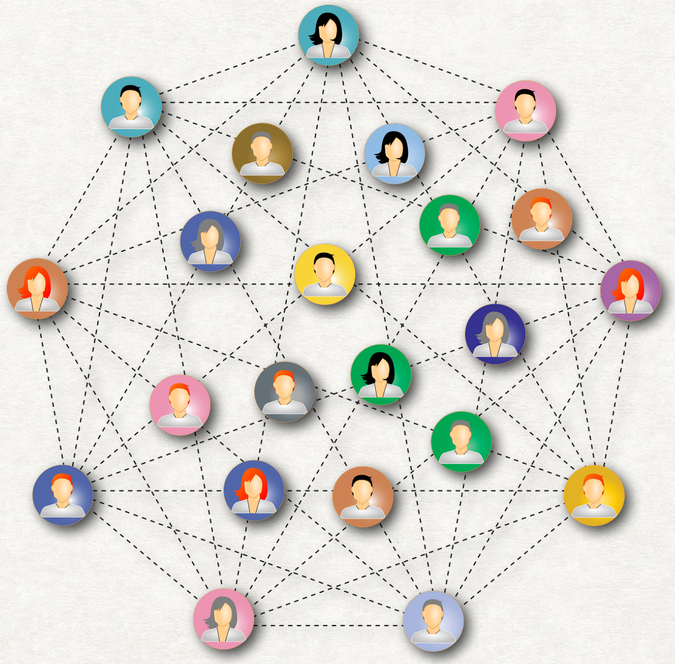
TOOLS FOR MISINFORMATION CONTROL ON SOCIAL NETWORKS



MIT MEDIA LAB

SWAPNEEL MEHTA
NYU CDS / CSMAP

ABOUT ME

swapneelm.github.io

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

Social Networks, Probabilistic Inference, Causal Discovery

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 —

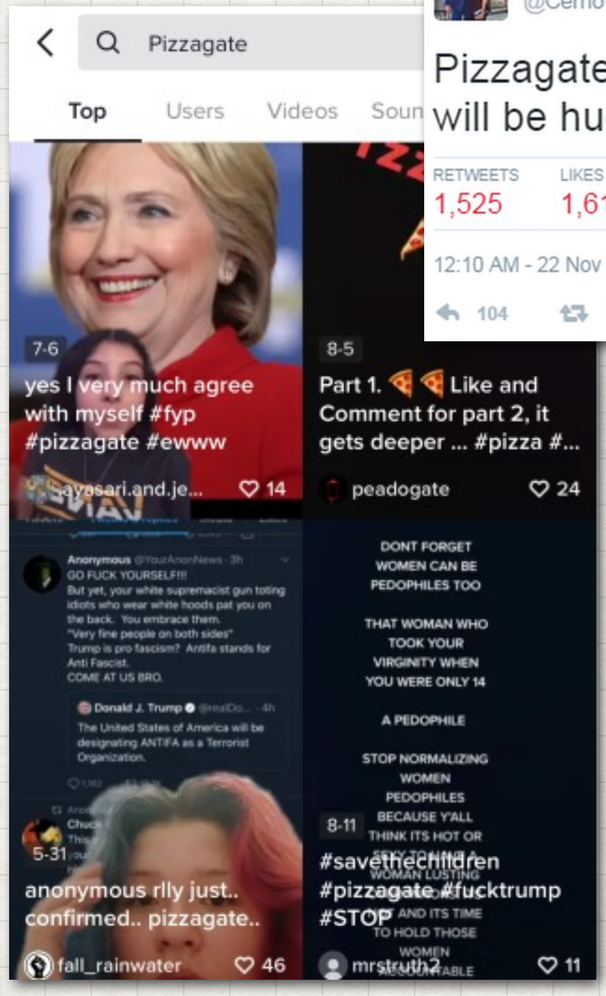
"Hacker X"—the American who built a pro-Trump fake news empire—unmasks himself

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

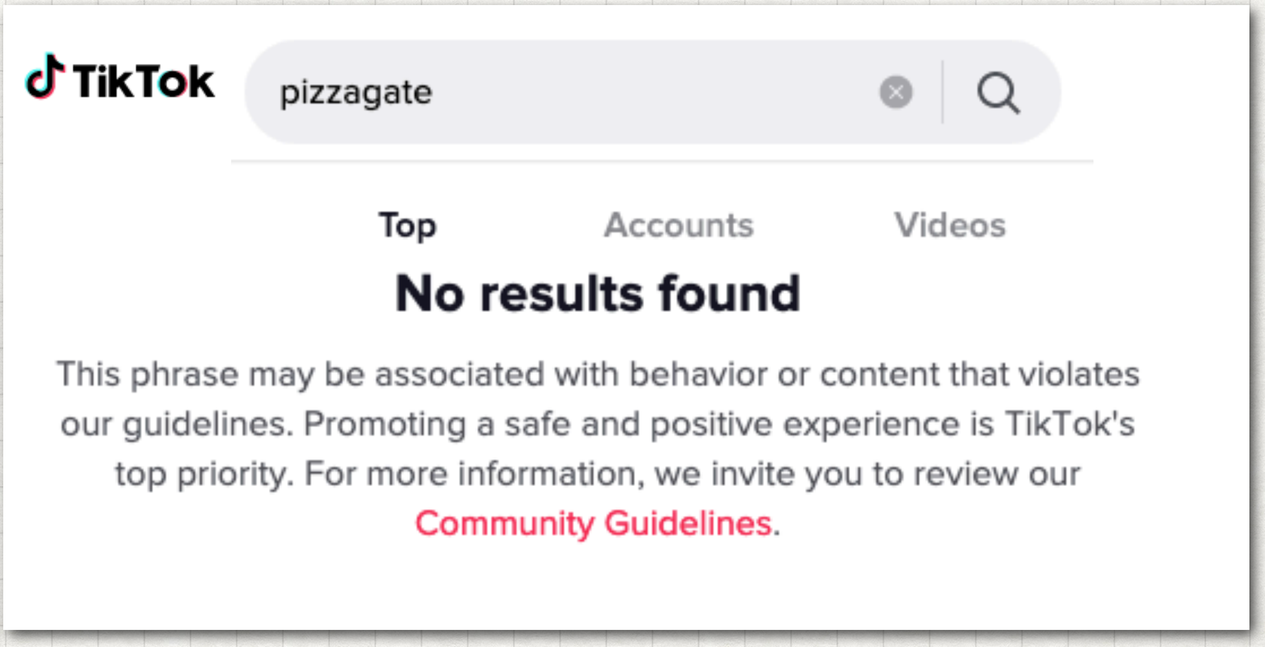
Gary King and Nathaniel Persily

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?

DEBUG POLICY INTERVENTIONS!

DEBUGGING POLICY INTERVENTIONS

I. Problems with How Interventions are Studied

II. Measuring Harms on Social Networks

III. Applying Techniques in Practice

DEBUGGING POLICY INTERVENTIONS

I. **Problems with How Interventions are Studied**

II. Measuring Harms on Social Networks

III. Applying Techniques in Practice

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

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

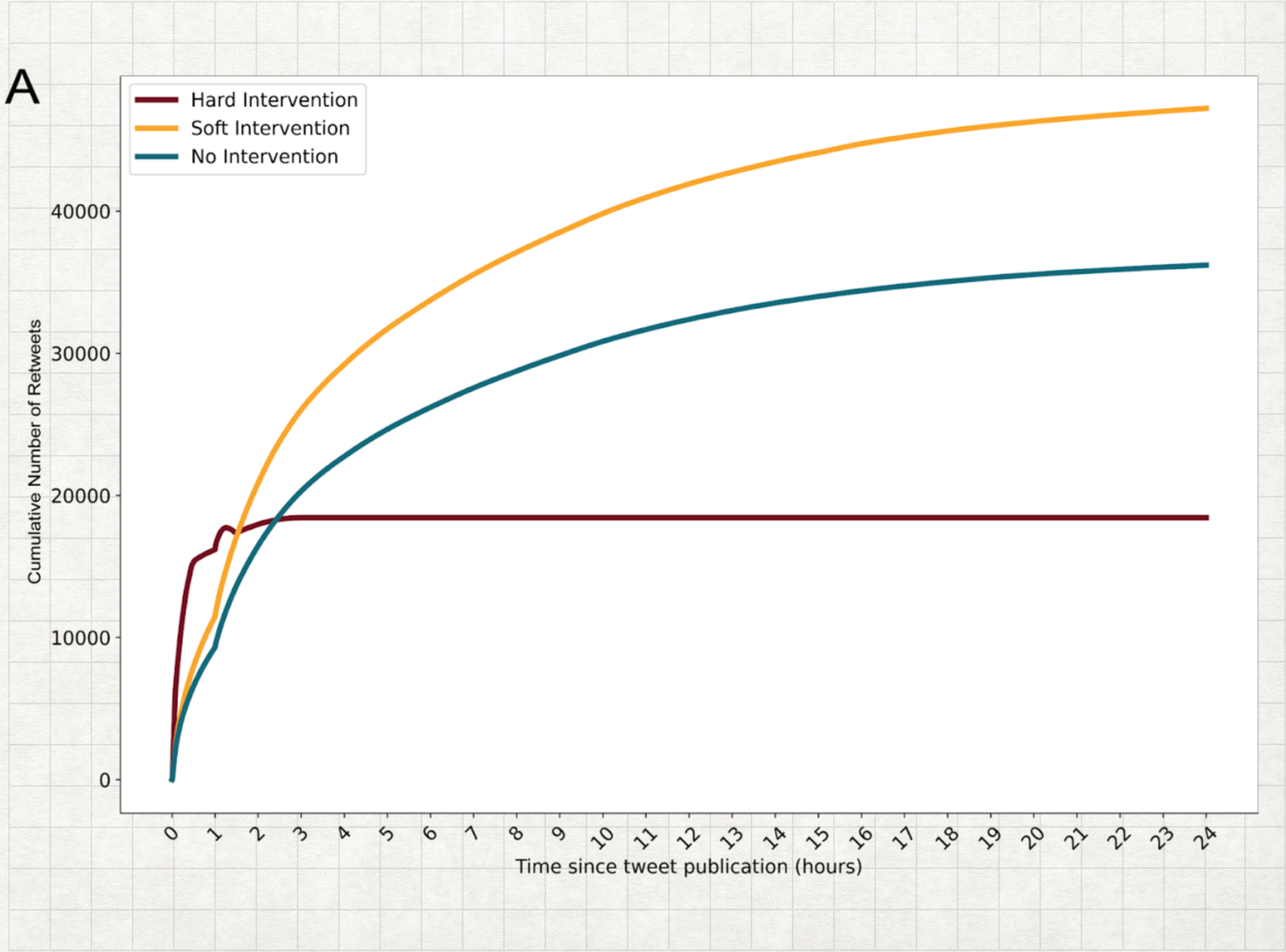
CAUSAL EFFECTS

- Estimating the causal effect of Twitter's Interventions on Trump's Tweets
- Two types of interventions: **warning labels** and **removal**
- Looking at multiple platforms reveal surprising conclusions about the interventional effects

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

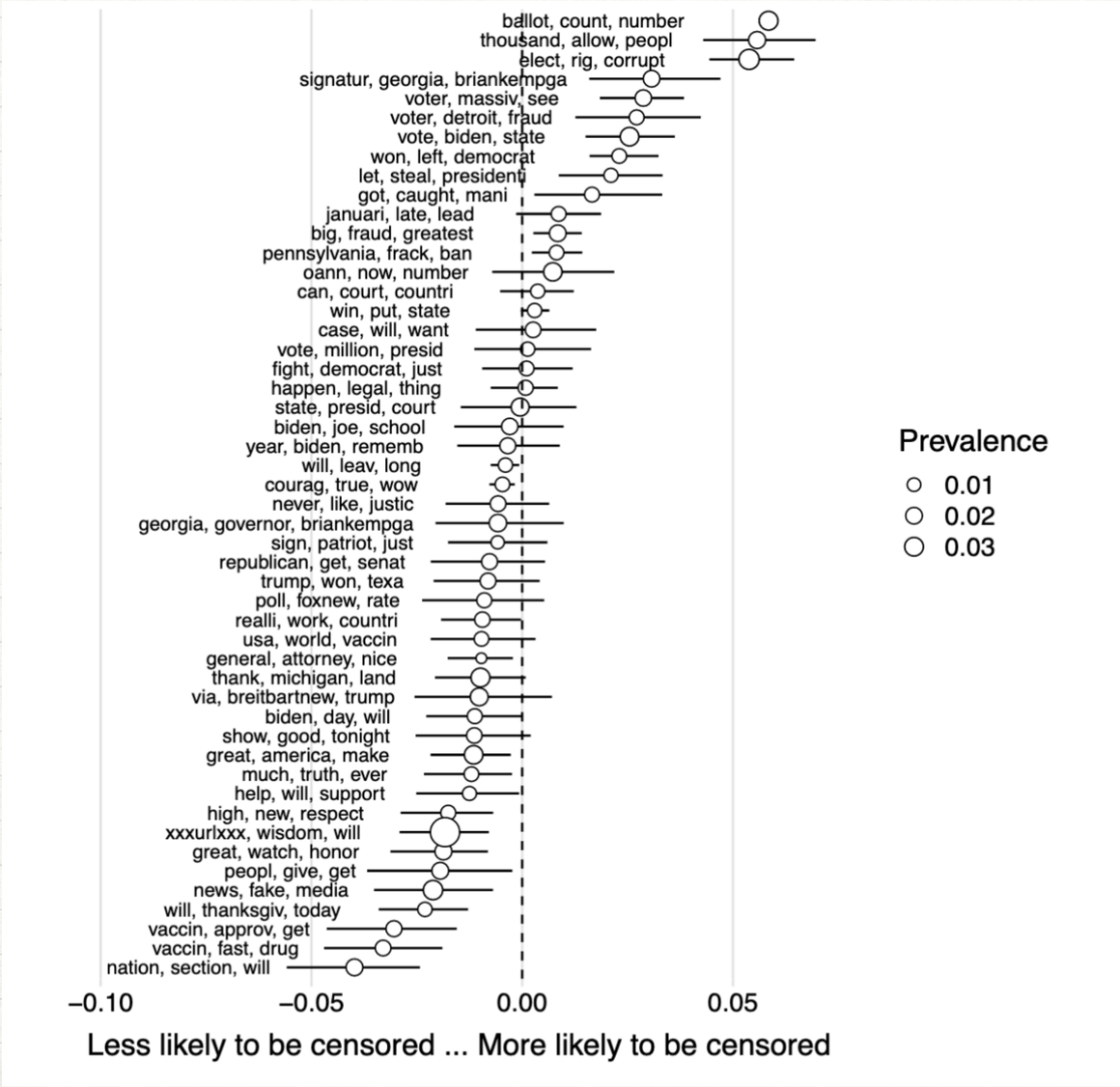
- Naive estimates by Sanderson et. al (2021), indicate strong Streisand effect!
- **Wait does this mean interventions are bad?!**
- What tweets are we really comparing here and what are their features like?



HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

- 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

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

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**
  - **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)

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

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)

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

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

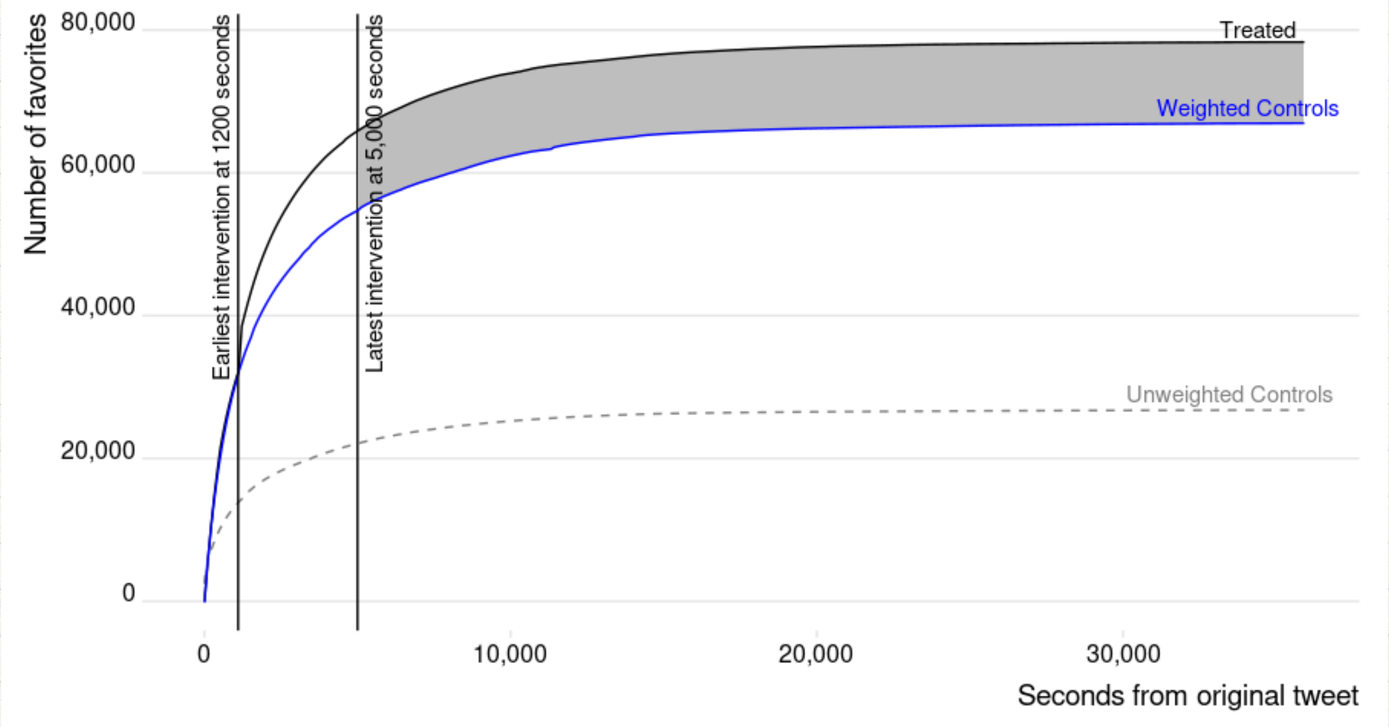
Relaxes Linearity in Prior Outcomes (LPO)

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

HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

CAUSAL EFFECTS

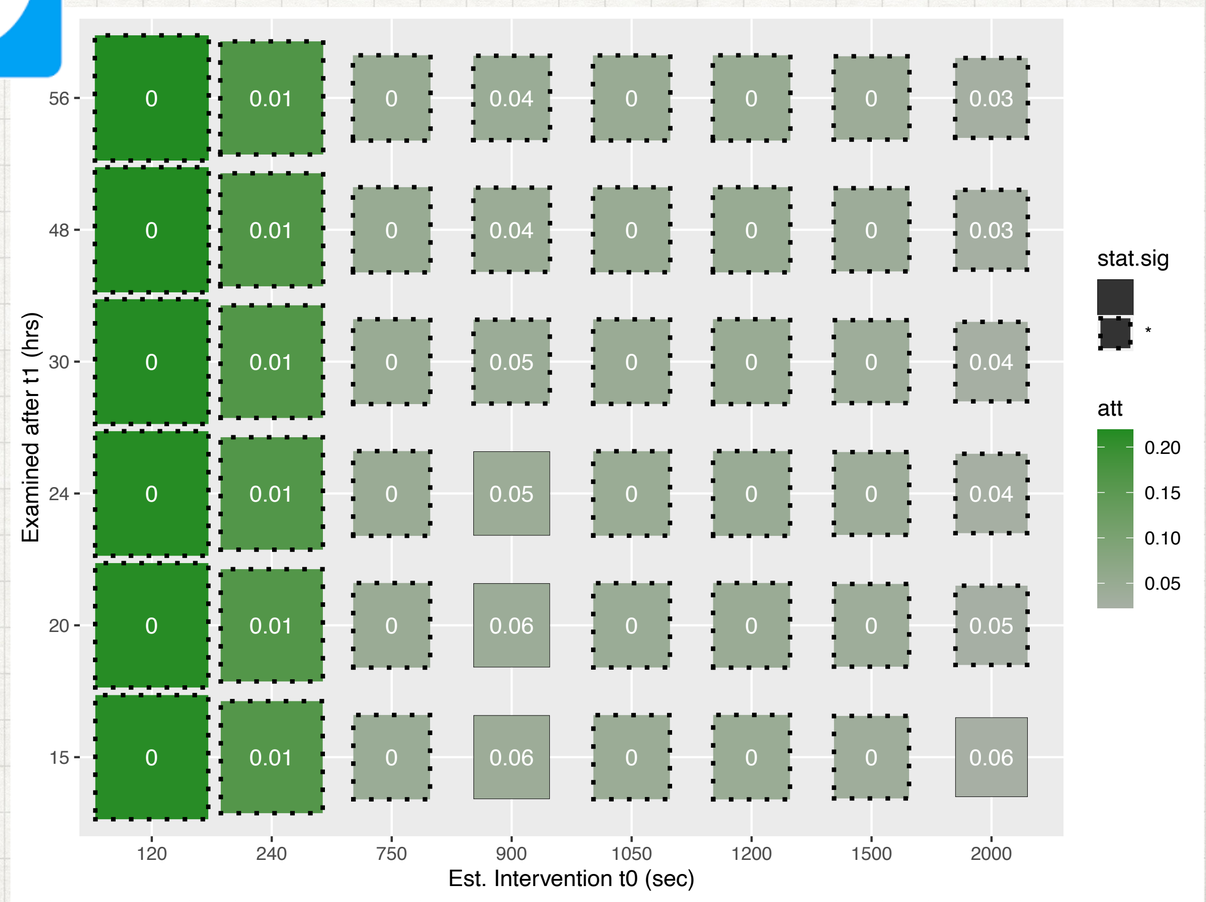
- Robustness check over assumptions of earliest to latest time they could have intervened!
- T0 = earliest possible intervention
- T1 = latest possible intervention
- Kernel balanced estimates are most robust to changes



RESULTS

WHAT DID INTERVENTIONS CAUSE?

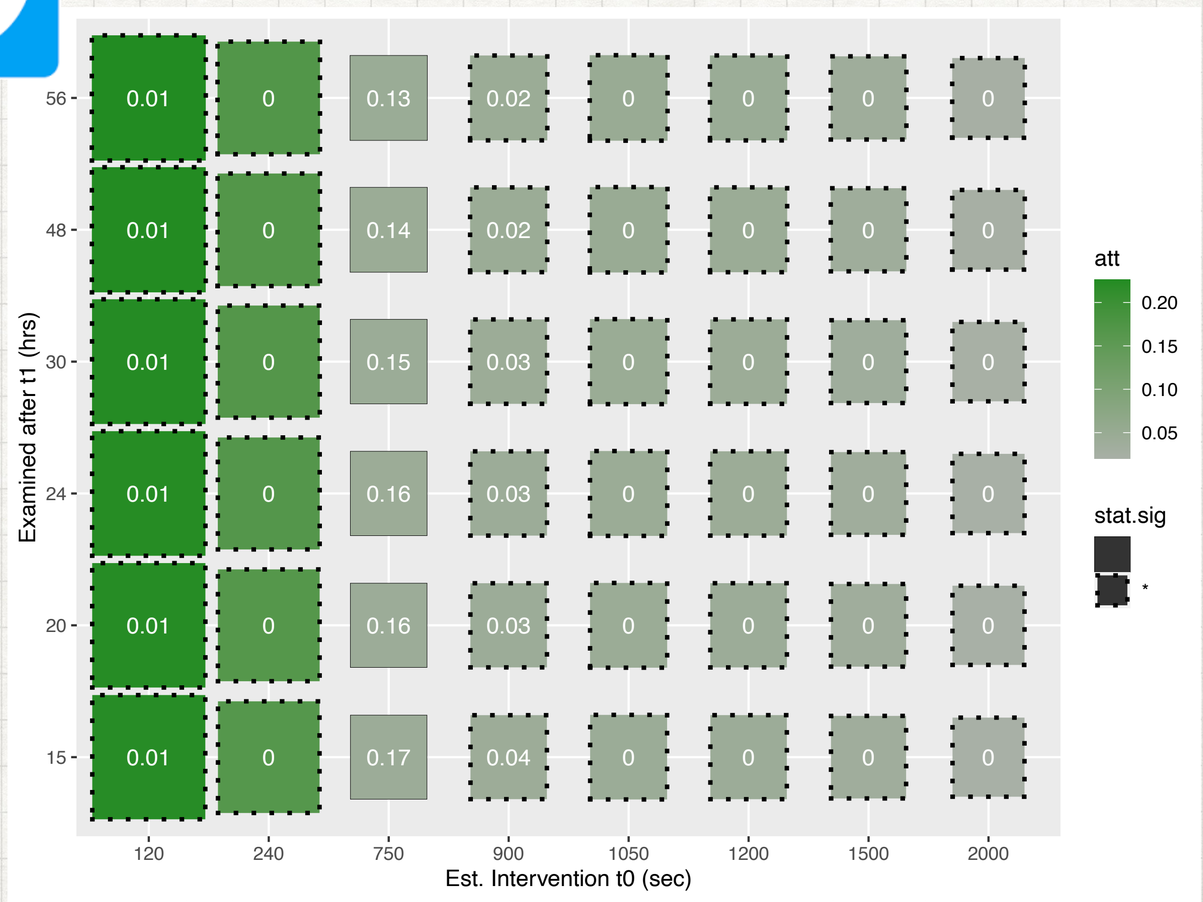
- Streisand effect present but milder on Twitter
- Interventions on other platforms:
  - Reduced public discussions of Trump's tweets
  - Marginal increase in private discussions but needs analysis of content
- Interventions have distinct cross-platform effects on public discourse!



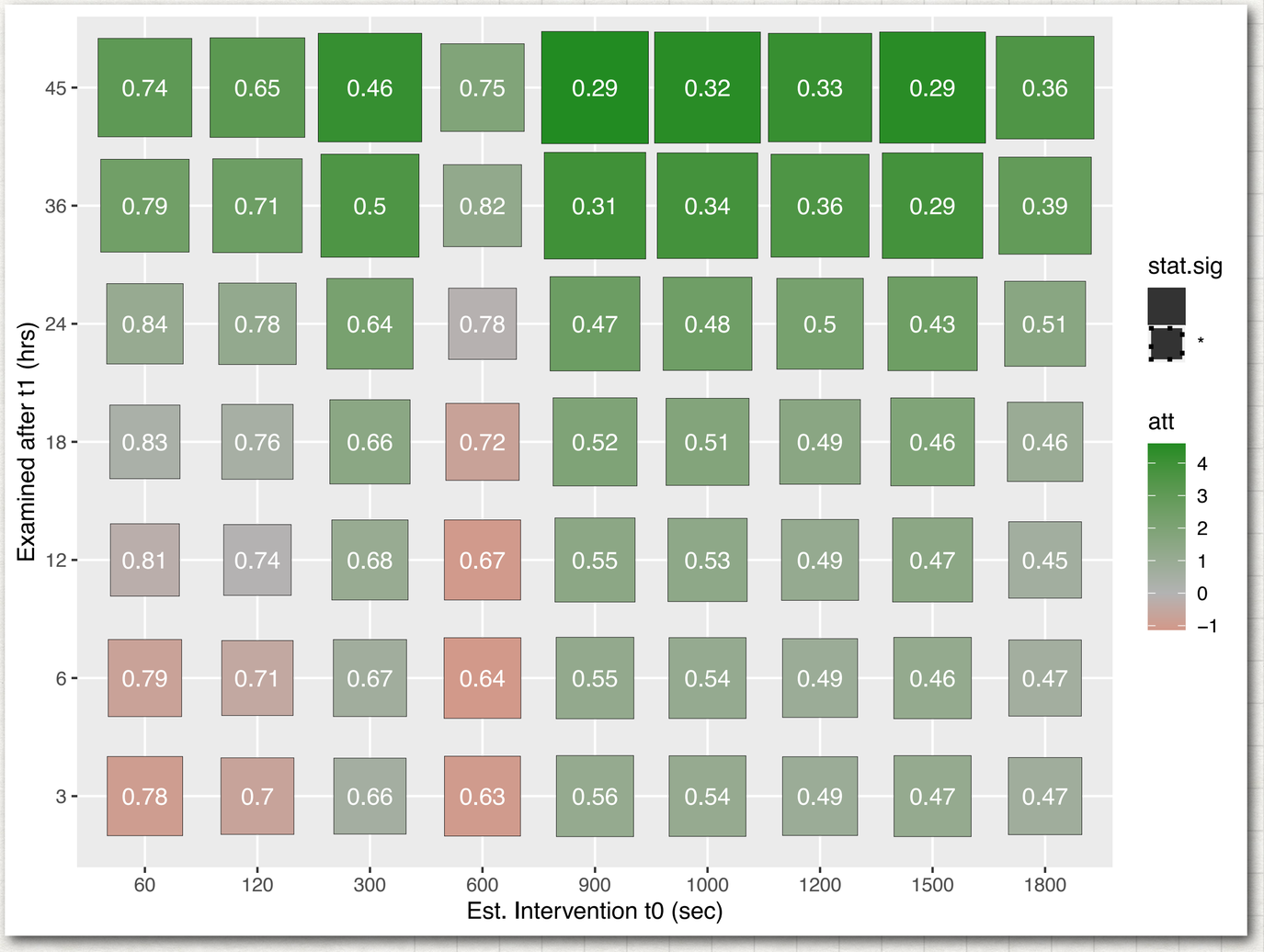
RESULTS

WHAT DID INTERVENTIONS CAUSE?

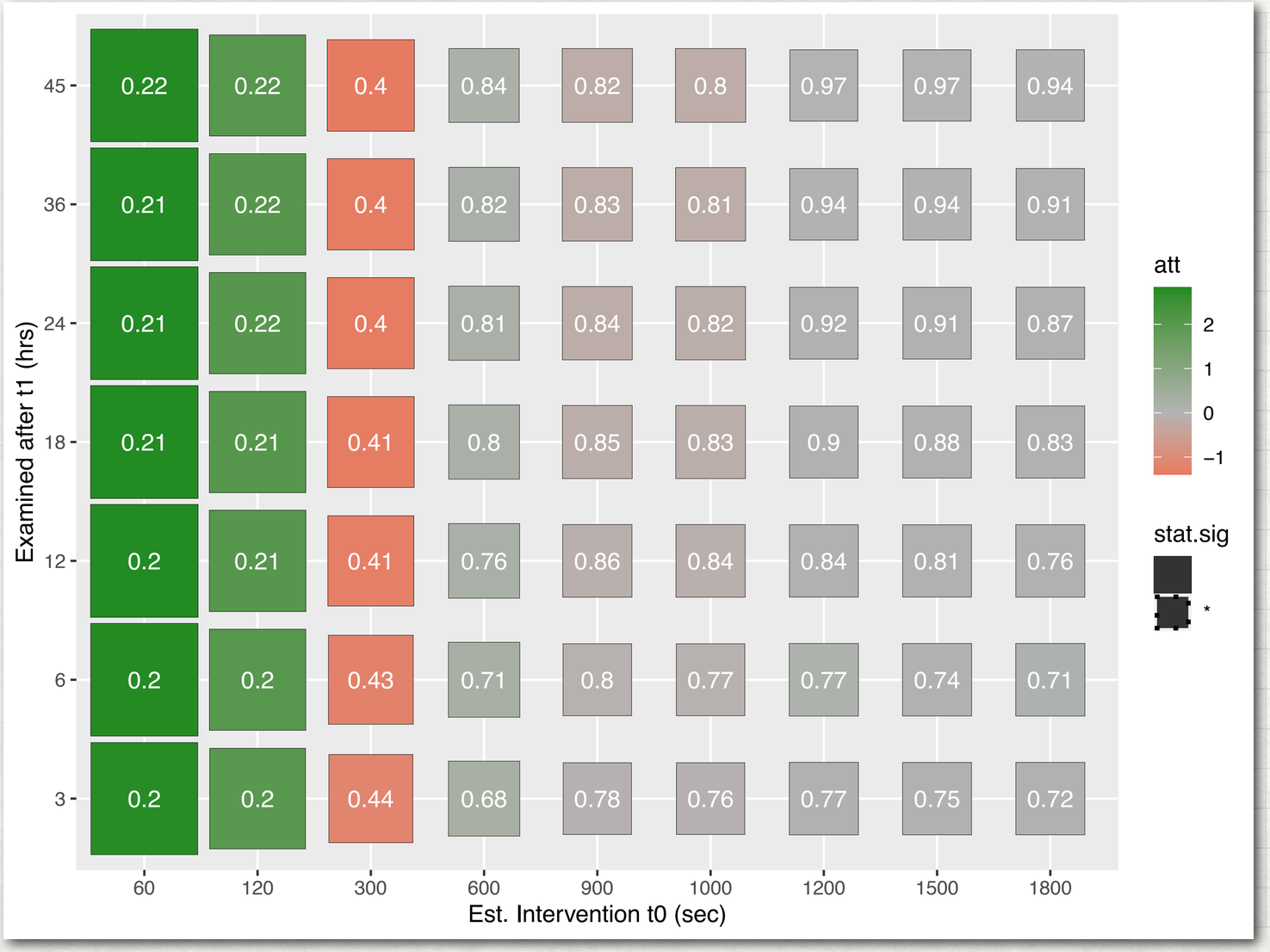
- Streisand effect present but milder on Twitter
- Interventions on other platforms:
  - Reduced public discussions of Trump's tweets
  - Marginal increase in private discussions but needs analysis of content
- Interventions have distinct cross-platform effects on public discourse!



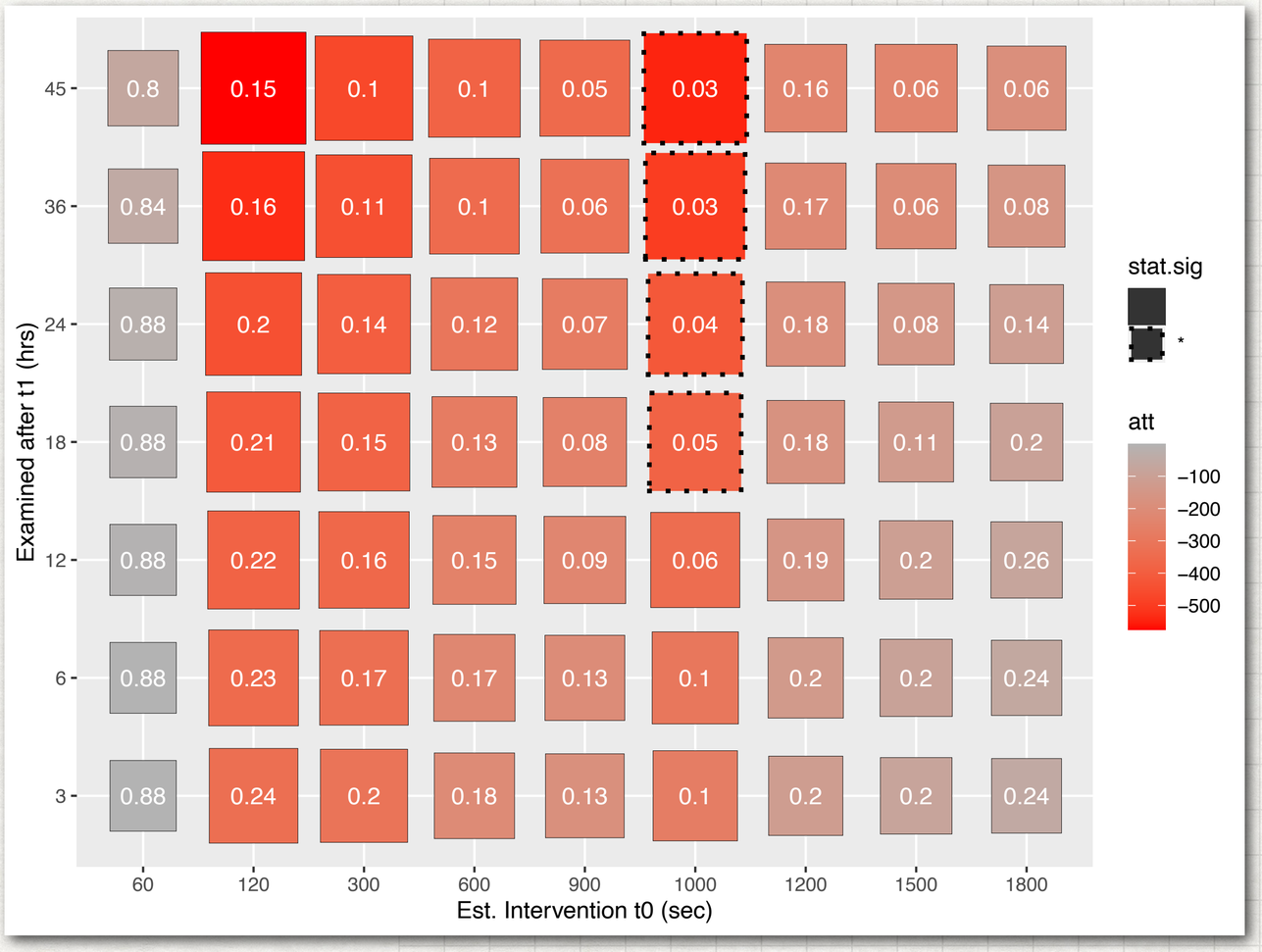
REDDIT - HARD


45
36
24
18
12
6
3
0.74 0.65 0.46 0.75 0.29 0.32 0.33 0.29 0.36
0.79 0.71 0.5 0.82 0.31 0.34 0.36 0.29 0.39
0.84 0.78 0.64 0.78 0.47 0.48 0.5 0.43 0.51
0.83 0.76 0.66 0.72 0.52 0.51 0.49 0.46 0.46
0.81 0.74 0.68 0.67 0.55 0.53 0.49 0.47 0.45
0.79 0.71 0.67 0.64 0.55 0.54 0.49 0.46 0.47
0.78 0.7 0.66 0.63 0.56 0.54 0.49 0.47 0.47
Examined after t1 (hrs)
60 120 300 600 900 1000 1200 1500 1800
Est. Intervention t0 (sec)
stat.sig
*
att
4
3
2
1
0
−1

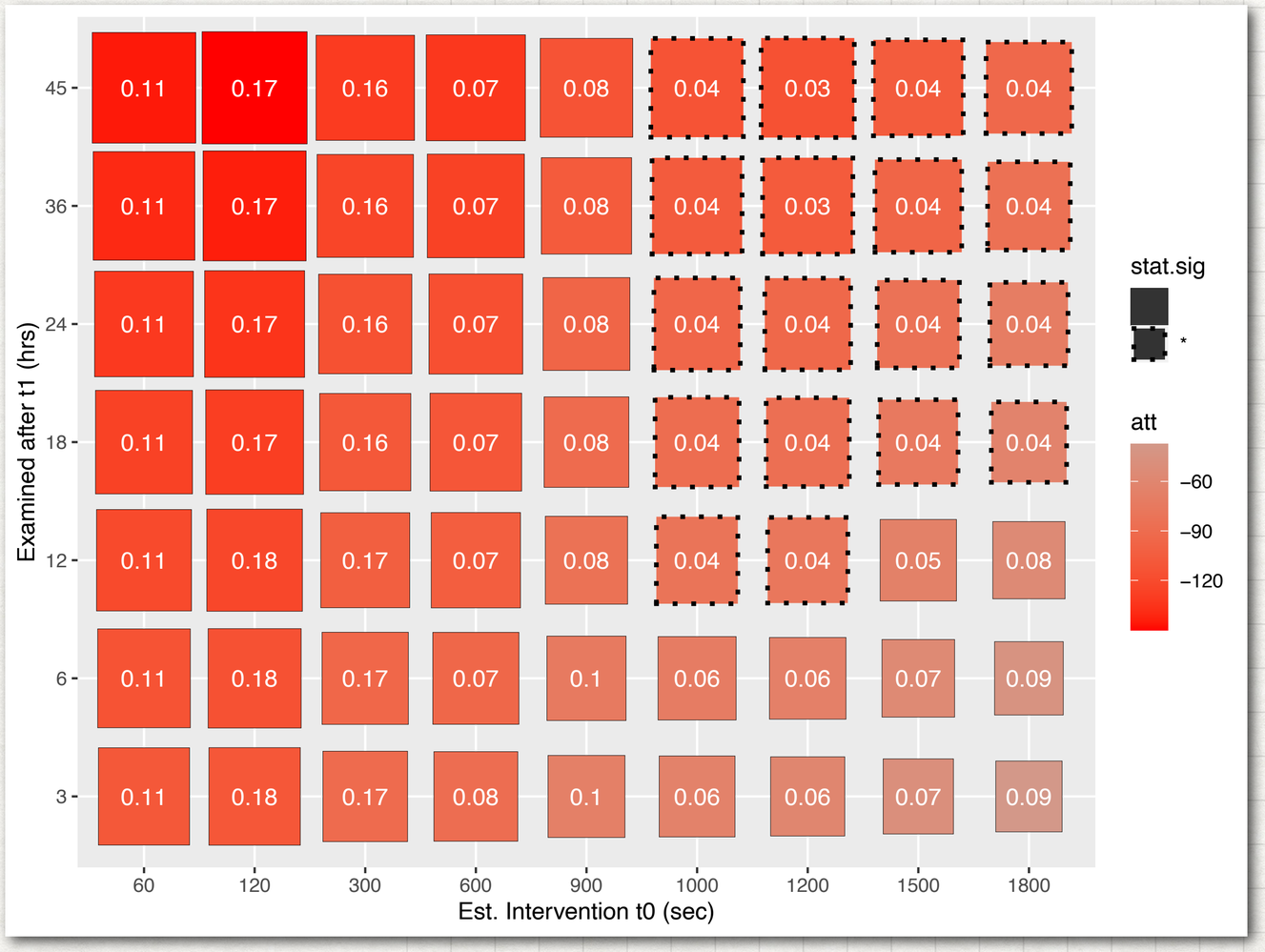
REDDIT - SOFT


Examined after t1 (hrs)
45
36
24
18
12
6
3
0.22 0.22 0.4 0.84 0.82 0.8 0.97 0.97 0.94
0.21 0.22 0.4 0.82 0.83 0.81 0.94 0.94 0.91
0.21 0.22 0.4 0.81 0.84 0.82 0.92 0.91 0.87
0.21 0.21 0.41 0.8 0.85 0.83 0.9 0.88 0.83
0.2 0.21 0.41 0.76 0.86 0.84 0.84 0.81 0.76
0.2 0.2 0.43 0.71 0.8 0.77 0.77 0.74 0.71
0.2 0.2 0.44 0.68 0.78 0.76 0.77 0.75 0.72
60
120
300
600
900
1000
1200
1500
1800
Est. Intervention t0 (sec)
att
2
1
0
−1
stat.sig
*

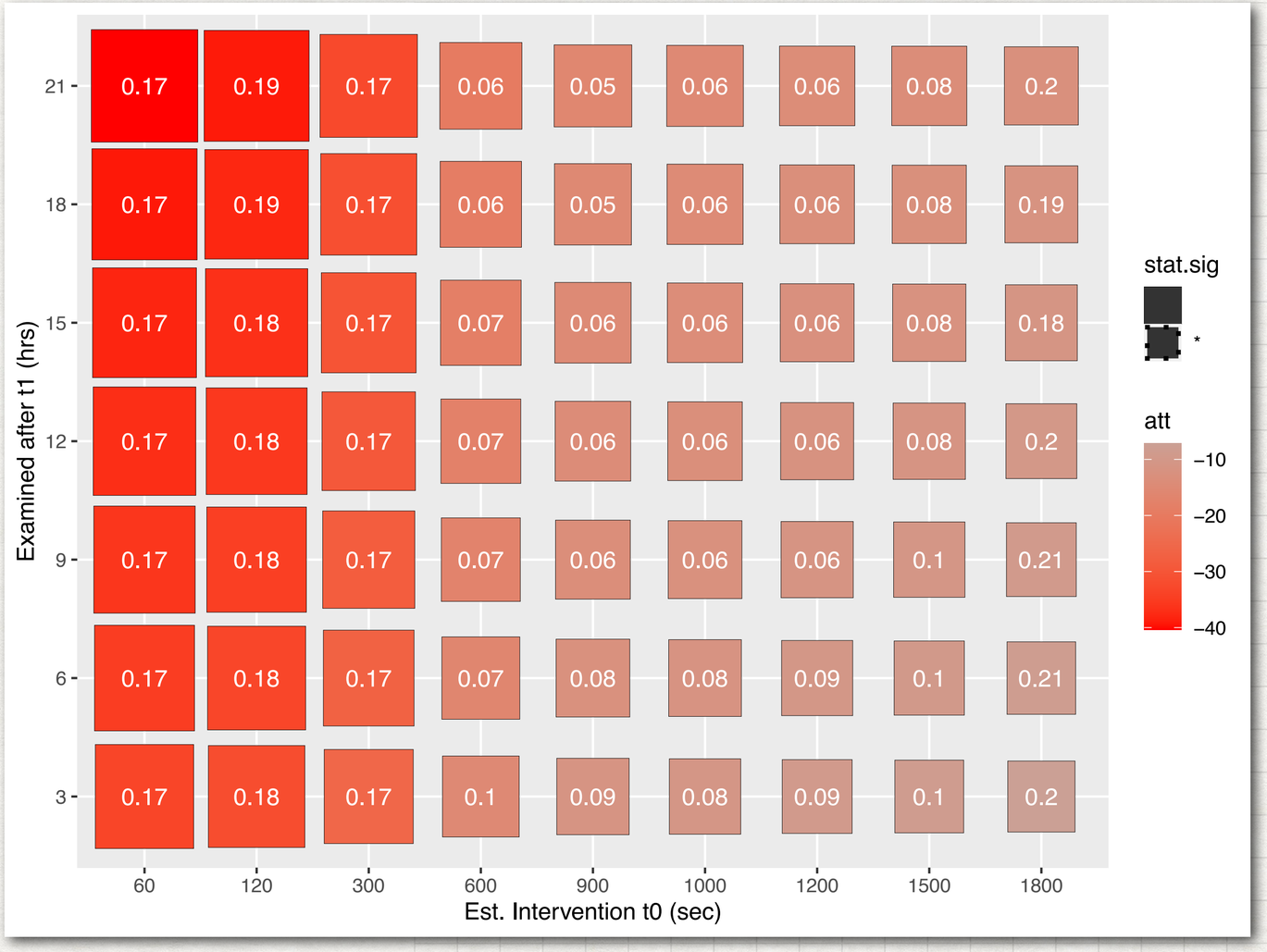
FACEBOOK - HARD



FACEBOOK - SOFT



INSTAGRAM - SOFT



DEBUGGING POLICY INTERVENTIONS

I. Problems with How Interventions are Studied

II. Measuring Harms on Social Networks

III. Applying Techniques in Practice

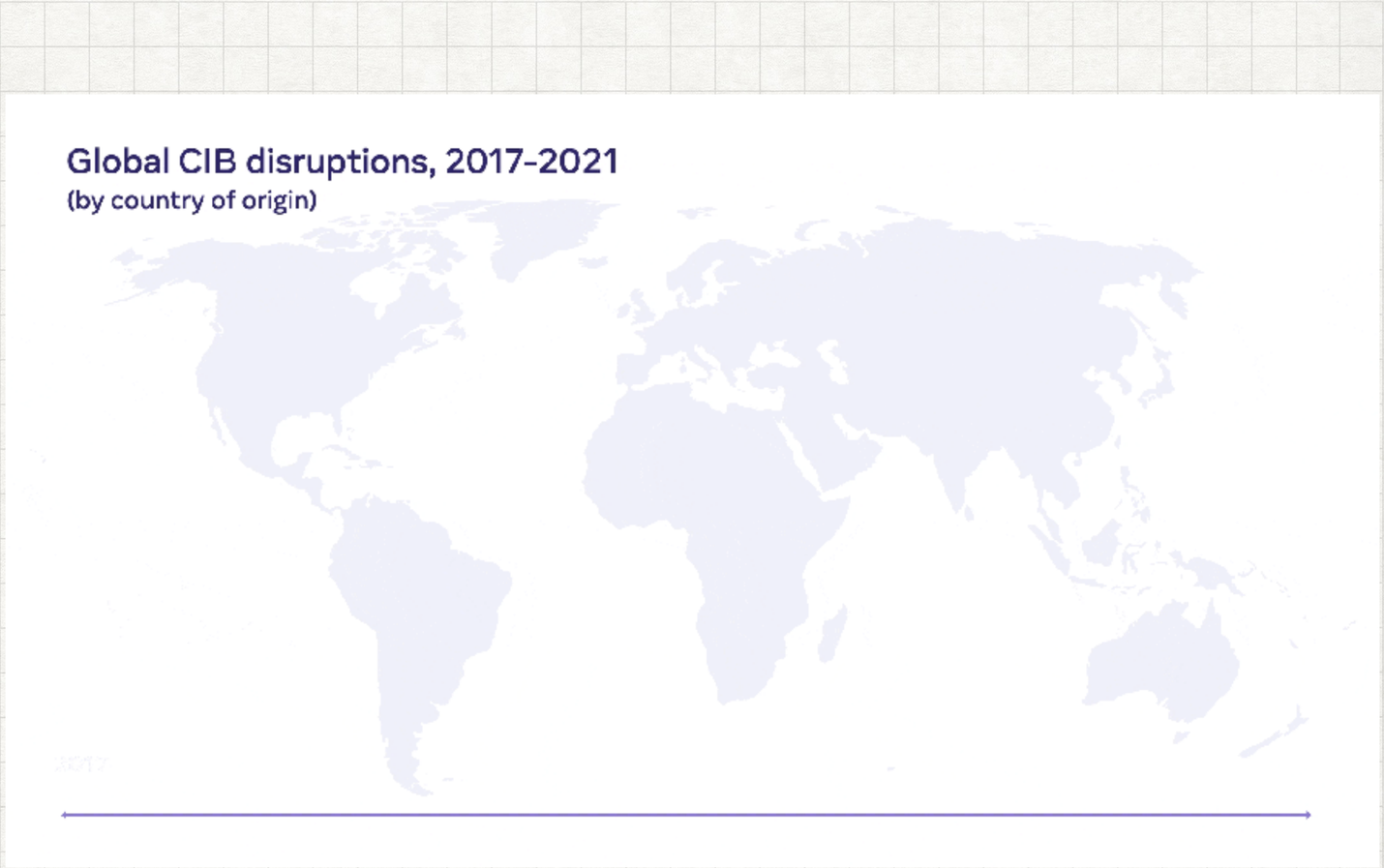
ESTIMATING THE IMPACT OF COORDINATED INAUTHENTIC BEHAVIOR ON CONTENT RECOMMENDATIONS IN SOCIAL NETWORKS

COORDINATED INAUTHENTIC BEHAVIOR

- Inauthentic behavior is defined as the use of assets (accounts, Pages, Groups, or Events), **to mislead people.**
- There are global coordinated networks of accounts promoting disinformation on social networks.

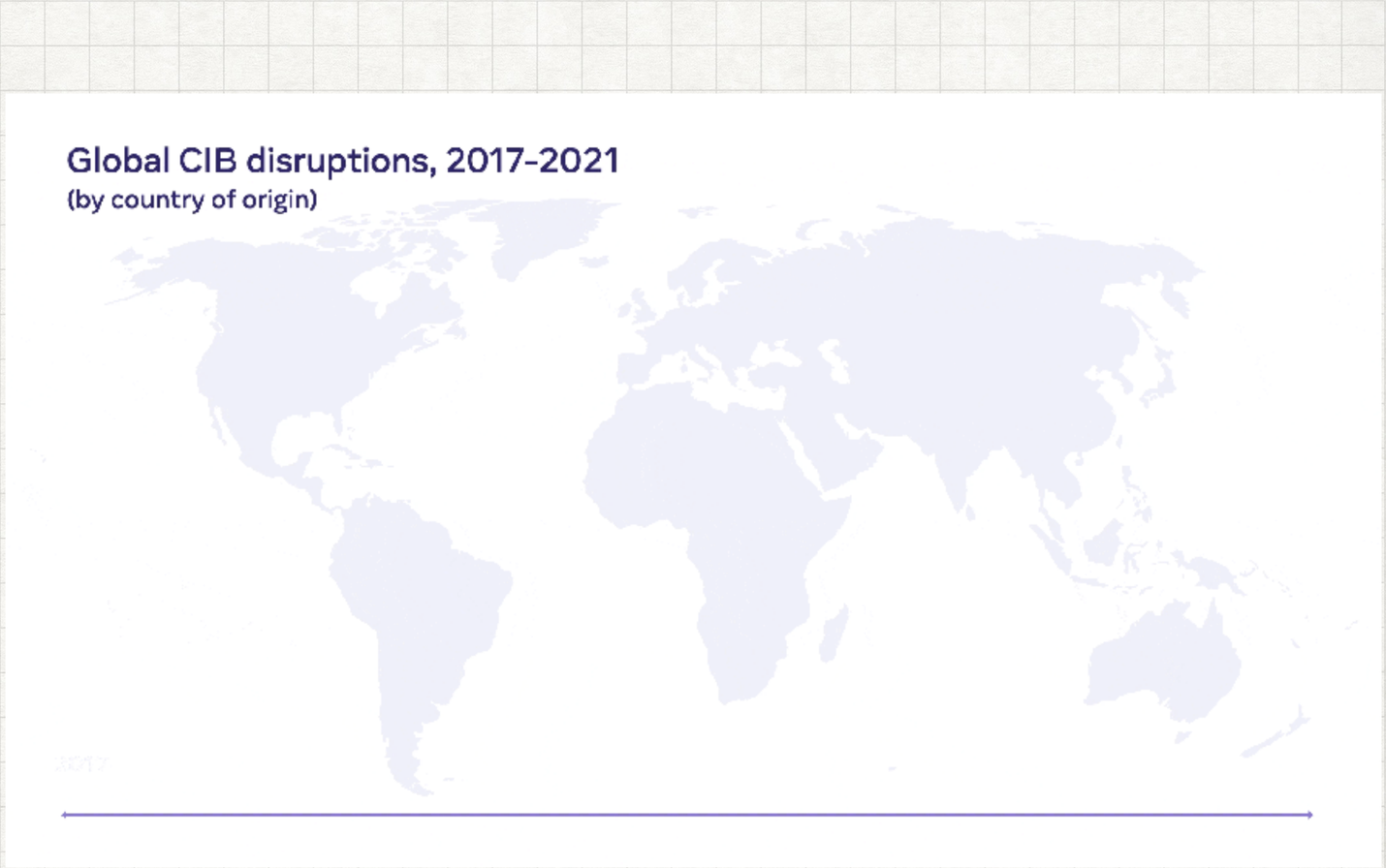
COORDINATED INAUTHENTIC BEHAVIOR

- Inauthentic behavior is defined as the use of assets (accounts, Pages, Groups, or Events), **to mislead people.**
- There are global coordinated networks of accounts promoting disinformation on social networks.
- Meta and Twitter release transparency reports about them a while after taking them down.



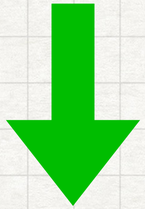
COORDINATED INAUTHENTIC BEHAVIOR

- Inauthentic behavior is defined as the use of assets (accounts, Pages, Groups, or Events), **to mislead people.**
- There are global coordinated networks of accounts promoting disinformation on social networks.
- Meta and Twitter release transparency reports about them a while after taking them down.
- No real-time solution nor verified damage assessment because effects are hard to quantify externally!



(COST TO)
MITIGATE THESE INFLUENCE OPS!

MEASURING THE HARMS DUE TO COORDINATED INAUTHENTIC BEHAVIOR



(COST TO) MITIGATE THESE INFLUENCE OPS!

RESEARCH GOALS

- RQ: Quantify the relative impact of different algorithmic choices that a platform makes, for instance:

RESEARCH GOALS

- RQ: Quantify the relative impact of different algorithmic choices that a platform makes, for instance:
  - To maximize engagement or not to maximize engagement?
  - Should we promote diverse content that isn’t getting early views?

RESEARCH GOALS

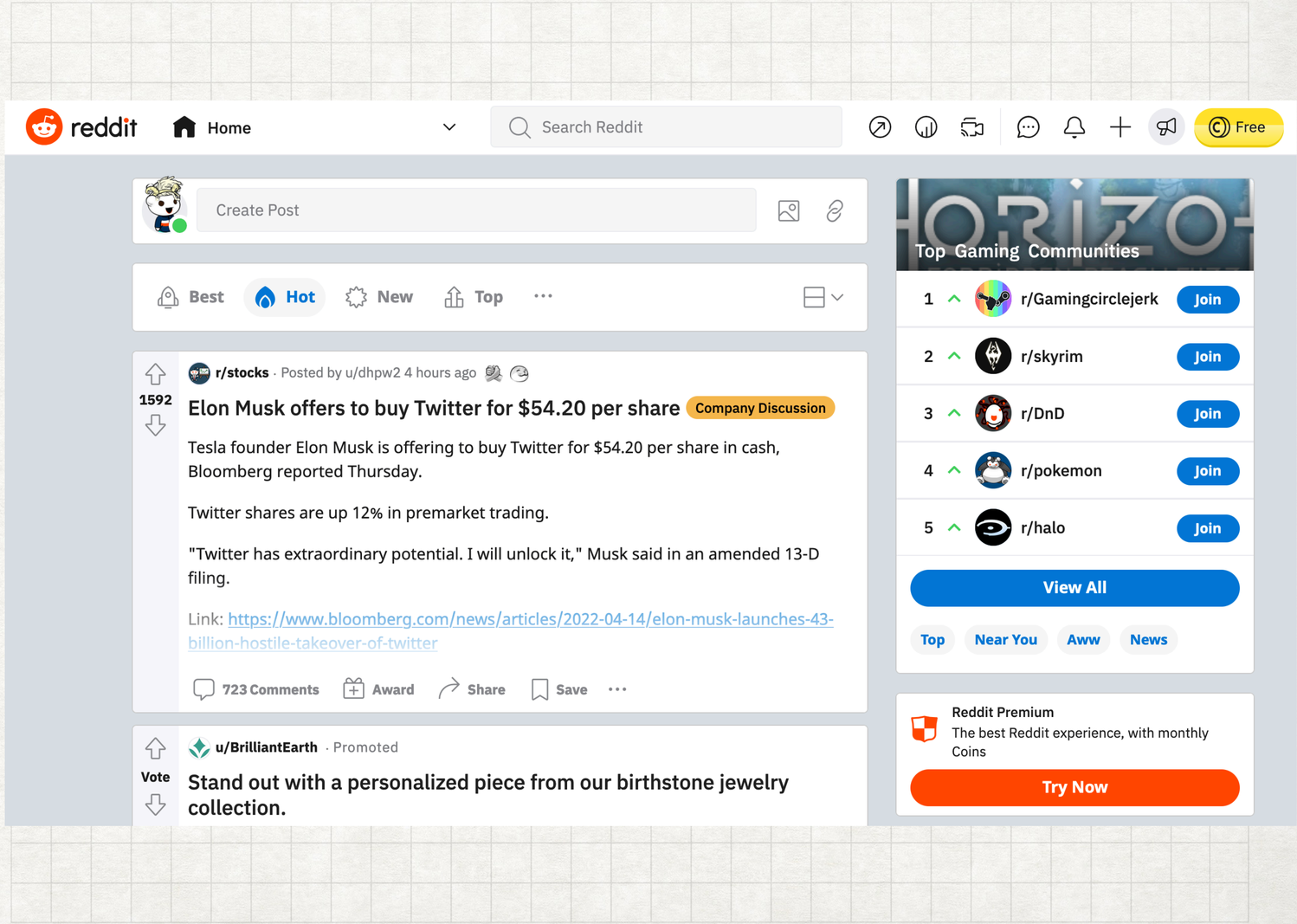
- RQ: Quantify the relative impact of different algorithmic choices that a platform makes, for instance:
  - To maximize engagement or not to maximize engagement?
  - Should we promote diverse content that isn't getting early views?
  - Dealing with "controversial" opinions?

RESEARCH GOALS

- RQ: Quantify the relative impact of different algorithmic choices that a platform makes, for instance:
  - To maximize engagement or not to maximize engagement?
  - Should we promote diverse content that isn't getting early views?
  - Dealing with "controversial" opinions?
  - Are penalties fair — what about false positives?

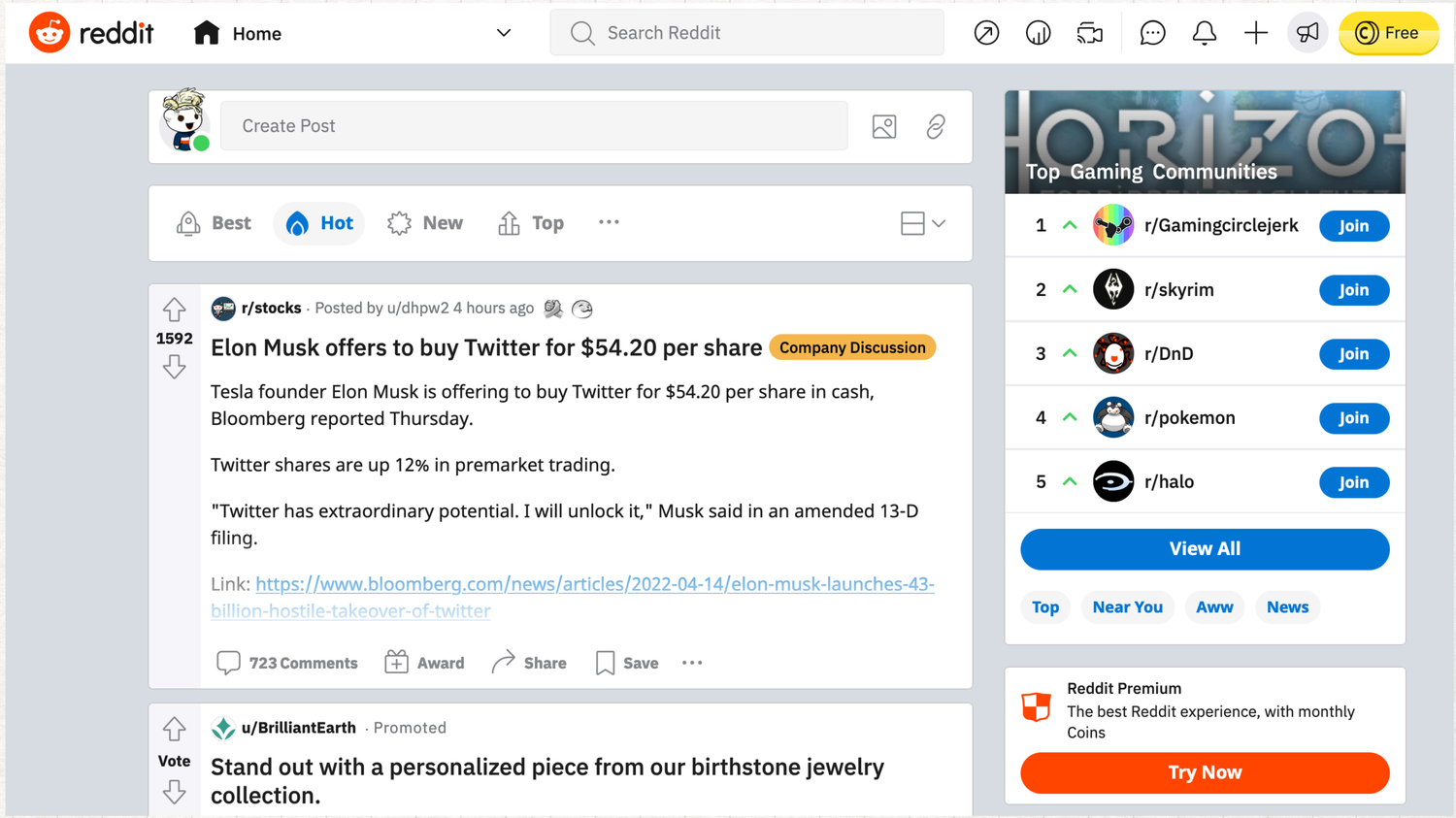
SIMULATE A SOCIAL NETWORK

- Reddit is a pseudonymous social network comprising users who are part of like-minded groups or subreddits
- It has a community-based structure



SIMULATE A SOCIAL NETWORK

- Reddit is a pseudonymous social network comprising users who are part of like-minded groups or subreddits
- It has a community-based structure
- The state-action space for a user includes:
  - Create a post/comment
  - Upvote a post/comment
  - Downvote a post/comment
  - Cross-post an existing post



SIMULATING SOCIAL NETWORKS

REDDIT

- Developing a model for a user's posting behavior
- Creating a story of how a user interacts with Reddit

Algorithm 6: Simulating User Activity on Reddit

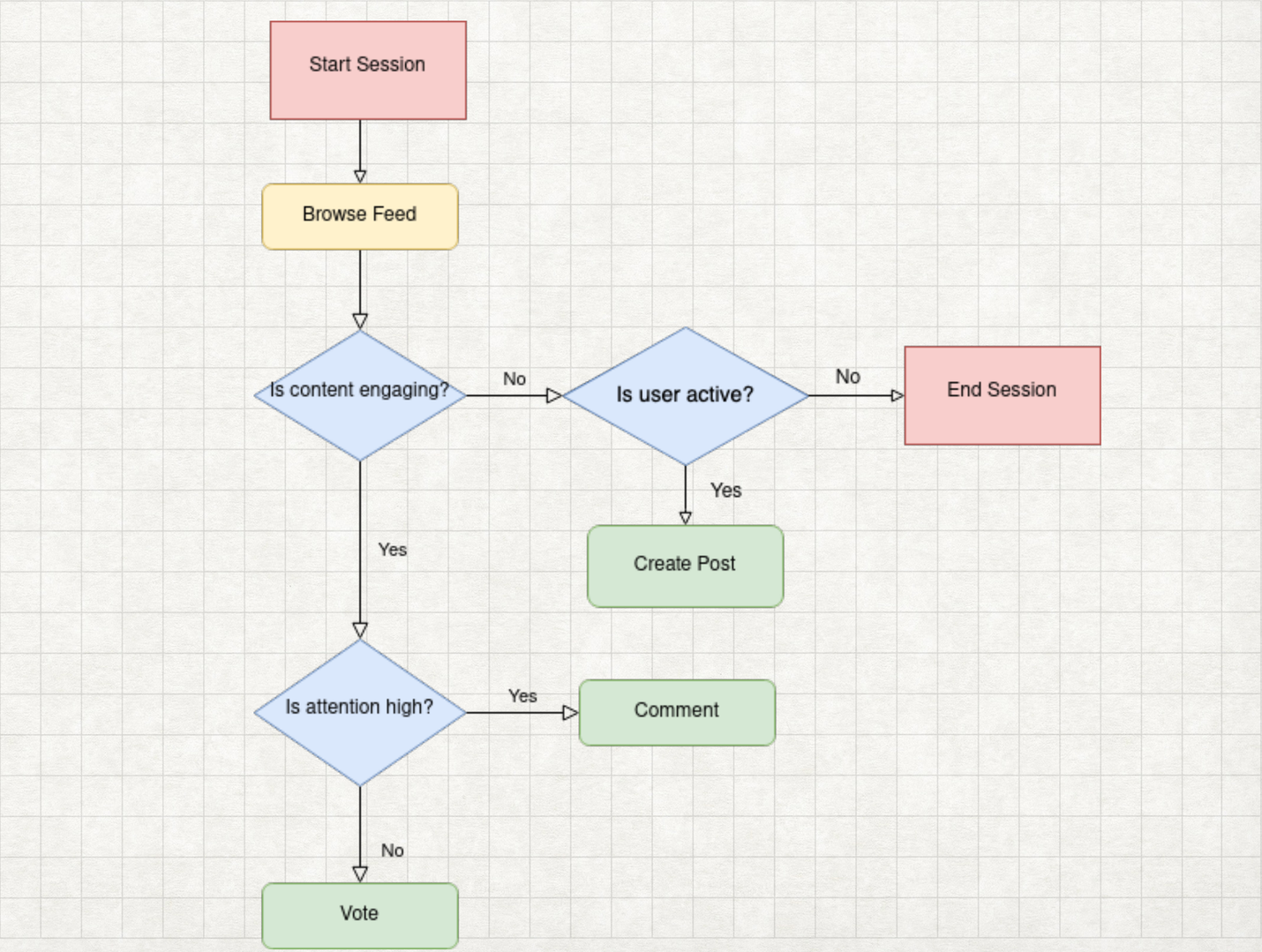
$N \in \mathbb{N}$: Number of users
$T \in \mathbb{N}$: Time steps
$S \in \mathbb{N}$: Number of sub-reddit categories
$\{\pi_i \in \text{Uniform}(0,1)\}_{i=1,j=1}^{N,S}$ ▷ Interaction frequency

procedure REDDIT
Sample latents:
$\boldsymbol{\upsilon} \leftarrow \{\upsilon_i \sim \text{Uniform}(0,1)\}_{i=1,j=1}^{N,S}$ ▷ Interaction propensity over subreddit categories
Simulate:
    $\Phi_{1:N} \leftarrow \langle\rangle$ ▷ User Activity
    for $t = 1 : T$ **do**
        for $i = 1 : N$ **do**
            $\gamma \sim \text{Categorical}(\pi_i)$ ▷ Choose Subreddit (category)
            $\tau \sim \text{Bernoulli}(\upsilon_{i,\gamma})$ ▷ Interact with Subreddit (category)
            $\Phi_i \leftarrow \Phi_i + \langle\tau\rangle$ ▷ Append to user's activity
    return $\Phi_{1:N}$ ▷ All user activity

SIMULATING SOCIAL NETWORKS

REDDIT

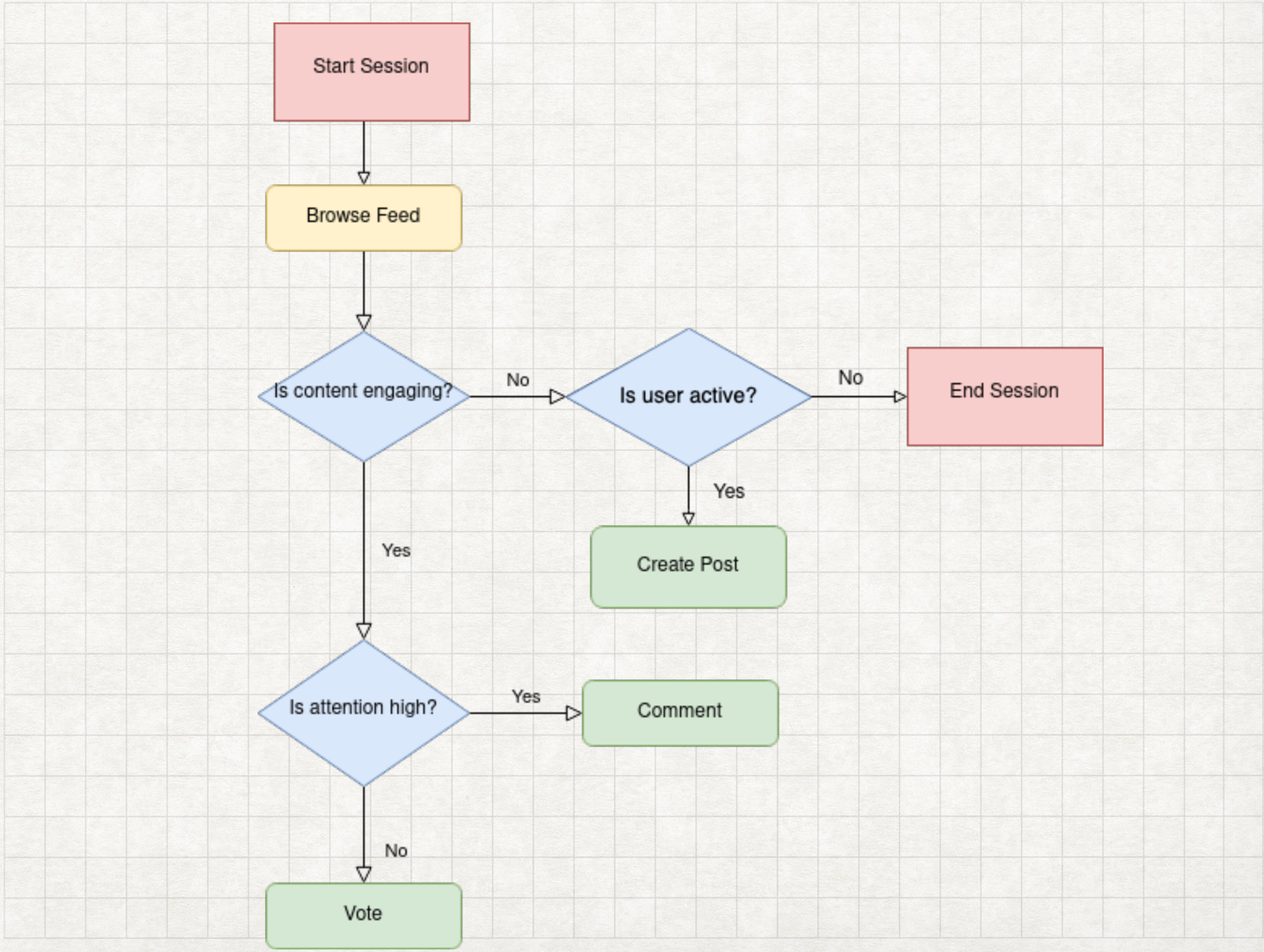
- Developing a model for a user's posting behavior
- Creating a story of how a user interacts with Reddit



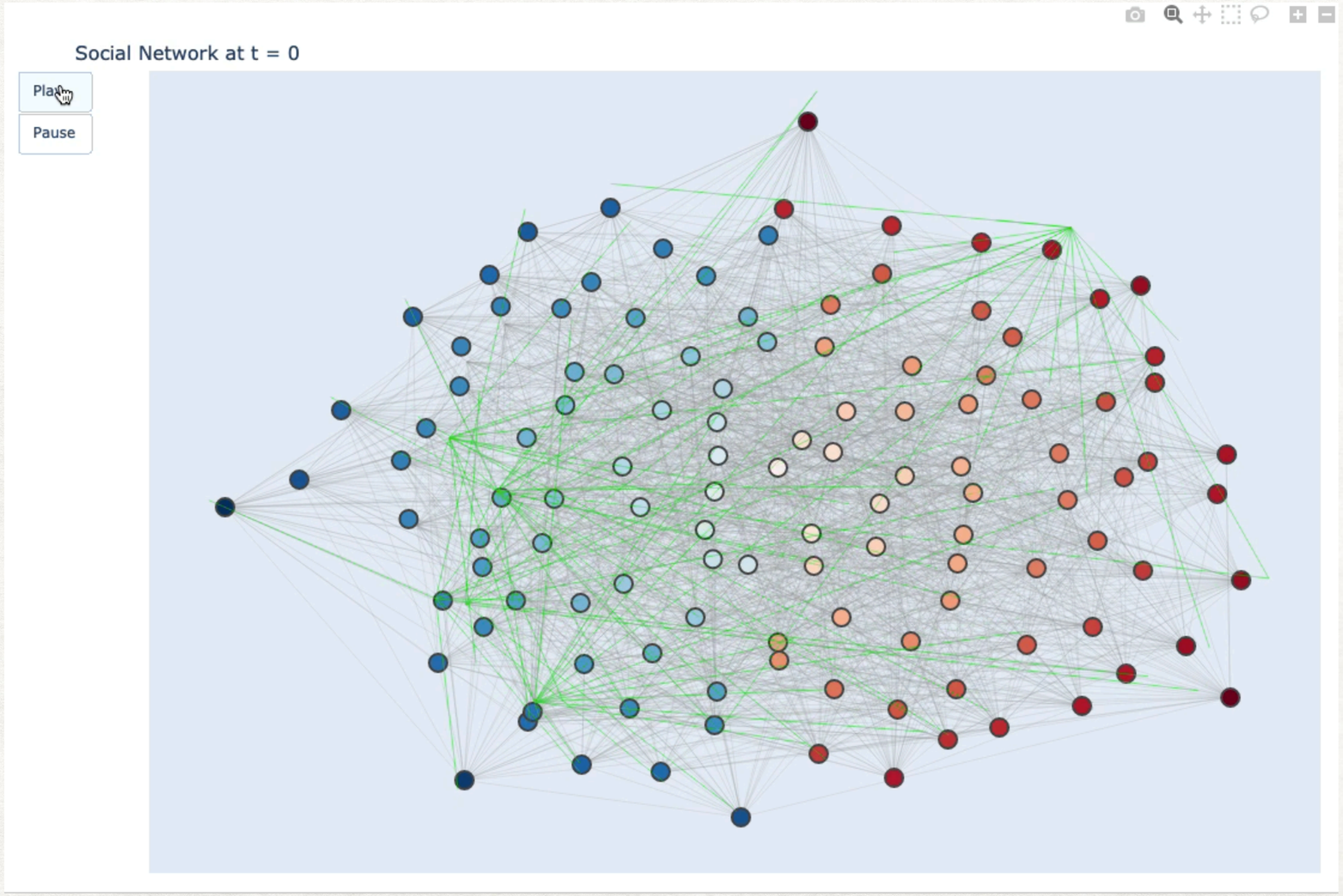
SIMULATING SOCIAL NETWORKS

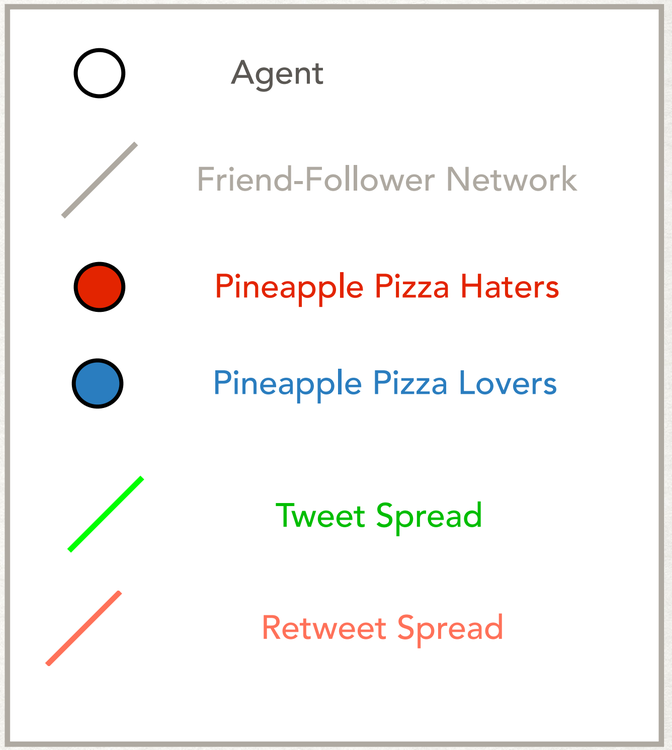
REDDIT

- Developing a model for a user's posting behavior
- Creating a story of how a user interacts with Reddit
- Use the data to set priors on interaction frequency
- Simulate counterfactual outcomes!



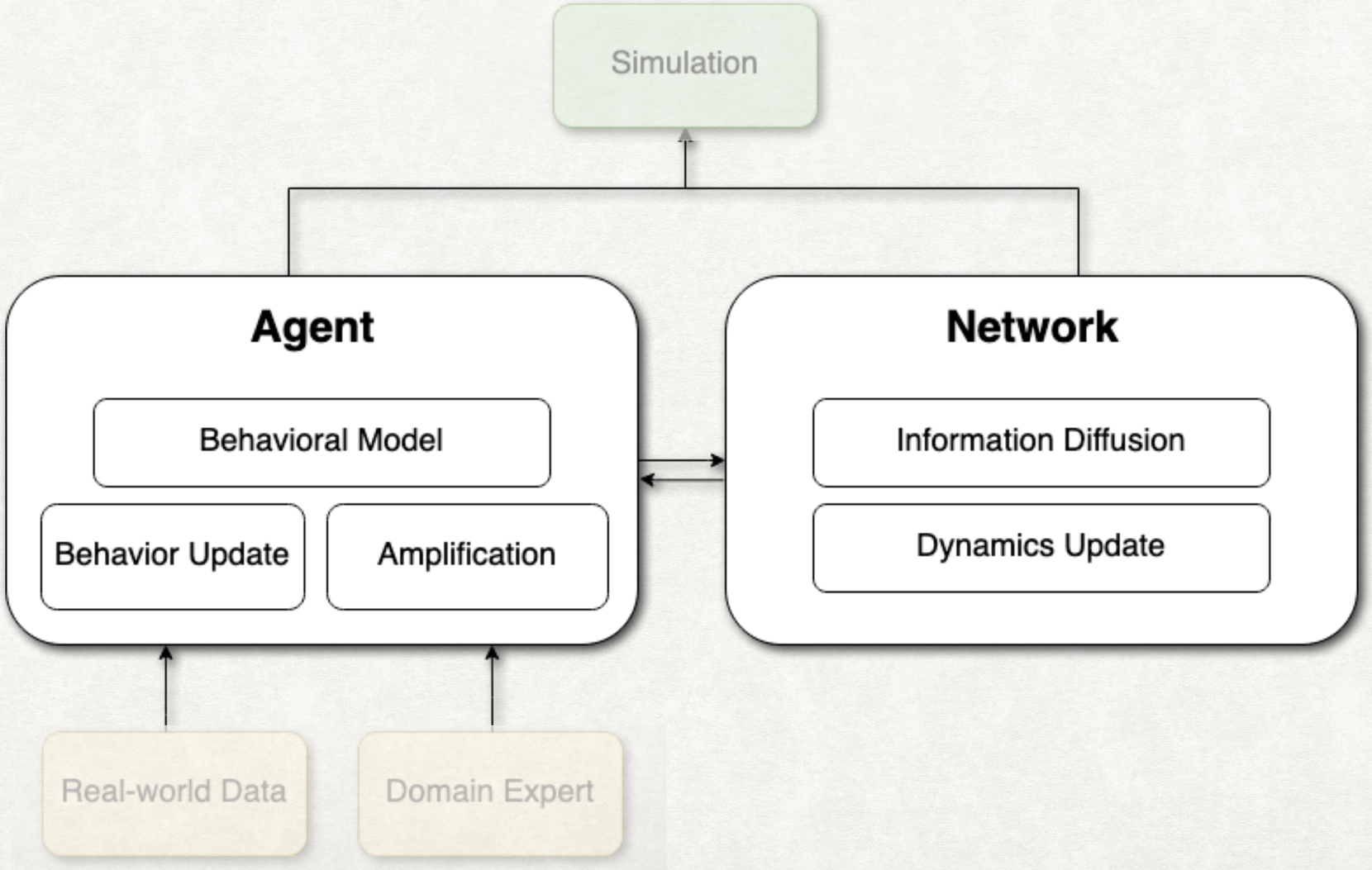
SIMULATING SOCIAL NETWORKS



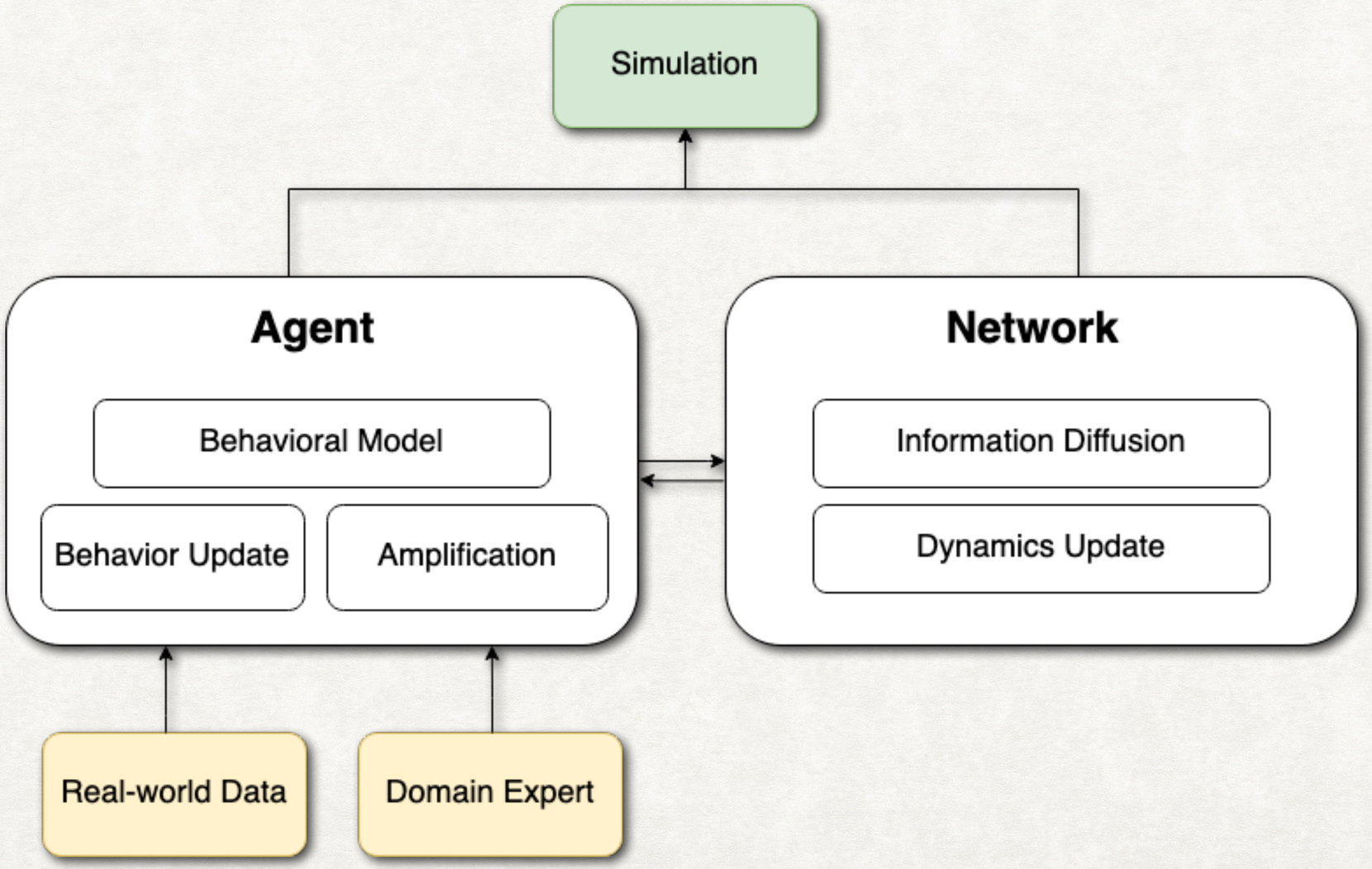


https://youtu.be/GV5XuftiD7s

SIMPPL


Simulation
Agent
Behavioral Model
Behavior Update
Amplification
Network
Information Diffusion
Dynamics Update
Real-world Data
Domain Expert

SIMPPL

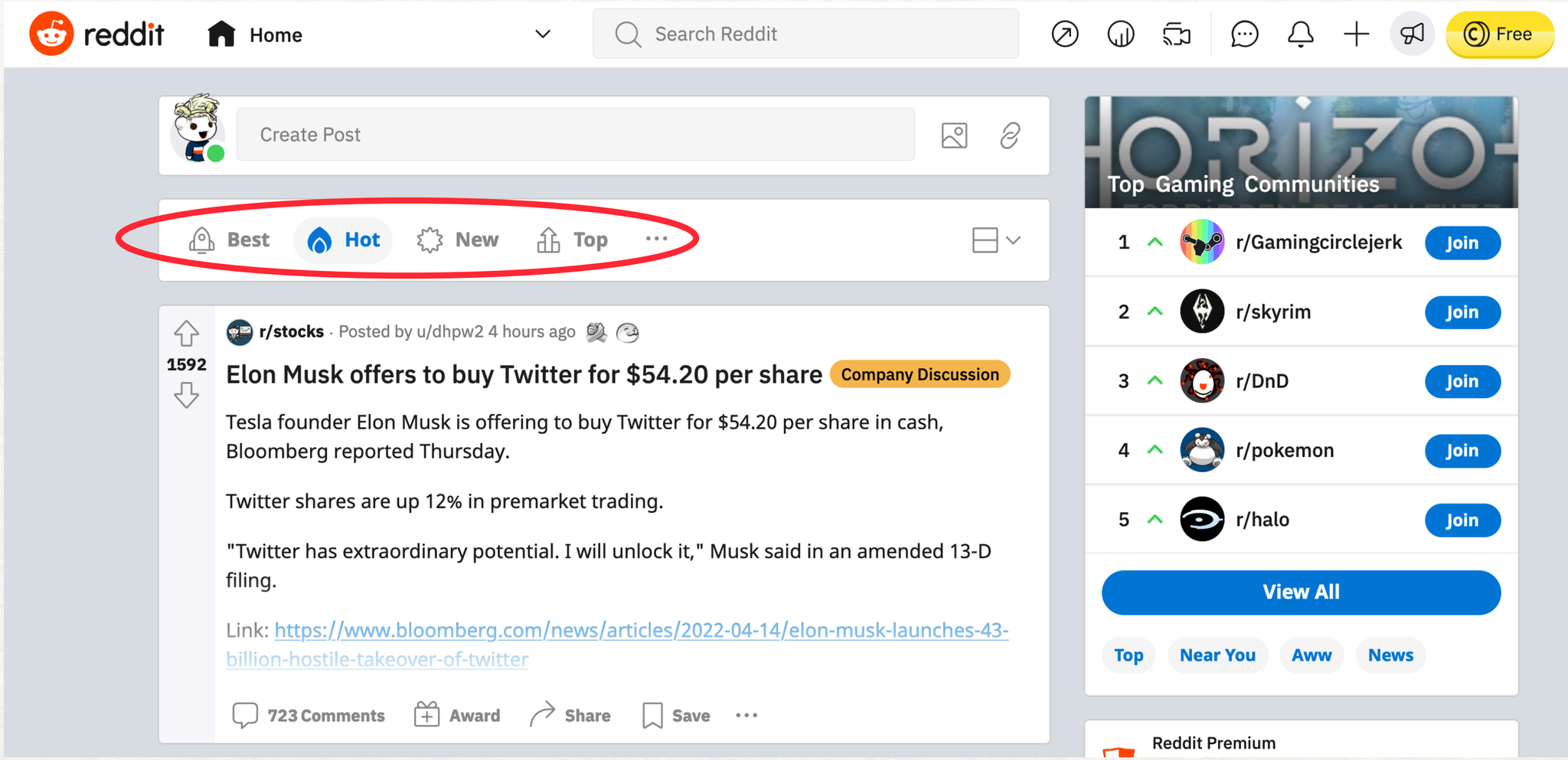

Simulation
Agent
Behavioral Model
Behavior Update
Amplification
Network
Information Diffusion
Dynamics Update
Real-world Data
Domain Expert

DISINFORMATION MANOEUVRES

	Manipulating the narrative		Manipulating the social network	
Positive	**Engage**	Messages that bring up a related but relevant topic	**Back**	Actions that increase the importance of the opinion leader or create a new opinion leader
	Explain	Messages that provides details on or elaborate the topic	**Build**	Actions that create a group or the appearance of a group
	Excite	messages that elicit a positive emotion such as joy or excitement	**Bridge**	Actions that build a connection between two or more groups
	Enhance	Messages that encourage the topic-group to continue with the topic	**Boost**	Actions that grow the size of the group or make it appear that it has grown
Negative	**Dismiss**	Messages about why the topic is not important	**Neutralize**	Actions decrease the importance of the opinion leader
	Distort	Messages that alter the main message of the topic	**Nuke**	Actions that lead to a group being dismantled or breaking up, or appearing to be broken up
	Dismay	Messages that elicit a negative emotion such as sadness or anger	**Narrow**	Actions that lead to a group becoming sequestered from other groups or marginalized
	Distract	Discussion about a totally different topic and irrelevant	**Neglect**	Actions that reduce the size of the group or make it appear that the group has grown smaller

K. Carley, 2020

REDDIT RECOMMENDER SYSTEMS

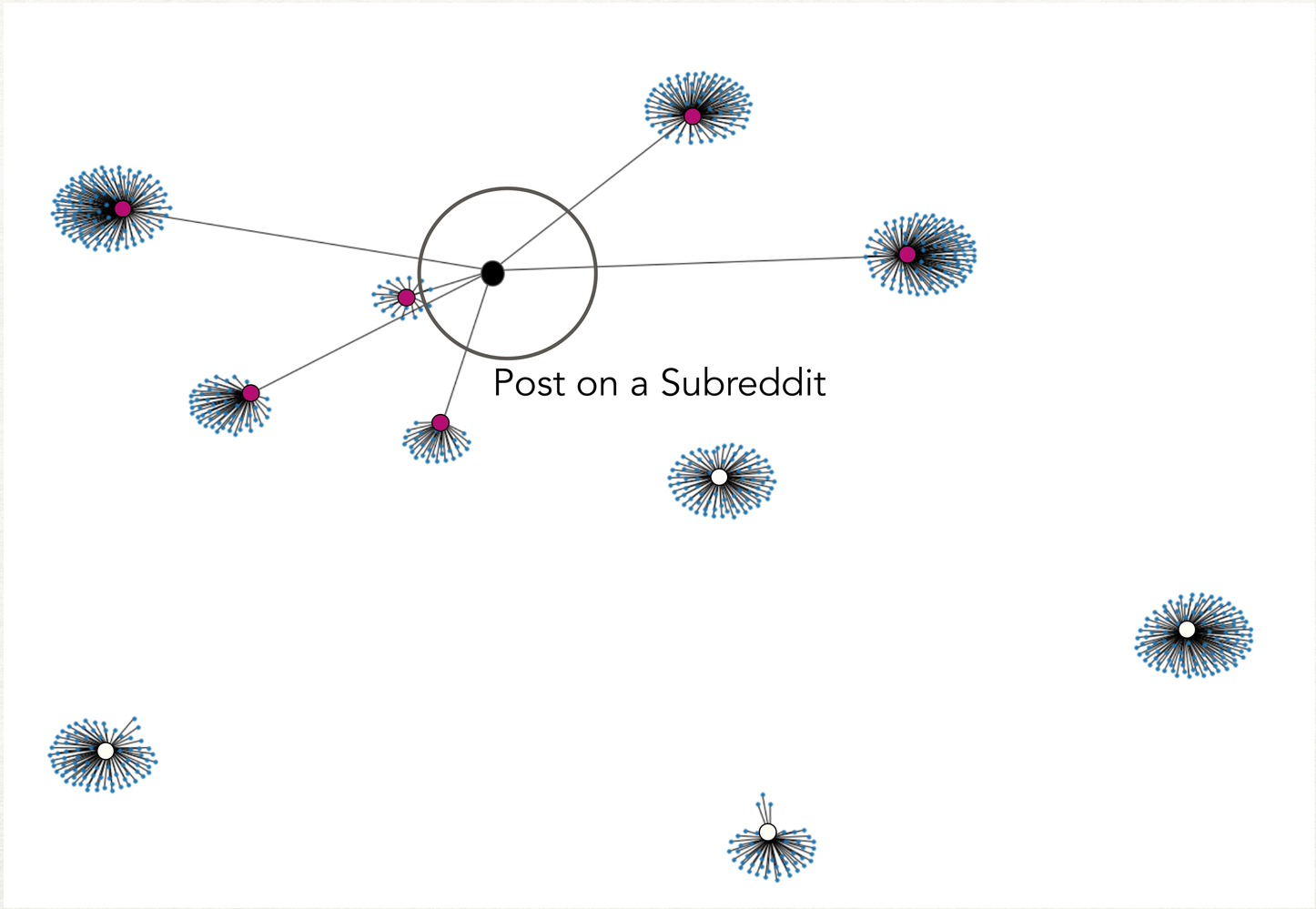


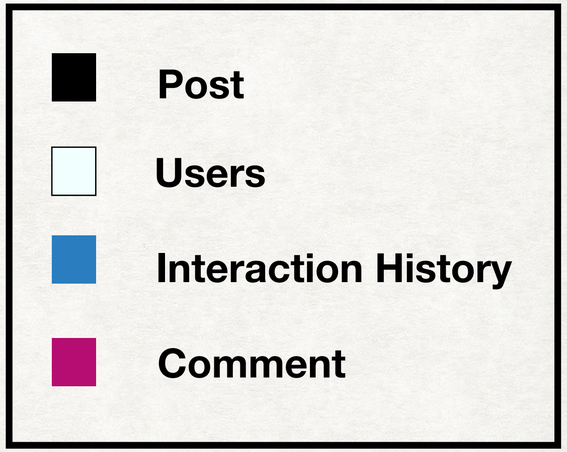
RANKING AND RECOMMENDATION ALGORITHMS

New	Top	Rising	Controversial	Best (Personalized)
Age of Post	Age of Post			Age of Post
	# of Upvotes	# of Upvotes	# of Upvotes	# of Upvotes
			# of Downvotes	
		Age of Votes		
		Age of Comments		
				Relevance to User
				Subreddit Membership

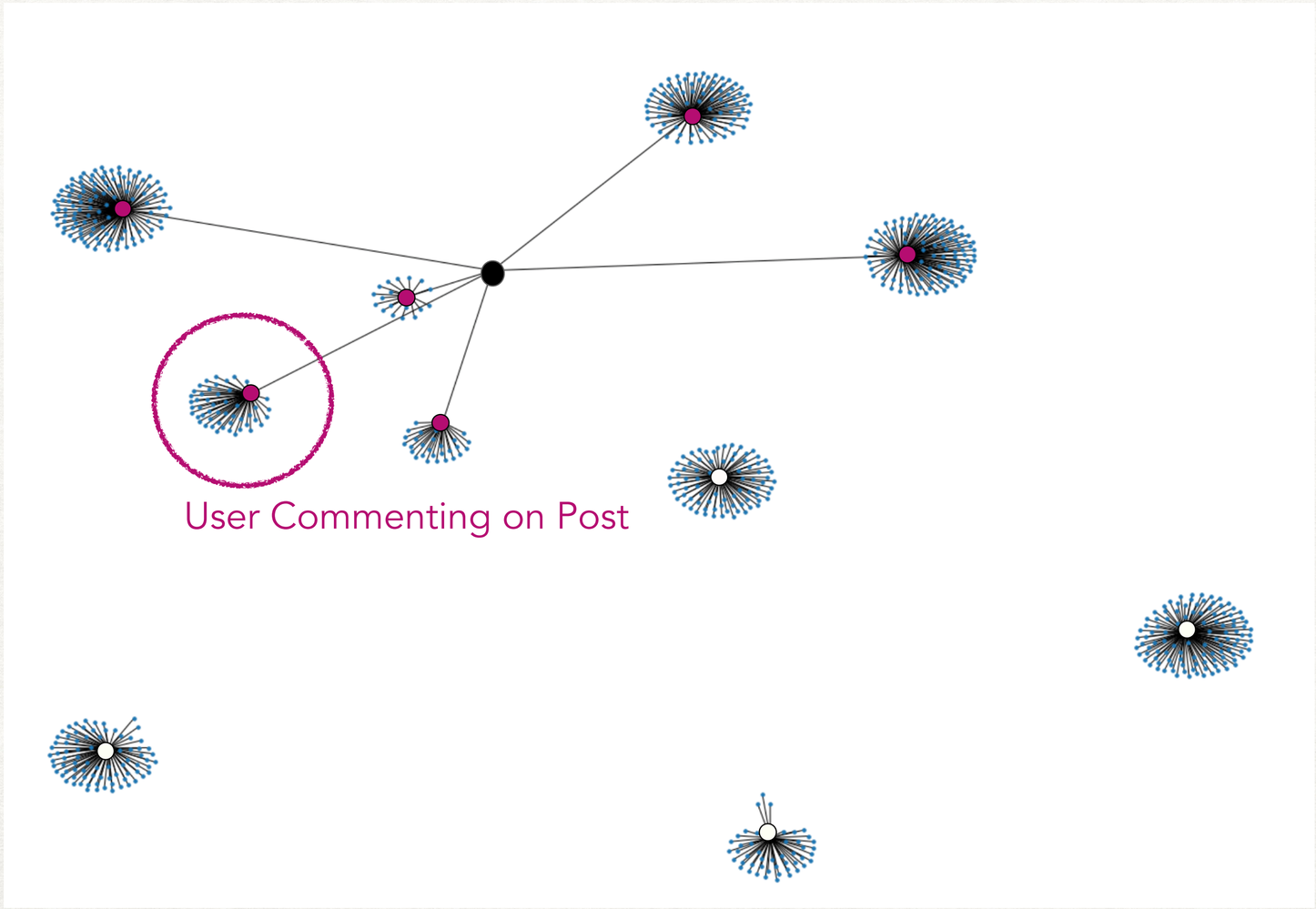
USING REAL-WORLD DATA TO DRIVE THE SIMULATIONS

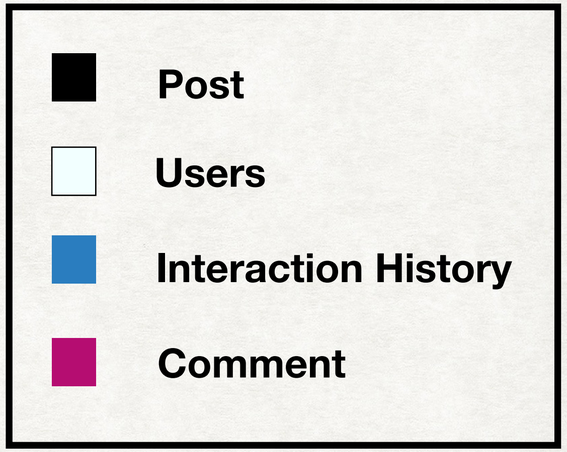
REDDIT POST


Post on a Subreddit

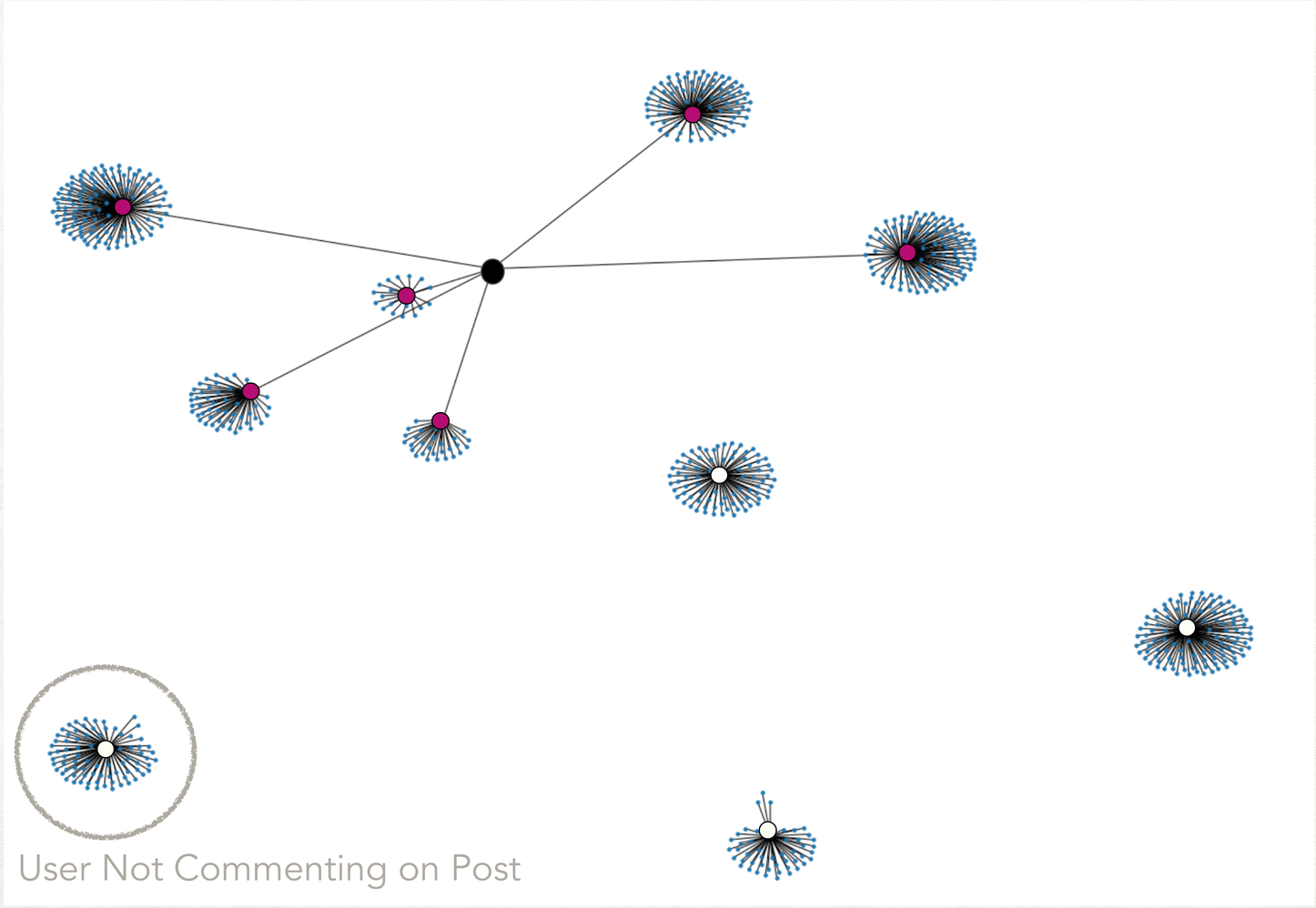

Post
Users
Interaction History
Comment

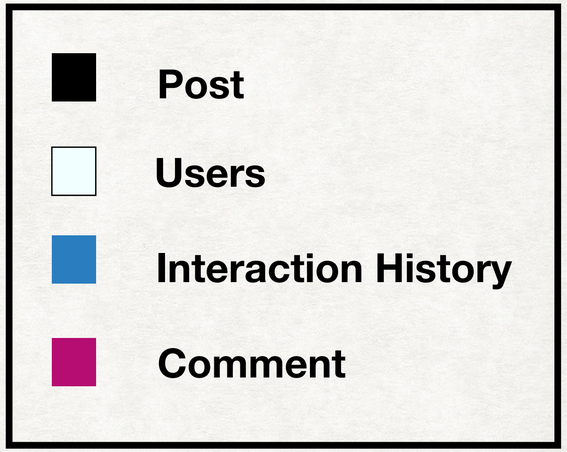
REDDIT POST


User Commenting on Post

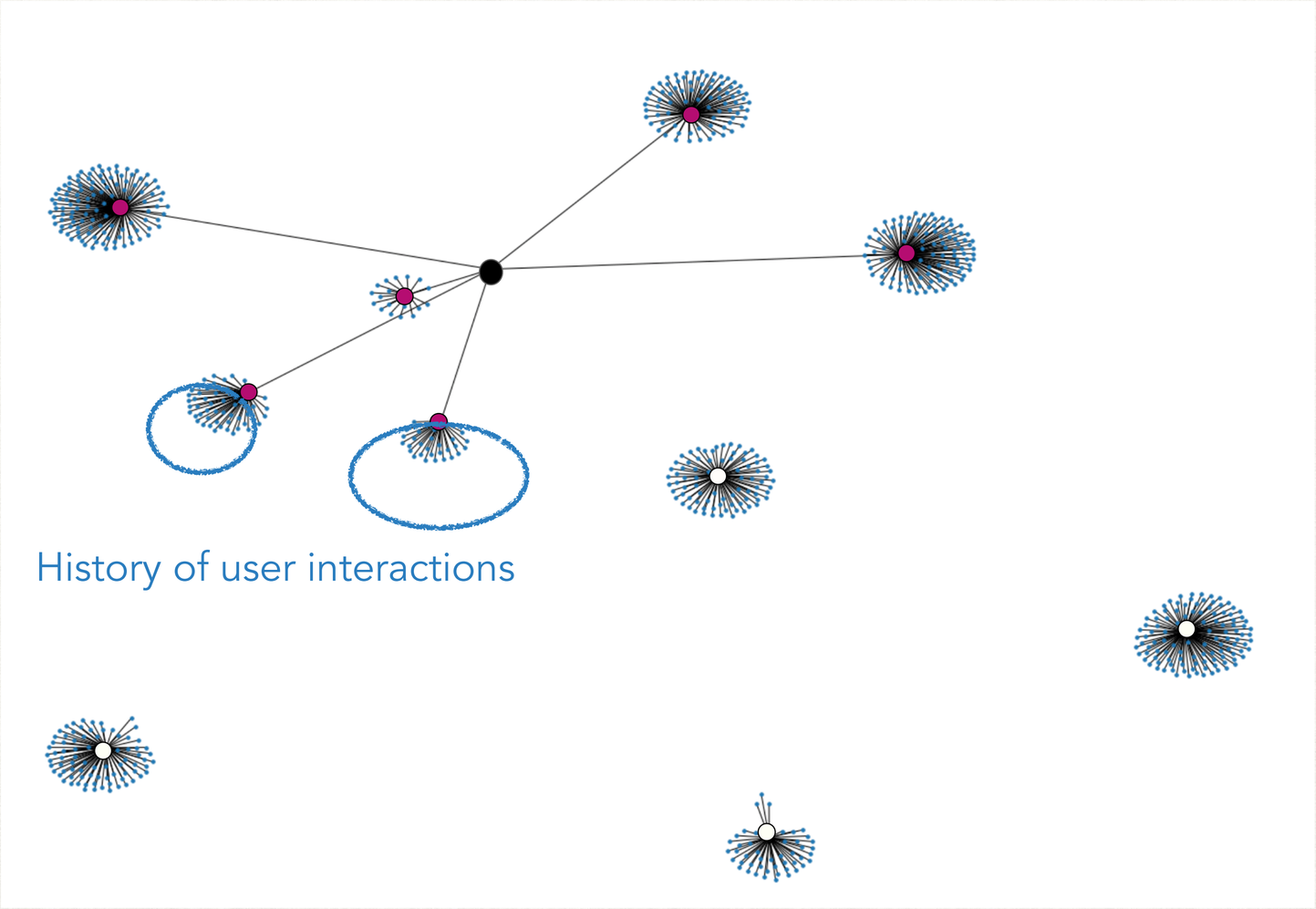

Post
Users
Interaction History
Comment

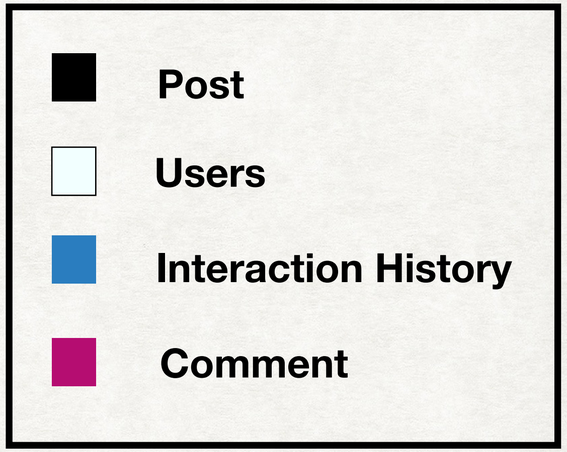
REDDIT POST


User Not Commenting on Post

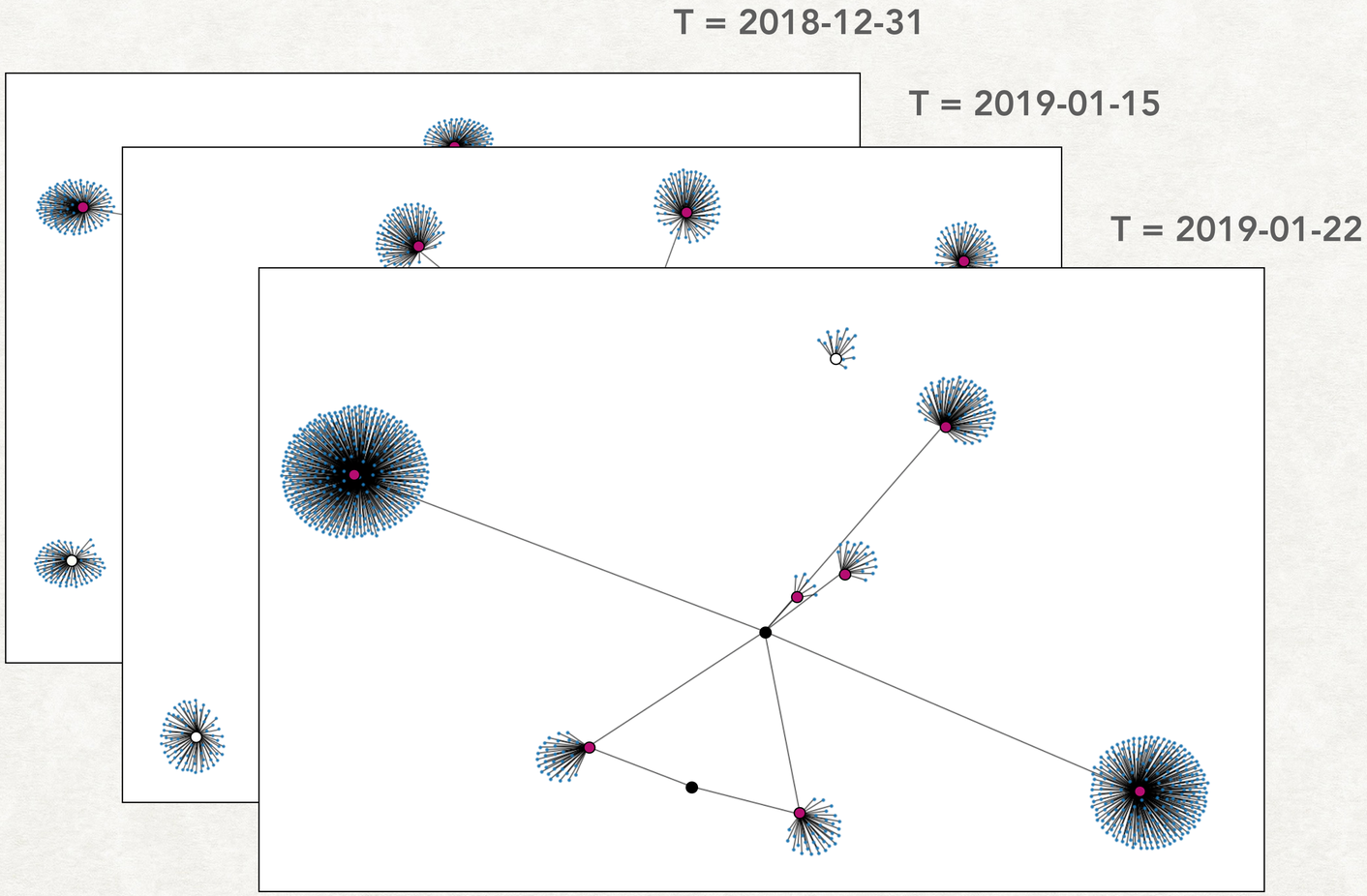

Post
Users
Interaction History
Comment

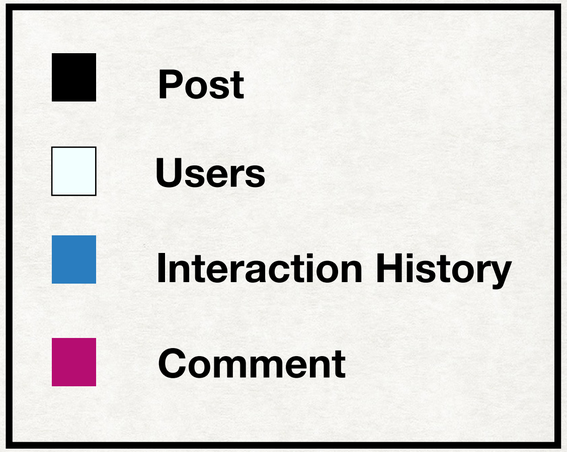
REDDIT POST


History of user interactions


Post
Users
Interaction History
Comment

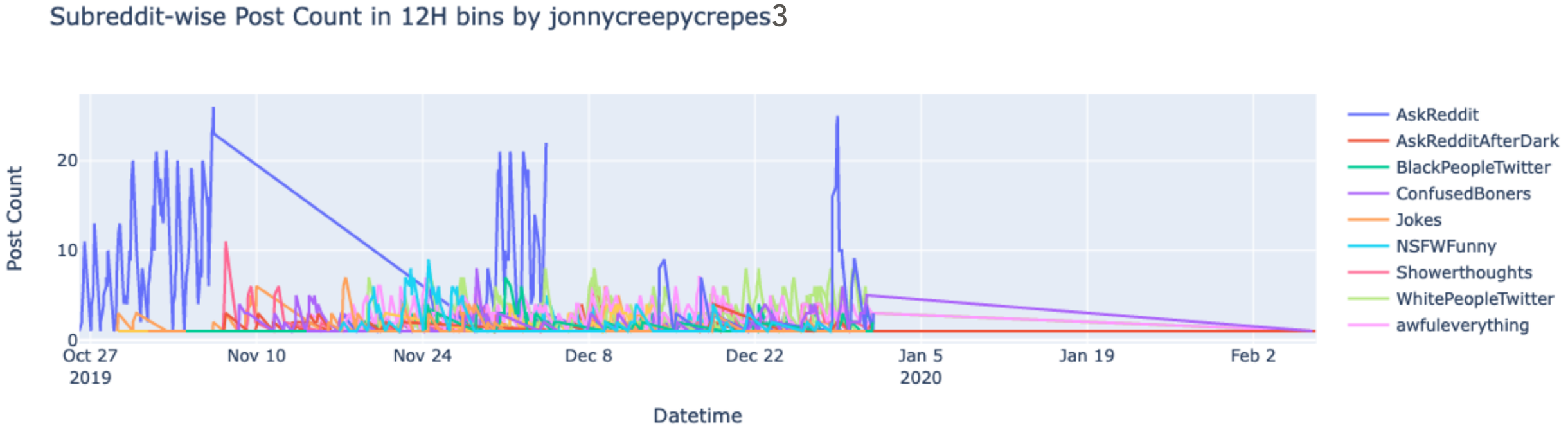
REDDIT POSTS



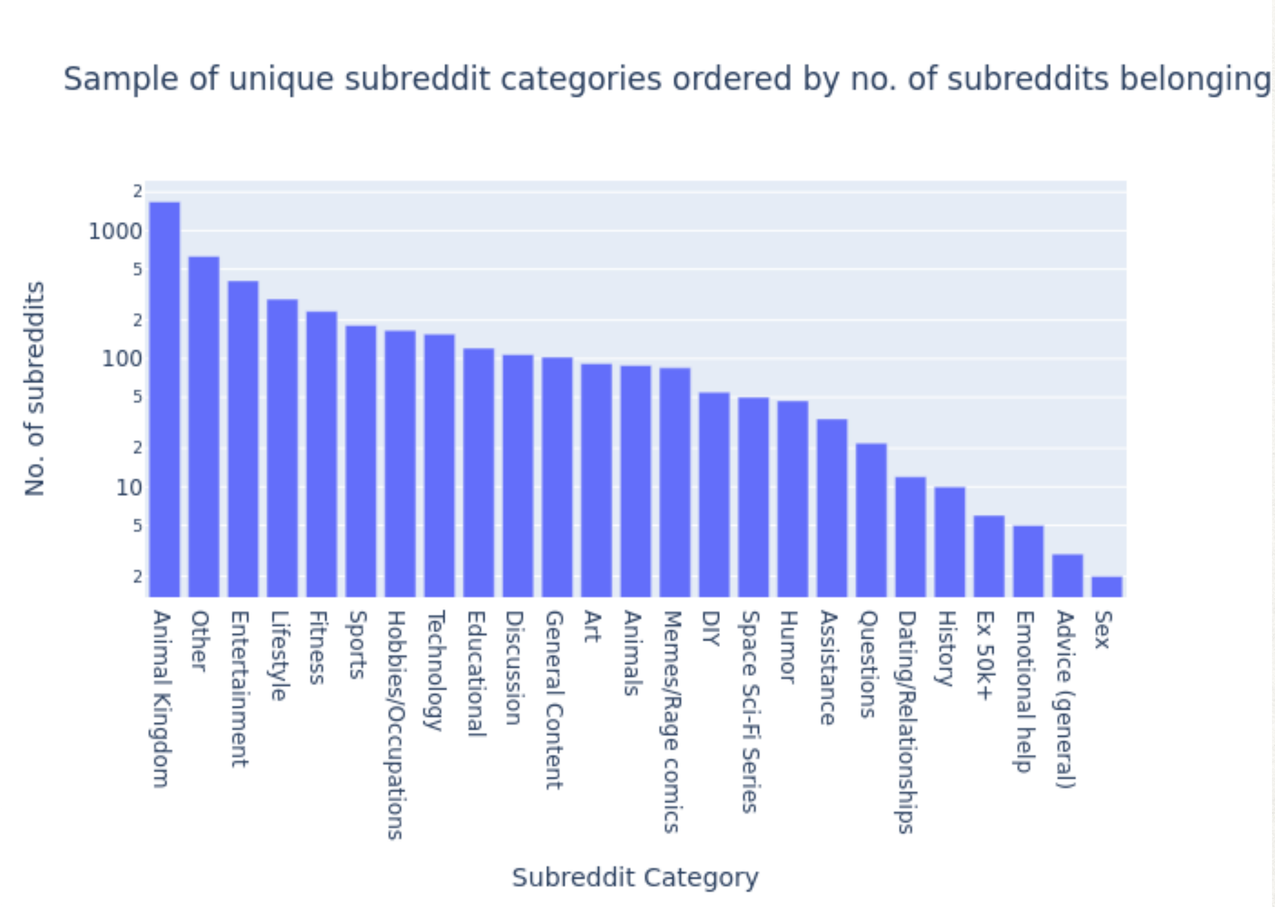


REAL-WORLD MODELING

REDDIT USER 'JONNYCREEPYCREPES3'



CATEGORIZE POSTS ACROSS SUBREDDITS



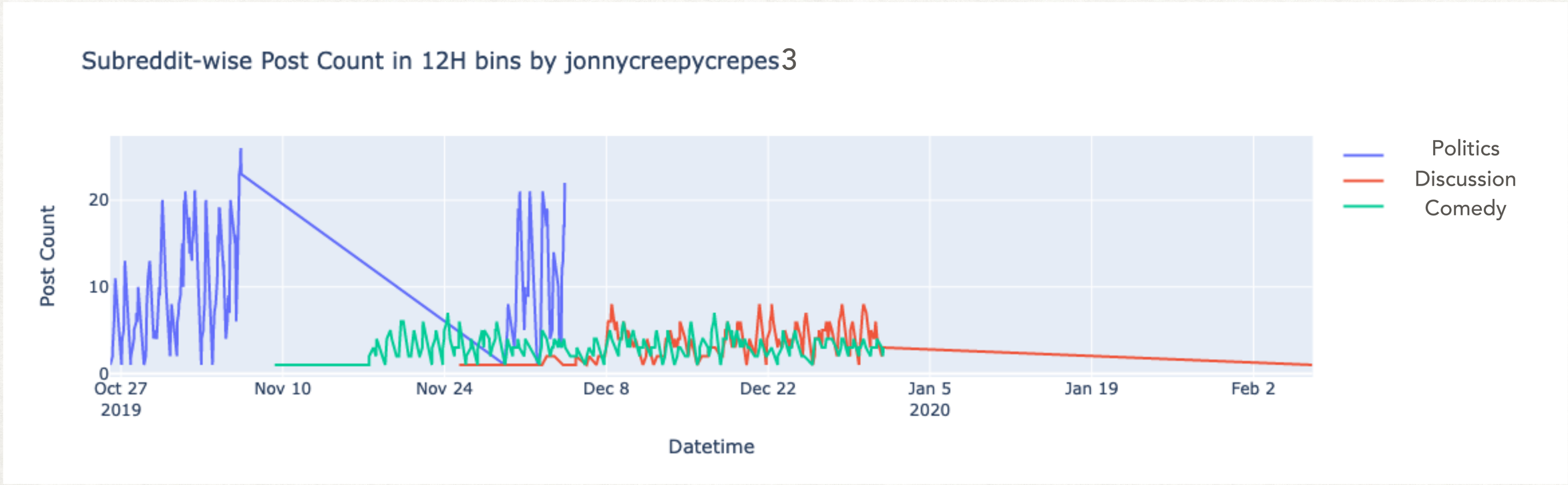
['mormon', 'politics']

Similar Subreddits:
['mormonhistory', 'ldshistory', 'christianhistory', 'jewishhistory', 'historicalreligion']

Predicted Category:
[[None, 'History of People', None, None, None],

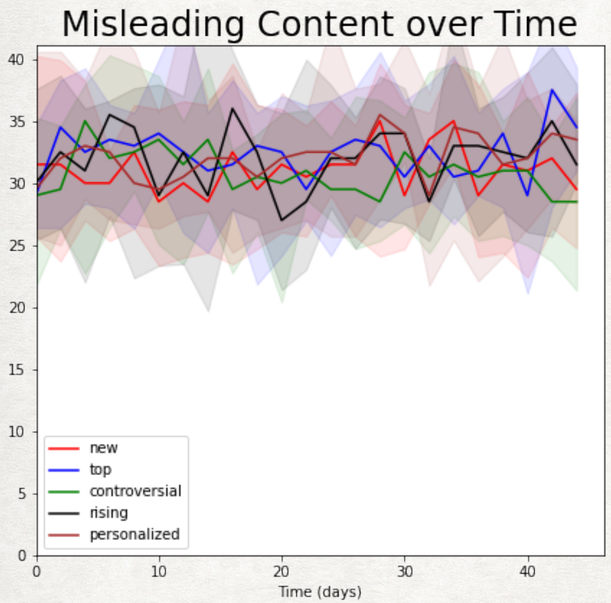
MODELING CATEGORIES

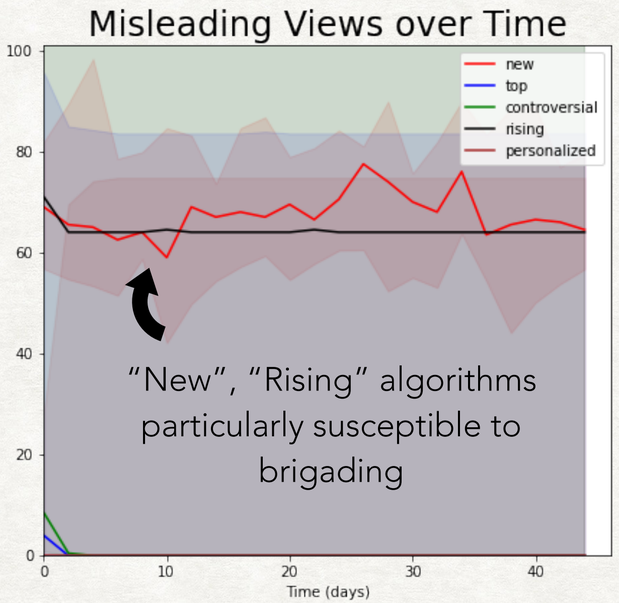
REDUCING SUBREDDIT DIMENSIONALITY

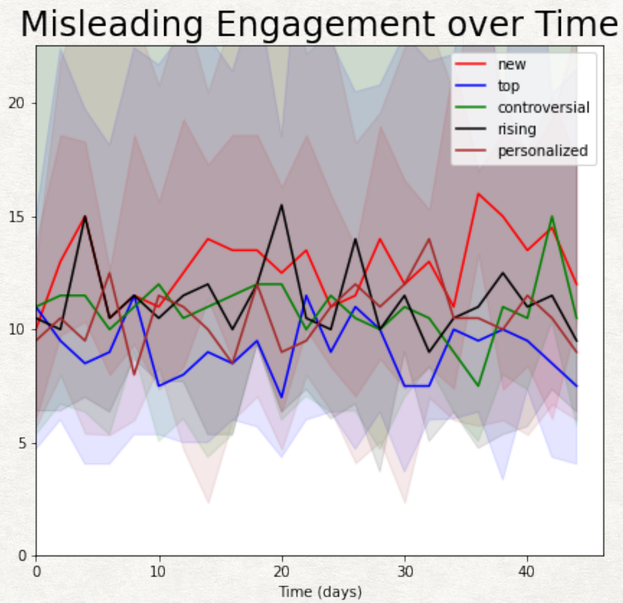


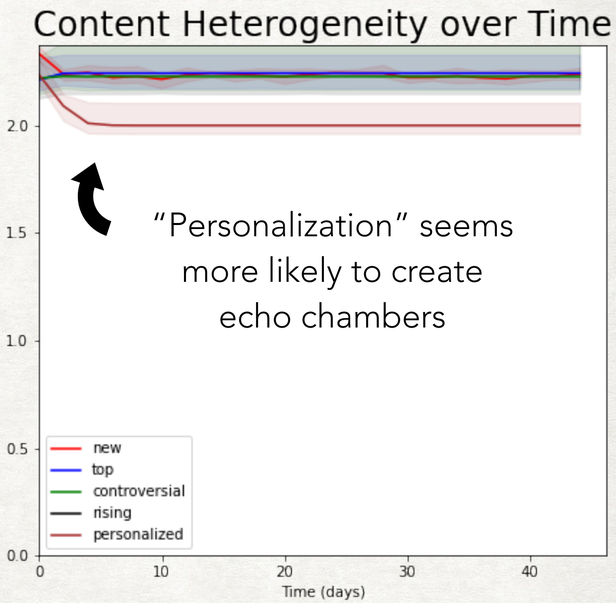
HOW SUSCEPTIBLE ARE ALGORITHMS TO COORDINATED INAUTHENTIC BEHAVIOR?

WHAT ARE THE EFFECTS OF BRIGADING?



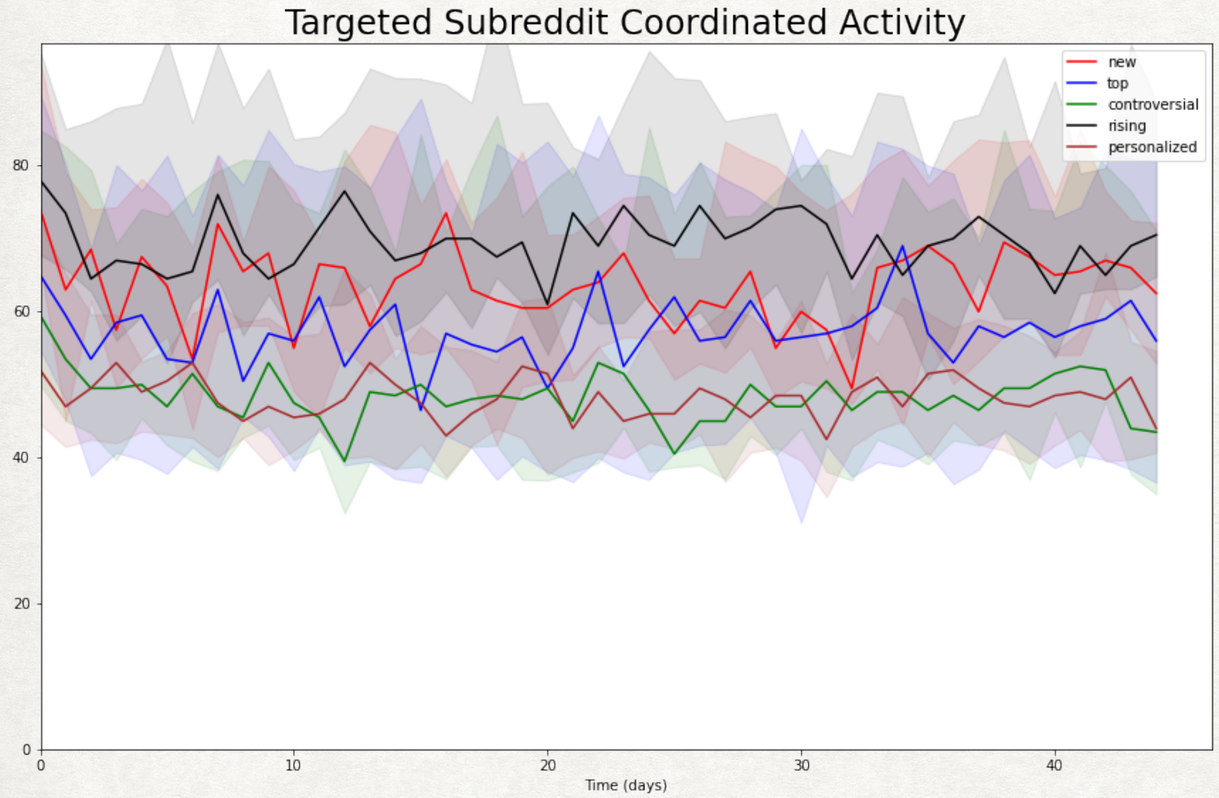


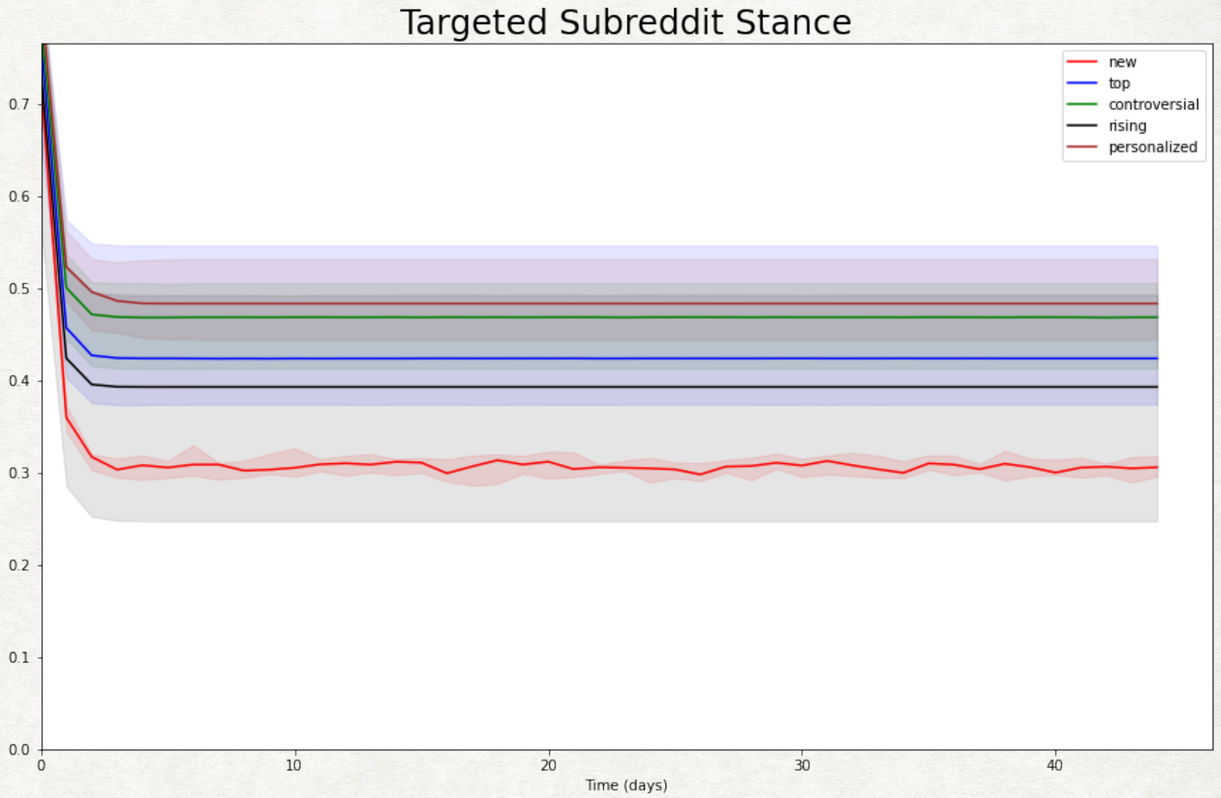




Similar levels of misleading content leads to different emergent dynamics

WHAT ARE THE EFFECTS OF BRIGADING?





Despite similar levels of activity, there is a significant drop in the positive opinions expressed on the target subreddit for the “New” ranking algorithm

CONTENT DISTRIBUTION CHOICES

APPS | HOW-TO | REVIEWS

How to switch your Twitter feed to a chronological timeline

Look for the sparkle

By Natt Garun | @nattgarun | Mar 6, 2020, 11:47am EST

Facebook's new 'Feeds' tab chronologically displays posts from your friends and groups

Aisha Malik @aiishamalik1 / 10:13 AM EDT • July 21, 2022 Comment

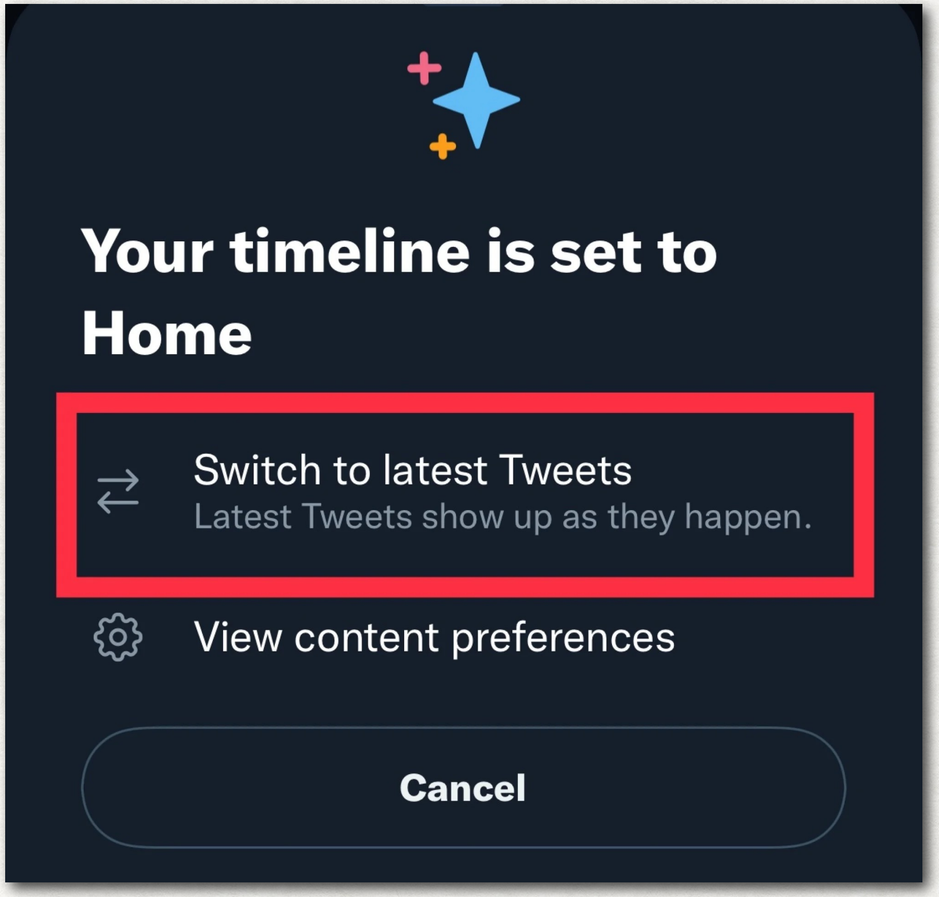
TECH • BIG TECH

Facebook Is Finally Giving People A Non-Algorithmic News Feed

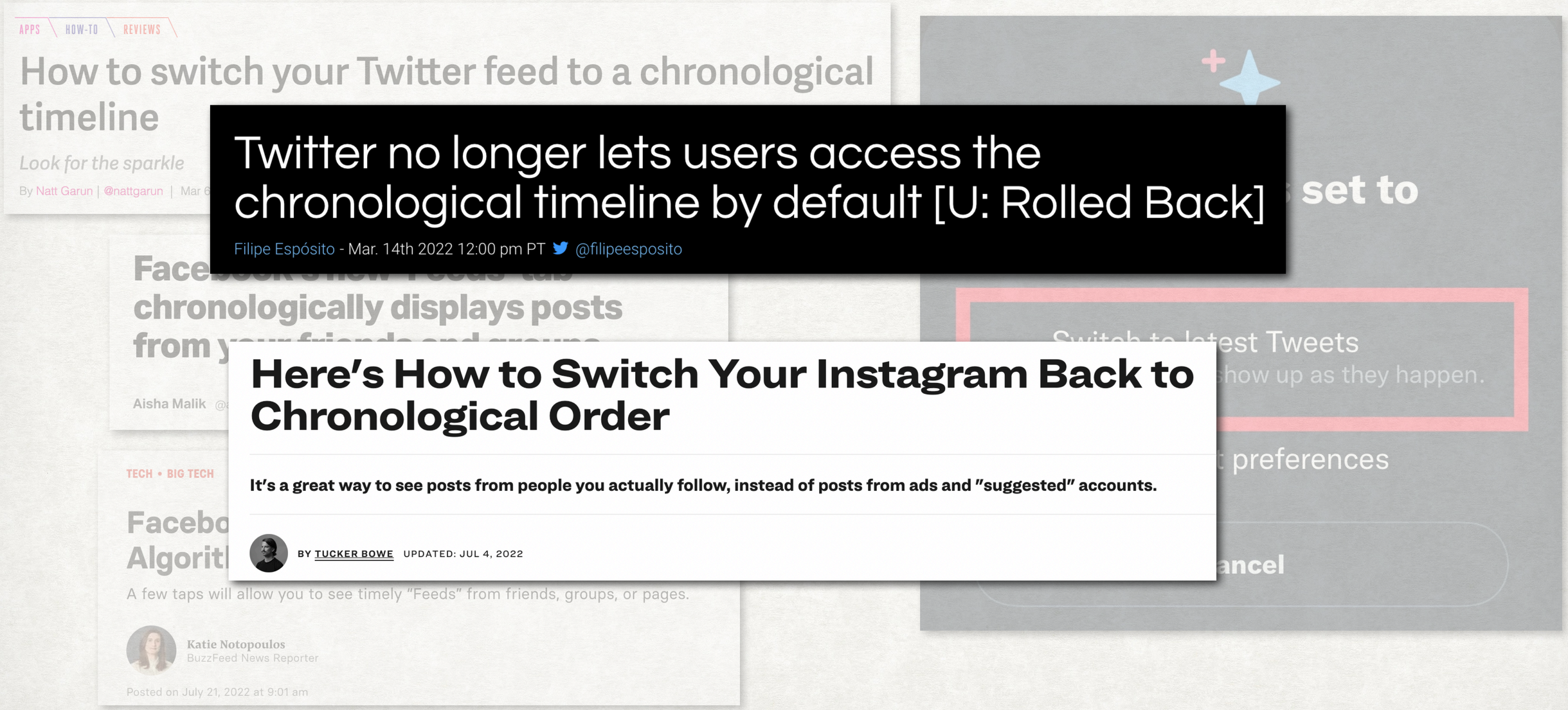
A few taps will allow you to see timely "Feeds" from friends, groups, or pages.

Katie Notopoulos
BuzzFeed News Reporter

Posted on July 21, 2022 at 9:01 am



CONTENT DISTRIBUTION CHOICES



DEBUGGING POLICY INTERVENTIONS

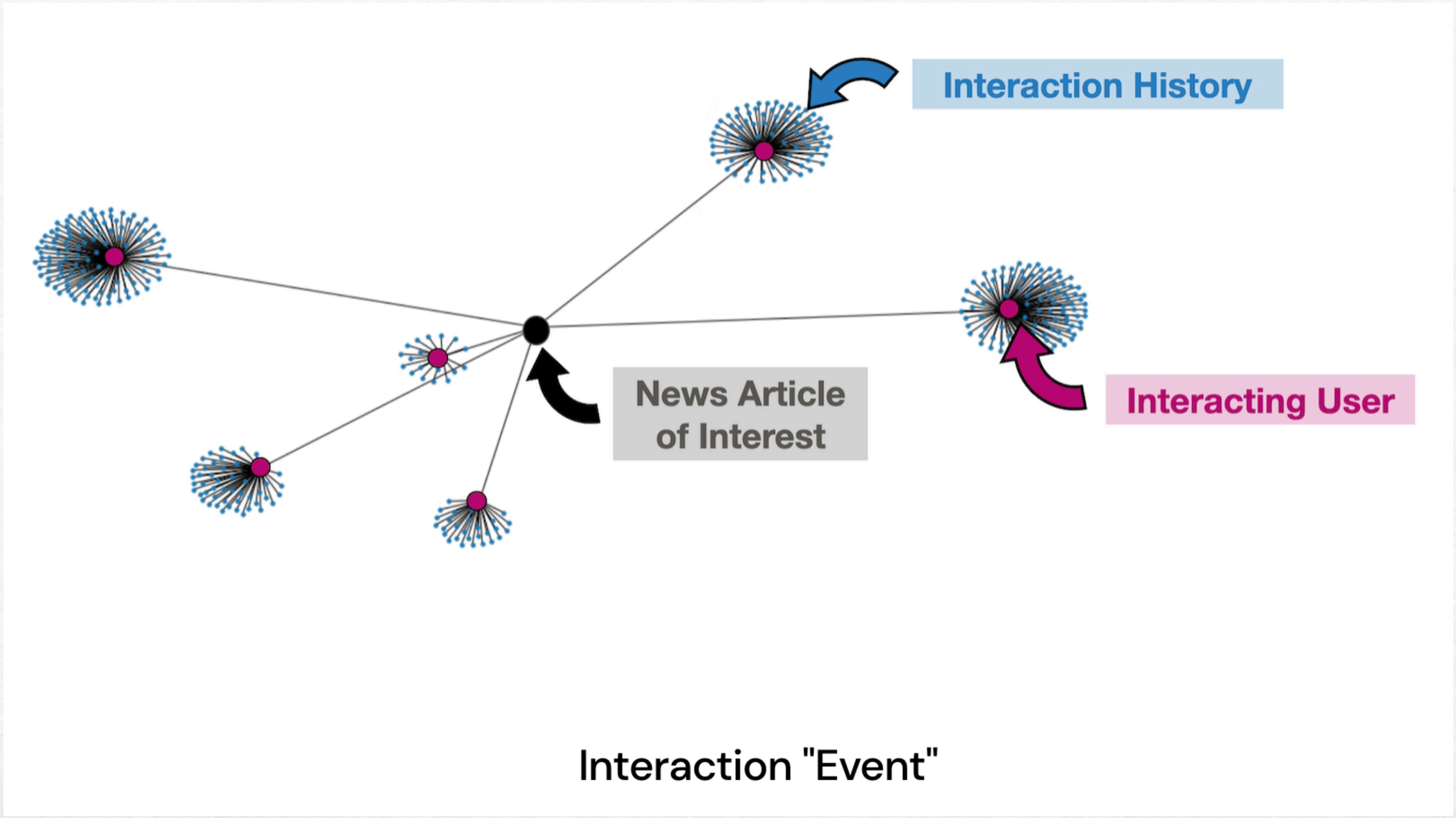
I. Problems with How Interventions are Studied

II. Measuring Harms on Social Networks

III. Applying Techniques in Practice

HELPING LOCAL NEWS UNDERSTAND THEIR ONLINE AUDIENCES WITH AI

APPLYING SIMPPL'S TECH TO NEWS ARTICLES



HELPING LOCAL NEWS UNDERSTAND AUDIENCES

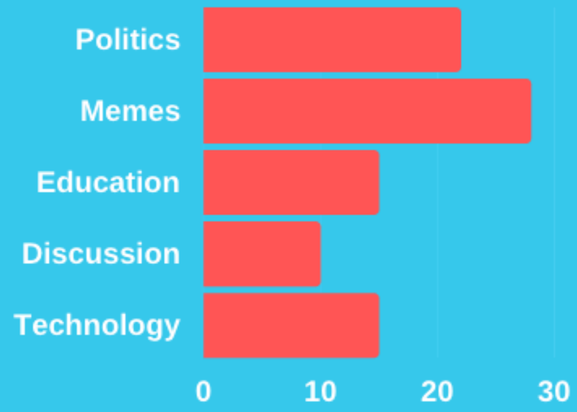
You're Getting Upvotes!

5	9	47↑
r/VTGuns	r/VoteDEM	r/BernieSanders

Users belong to Subreddits

r/politics	10	8.1m
r/blm	8	4.8k
r/AskReddit	4	35.9m
r/ragecomics	3	48.1k

Users also Engaged With



Related Engagement

151	100	1400↑
r/CRT_so_scary	r/Enough_Sanders_Spam	r/Residency

sevendaysvt.com

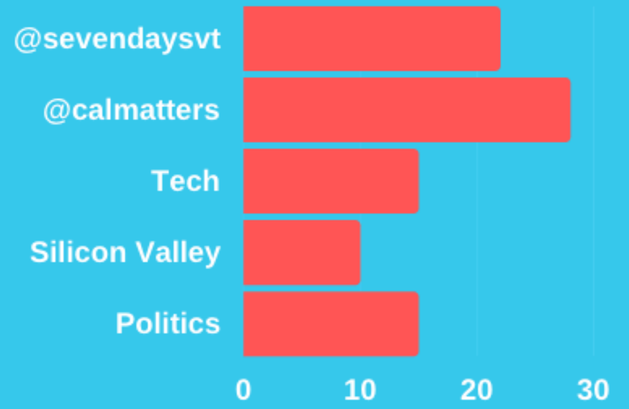
HELPING LOCAL NEWS UNDERSTAND AUDIENCES

You're Getting Favorites!

4	5	20↑
The CDC recommends that people in high-level ...	With the Senate Divided 50-50 and Republicans united against ...	"The current Act 250 bill would actually make ...

You're Getting Traffic From

Engaged Users also Follow



Related Engagement

14	23	86 ↑
Politics	Elections	Bernie Sanders

@sevendaysvt

EXAMPLE

VTDigger

Analyzing VTDigger Posts shared to Reddit.

r/VoteDEM · Posted by u/kittehgoesmeow MD-08 15 days ago

At VTDigger debate, Democratic candidates for US House carve out distinct policy positions [Vermont At Large]
vtdigger.org/2022/0...

9 upvotes 2 comments 0 awards

r/worldnewgame · Posted by u/shawn19 1 hour ago

Vermont game wardens investigate fatal dog shooting in Tunbridge - VTDigger
vtdigger.org/2022/0...

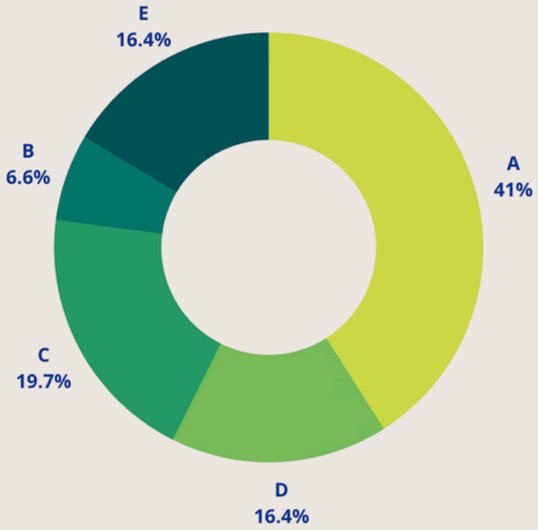
0 upvotes 0 comments 0 awards

r/VTGuns · Posted by u/PROPAGANDA-DESTROYER 8 days ago

https://vtdigger.org/2022/04/13/christopher-herrick-conserving-wildlife-requires-respecting-differences/

5 upvotes 4 comments 0 awards

A. **Politics**
B. Education
C. **Elections**
D. Shooting
E. Jobs



EXAMPLE

VTDigger

Predicting where to share the next post!

r/VoteDEM · Posted by u/kittehgoesmeow MD-08 15 days ago

At VTDigger debate, Democratic candidates for US House carve out distinct policy positions [Vermont At Large]
vtdigger.org/2022/0...

9 upvotes 2 comments 0 awards

r/worldnewgame · Posted by u/shawn19 1 hour ago

Vermont game wardens investigate fatal dog shooting in Tunbridge - VTDigger
vtdigger.org/2022/0...

0 upvotes 0 comments 0 awards

r/VTGuns · Posted by u/PROPAGANDA-DESTROYER 8 days ago

https://vtdigger.org/2022/04/13/christopher-herrick-conserving-wildlife-requires-respecting-differences/

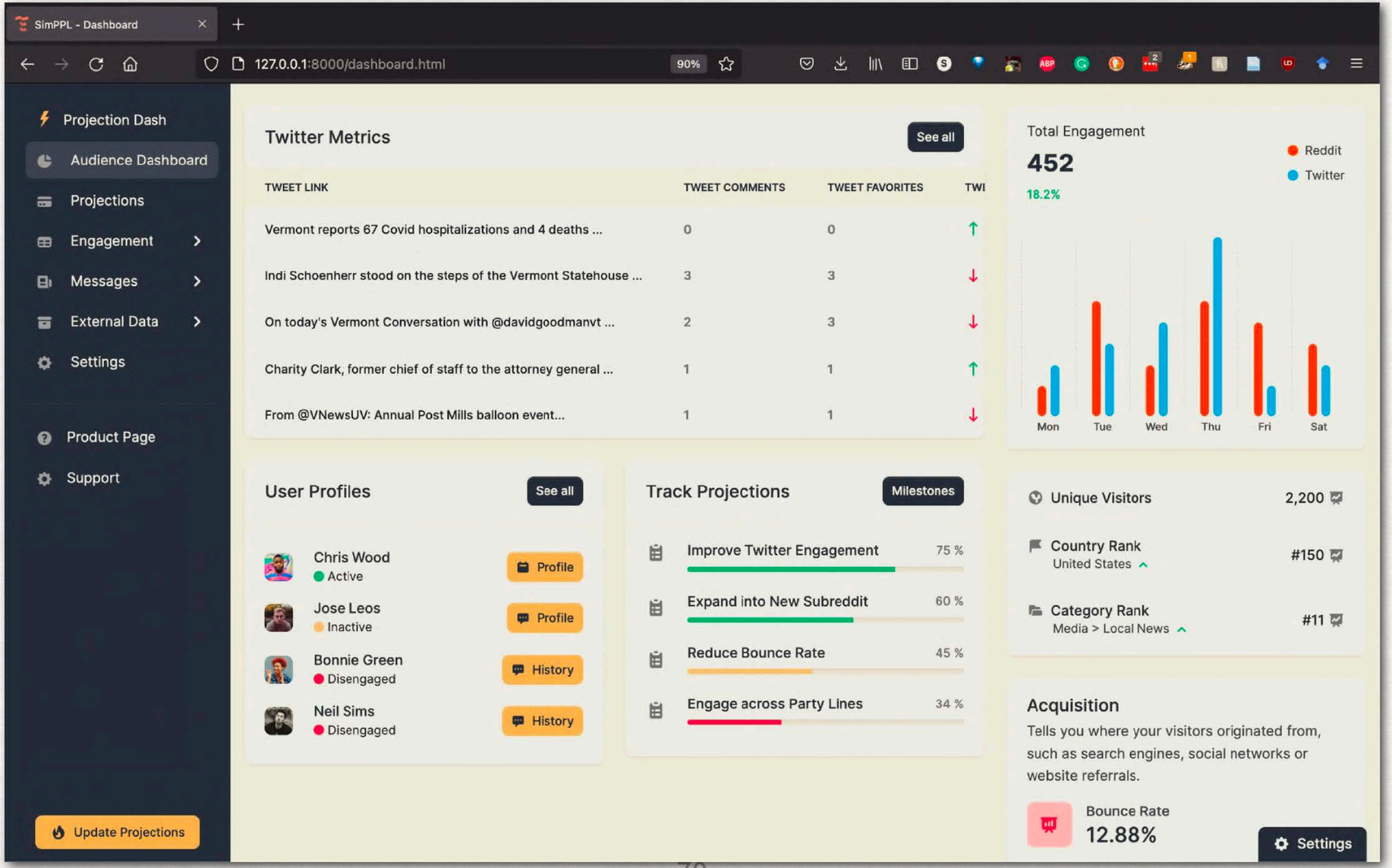
5 upvotes 4 comments 0 awards

Leahy, Sanders vote to change filibuster rules in failed attempt to advance voting rights bill - vtdigger.org
vtdigger.org/2022/0...

47 upvotes 2 comments 0 awards

DEMO

https://simppl.org



PRODUCT

Track Engagement and KPIs

Monitor Subscriber Preferences

Predict a Post's Reach

Automate Digital Strategy

COLLABORATORS

Jonathan Nagler

Richard Bonneau

Philip Torr

Joshua Tucker

James Bisbee

Atılım Güneş Baydin

Bogdan State

UNICODE RESEARCH - ML COLLABORATION

Mentors

Swapneel Mehta

Researchers

Aryan Chouhan

Deep Gandhi

Devang Shah

Fenil Doshi

Gaurang Raje

Jash Mehta

Jay Gala

Jhagrut Lalwani

Nemil Shah

Priyambi Hiran

Sarthak Ojha

Shrey Parekh

Shwe Han

https://unicode-research.netlify.app/people/

SUMMARY

I. We can study interventions better, causally

II. We should track harms from CIB

III. Build more tools for policymakers and the public

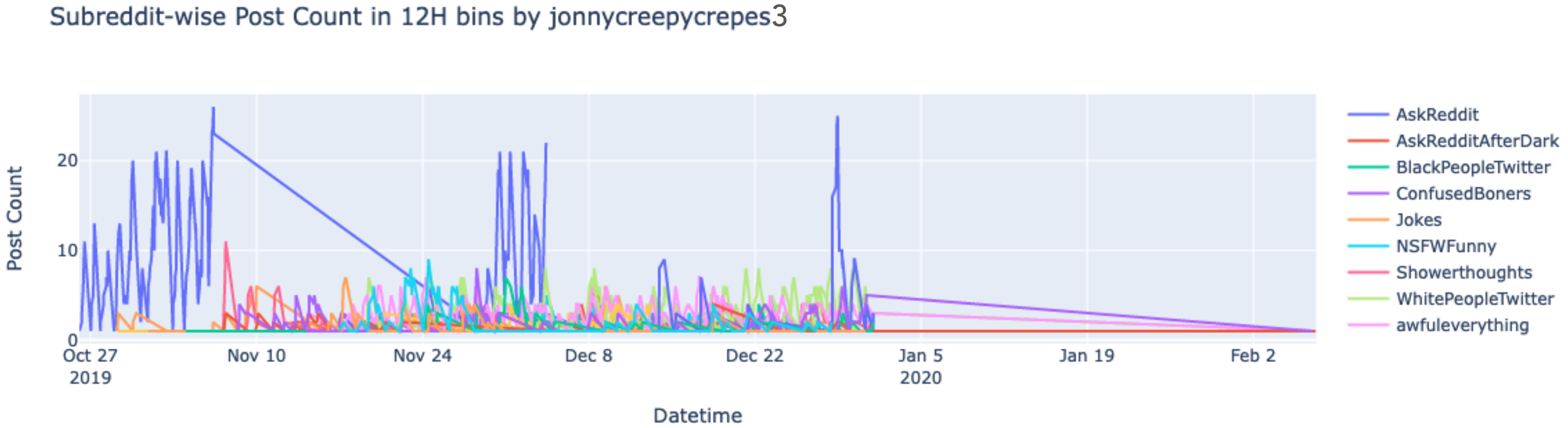
LET'S TALK!

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

Shoutout to ai4abm.org

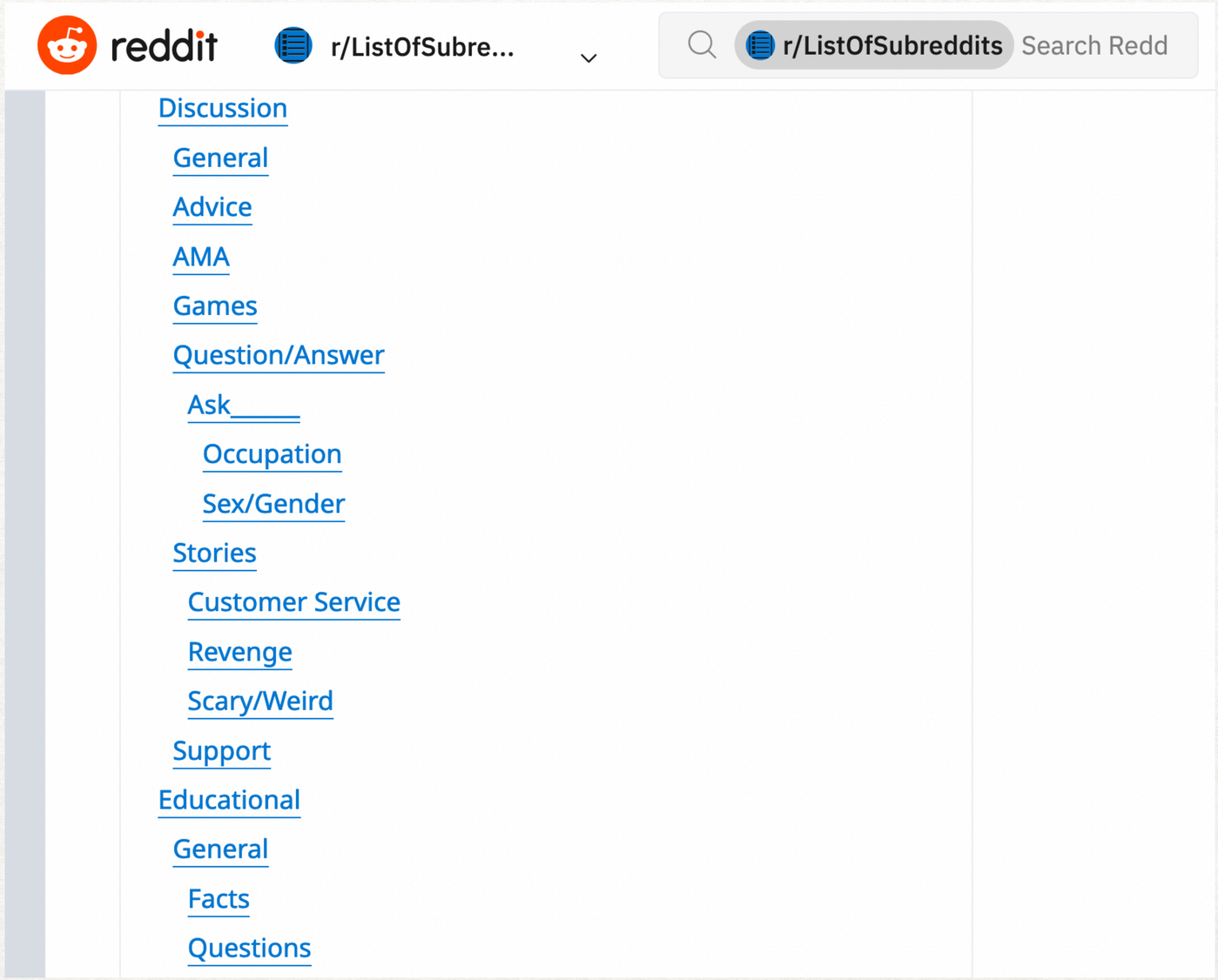
REAL-WORLD MODELING

REDDIT USER 'JONNYCREEPYCREPES3'



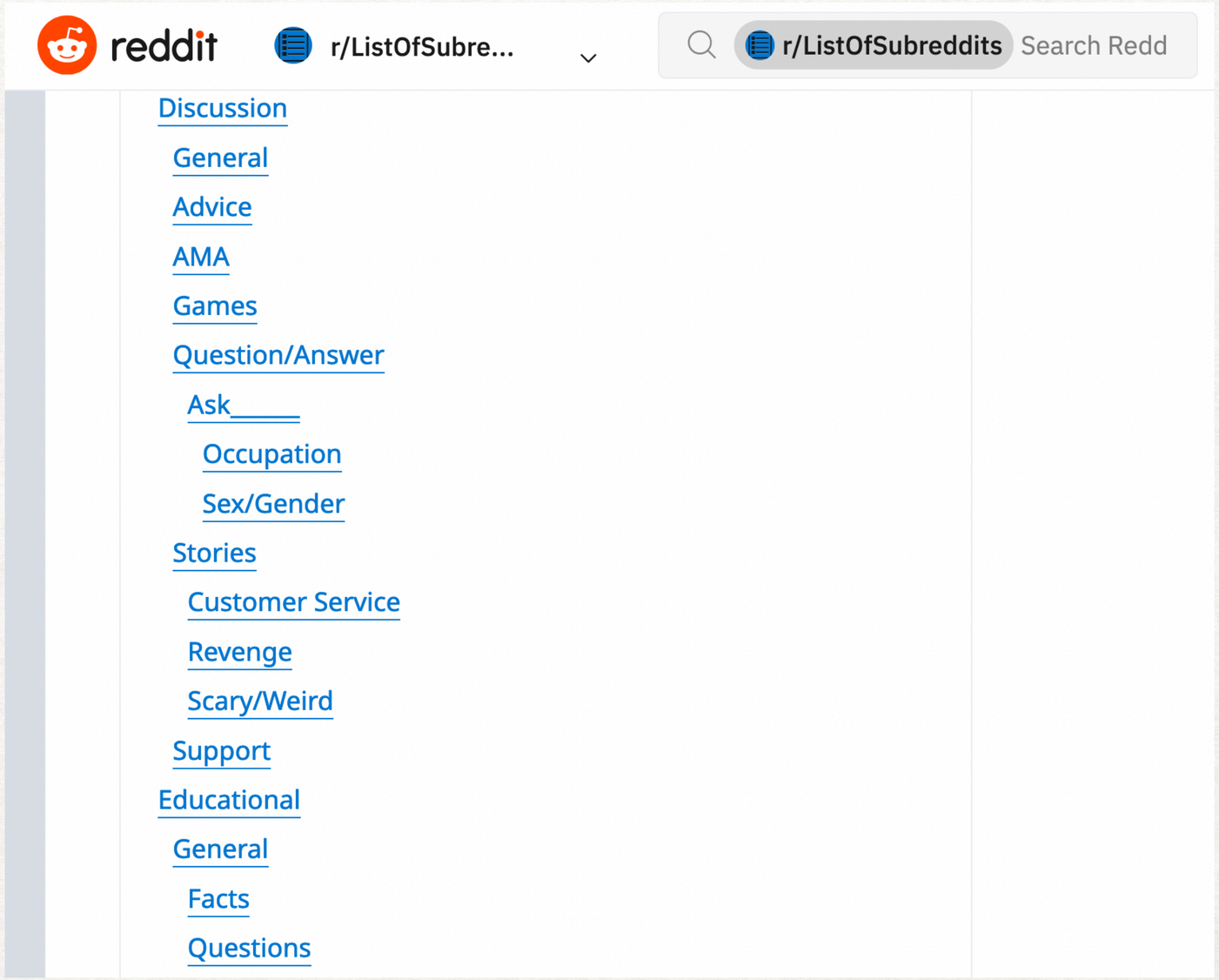
REDDIT WIKI PAGES

- User-driven hierarchical categorization of subreddits
- Discussion > Stories > Customer Service
- 4998 subreddits
- 5-level hierarchy of categories

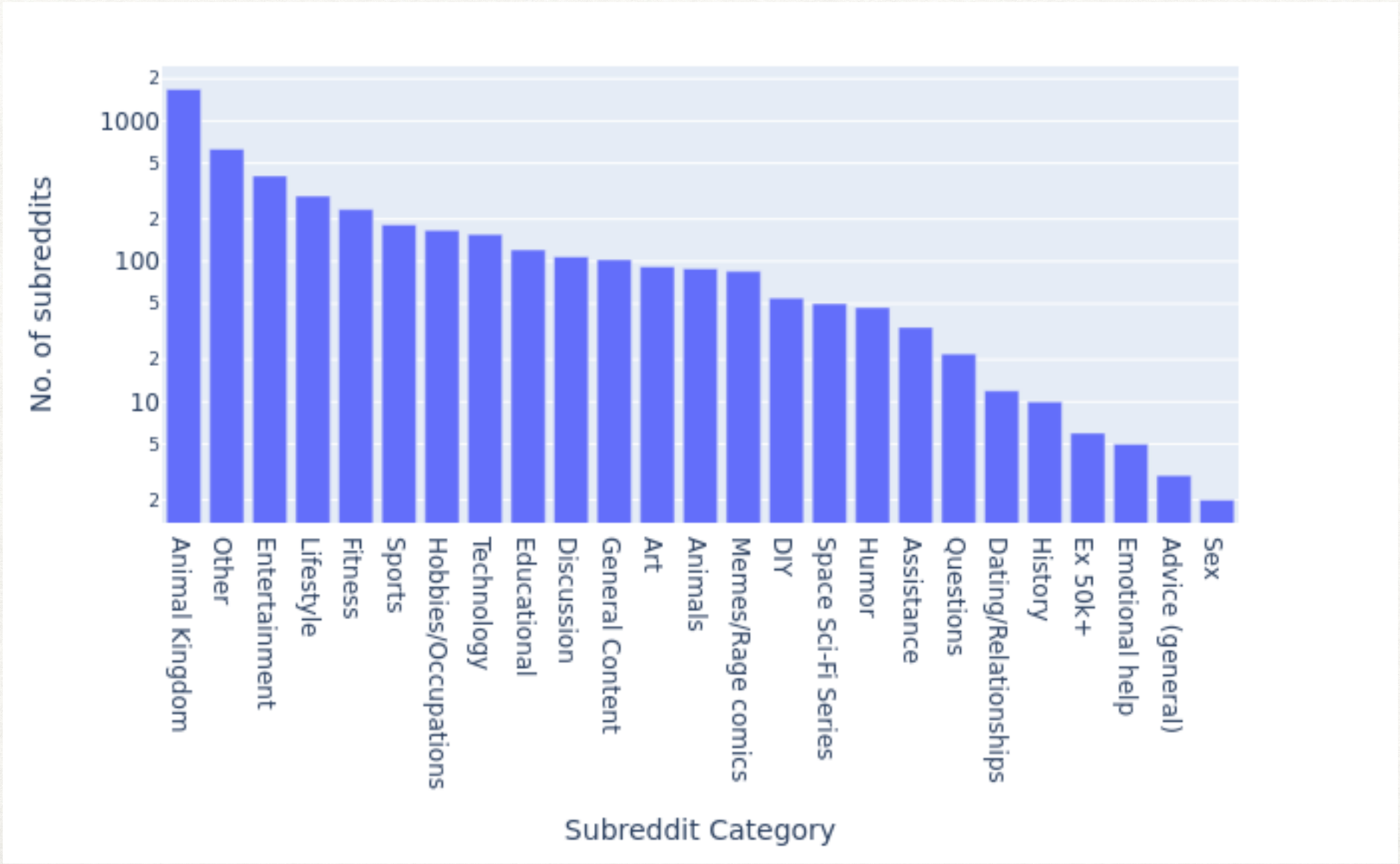


REDDIT WIKI PAGES

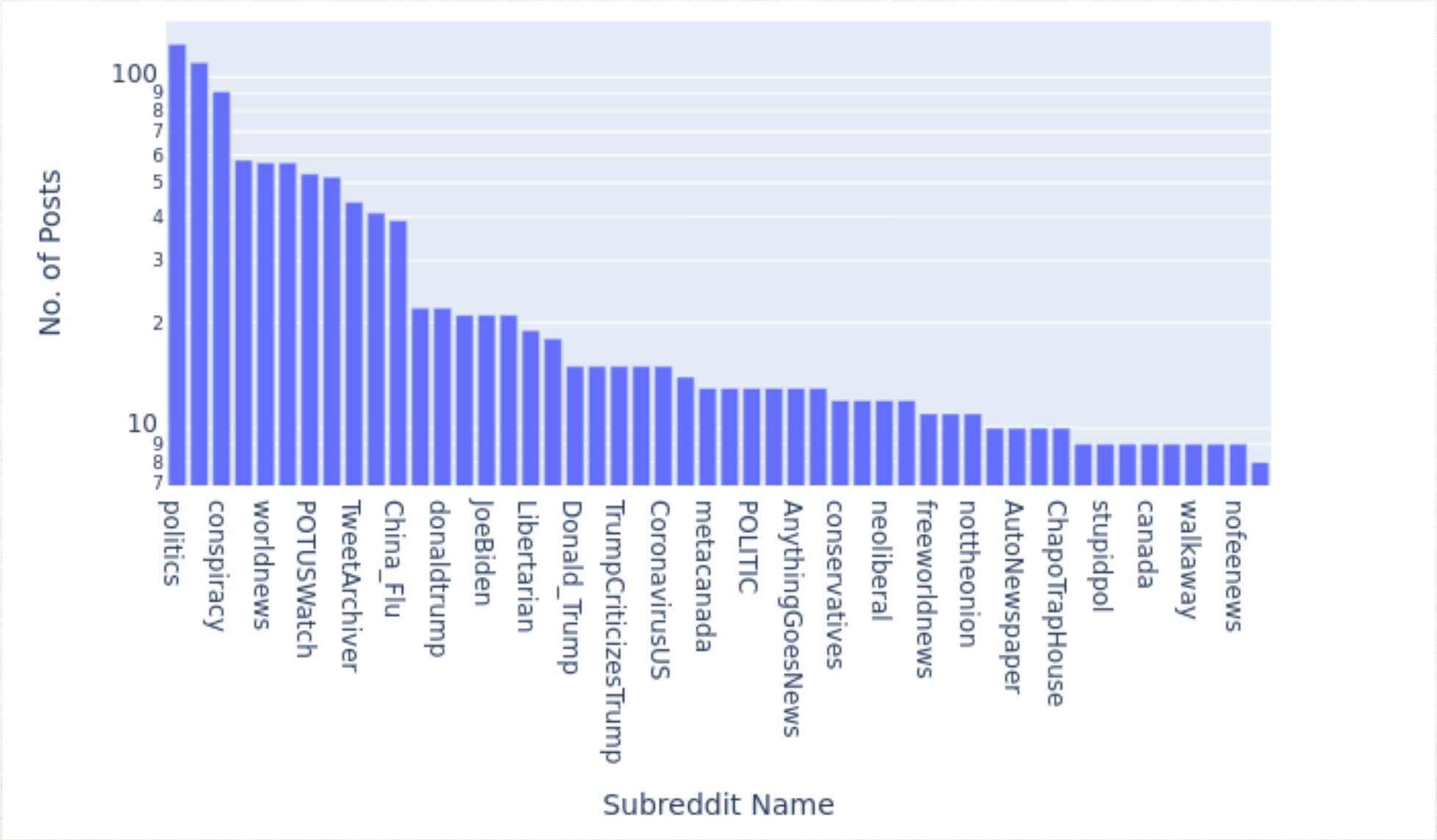
- User-driven hierarchical categorization of subreddits
- Discussion > Stories > Customer Service
- 4998 subreddits
- 5-level hierarchy of categories
- For each new subreddit
  - Split words > match embeddings > associate category



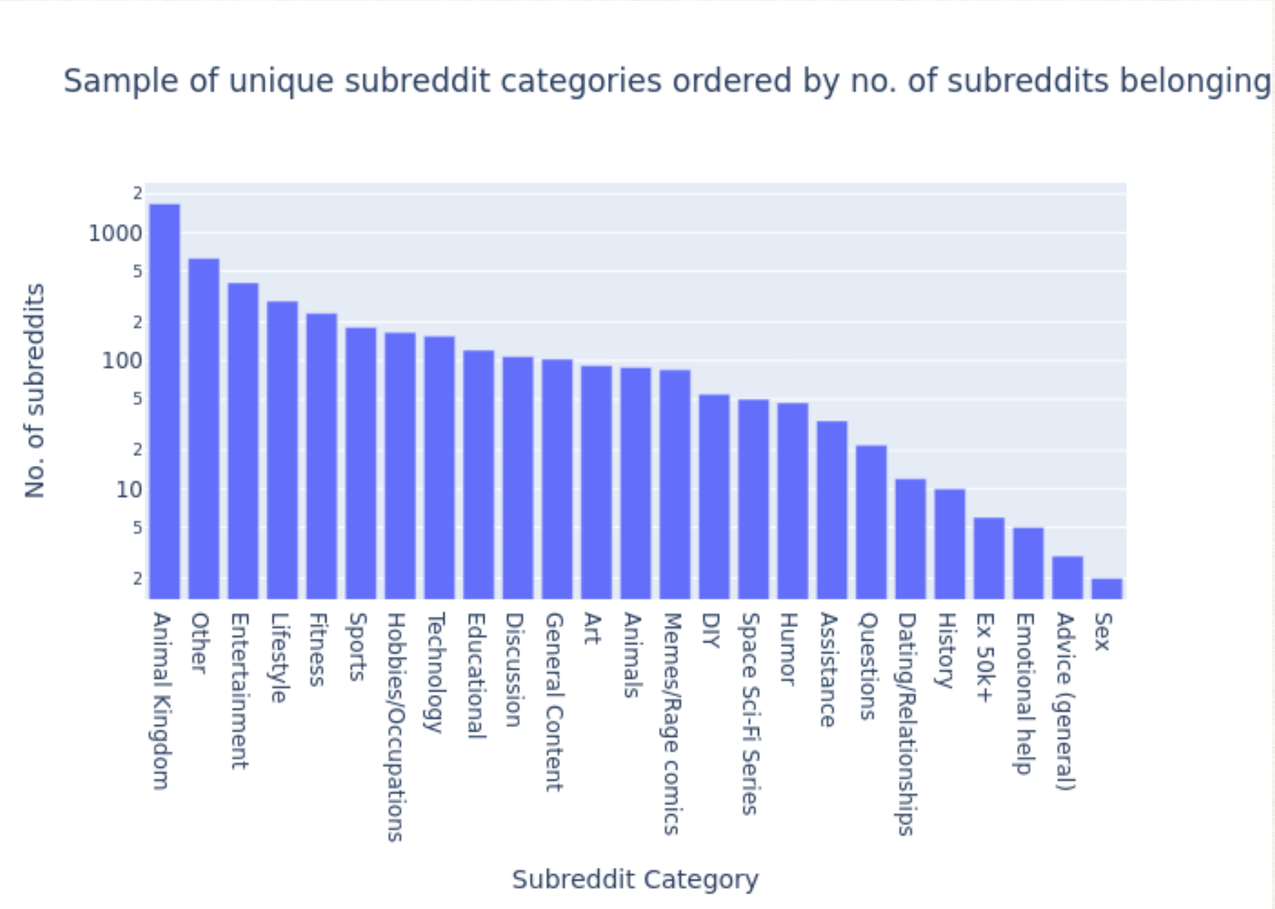
CATEGORIZE POSTS ACROSS SUBREDDITS



CATEGORIZE POSTS ACROSS SUBREDDITS



CATEGORIZE POSTS ACROSS SUBREDDITS



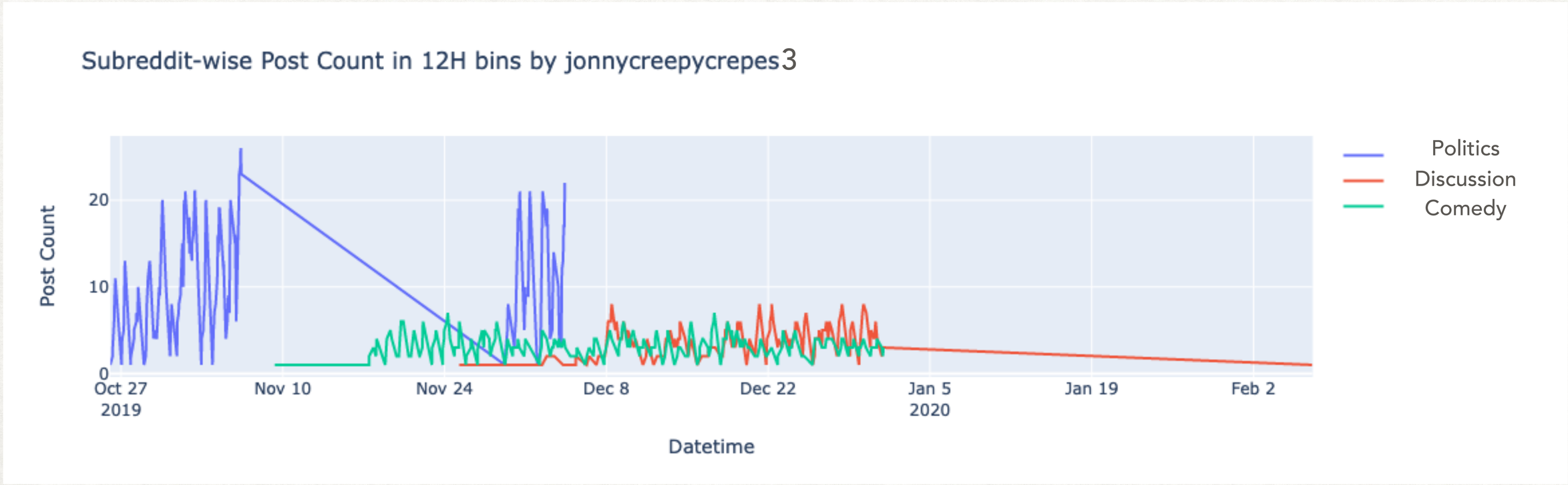
['mormon', 'politics']

Predicted Subreddits:
['mormonhistory', 'ldshistory', 'christianhistory', 'jewishhistory', 'historicalreligion']

Predicted Category:
[[None, 'History of People', None, None, None],

MODELING CATEGORIES

REDUCING SUBREDDIT DIMENSIONALITY



REAL-WORLD INTERVENTIONS

1. **AGENT - LEVEL**

   Awareness campaigns, training, ideological change

2. **NETWORK - LEVEL**

   Reduced sharing, visibility, confirmation of retweets

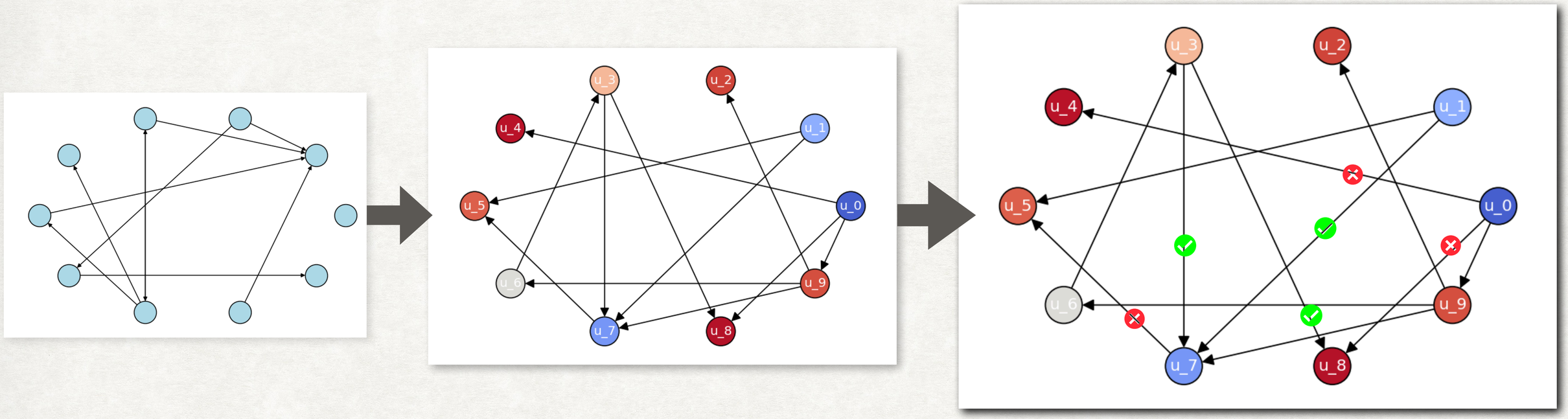
3. **HYBRID**

   Blocking/Temporarily suspending users, articles, links

4. **ADAPTIVE**

   Time-limited blocking and reductions in sharing, visibility

INTERVENTIONS TO LIMIT DISINFORMATION



Agents + Networks

Agents + Networks + Behaviors

Agents + Networks + Behaviors + Interventions