods:
  - **Mean Balancing**
  - **Kernel Balancing**

$$ATT = \mathbb{E}[Y_{it}^1 - Y_{it}^0 \mid G_i = 1]$$

Requires Linearity in Prior Outcomes (LPO)

$$\mathbb{E}[Y_{it}^0 \mid \mathbf{Y}_{i,pre}] = (1\ \mathbf{Y}_{i,pre})^T\ \theta_t$$

$$\frac{1}{N_{tr}} \sum_{G_i=1} \mathbf{Y}_{i,pre} = \sum_{G_i=0} w_i \mathbf{Y}_{i,pre}$$

Relaxes Linearity in Prior Outcomes (LPO)

$$\mathbb{E}[Y_{it}^0 \mid \mathbf{Y}_{i,pre}] = \phi(\mathbf{Y}_{i,pre})^T\ \theta_t \ \ ; \ \ t > T_0$$

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- Trajectory Balancing for matching the tweets using two methods:
  - **Mean Balancing**
  - **Kernel Balancing**
- Also balanced on toxicity scores, topics, sharing by elite users

$$ATT = \mathbb{E}[Y_{it}^1 - Y_{it}^0 \mid G_i = 1]$$

Requires Linearity in Prior Outcomes (LPO)

$$\mathbb{E}[Y_{it}^0 \mid \mathbf{Y}_{i,pre}] = (1\ \mathbf{Y}_{i,pre})^T\ \theta_t$$

$$\frac{1}{N_{tr}} \sum_{G_i=1} \mathbf{Y}_{i,pre} = \sum_{G_i=0} w_i \mathbf{Y}_{i,pre}$$

Relaxes Linearity in Prior Outcomes (LPO)

$$\mathbb{E}[Y_{it}^0 \mid \mathbf{Y}_{i,pre}] = \phi(\mathbf{Y}_{i,pre})^T\ \theta_t \ \ ; \ \ t > T_0$$

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- Trajectory Balancing for matching the tweets using two methods:
  - **Mean Balancing**
  - **Kernel Balancing**
- Also balanced on toxicity scores, topics, sharing by elite users
- But we don't know intervention time so we need to guess when Twitter intervened…

$$ATT = \mathbb{E}[Y_{it}^1 - Y_{it}^0 \,|\, G_i = 1]$$

Requires Linearity in Prior Outcomes (LPO)

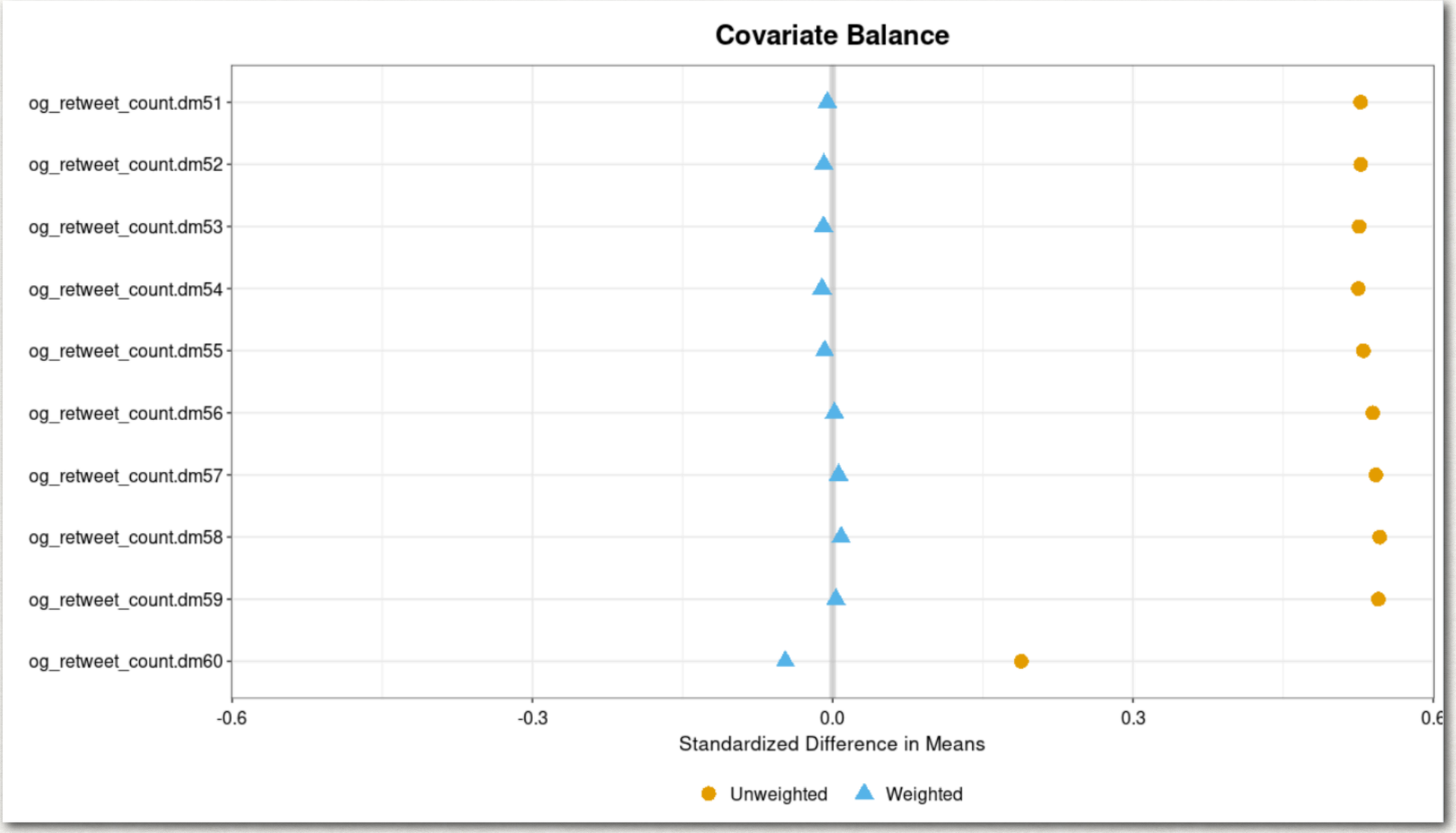
$$\mathbb{E}[Y_{it}^0 \,|\, \mathbf{Y}_{i,pre}] = (1\ \mathbf{Y}_{i,pre})^T\ \theta_t$$

$$\frac{1}{N_{tr}} \sum_{G_i=1} \mathbf{Y}_{i,pre} = \sum_{G_i=0} w_i \mathbf{Y}_{i,pre}$$

Relaxes Linearity in Prior Outcomes (LPO)

$$\mathbb{E}[Y_{it}^0 \,|\, \mathbf{Y}_{i,pre}] = \phi(\mathbf{Y}_{i,pre})^T\ \theta_t \ \ ; \ \ t > T_0$$

CHECKING PRE-TREATMENT BALANCE



TWITTER

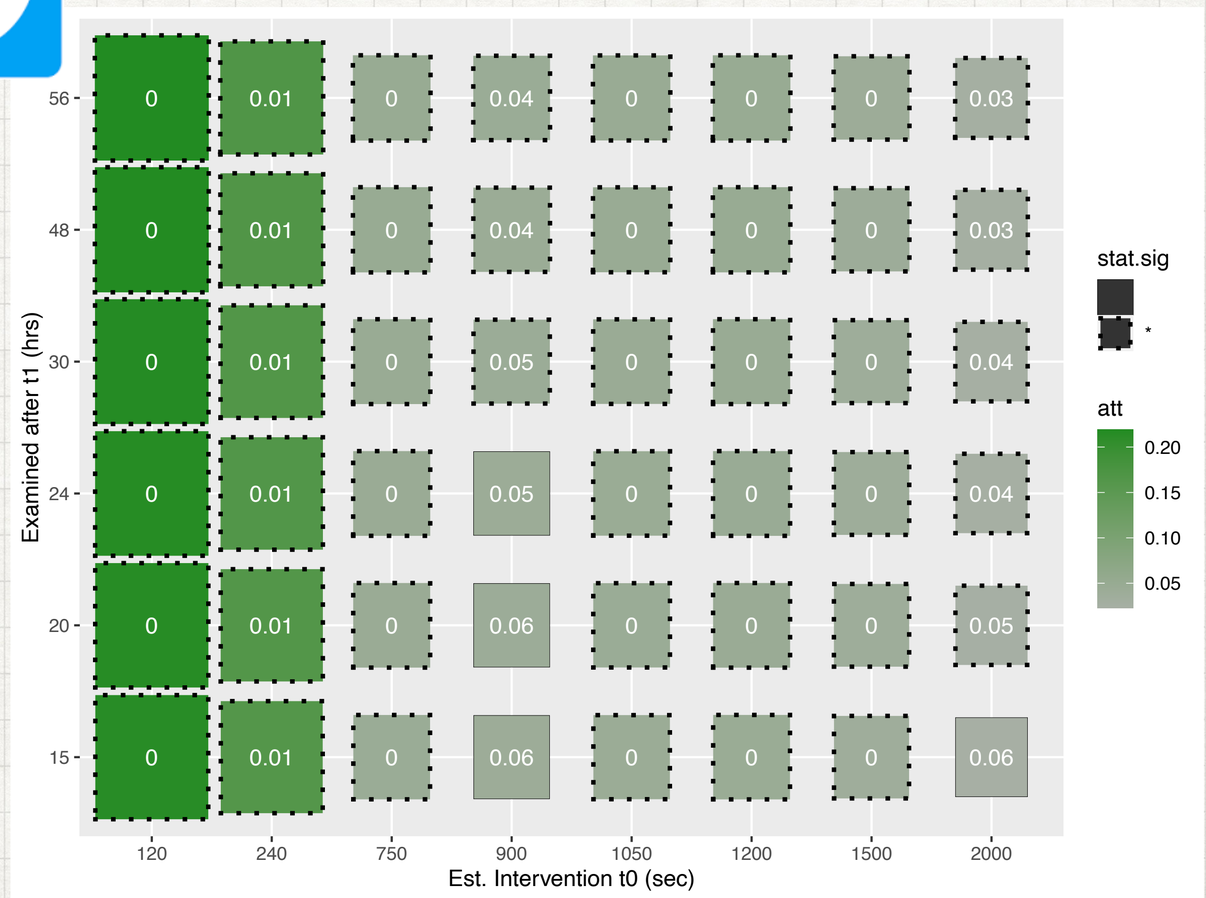
WHAT DID INTERVENTIONS CAUSE?

TWITTER

WHAT DID INTERVENTIONS CAUSE?

TWITTER

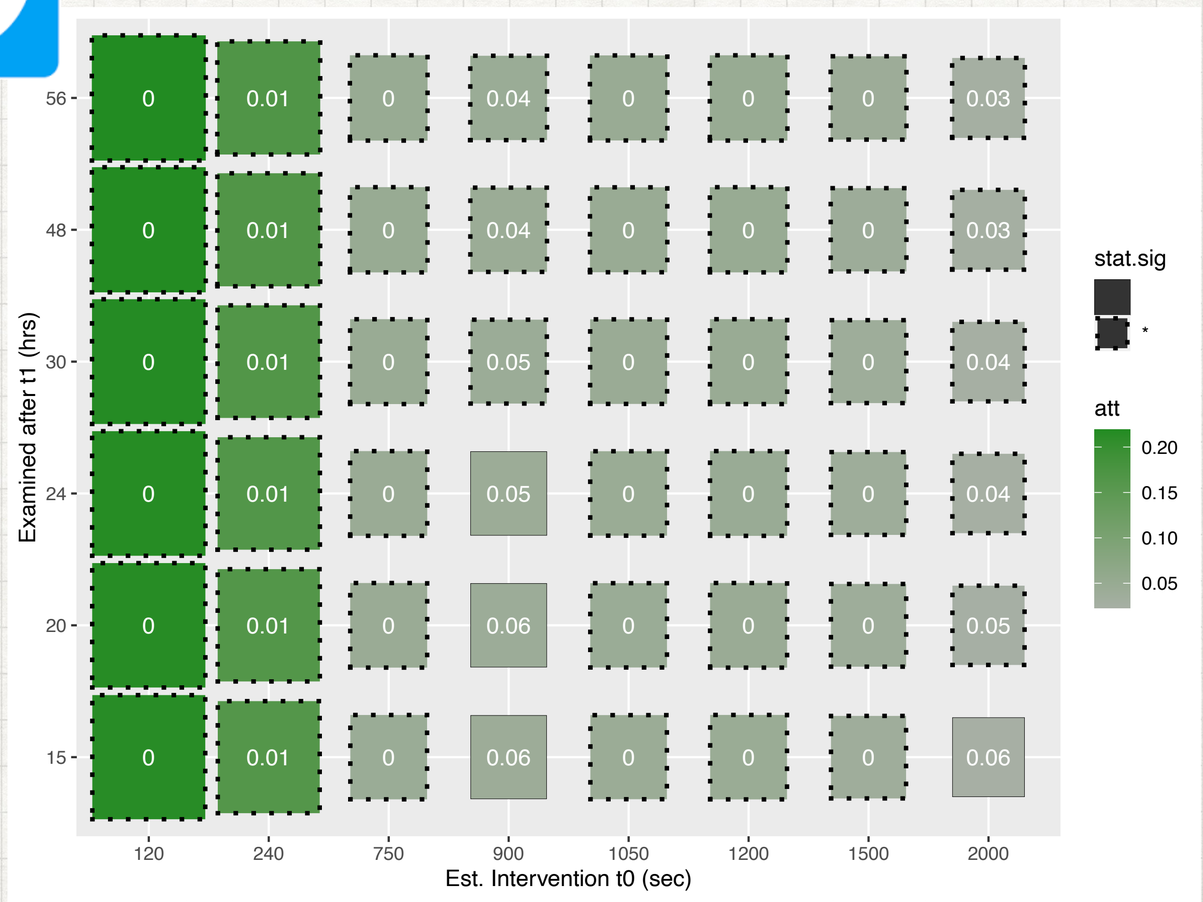
WHAT DID INTERVENTIONS CAUSE?



TWITTER

WHAT DID INTERVENTIONS CAUSE?

- Streisand effect present but milder on Twitter (retweets)



TWITTER

WHAT DID INTERVENTIONS CAUSE?

TWITTER

WHAT DID INTERVENTIONS CAUSE?

- Streisand effect present but milder on Twitter (favorites)
- Interventions on other platforms:
  - Facebook
  - Reddit
  - Instagram
- But Interventions have cross-platform effects!

TWITTER

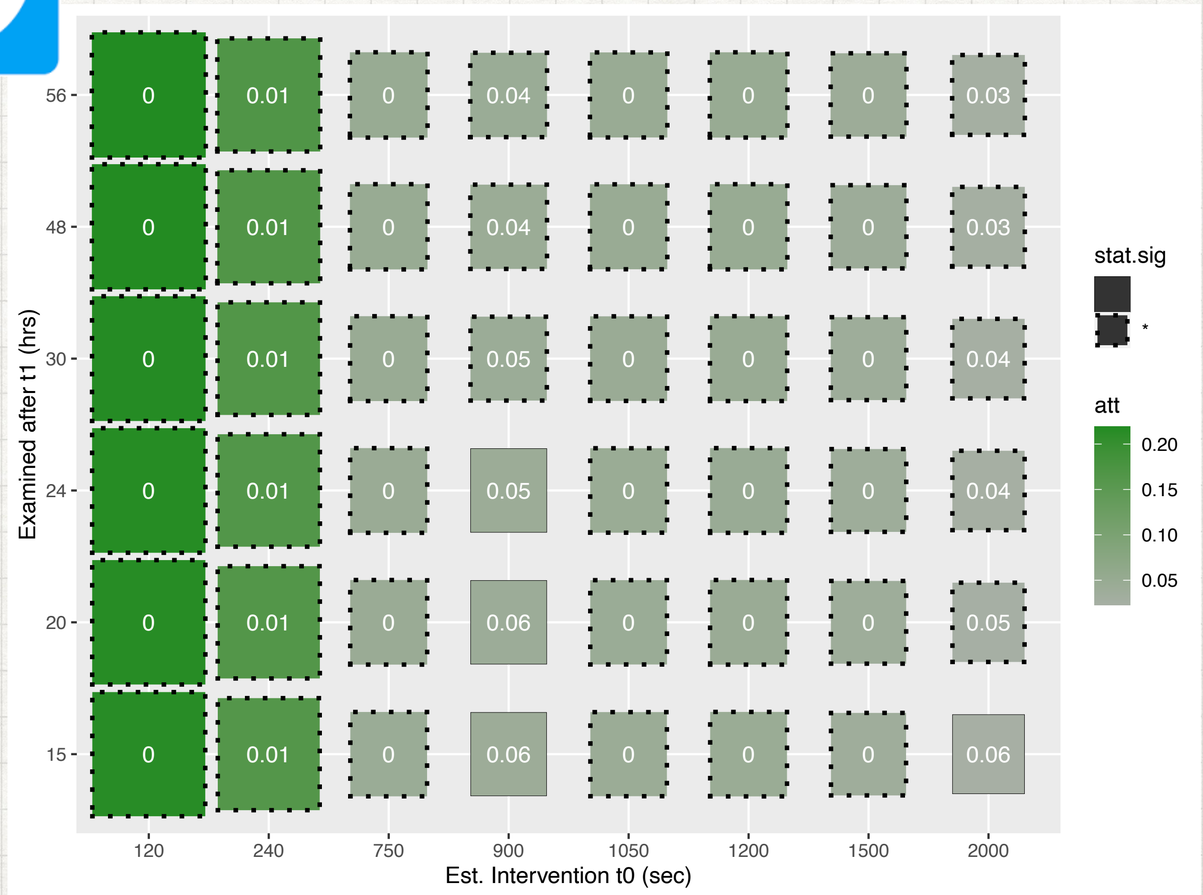
WHAT DID INTERVENTIONS CAUSE?

- Streisand effect present but milder on Twitter (favorites)
- Interventions on other platforms:
  - Facebook
  - Reddit
  - Instagram
- But Interventions have cross-platform effects!

TWITTER

WHAT DID INTERVENTIONS CAUSE?

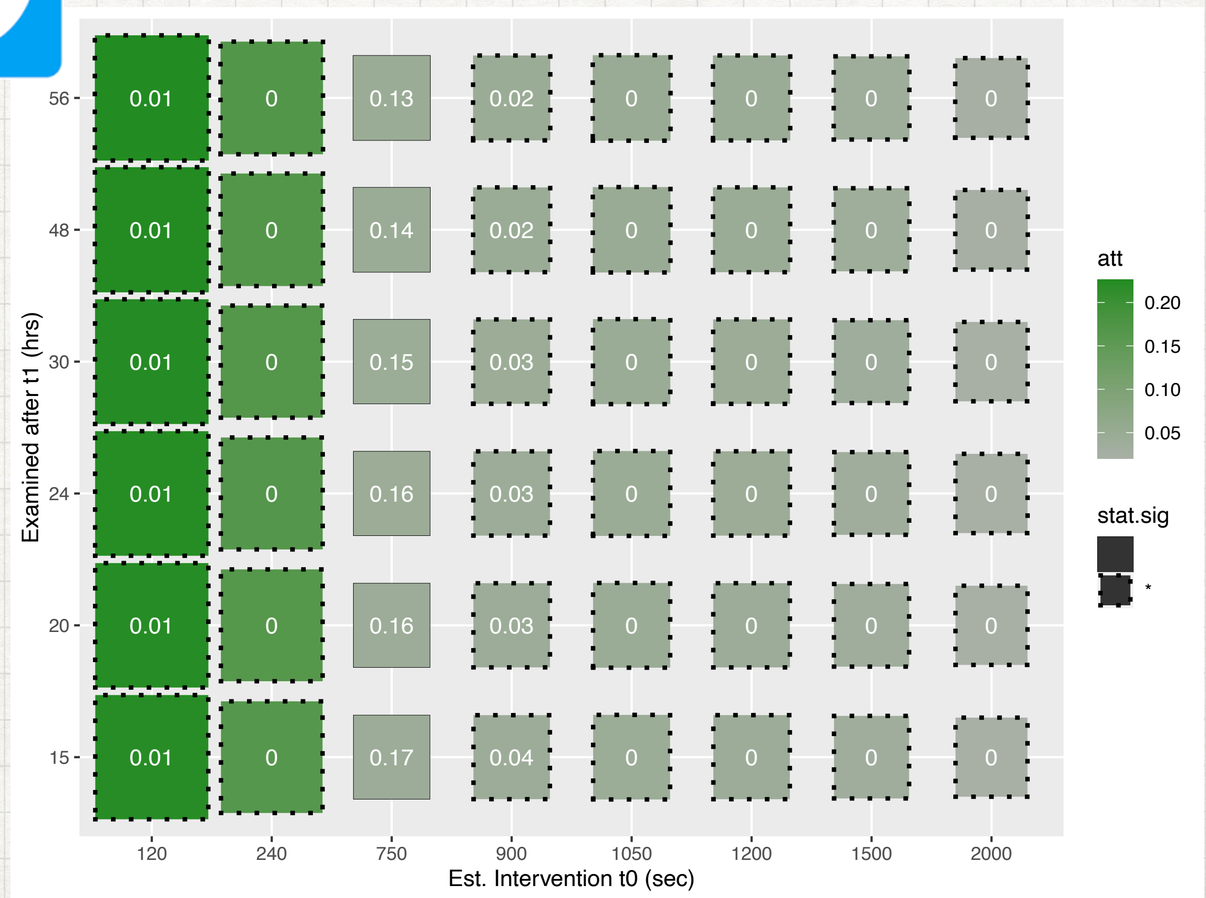
- Streisand effect present but milder on Twitter (favorites)
- Interventions on other platforms:
  - Facebook
  - Reddit
  - Instagram
- But Interventions have cross-platform effects!



TWITTER

WHAT DID INTERVENTIONS CAUSE?

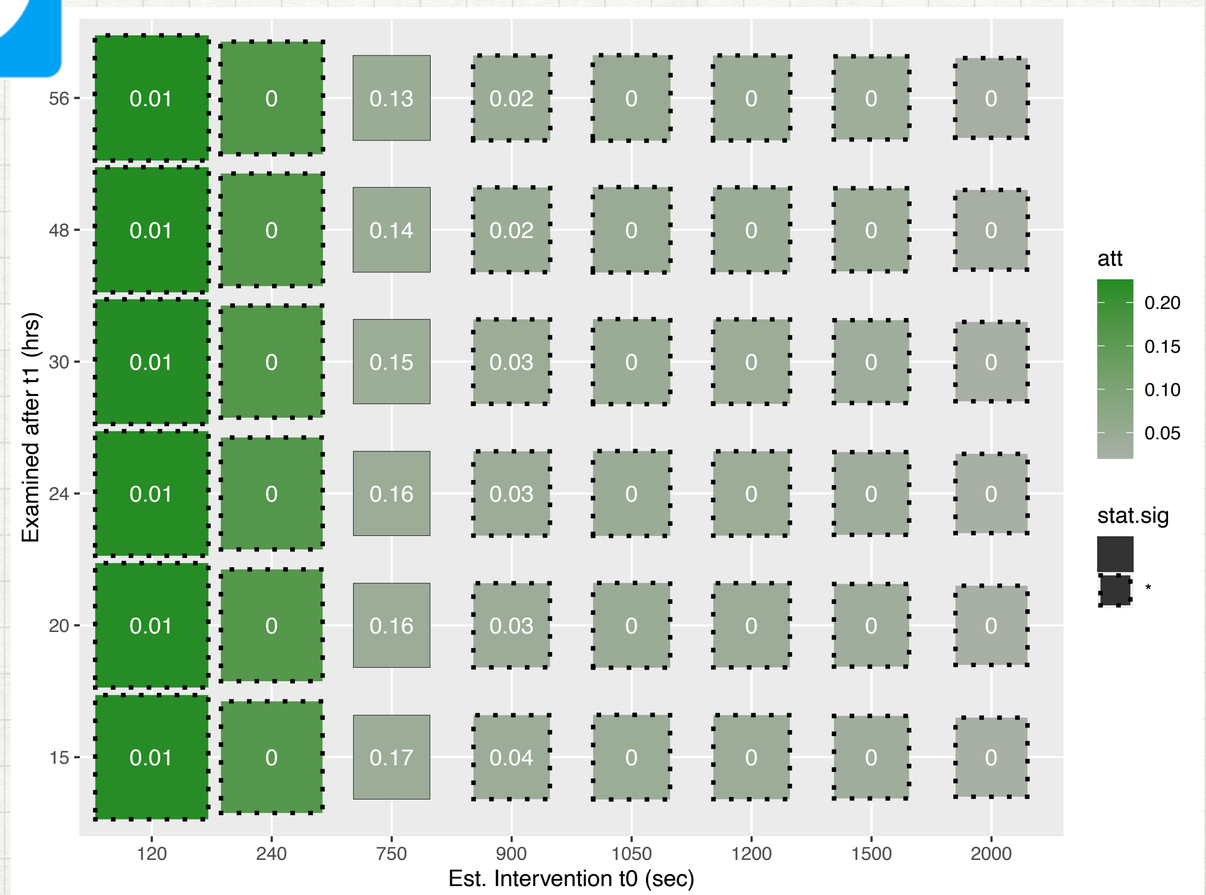
- Streisand effect present but milder on Twitter (favorites)
- Interventions on other platforms:
  - Facebook
  - Reddit
  - Instagram
- But Interventions have cross-platform effects!



TWITTER

WHAT DID INTERVENTIONS CAUSE?

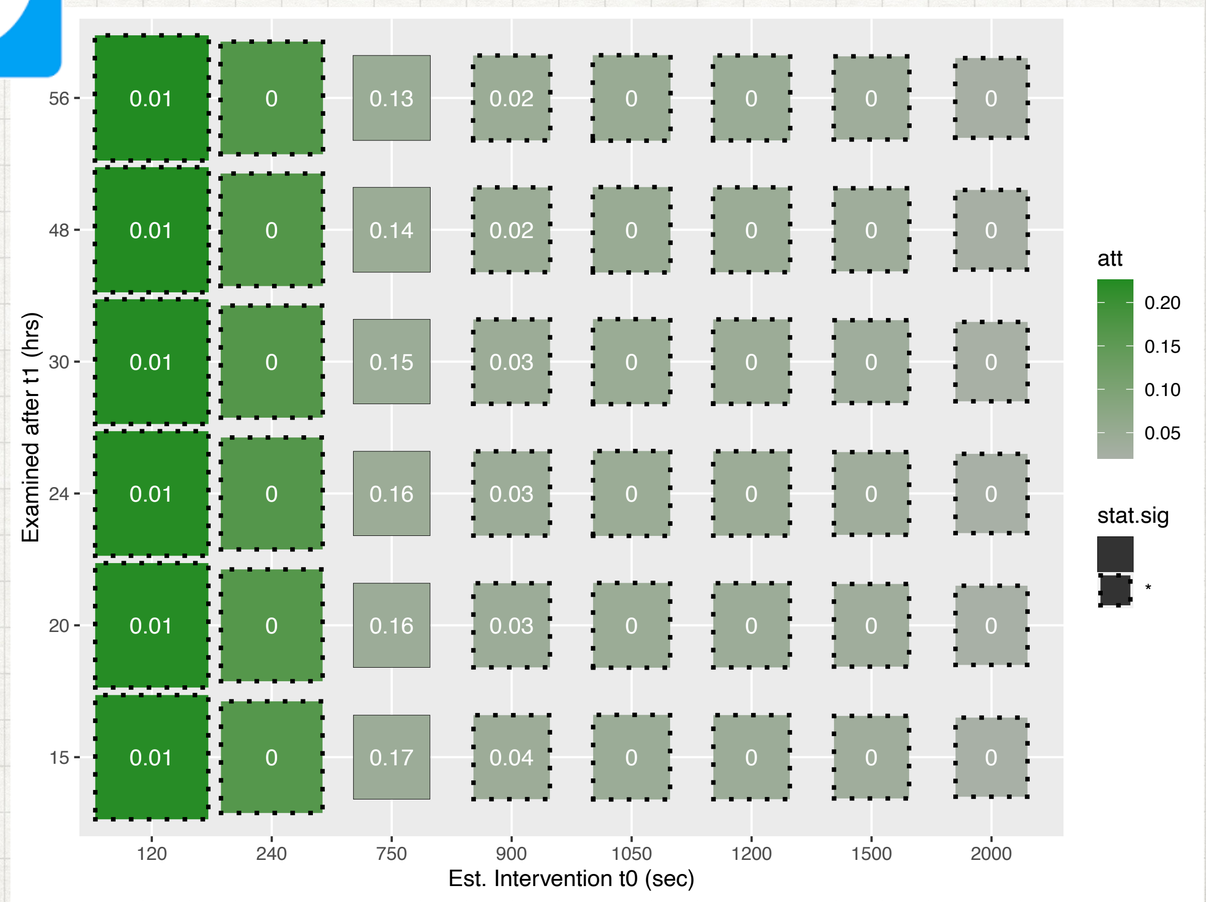
- Streisand effect present but milder on Twitter (favorites)
- Interventions on other platforms:
  - Facebook
  - Reddit
  - Instagram
- But Interventions have cross-platform effects!



TWITTER

WHAT DID INTERVENTIONS CAUSE?

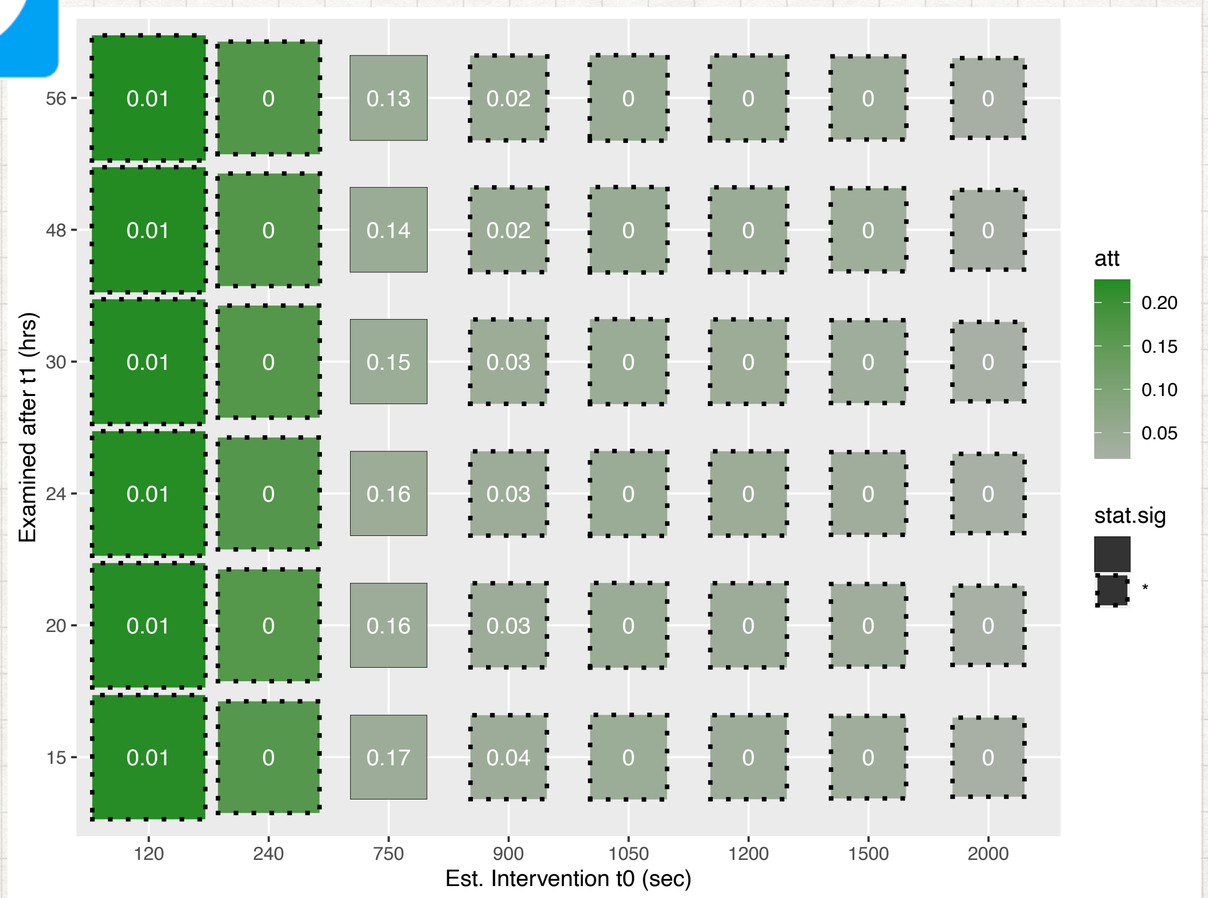
- Streisand effect present but milder on Twitter (favorites)
- Interventions on other platforms:
  - Facebook
  - Reddit
  - Instagram
- But Interventions have cross-platform effects!



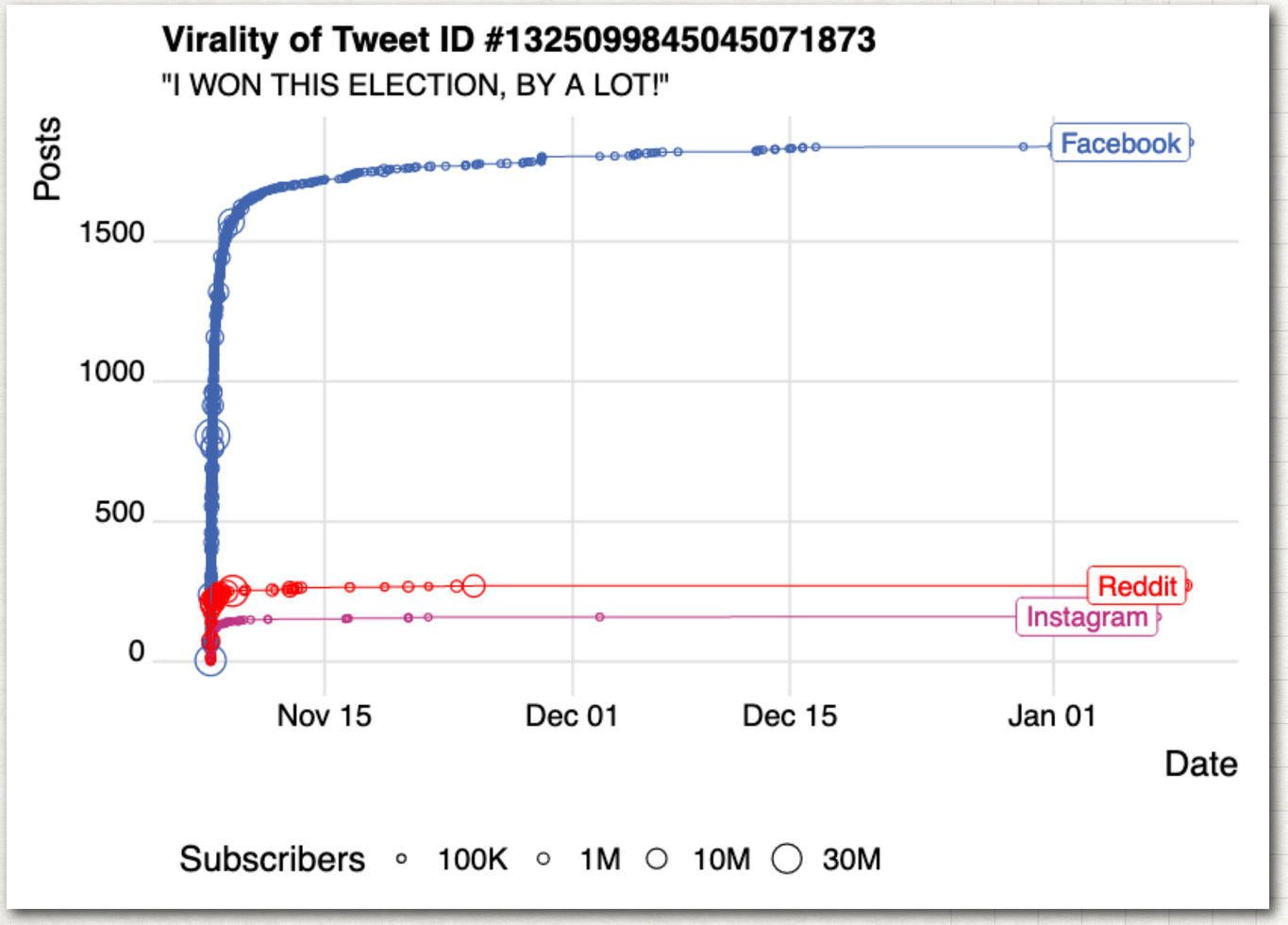
TWITTER

WHAT DID INTERVENTIONS CAUSE?

- Streisand effect present but milder on Twitter (favorites)
- Interventions on other platforms:
  - Facebook
  - Reddit
  - Instagram
- But Interventions have cross-platform effects!

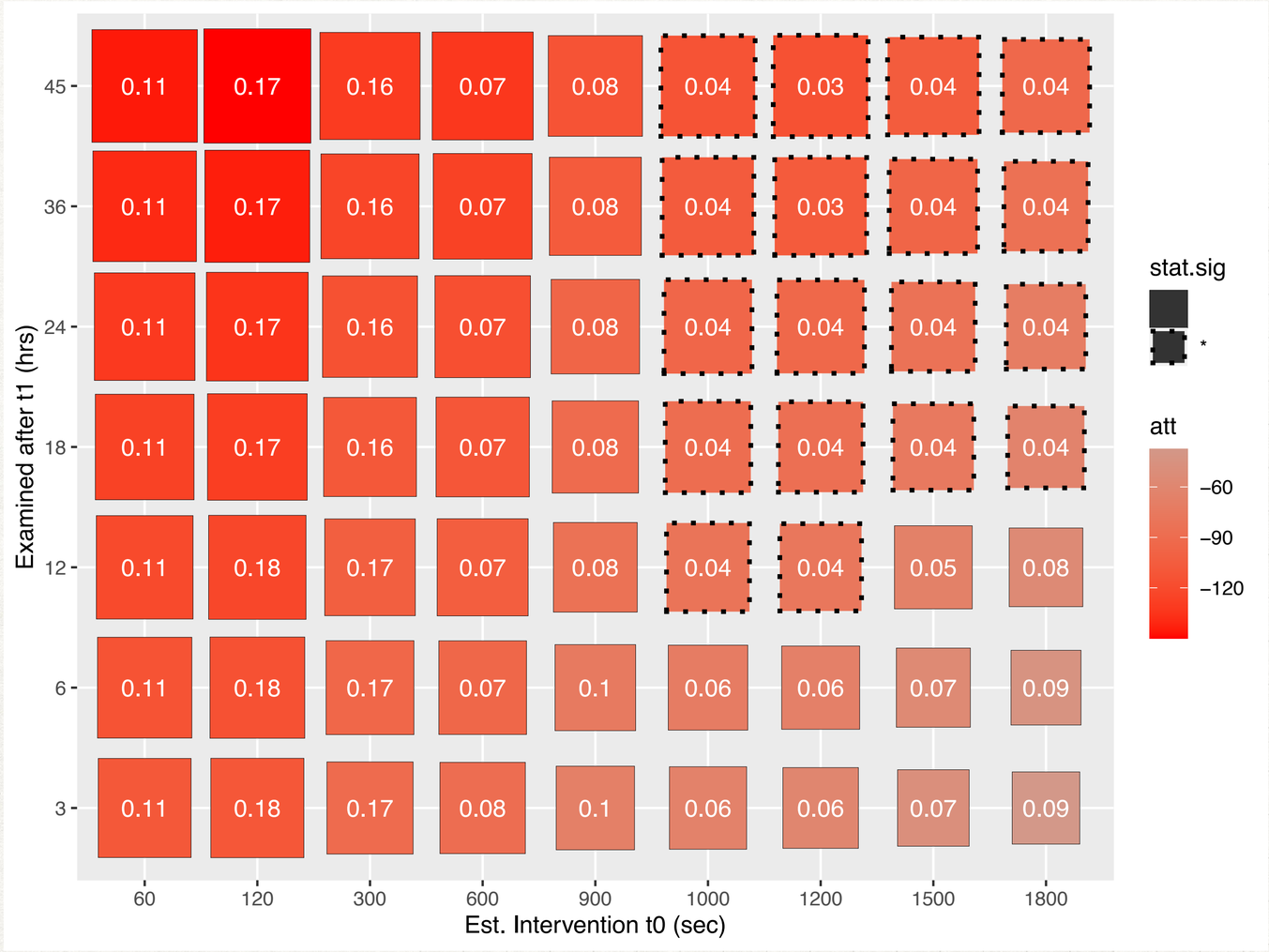


CROWDTANGLE RESULTS - CROSS-PLATFORM MATCHES

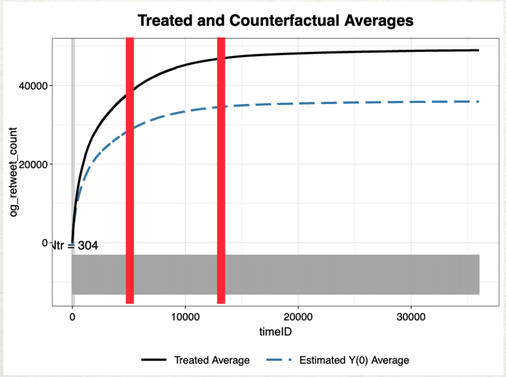


- Tweets are shared beyond just platform of origin, Twitter
- What are the effects of interventions across different platforms?
- **Use CrowdTangle to collect cross-platform data re: tweets**
- **Can track "hard" interventions!**

FACEBOOK - SOFT INTERVENTIONS

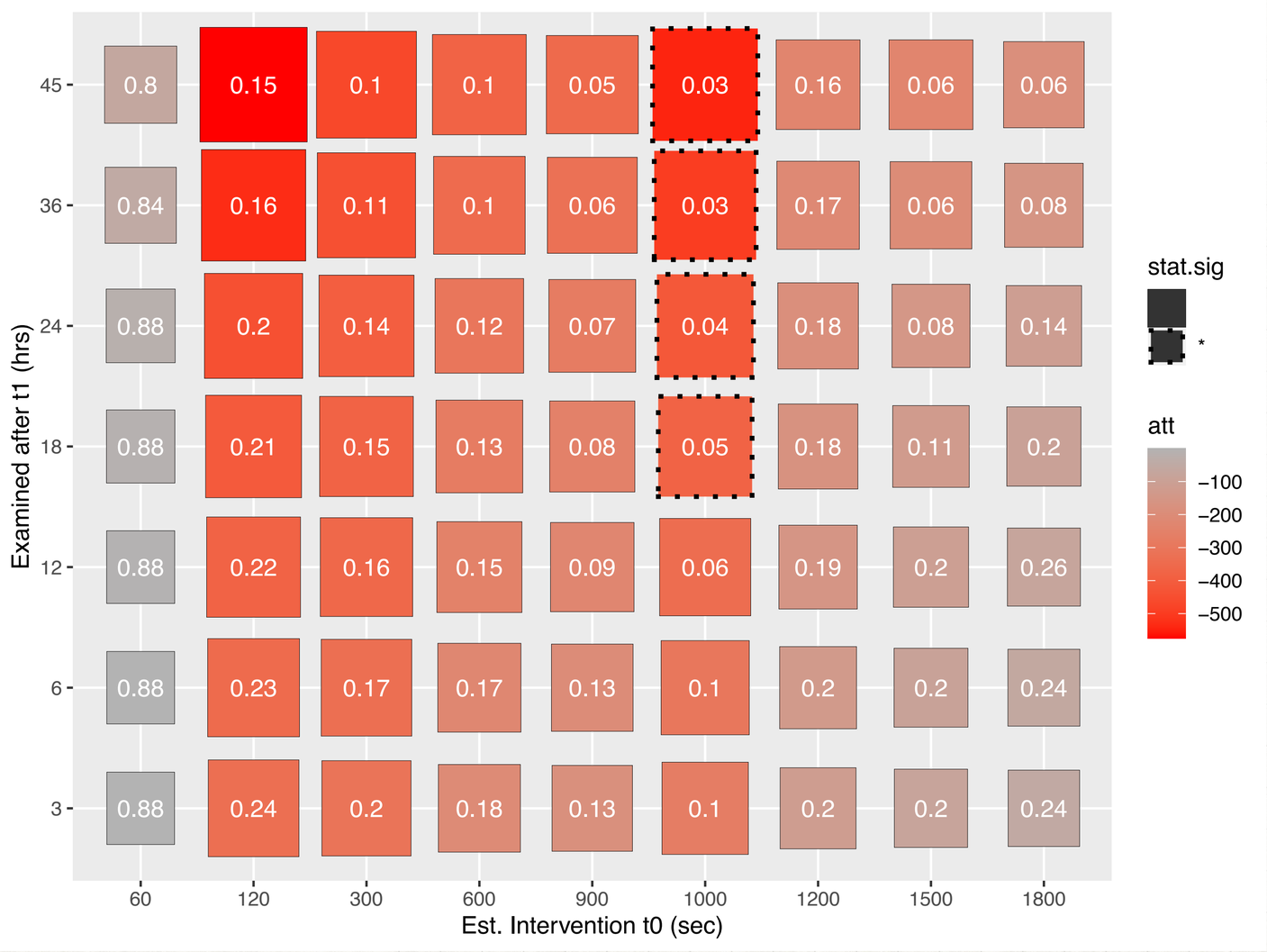


t0 t1

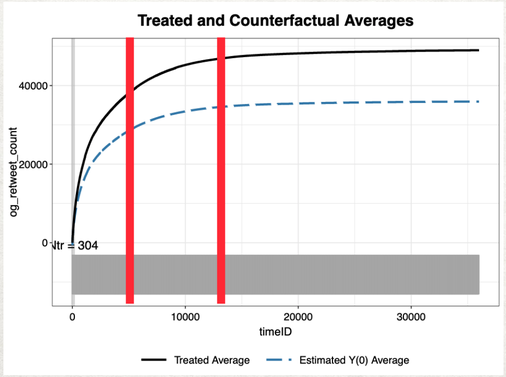


Soft Interventions cause a decrease in FB posts!

FACEBOOK - HARD INTERVENTIONS

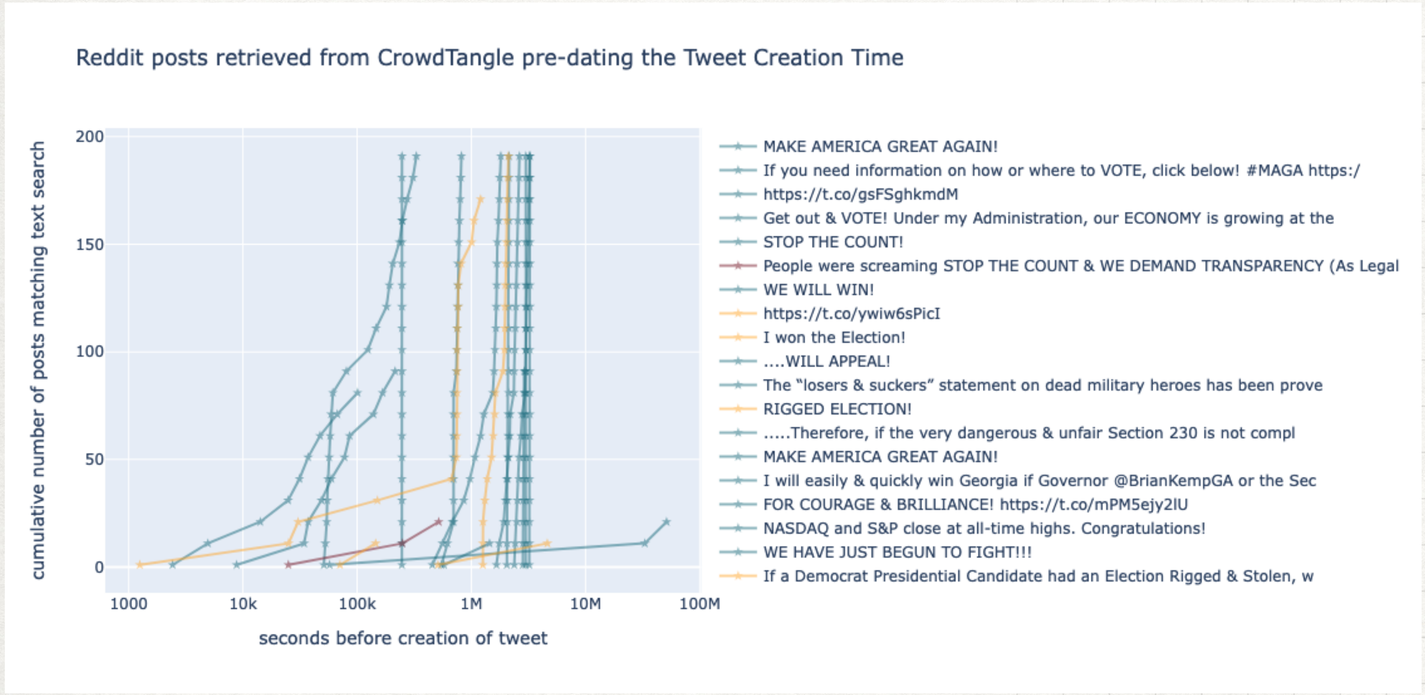


t0 t1



Hard Interventions cause a decrease in FB posts too!

PROBLEMS WITH CROWDTANGLE DATA COLLECTION

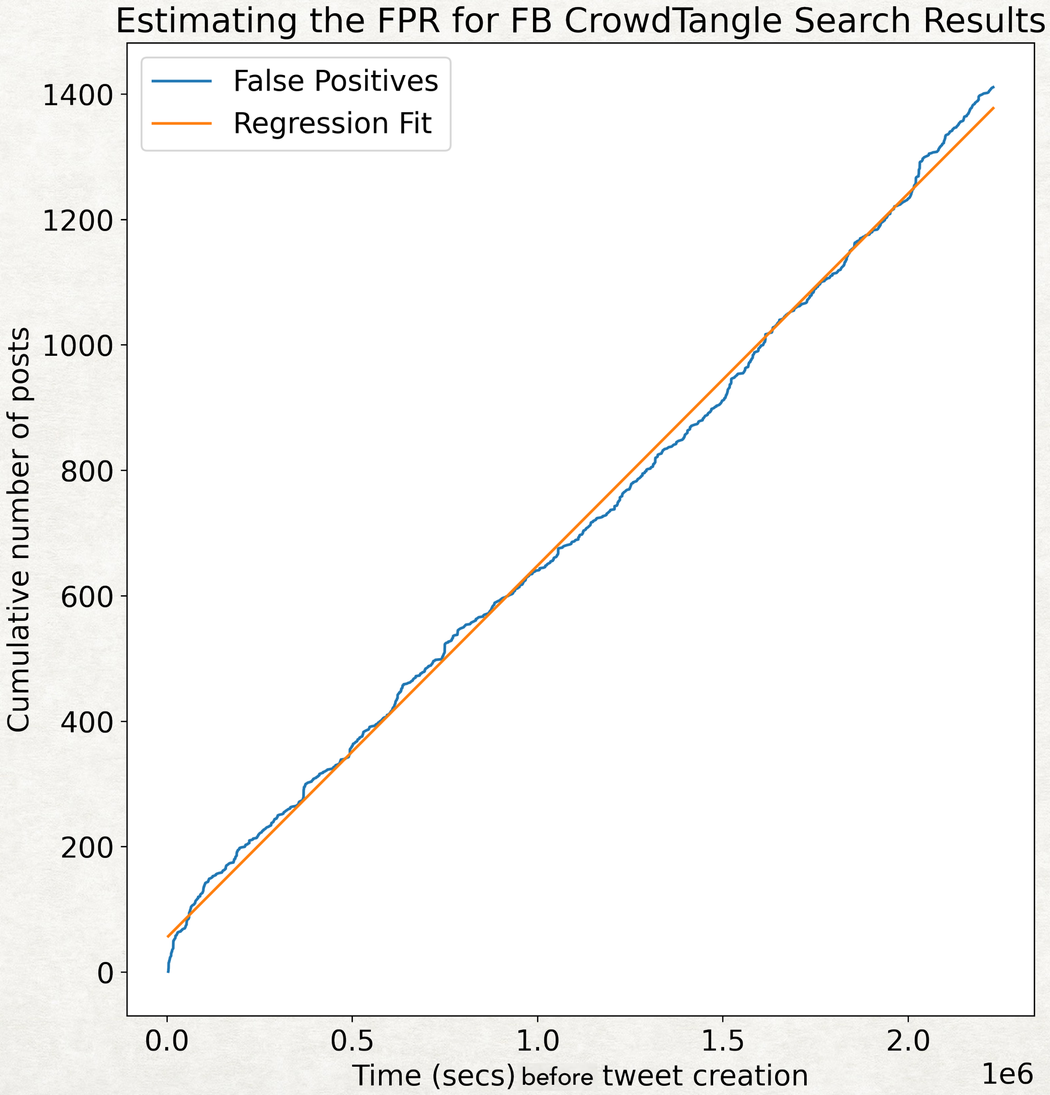


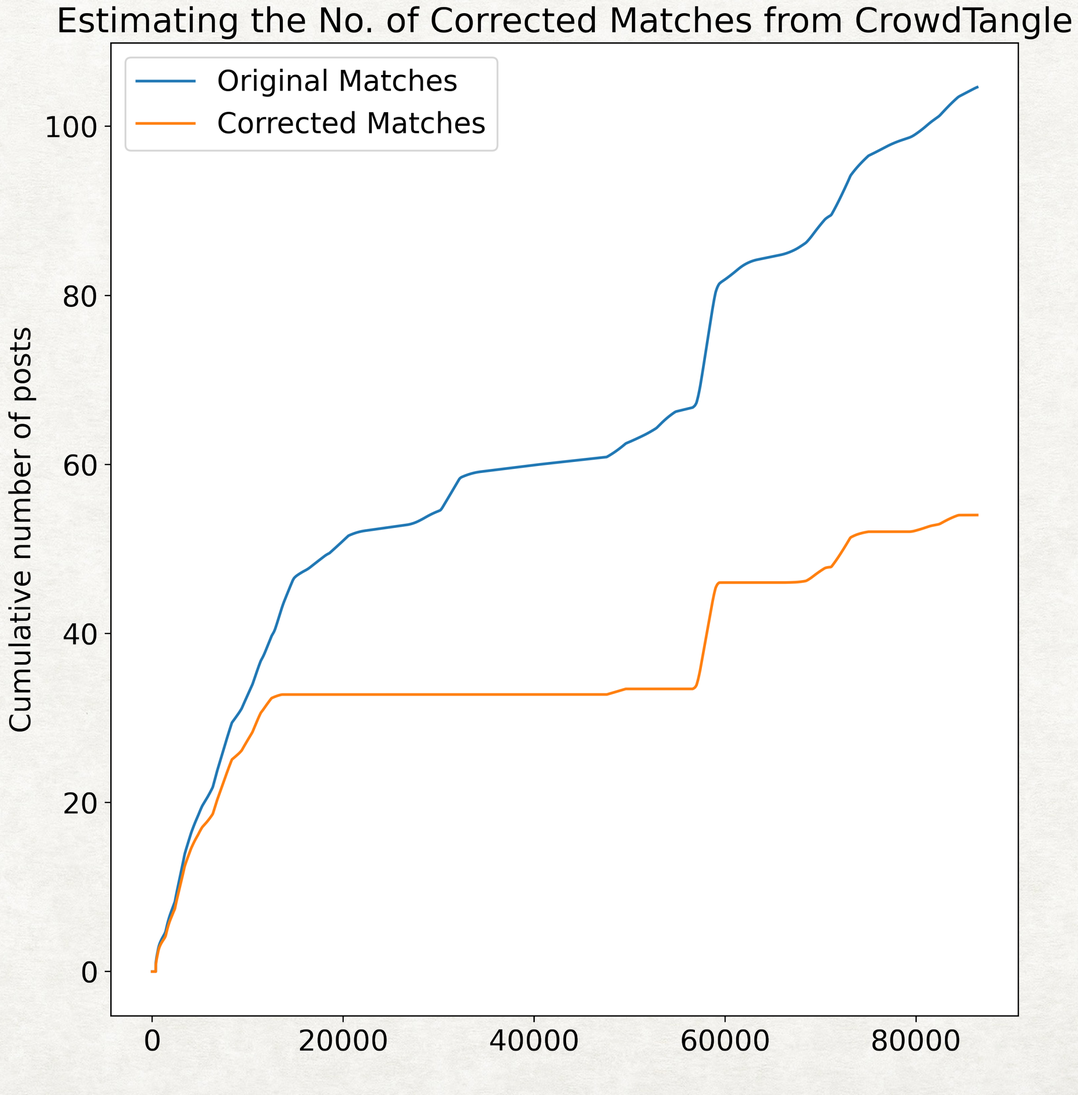
- Collected data by searching for tweet URL and exact text match for tweets
- Exact text match is not really exact…
- Returns posts as search results from before the tweet was actually created
- To top it off, we are required to delete locally collected post content that is banned on-platform – no local copies to qualitatively check post content!
- How to fix this?

PROBLEMS WITH CROWDTANGLE DATA COLLECTION

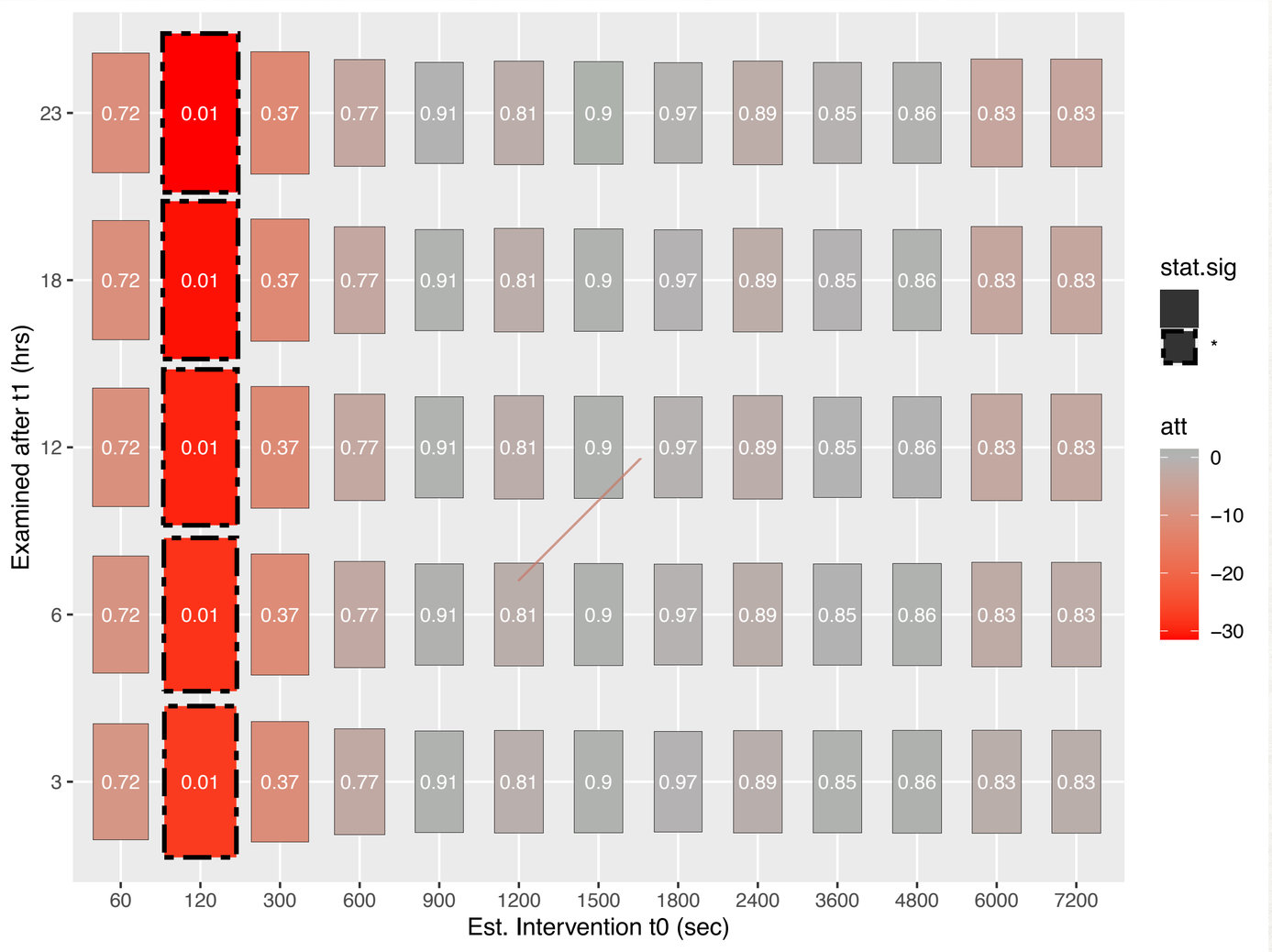
- Fit regression model to estimate fixed False Positive Rate for search results on each tweet
- Remove the false matched posts from the original tweets
- Reestimate trajectories based on correct matches
- Unfortunately not left with enough reliable data on hard interventions to estimate confidence intervals…

FIXING CROWD TANGLE DATA COLLECTION

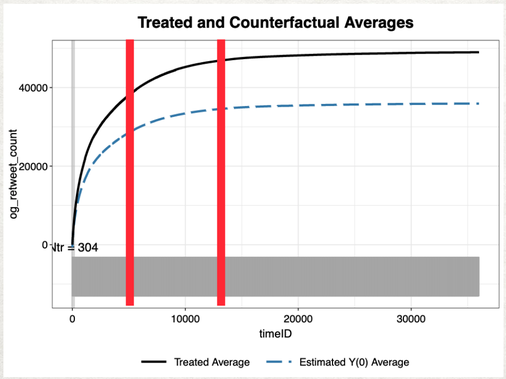




FACEBOOK - SOFT INTERVENTIONS

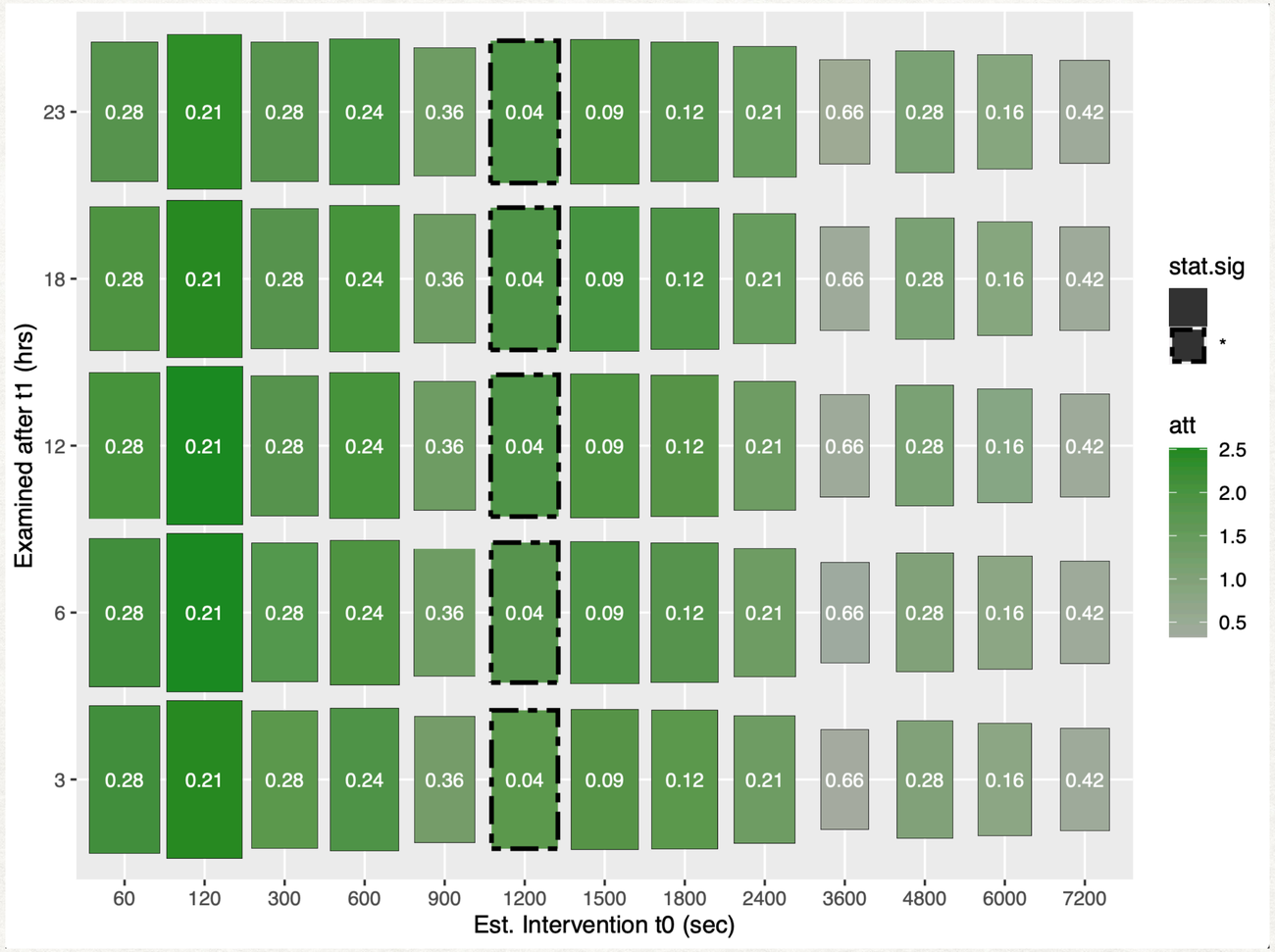


t0 t1

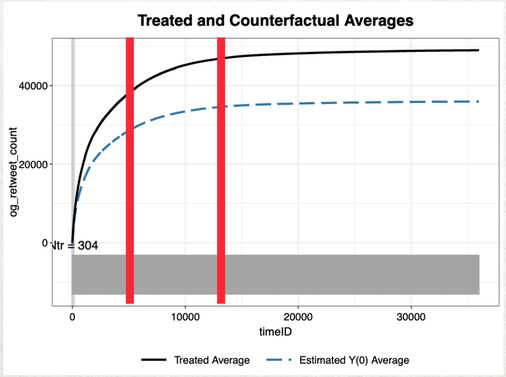


Soft Interventions cause a decrease in FB posts!

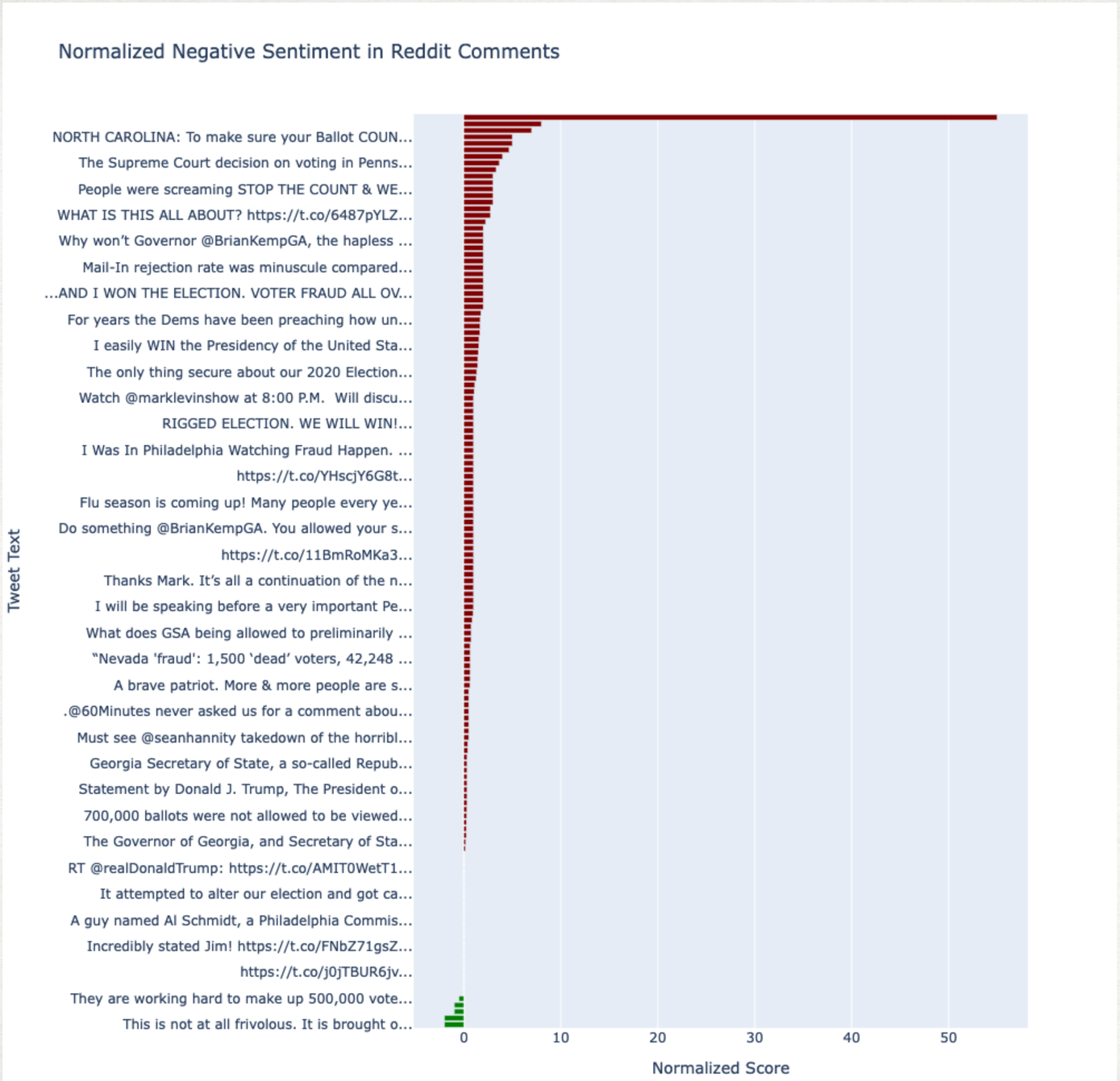
REDDIT - SOFT INTERVENTIONS



t0 t1



Soft Interventions cause an increase in Reddit posts…



...but with mostly negative sentiment!

$$s = \frac{\sum_{i=1}^{n} \mathbb{1} - \sum_{i=1}^{p} \mathbb{1}}{\sum_{i=1}^{u} \mathbb{1} + k}$$

normalized negative sentiment score

PROBLEMS WITH CROWD TANGLE DATA COLLECTION

- Fit regression model to estimate fixed False Positive Rate for search results on each tweet
- Remove the false matched posts from the original tweets
- Reestimate trajectories based on correct matches
- Unfortunately not left with enough reliable data on hard interventions to estimate confidence intervals… except for one thing!

FACEBOOK - HARD INTERVENTIONS

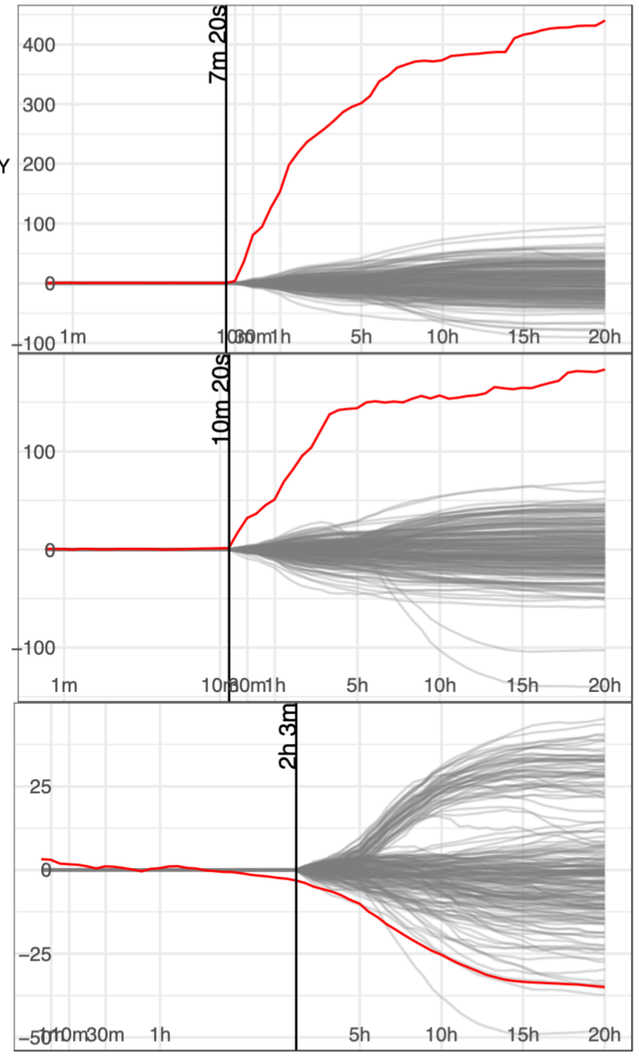
Facebook

ANY VOTE THAT CAME IN AFTER ELECTION DAY WILL NOT BE COUNTED!

.....there was a large number of secretly dumped ballots as has been widely reported!

***

https://t.co/11BmRoMKa3



- Run a "placebo test"
- Sample 30 unintervened tweets at random, pretend they are treated and estimate ATT through tjbal – null distribution
- Then do the same for the actual hard intervention since time of intervention is known!
- Able to estimate p-values and find that hard interventions do increase number of FB posts about the tweet!

***

SUMMARY

- There was a milder Streisand effect on Twitter than originally thought, but a statistically significant one nevertheless
- Interventions on a single platform have a downstream effect on the ecosystem; cross platform effects estimation has its own set of challenges
- Need more detailed analysis of post and comment contents to understand broader impact of interventions
- We need more evidence-based evaluation for policy interventions deployed by social platforms!

COLLABORATORS

James Bisbee

Jonathan Nagler

Richard Bonneau

Joshua Tucker

LET'S TALK!

@swapneel_mehta
swapneelm.github.io
swapneel.mehta@nyu.edu