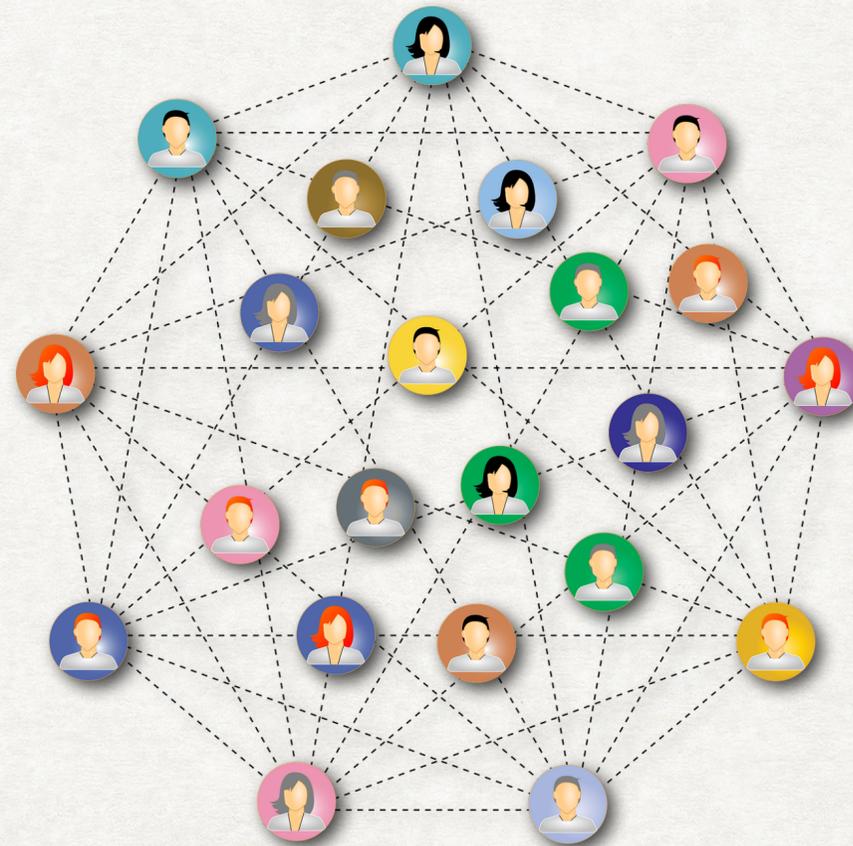


# TOOLS FOR MISINFORMATION CONTROL ON SOCIAL NETWORKS



# ABOUT ME



**2019 - 23** Ph.D. @ NYU Data Science, Center for Social Media + Politics  
Social Networks, Probabilistic Inference, Causal Discovery

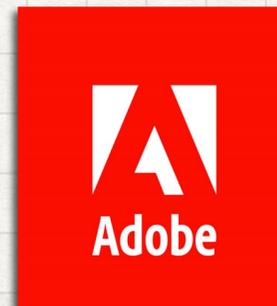
**2022** Ph.D. ML Engineering Intern at Twitter  
Civic Integrity, Misinformation

**2020 - 21** Data Science Research Intern at Adobe  
Trending Hashtag Recommendation for Videos

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



[swapneelm.github.io](https://swapneelm.github.io)



# 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

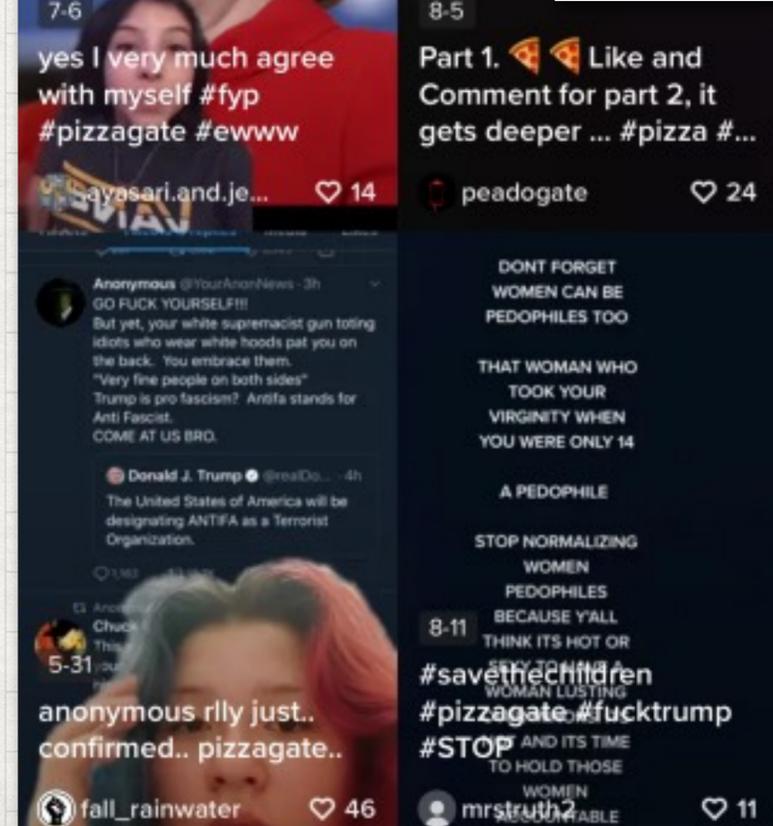
He was hired to build a fake news op but now wants to put things right.

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

# ...HAS REAL-WORLD CONSEQUENCES!



A screenshot of a Twitter post by Mike Cernovich (@Cernovich) dated 12:10 AM on November 22, 2016. The tweet reads: "Pizzagate is not going to go away, this story will be huge! [reddit.com/r/pizzagate/](https://reddit.com/r/pizzagate/)". The tweet has 1,525 retweets and 1,615 likes. The interface shows a search bar with "Pizzagate" and navigation options like "Top", "Users", "Videos", and "Sound".



A screenshot of a social media thread. The top post is by "yes I very much agree with myself #fyp #pizzagate #ewww" with 14 likes. Below it is a post by "Anonymous @YourAcornNews - 3h" with a text-heavy comment. Another post by "Donald J. Trump" is visible. The bottom part of the screenshot shows a post by "anonymous rilly just.. confirmed.. pizzagate.." with 46 likes.

## 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

TikTok

pizzagate



Top

Accounts

Videos

### No results found

This phrase may be associated with behavior or content that violates our guidelines. Promoting a safe and positive experience is TikTok's top priority. For more information, we invite you to review our

[Community Guidelines](#).

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](https://www.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

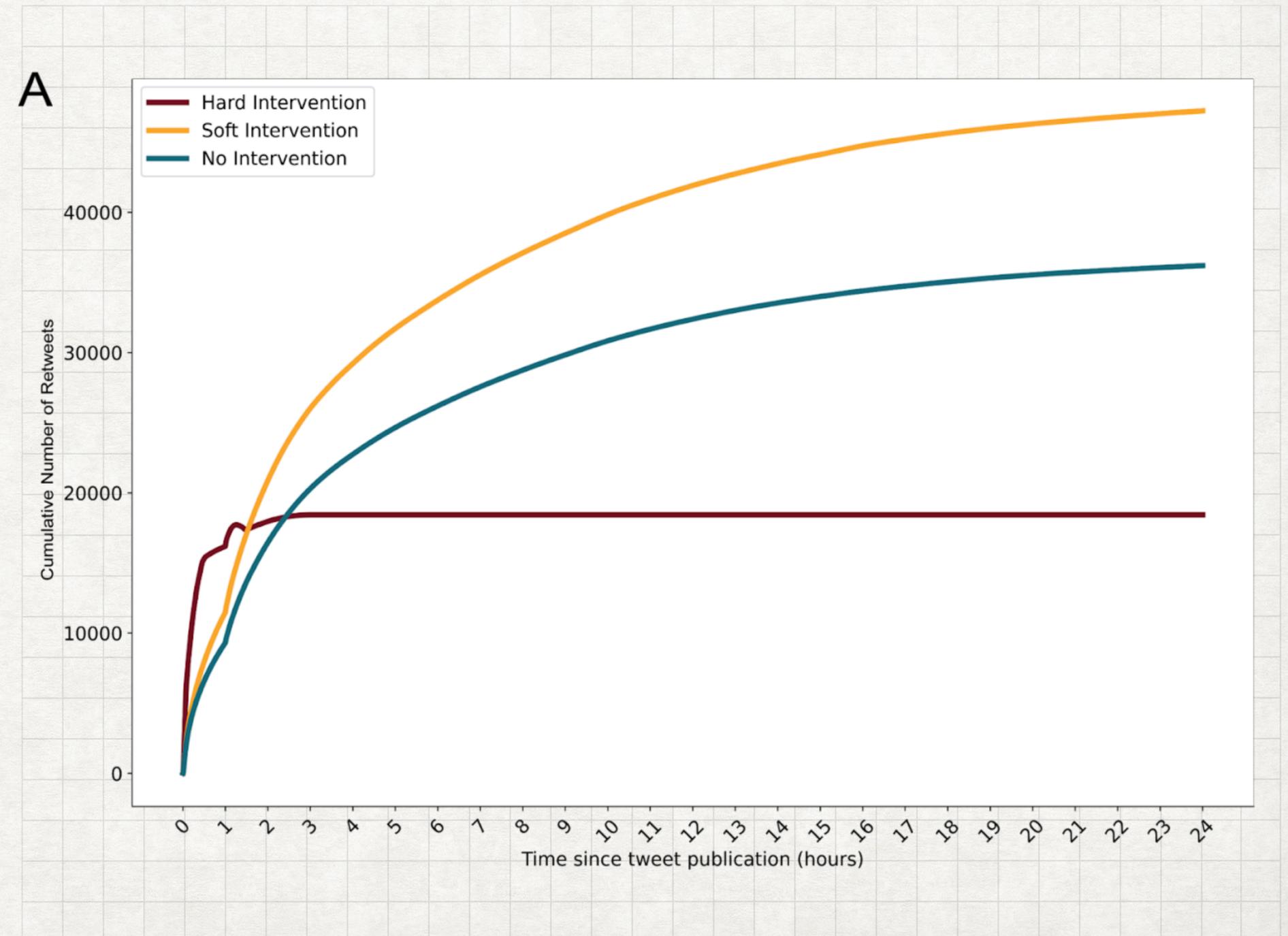
- Estimating the causal effect of Twitter's Interventions on Trump's Tweets
- Two types of interventions: **warning labels** and **removal**
- Looking at multiple platforms reveal surprising conclusions about the interventional effects



# HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

## CAUSAL EFFECTS

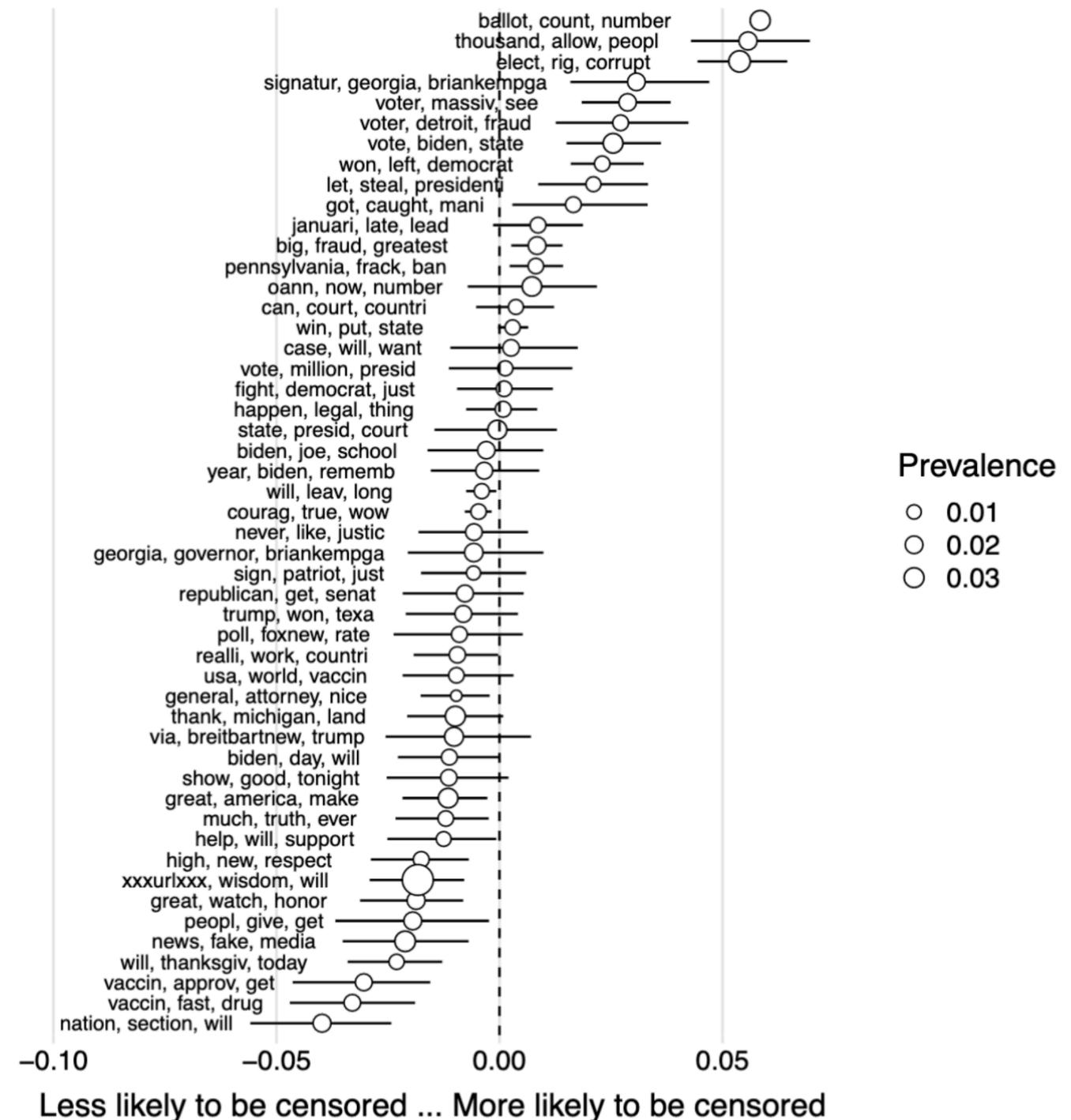
- Naive estimates by Sanderson et. al (2021), indicate strong Streisand effect!
- **Wait does this mean interventions are bad?!**
- What tweets are we really comparing here and what are their features like?



# 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

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

**Requires** Linearity in Prior Outcomes (LPO)

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

# HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

## CAUSAL EFFECTS

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

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

**Requires** Linearity in Prior Outcomes (LPO)

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

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

# HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

## CAUSAL EFFECTS

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

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

**Requires** Linearity in Prior Outcomes (LPO)

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

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

**Relaxes** Linearity in Prior Outcomes (LPO)

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

# 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

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

**Requires** Linearity in Prior Outcomes (LPO)

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

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

**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

- 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)

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

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

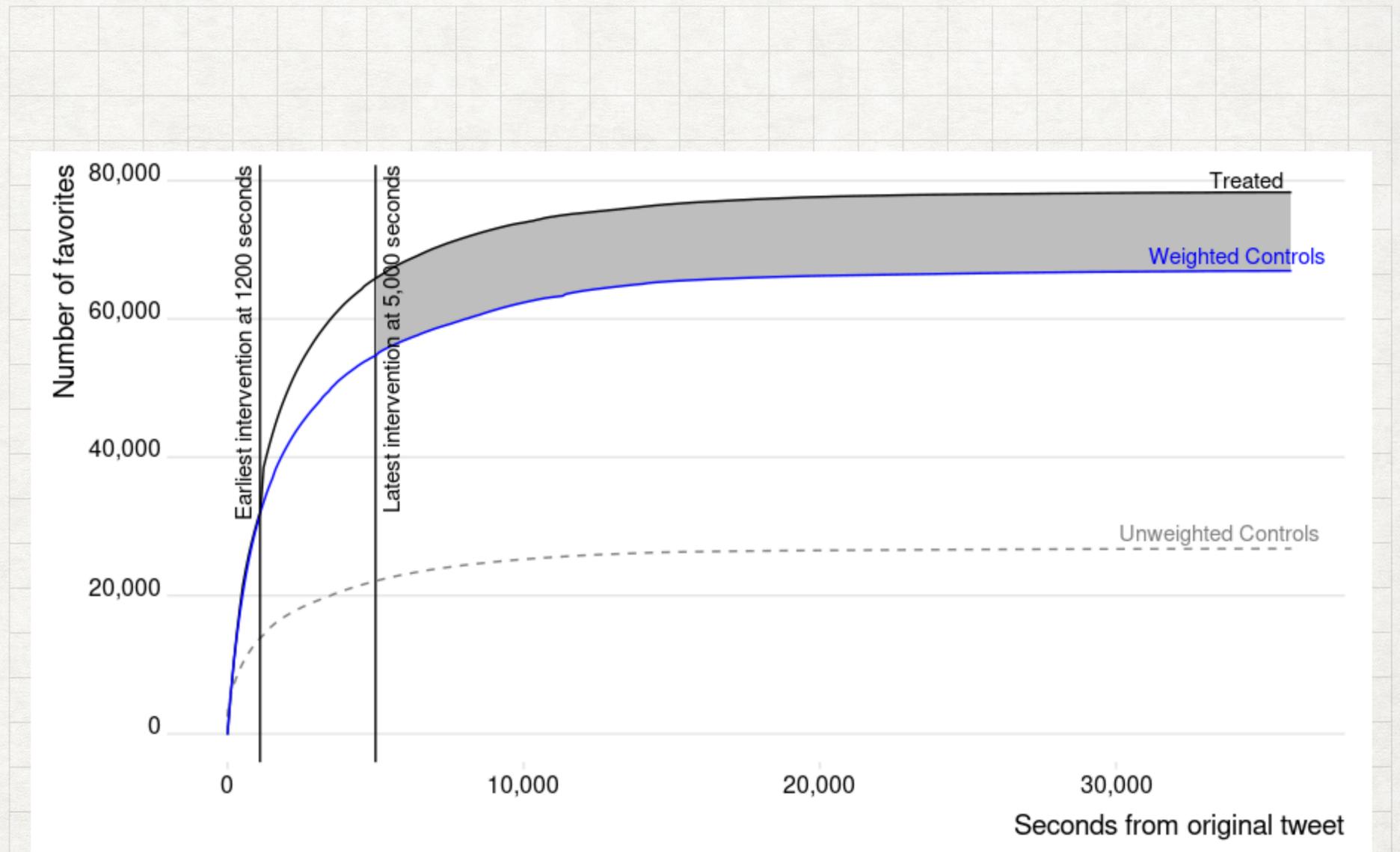
**Relaxes** Linearity in Prior Outcomes (LPO)

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

# HOW DID TWITTER'S INTERVENTIONS AFFECT TRUMP?

## CAUSAL EFFECTS

- Robustness check over assumptions of earliest to latest time they could have intervened!
- $T_0$  = earliest possible intervention
- $T_1$  = latest possible intervention
- Kernel balanced estimates are most robust to changes

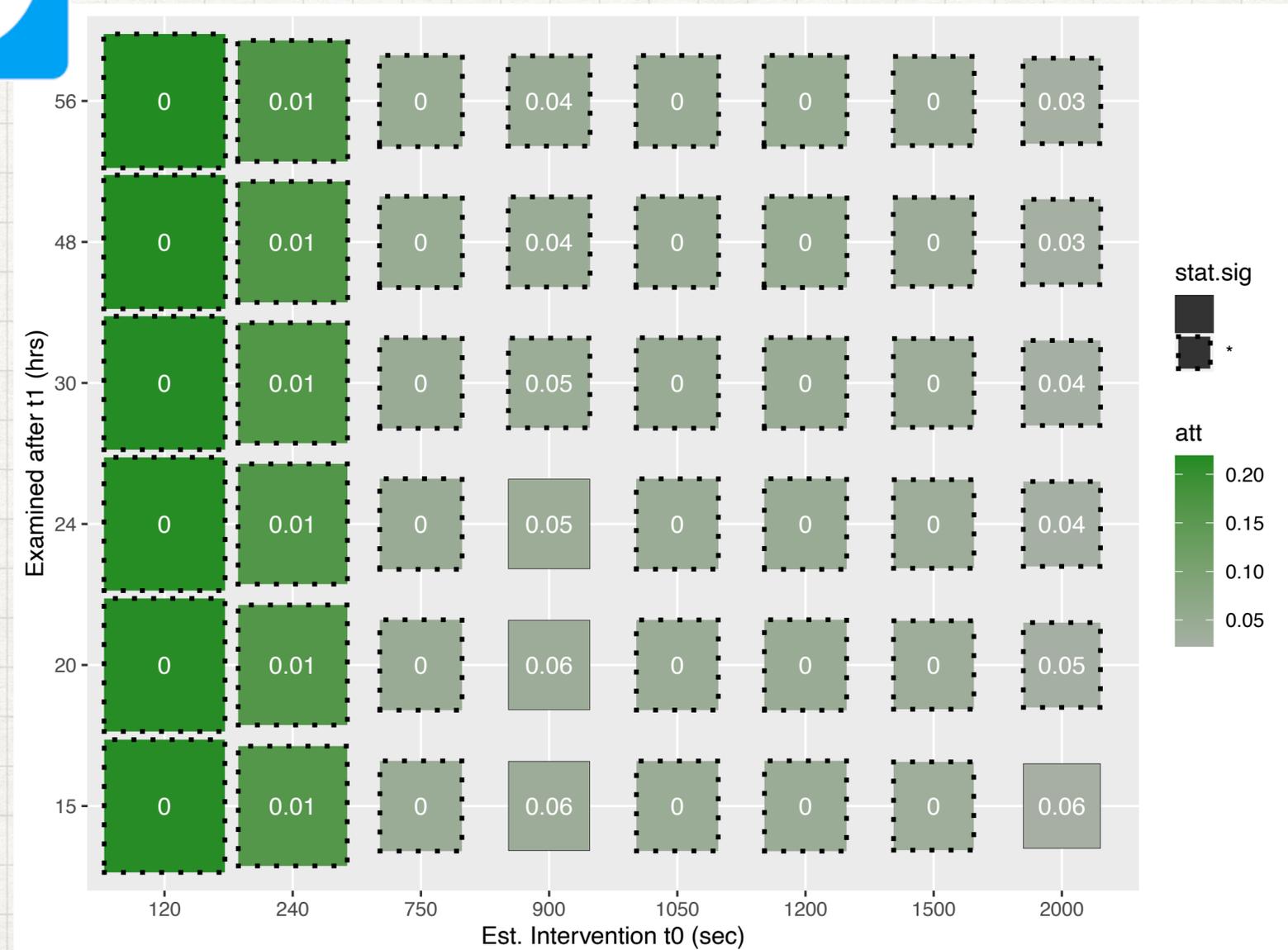


# RESULTS

## 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!

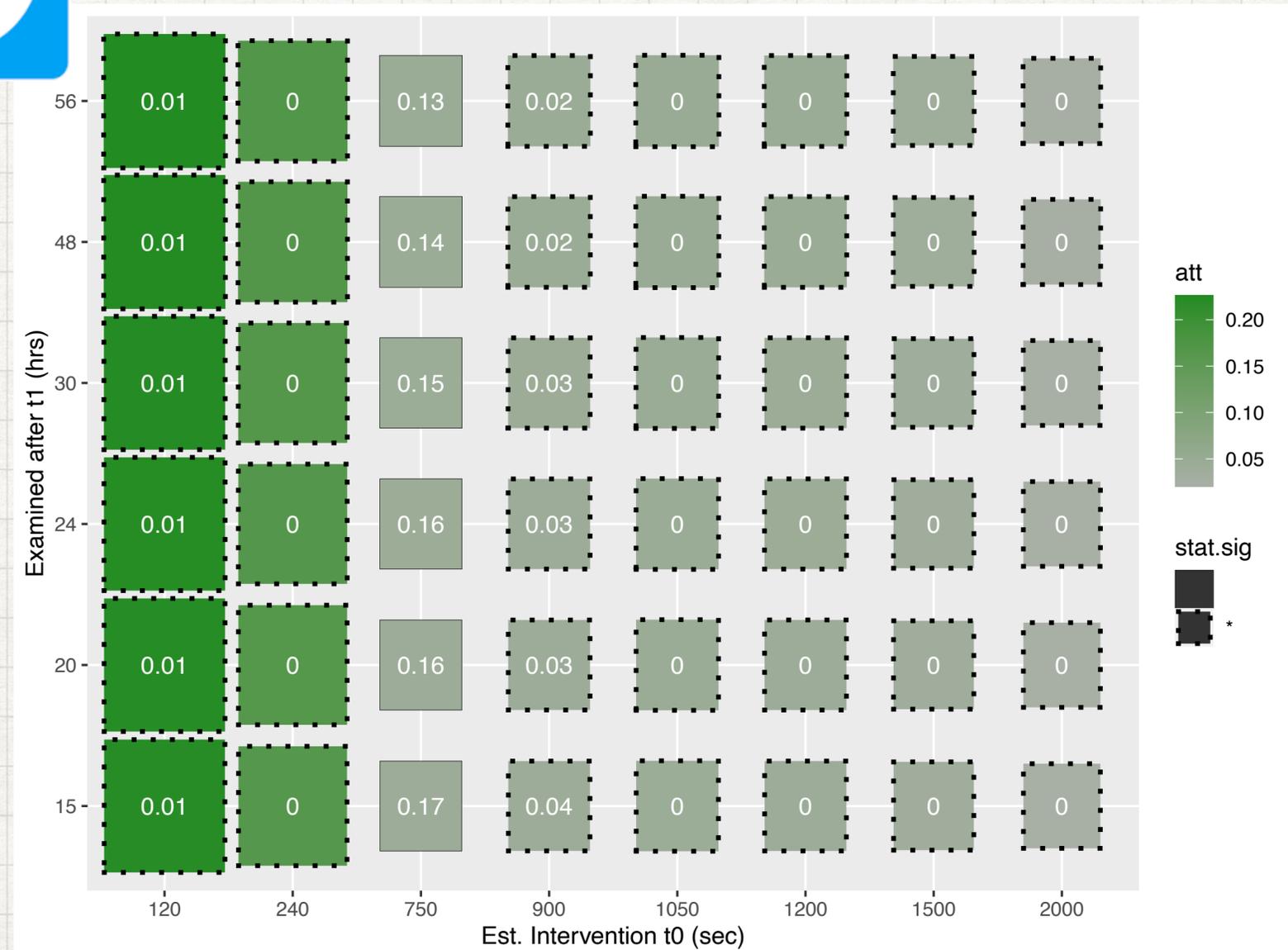


# RESULTS

## 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!



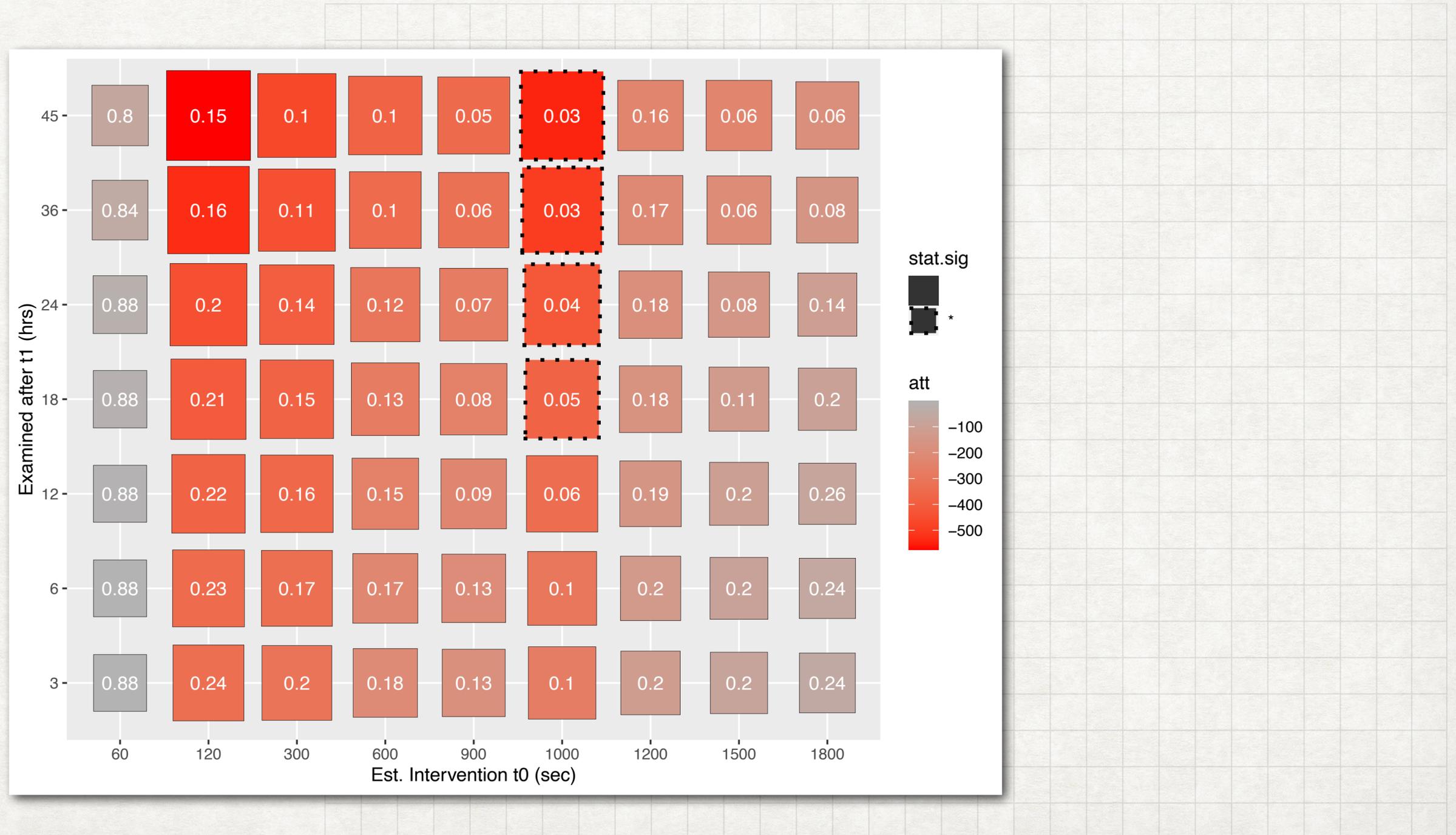
# REDDIT - HARD



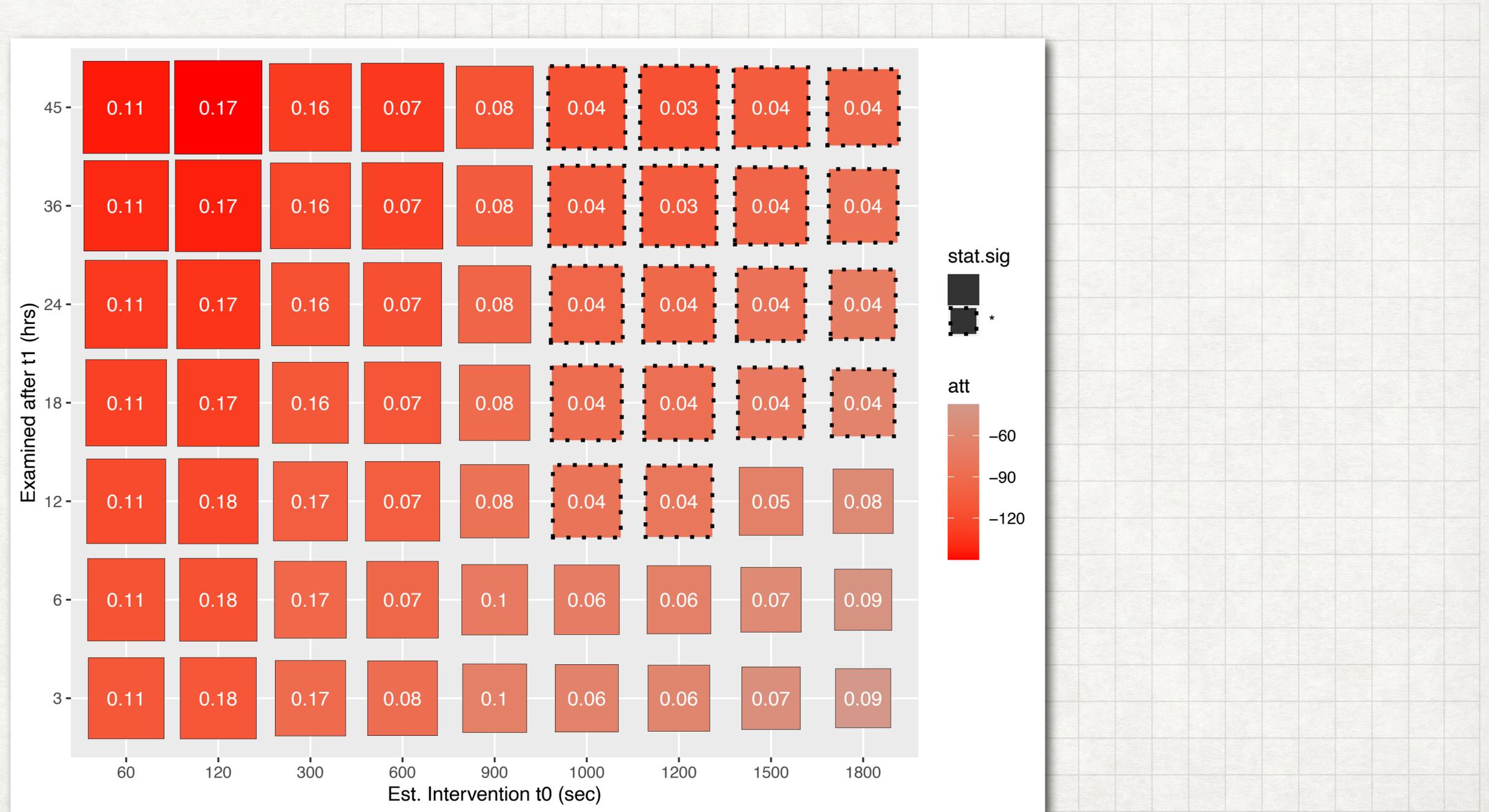
# REDDIT - SOFT



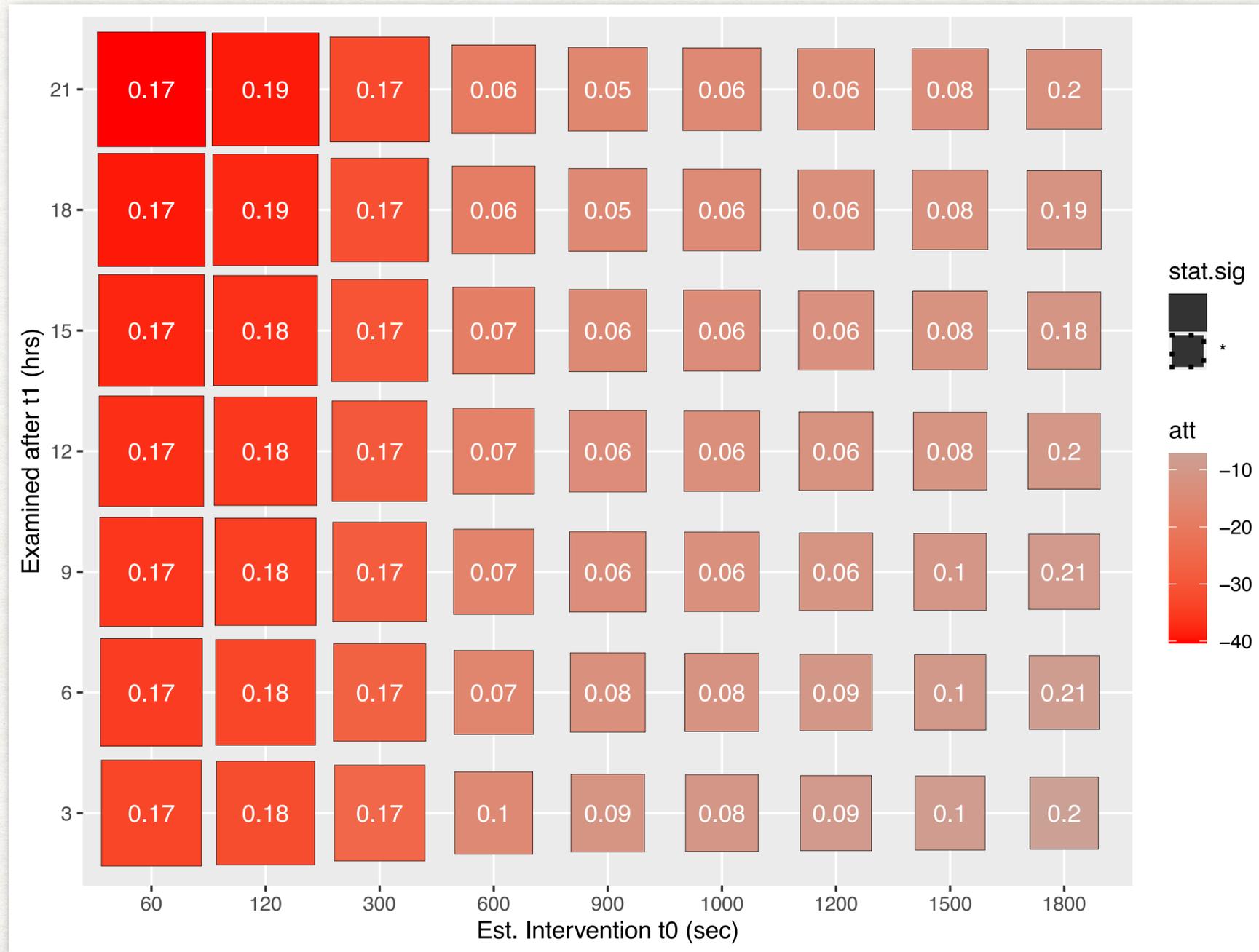
# FACEBOOK - HARD



# 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

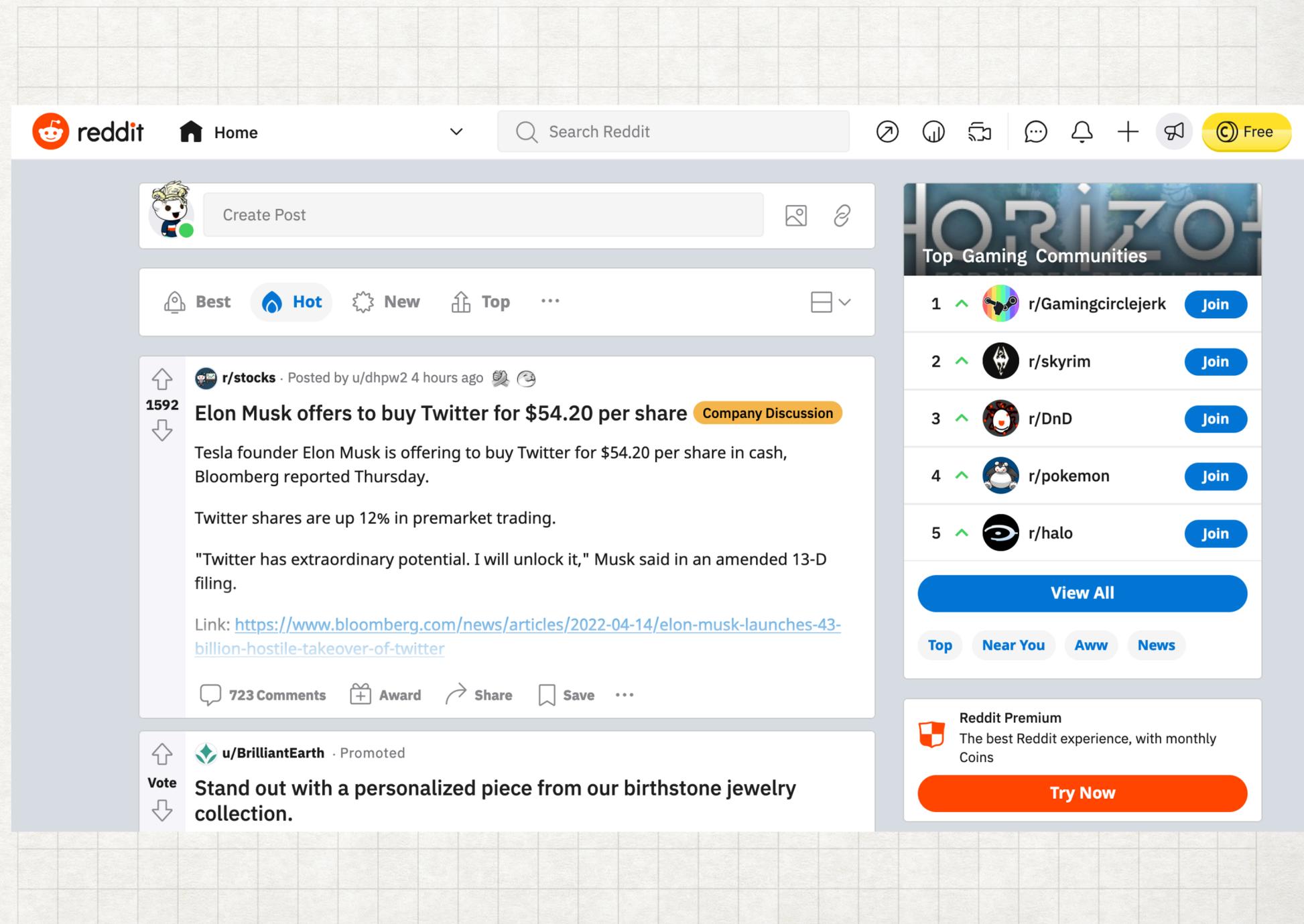
- RQ: Quantify the relative impact of different algorithmic choices that a platform makes, for instance:
  - To maximize engagement or not to maximize engagement?
  - Should we promote diverse content that isn't getting early views?
  - Dealing with "controversial" opinions?

# 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?

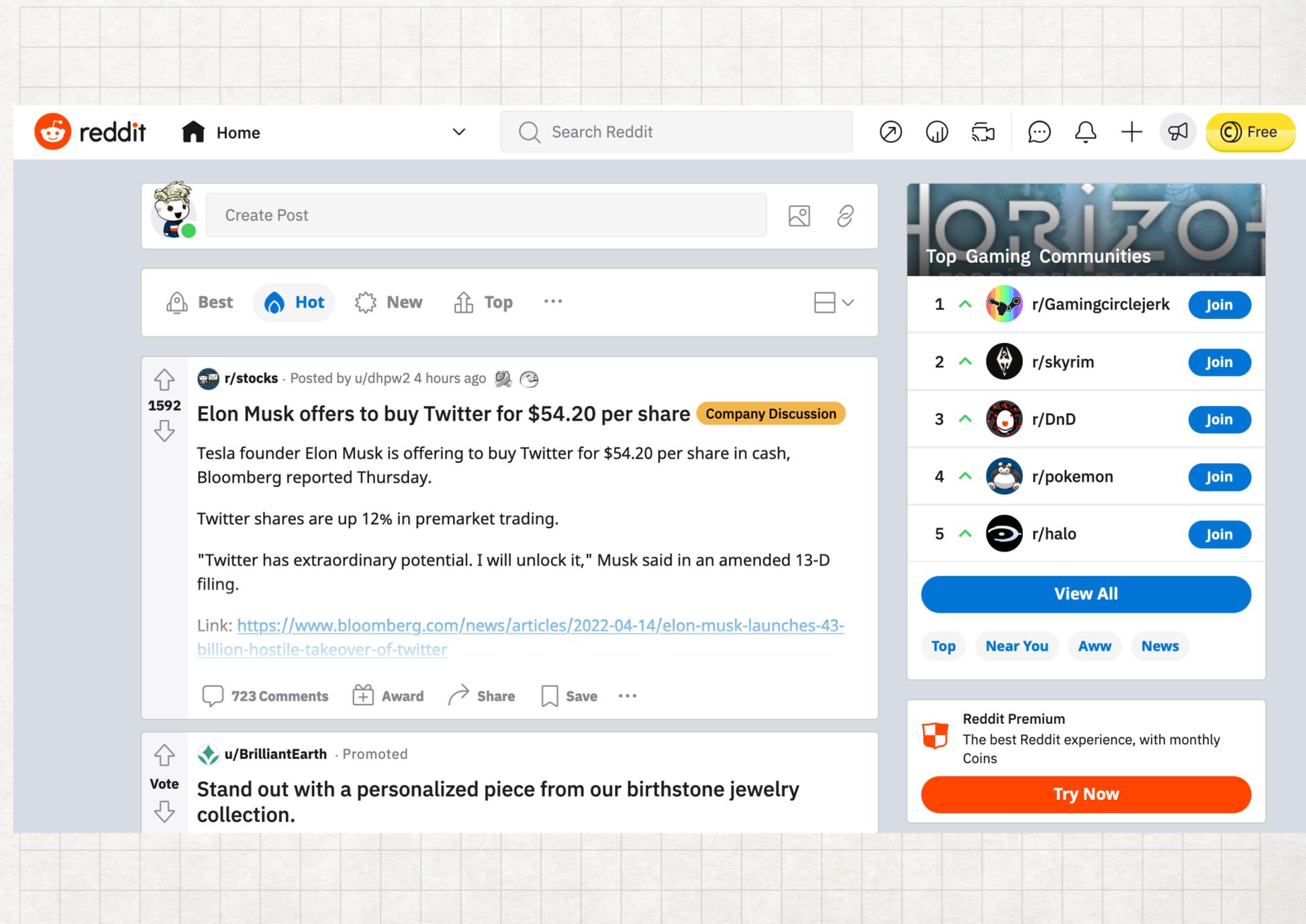
# SIMULATE A SOCIAL NETWORK

- 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

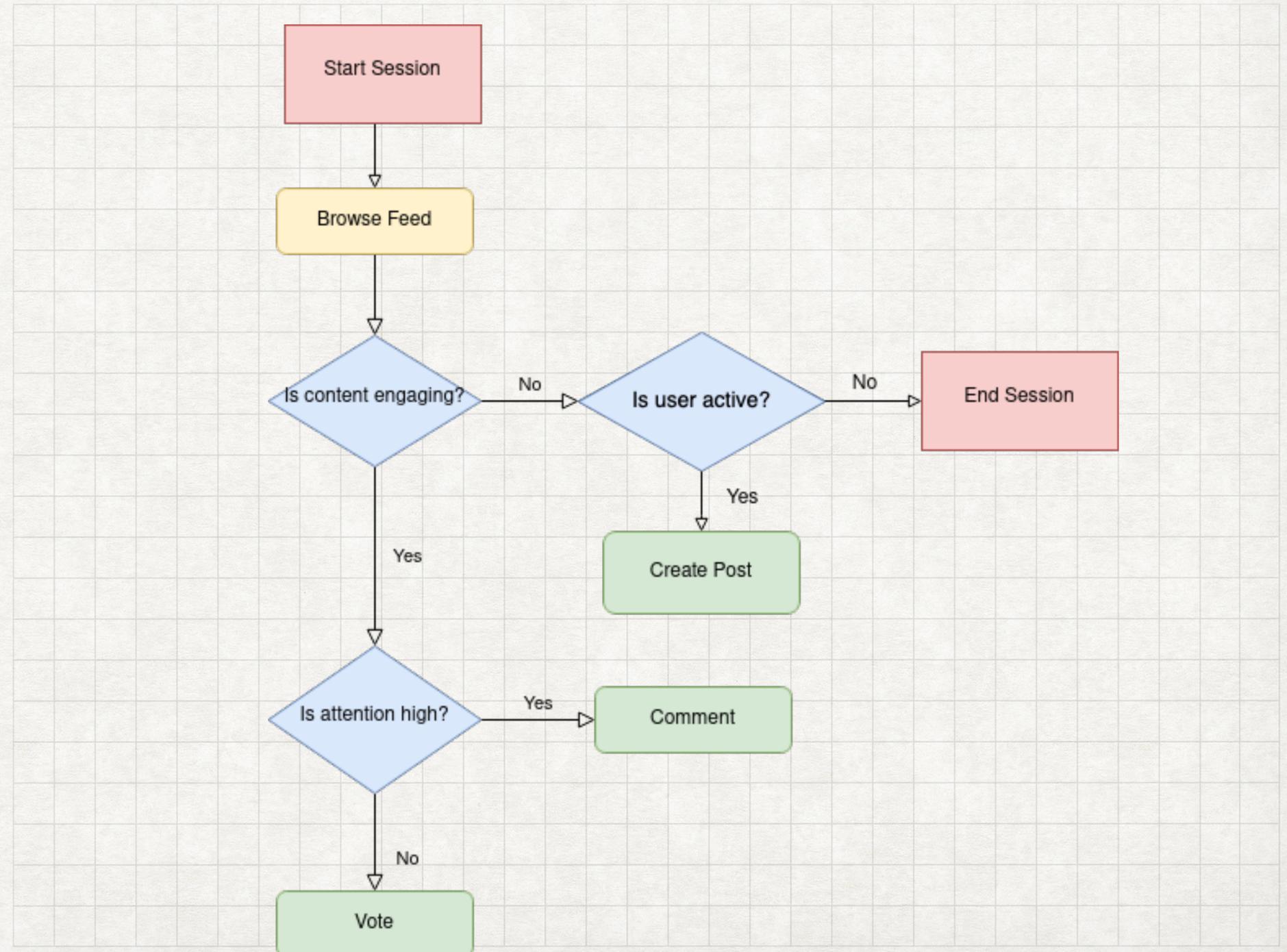
1: **procedure** REDDIT  
2: *Sample latents:*  
3:  $\mathbf{v} \leftarrow \{v_i \sim \text{Uniform}(0, 1)\}_{i=1, j=1}^{N, S}$  ▷ Interaction propensity over subreddit categories  
4: *Simulate:*  
5:  $\Phi_{1:N} \leftarrow \langle \rangle$  ▷ User Activity  
6: **for**  $t = 1 : T$  **do**  
7:     **for**  $i = 1 : N$  **do**  
8:          $\gamma \sim \text{Categorical}(\pi_i)$  ▷ Choose Subreddit (category)  
9:          $\tau \sim \text{Bernoulli}(v_i, \gamma)$  ▷ Interact with Subreddit (category)  
10:          $\Phi_i \leftarrow \Phi_i + \langle \tau \rangle$  ▷ Append to user's activity  
11: **return**  $\Phi_{1:N}$  ▷ All user activity



# SIMULATING SOCIAL NETWORKS

## REDDIT

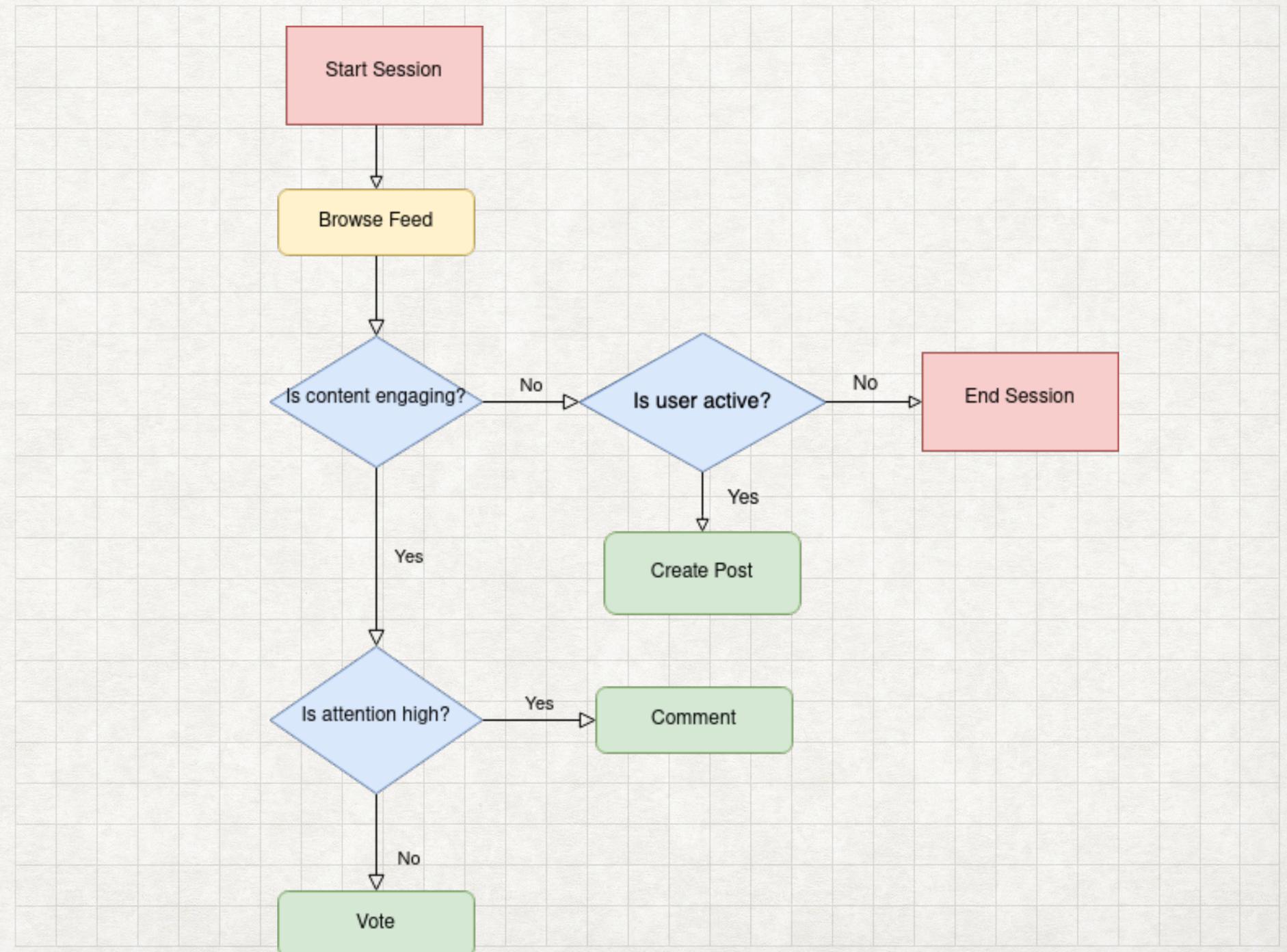
- Developing a model for a user's posting behavior
- Creating a story of how a user interacts with Reddit



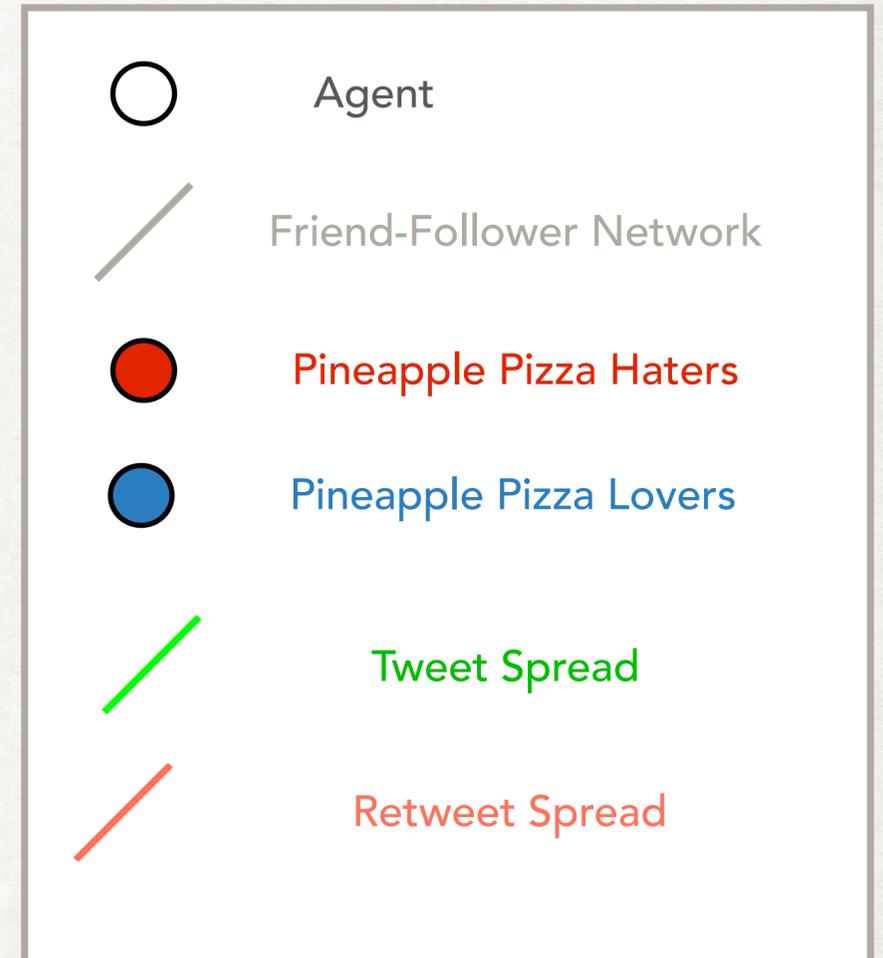
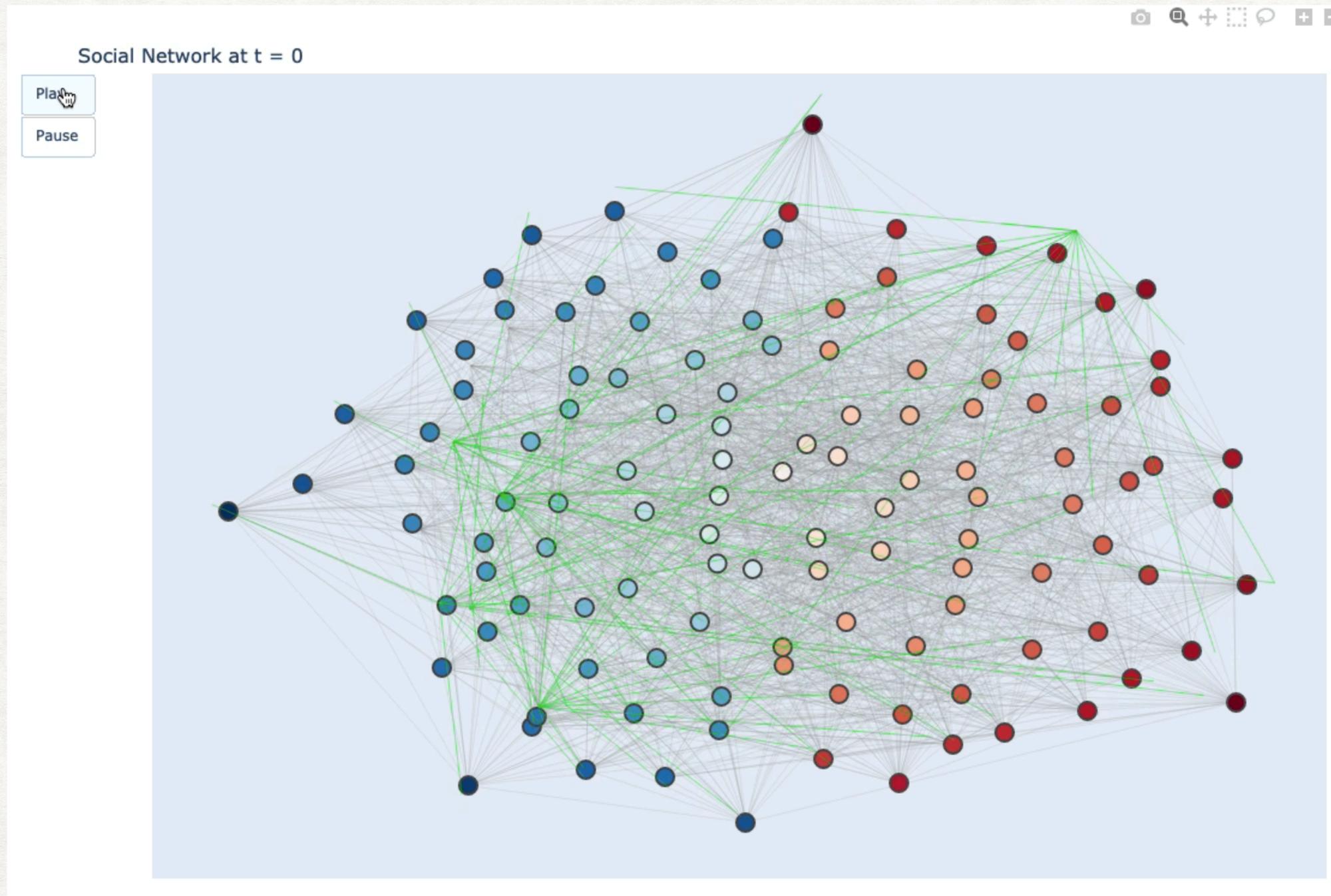
# SIMULATING SOCIAL NETWORKS

## 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!

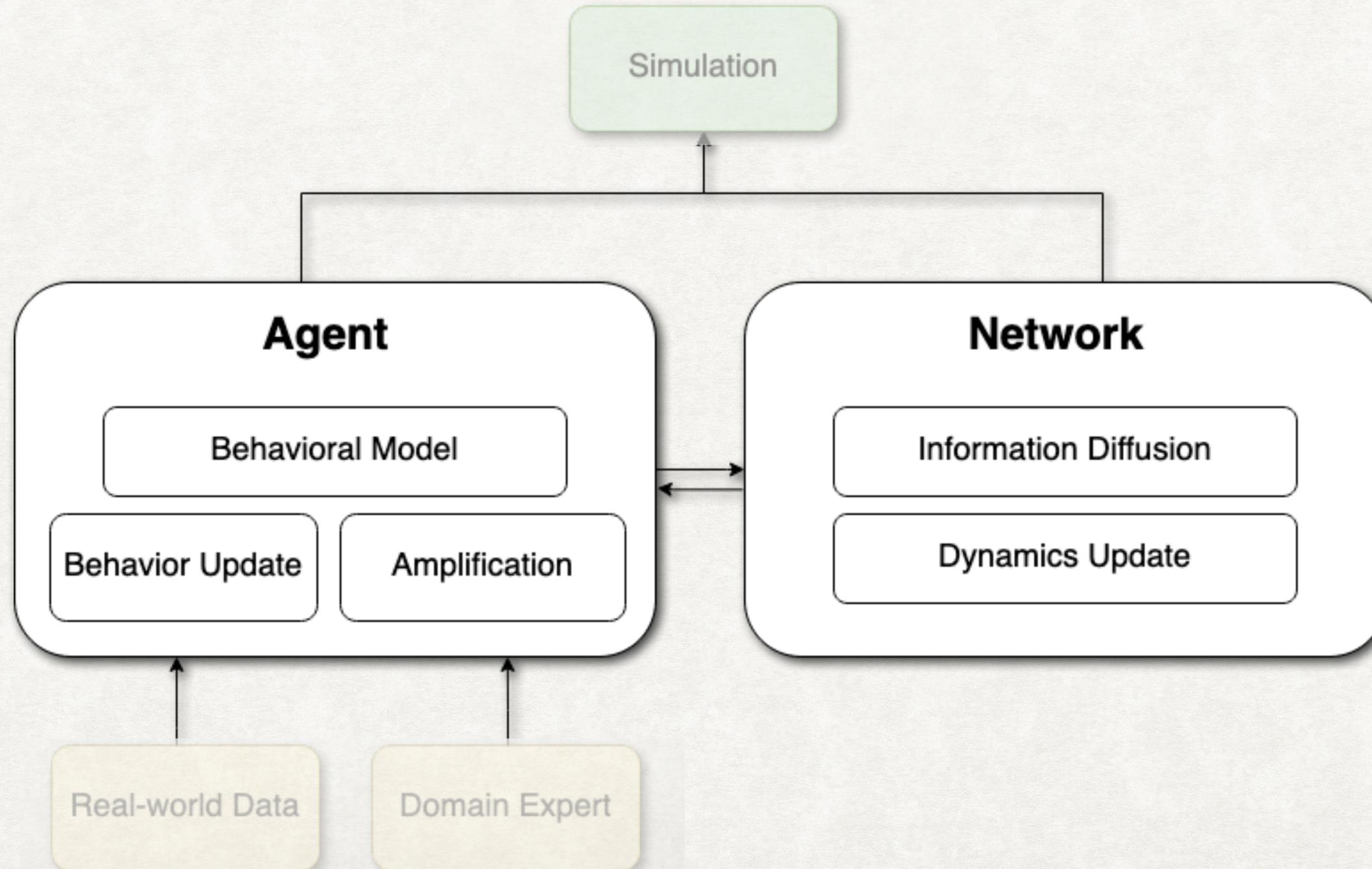


# SIMULATING SOCIAL NETWORKS

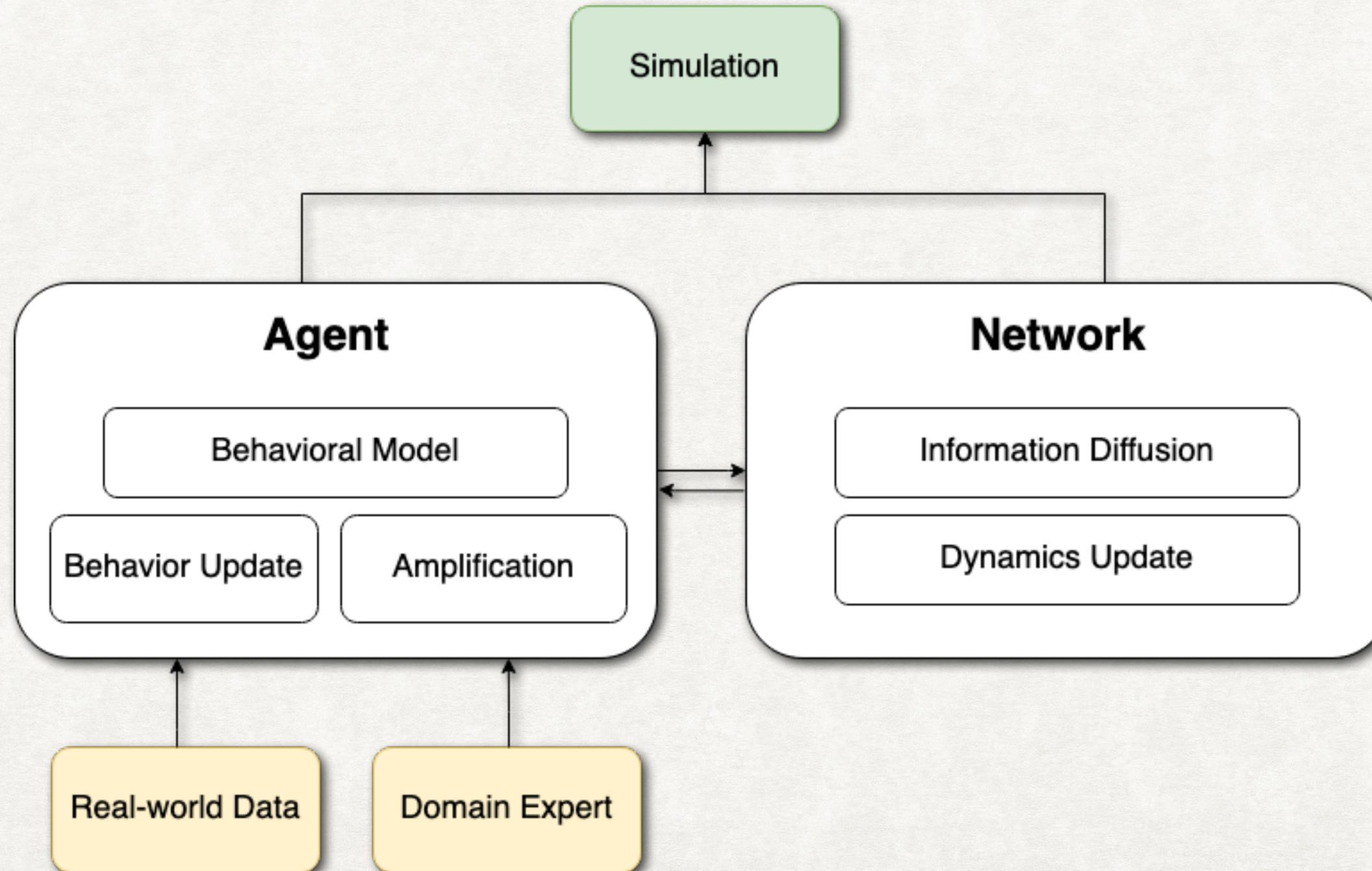


<https://youtu.be/GV5XuftiD7s>

# SIMPPL



# SIMPPL



# DISINFORMATION MANOEUVRES

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

K. Carley, 2020

# REDDIT RECOMMENDER SYSTEMS

The image shows a screenshot of the Reddit homepage. At the top, there is a navigation bar with the Reddit logo, a 'Home' button, a search bar, and various utility icons. Below the navigation bar is a 'Create Post' section. A red oval highlights the sorting options: 'Best', 'Hot', 'New', and 'Top'. The main content area features a post from the r/stocks subreddit, titled 'Elon Musk offers to buy Twitter for \$54.20 per share'. The post includes a link to a Bloomberg article and has 723 comments. On the right side, there is a 'Top Gaming Communities' section listing r/Gamingcirclejerk, r/skyrim, r/DnD, r/pokemon, and r/halo. At the bottom, there is a 'Reddit Premium' banner.

reddit Home Search Reddit

Create Post

Best Hot New Top

1592 ↑  
↓

r/stocks · Posted by u/dhpw2 4 hours ago

### Elon Musk offers to buy Twitter for \$54.20 per share Company Discussion

Tesla founder Elon Musk is offering to buy Twitter for \$54.20 per share in cash, Bloomberg reported Thursday.

Twitter shares are up 12% in premarket trading.

"Twitter has extraordinary potential. I will unlock it," Musk said in an amended 13-D filing.

Link: <https://www.bloomberg.com/news/articles/2022-04-14/elon-musk-launches-43-billion-hostile-takeover-of-twitter>

723 Comments Award Share Save

### Top Gaming Communities

- 1 r/Gamingcirclejerk Join
- 2 r/skyrim Join
- 3 r/DnD Join
- 4 r/pokemon Join
- 5 r/halo Join

View All

Top Near You Aww News

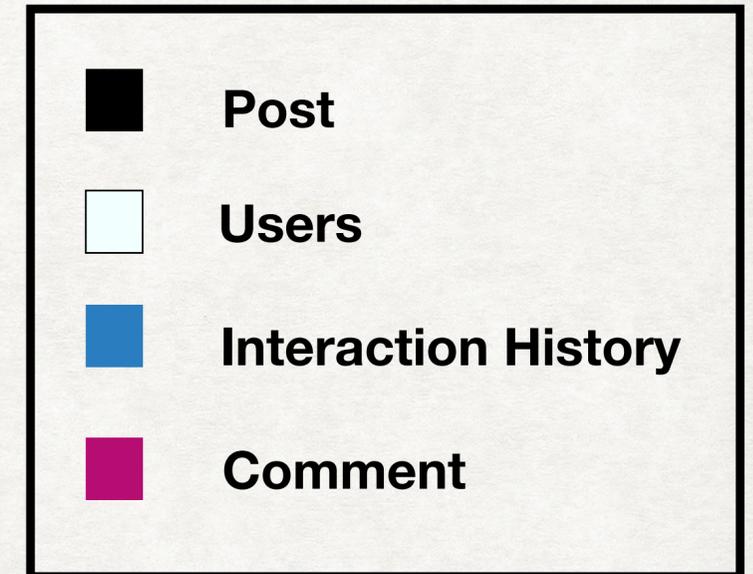
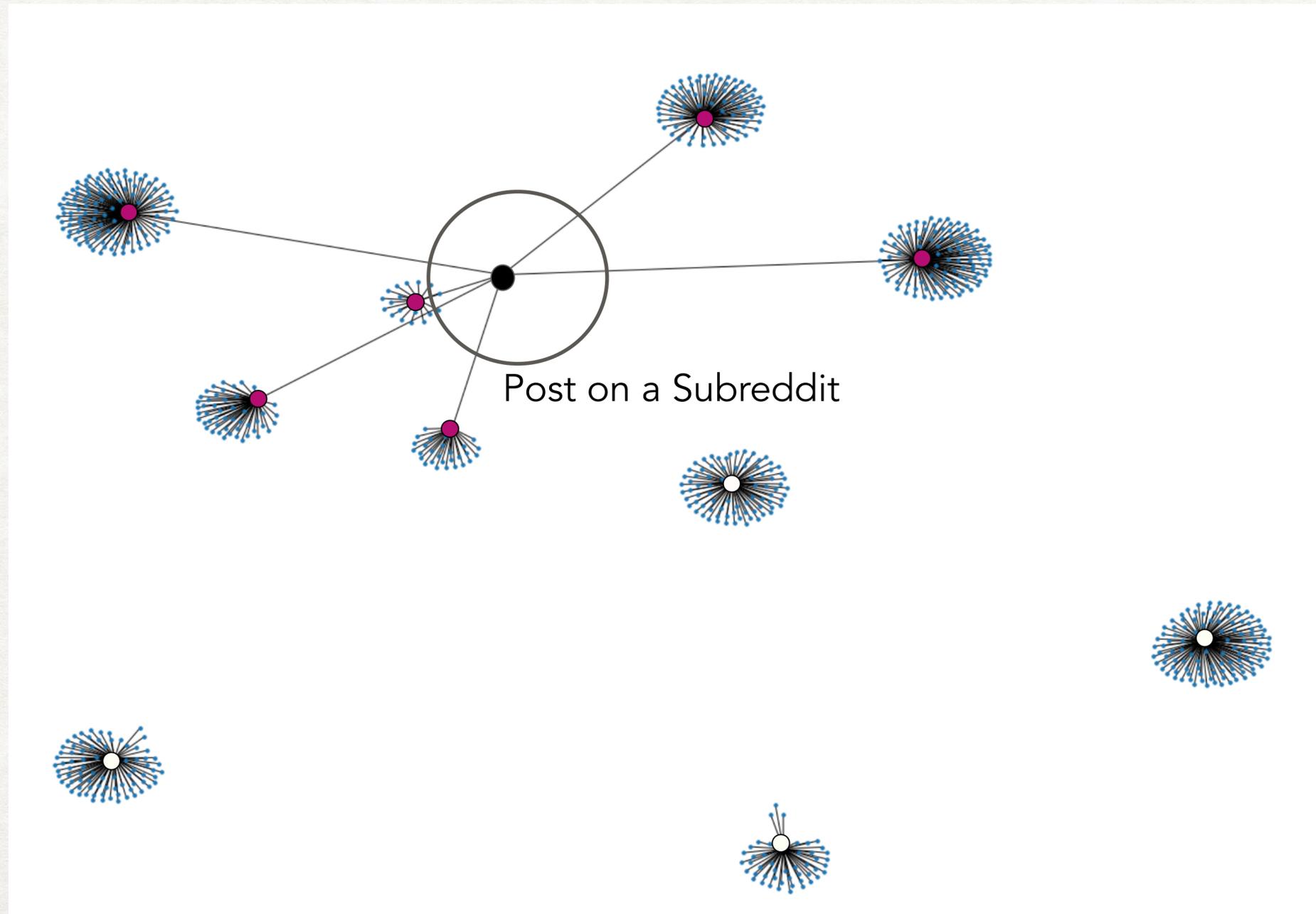
Reddit Premium

# RANKING AND RECOMMENDATION ALGORITHMS

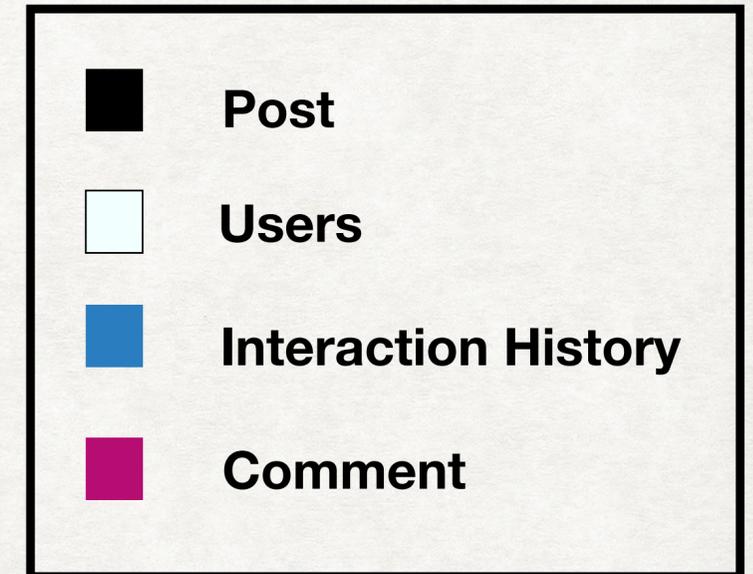
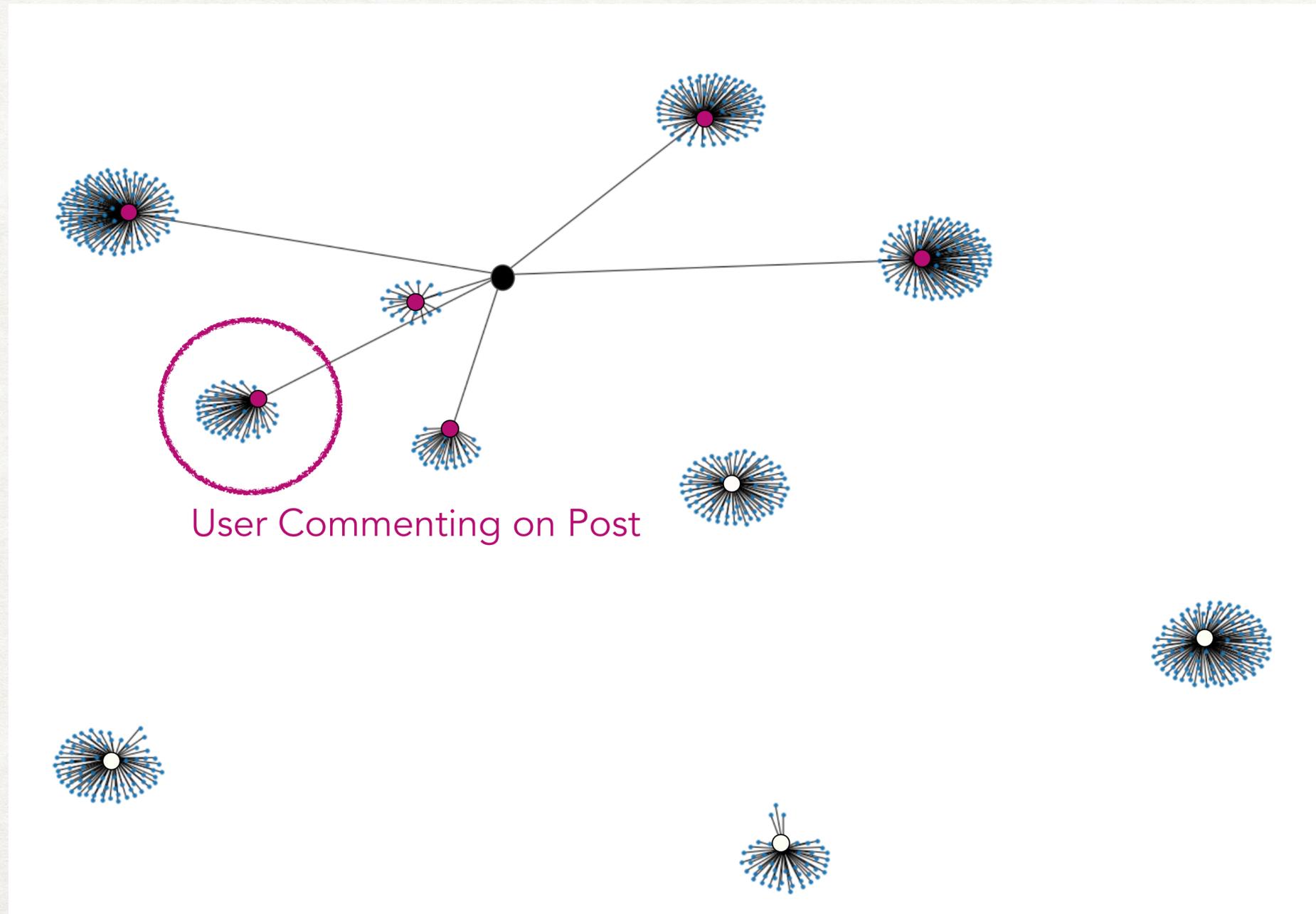
New	Top	Rising	Controversial	Best (Personalized)
Age of Post	Age of Post			Age of Post
	# of Upvotes	# of Upvotes	# of Upvotes	# of Upvotes
			# of Downvotes	
		Age of Votes		
		Age of Comments		
				Relevance to User
				Subreddit Membership

# USING REAL-WORLD DATA TO DRIVE THE SIMULATIONS

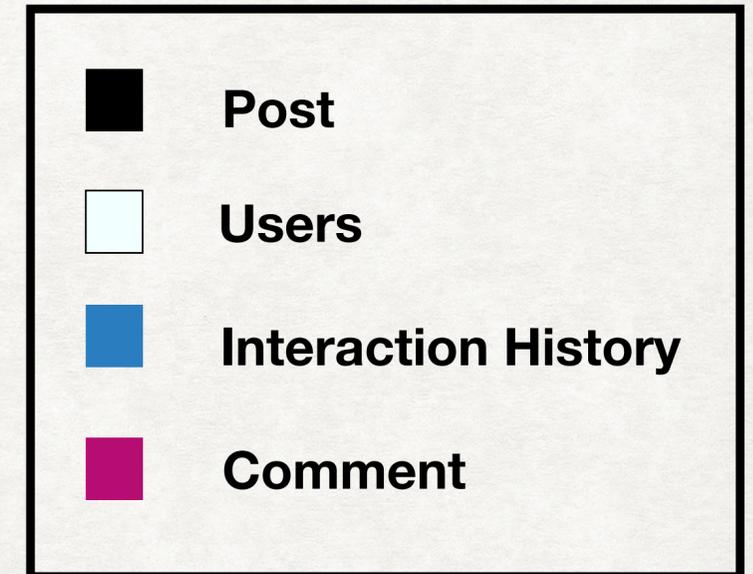
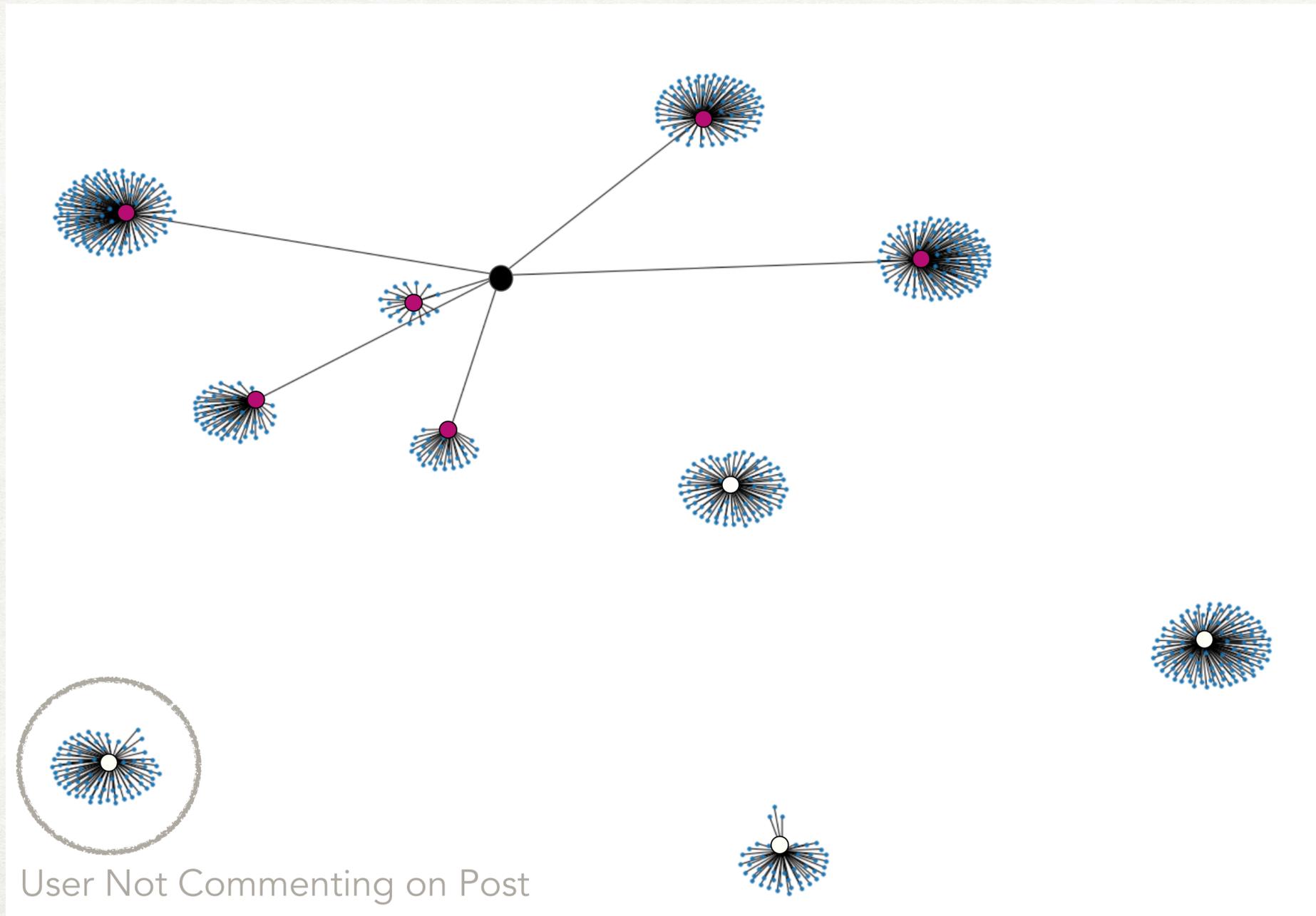
# REDDIT POST



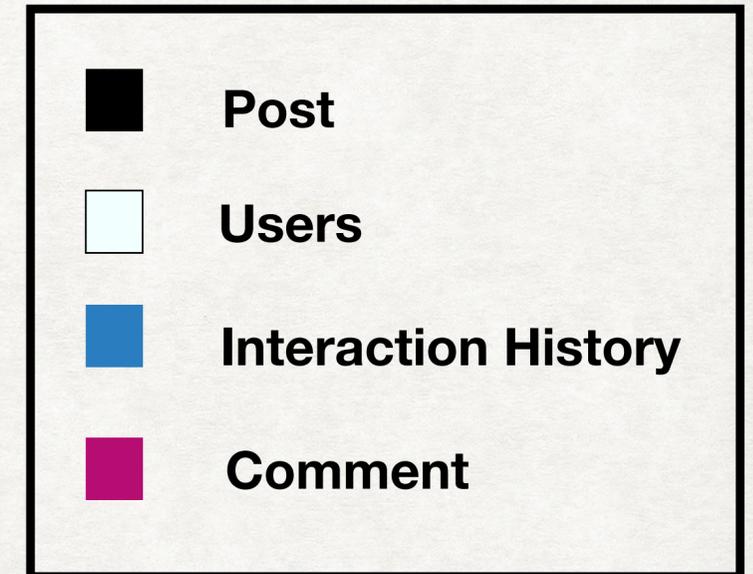
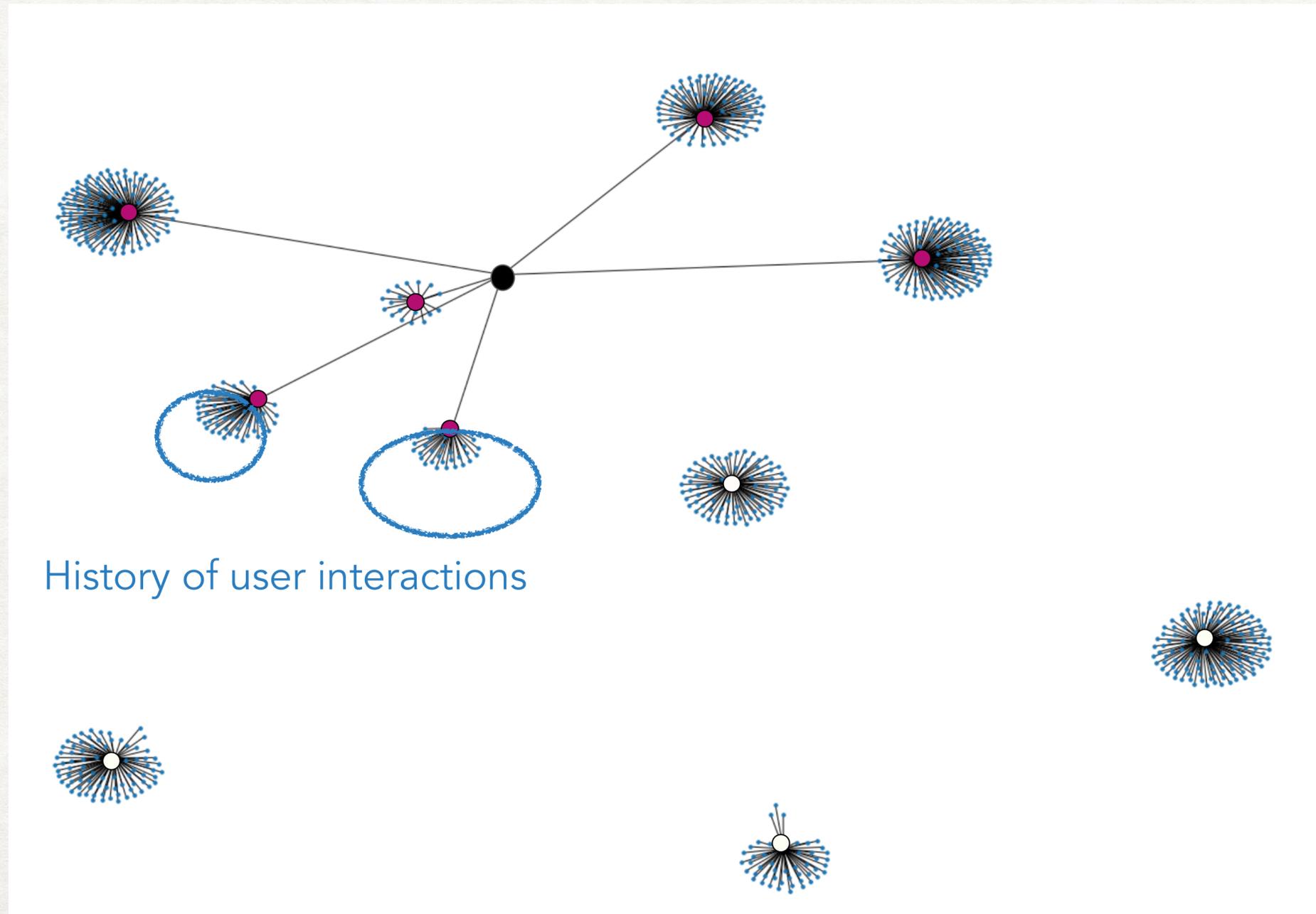
# REDDIT POST



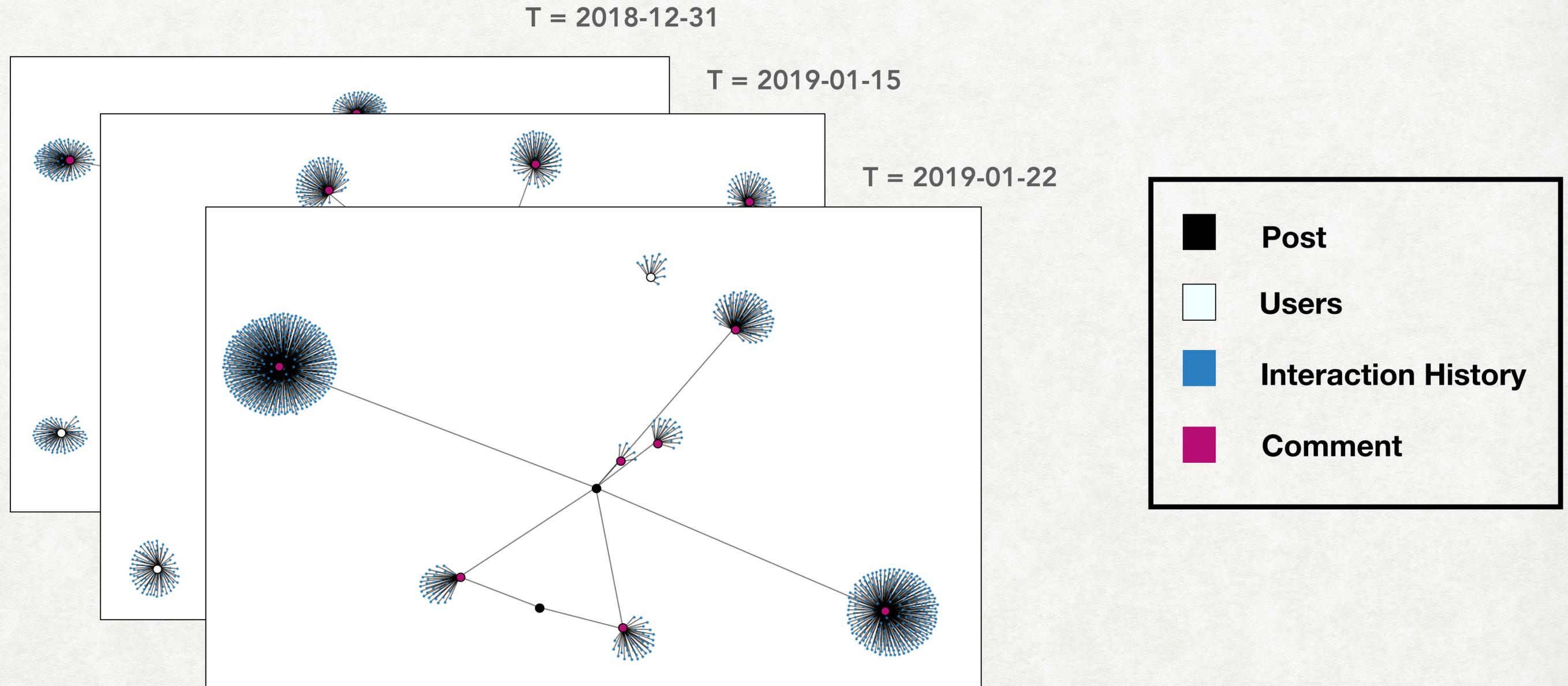
# REDDIT POST



# REDDIT POST



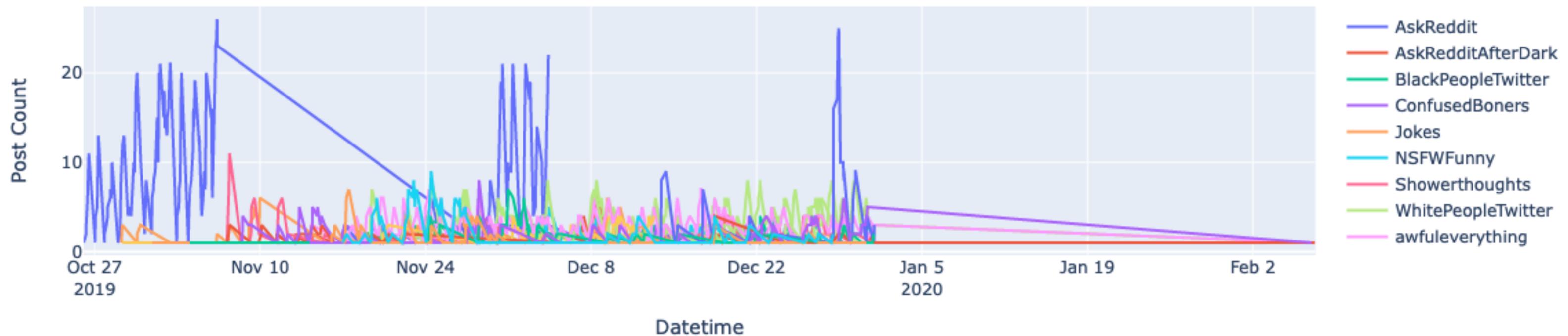
# REDDIT POSTS



# REAL-WORLD MODELING

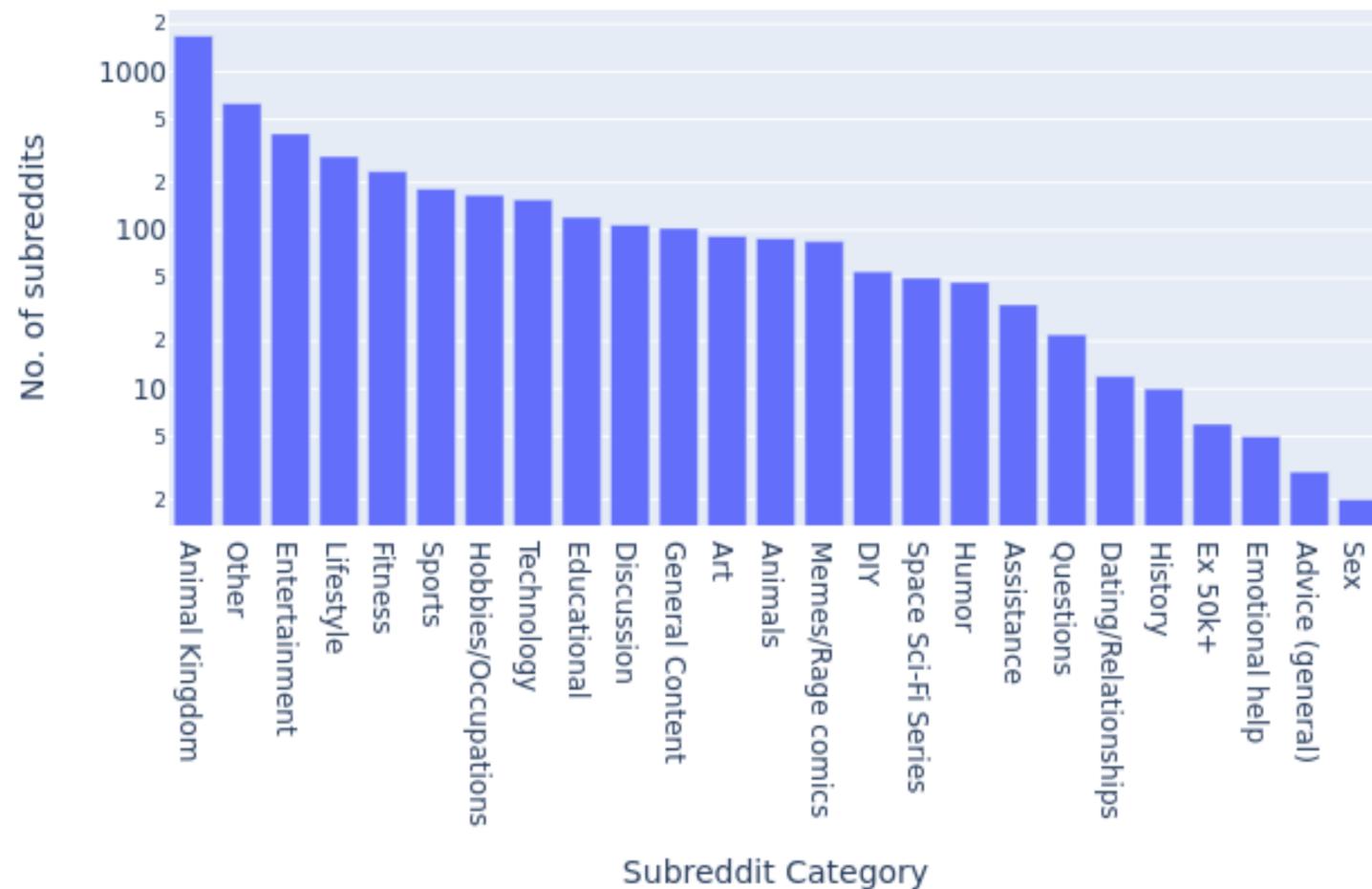
REDDIT USER 'JONNYCREEPYCREPES3'

Subreddit-wise Post Count in 12H bins by jonnycreepycrepes3



# CATEGORIZE POSTS ACROSS SUBREDDITS

Sample of unique subreddit categories ordered by no. of subreddits belonging



['mormon', 'politics']

## Similar Subreddits:

['mormonhistory', 'ldshistory', 'christianhistory', 'jewishhistory', 'historicalreligion']

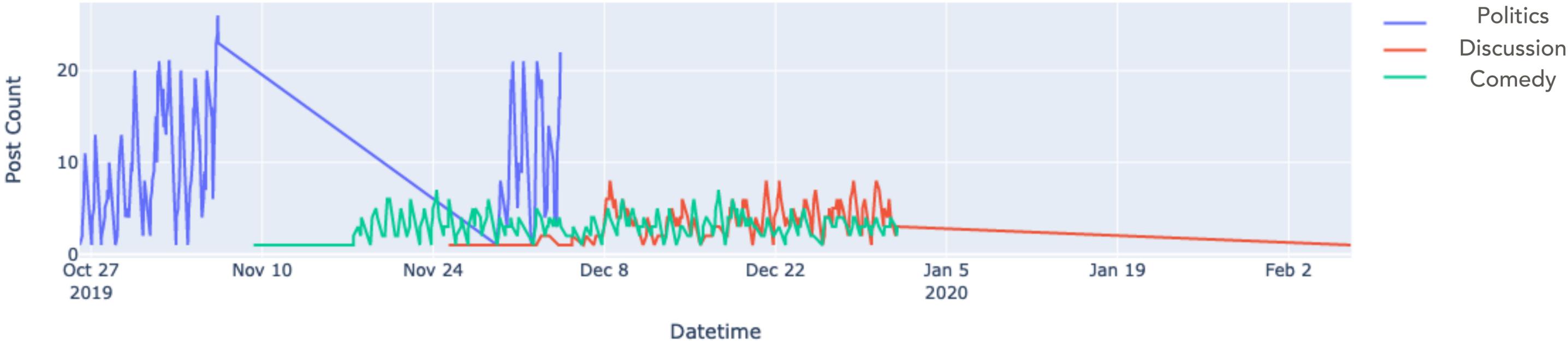
## Predicted Category:

[[None, 'History of People', None, None, None],

# MODELING CATEGORIES

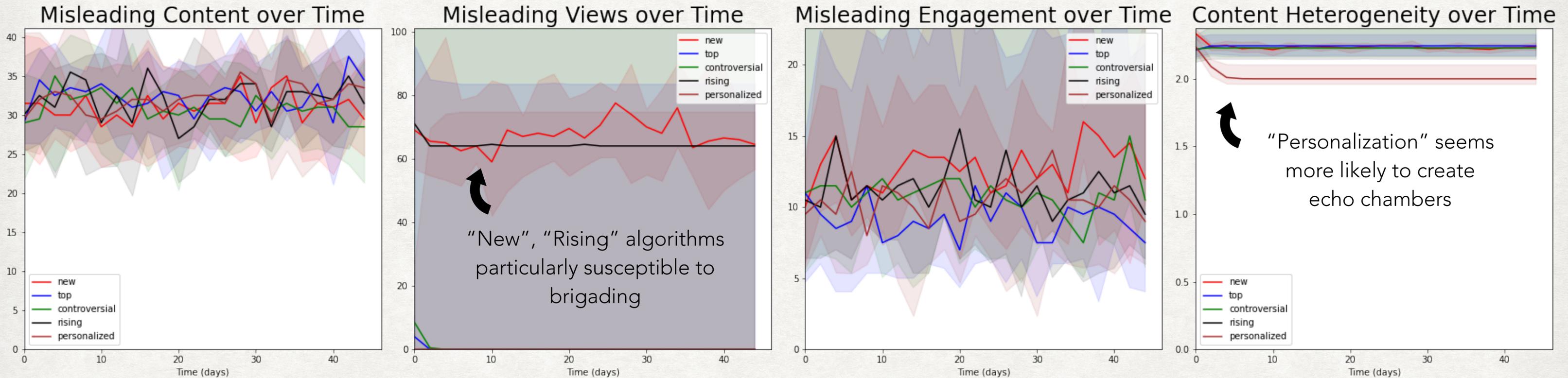
## REDUCING SUBREDDIT DIMENSIONALITY

Subreddit-wise Post Count in 12H bins by jonnycreepycrepes3



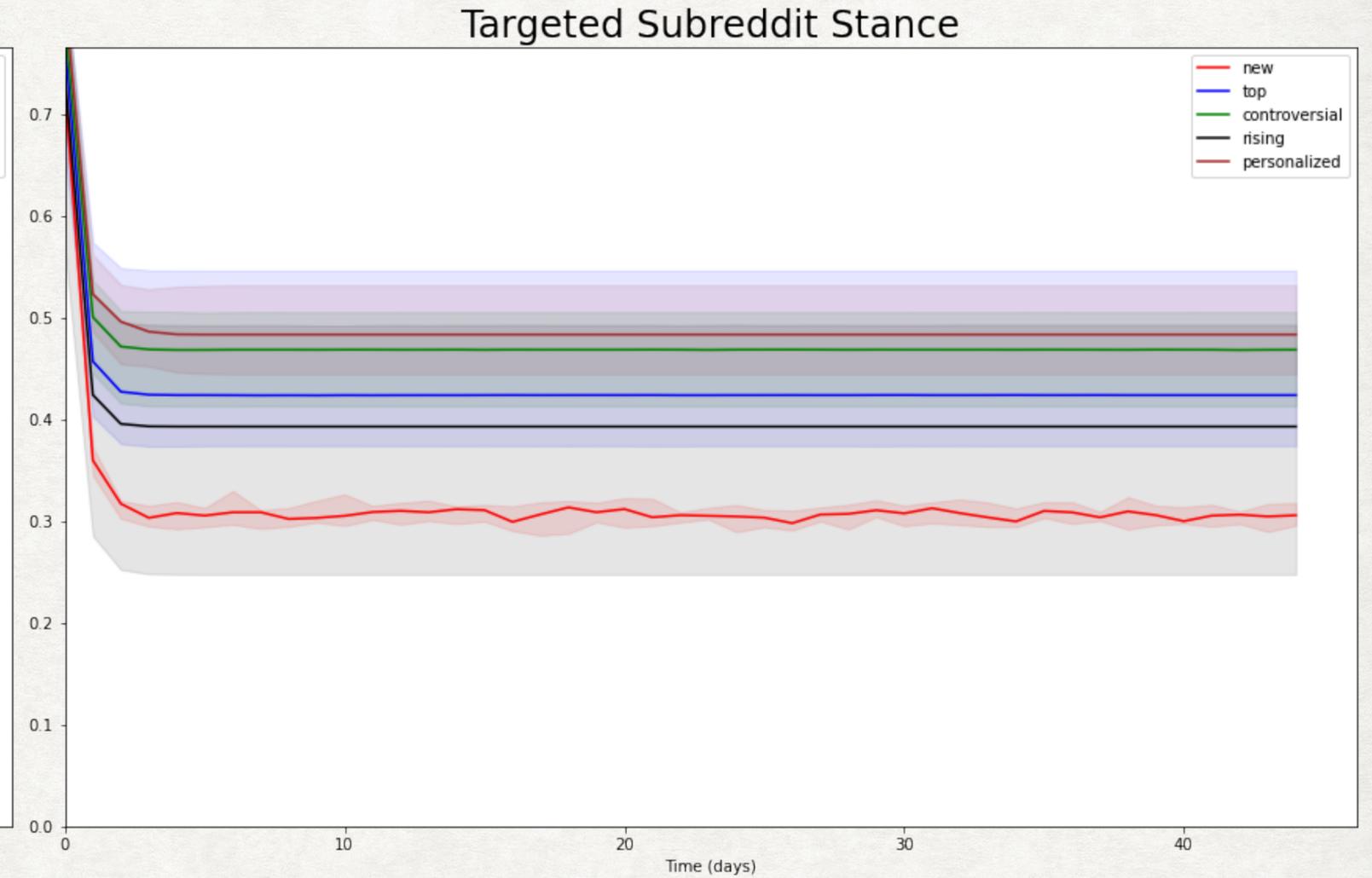
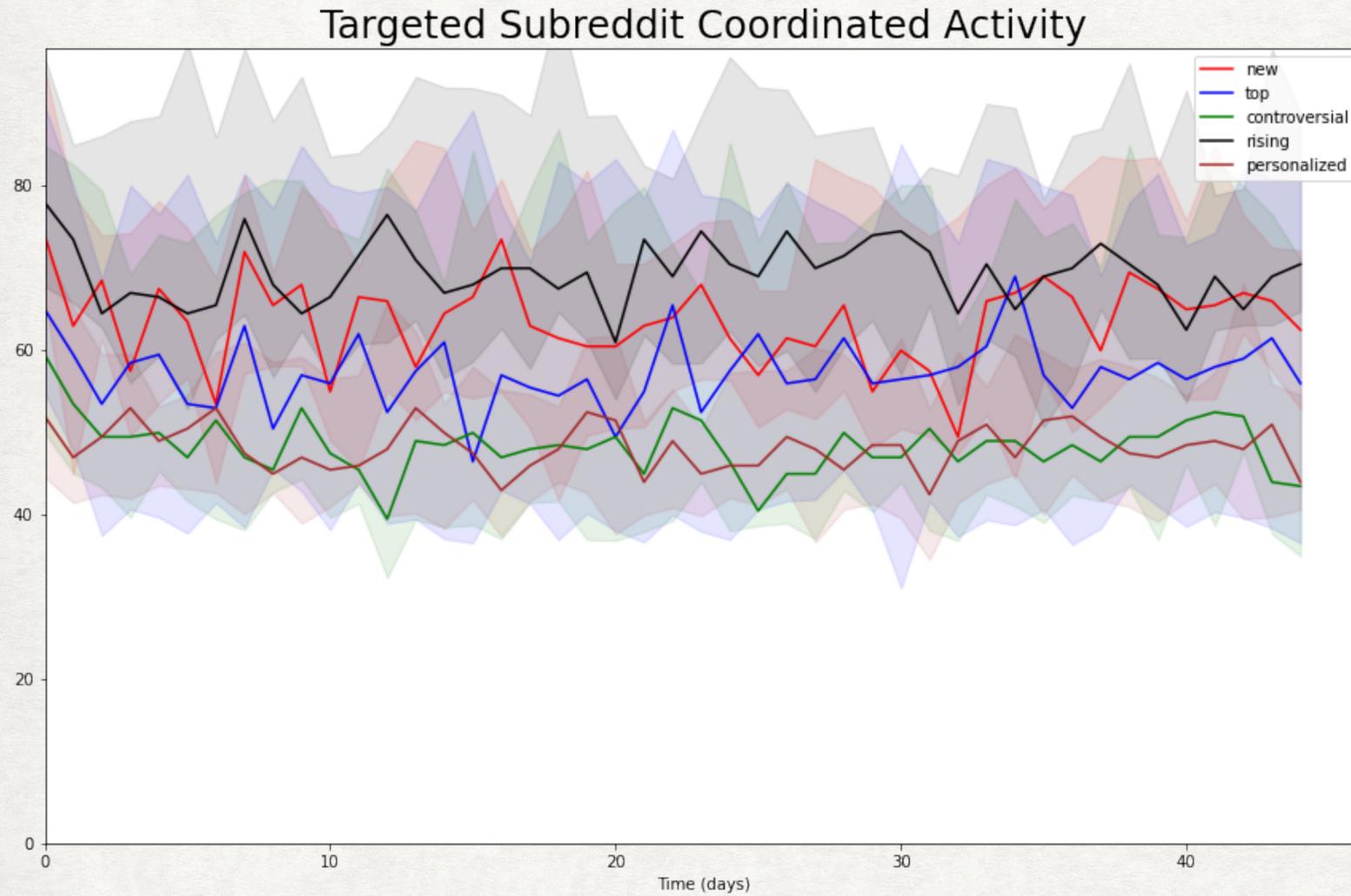
**HOW SUSCEPTIBLE ARE ALGORITHMS TO  
COORDINATED INAUTHENTIC BEHAVIOR?**

# 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



## Your timeline is set to Home



Switch to latest Tweets

Latest Tweets show up as they happen.



View content preferences

Cancel

# 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

Twitter no longer lets users access the chronological timeline by default [U: Rolled Back]

Filipe Espósito - Mar. 14th 2022 12:00 pm PT [@filipeesposito](#)

## Facebook now feeds up chronologically displays posts from your friends and groups

Aisha Malik @aishamalik

## Here's How to Switch Your Instagram Back to Chronological Order

It's a great way to see posts from people you actually follow, instead of posts from ads and "suggested" accounts.



BY [TUCKER BOWE](#) UPDATED: JUL 4, 2022

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



set to

Switch to latest Tweets

show up as they happen.

preferences

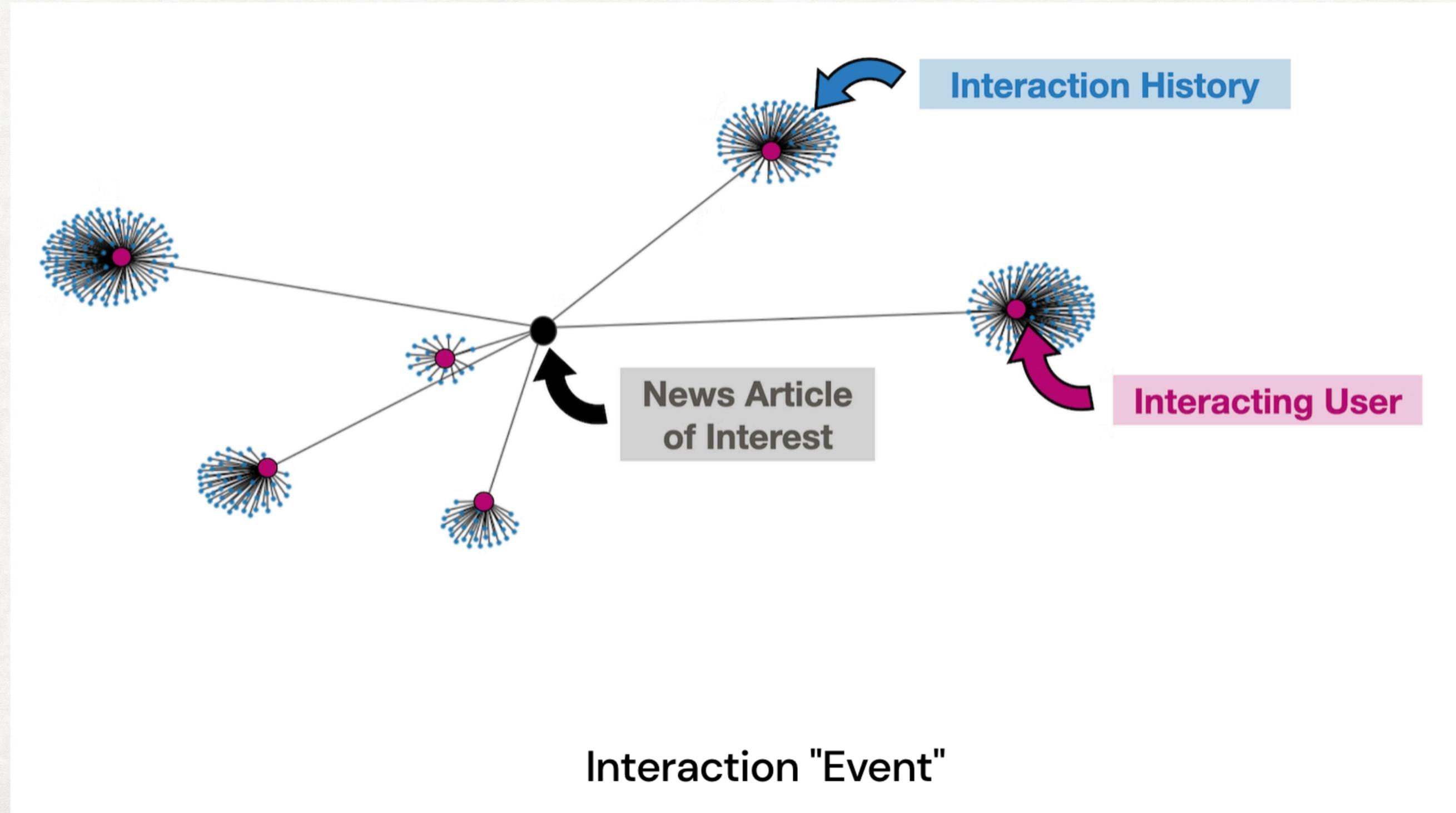
cancel

# 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

r/VTGuns

9

r/VoteDEM

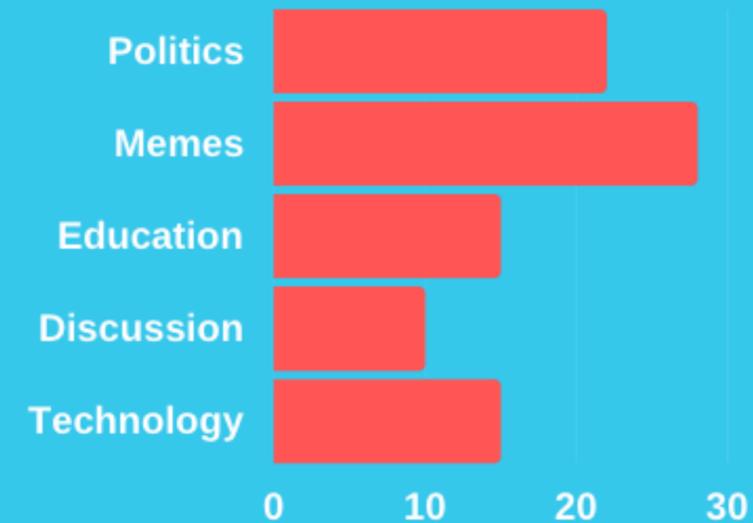
47<sup>↑</sup>

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

r/CRT\_so\_scary

100

r/Enough\_Sanders\_Spam

1400<sup>↑</sup>

r/Residency

sevendaysvt.com

# HELPING LOCAL NEWS UNDERSTAND AUDIENCES



## You're Getting Favorites!

4

The CDC recommends that people in high-level ...

5

With the Senate Divided 50-50 and Republicans united against ...

20↑

"The current Act 250 bill would actually make ...

## You're Getting Traffic From



## Engaged Users also Follow

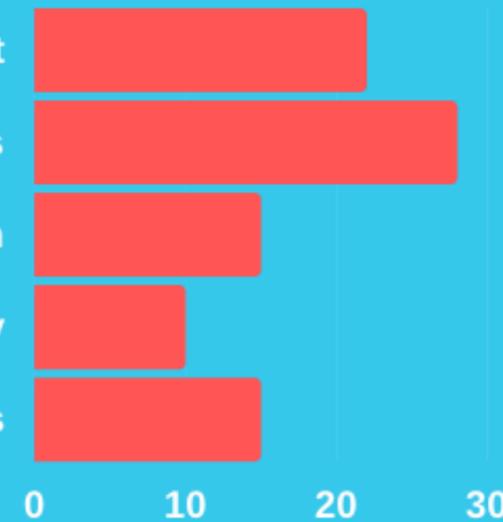
@sevendaysvt

@calmatters

Tech

Silicon Valley

Politics



## Related Engagement

14

Politics

23

Elections

86 ↑

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...](https://vtdigger.org/2022/0...)

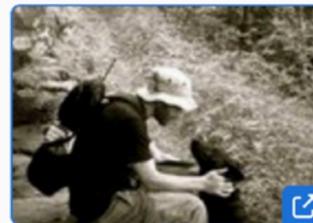


9 upvotes 2 comments 0 awards

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

Vermont game wardens investigate fatal dog shooting in Tunbridge - VTDigger

[vtdigger.org/2022/0...](https://vtdigger.org/2022/0...)



0 upvotes 0 comments 0 awards

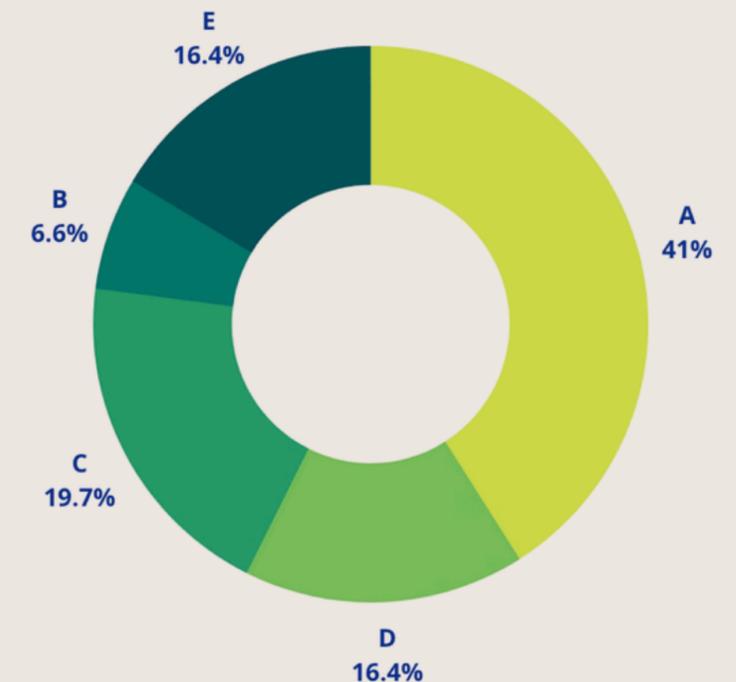
 **r/VTGuns** · Posted by u/PROPAGANDA-DESTROYER 8 days ago

<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...](https://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...](https://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...](https://vtdigger.org/2022/0...)



47 upvotes 2 comments 0 awards

# DEMO

<https://simppl.org>

The screenshot displays the SimPPL dashboard interface. On the left is a dark sidebar with navigation options: Projection Dash, Audience Dashboard, Projections, Engagement, Messages, External Data, Settings, Product Page, and Support. At the bottom of the sidebar is an 'Update Projections' button. The main content area is divided into several sections:

- Twitter Metrics:** A table with columns for Tweet Link, Tweet Comments, Tweet Favorites, and TWI. It lists five tweets with their respective engagement metrics and trend indicators.
- Total Engagement:** A summary card showing 452 total engagements, with a 18.2% increase. A bar chart compares engagement from Reddit (red) and Twitter (blue) across the days of the week (Mon-Sat).
- User Profiles:** A list of four users: Chris Wood (Active), Jose Leos (Inactive), Bonnie Green (Disengaged), and Neil Sims (Disengaged). Each user has a 'Profile' or 'History' button.
- Track Projections:** A list of four projection goals with progress bars: 'Improve Twitter Engagement' (75%), 'Expand into New Subreddit' (60%), 'Reduce Bounce Rate' (45%), and 'Engage across Party Lines' (34%).
- Key Performance Indicators:** Three cards showing 'Unique Visitors' (2,200), 'Country Rank' (United States, #150), and 'Category Rank' (Media > Local News, #11).
- Acquisition:** A section explaining visitor sources, featuring a 'Bounce Rate' of 12.88% and a 'Settings' button.

# 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



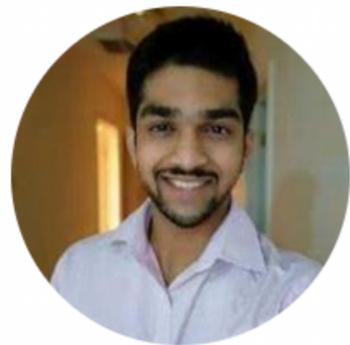
Atılım Güneş Baydin



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!

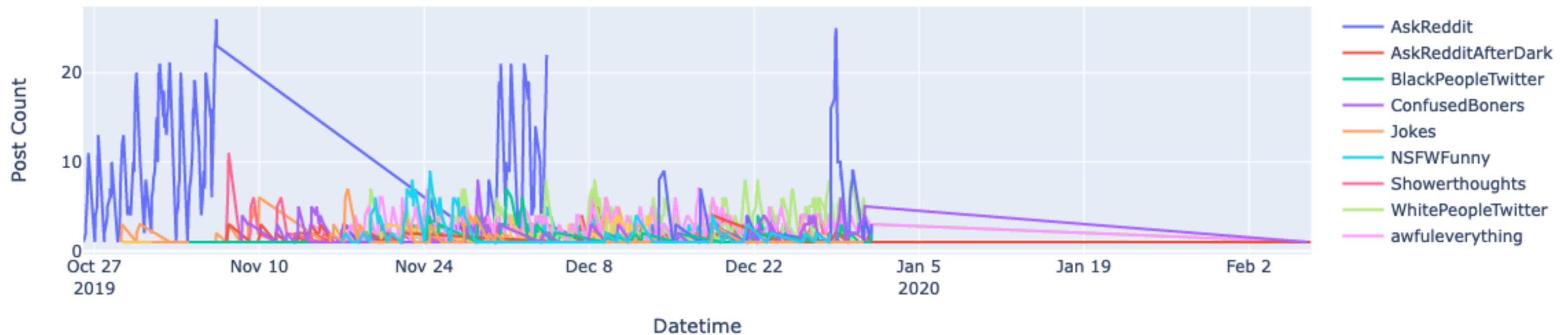
[@swapneel\\_mehta](#)  
[swapneelm.github.io](#)  
[swapneel.mehta@nyu.edu](mailto:swapneel.mehta@nyu.edu)

Shoutout to [ai4abm.org](#)

# REAL-WORLD MODELING

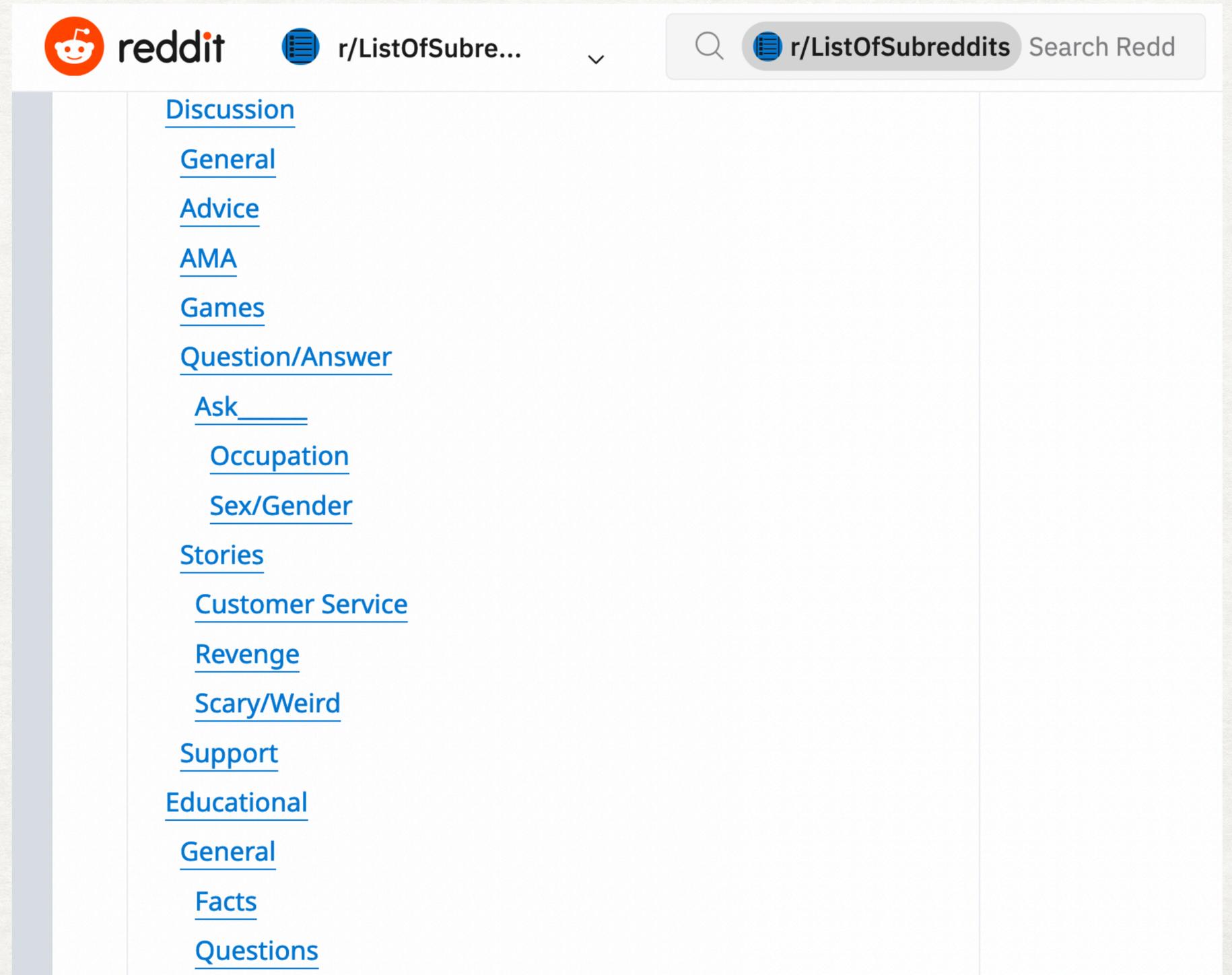
REDDIT USER 'JONNYCREEPYCREPES3'

Subreddit-wise Post Count in 12H bins by jonnycreepycrepes3



# REDDIT WIKI PAGES

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

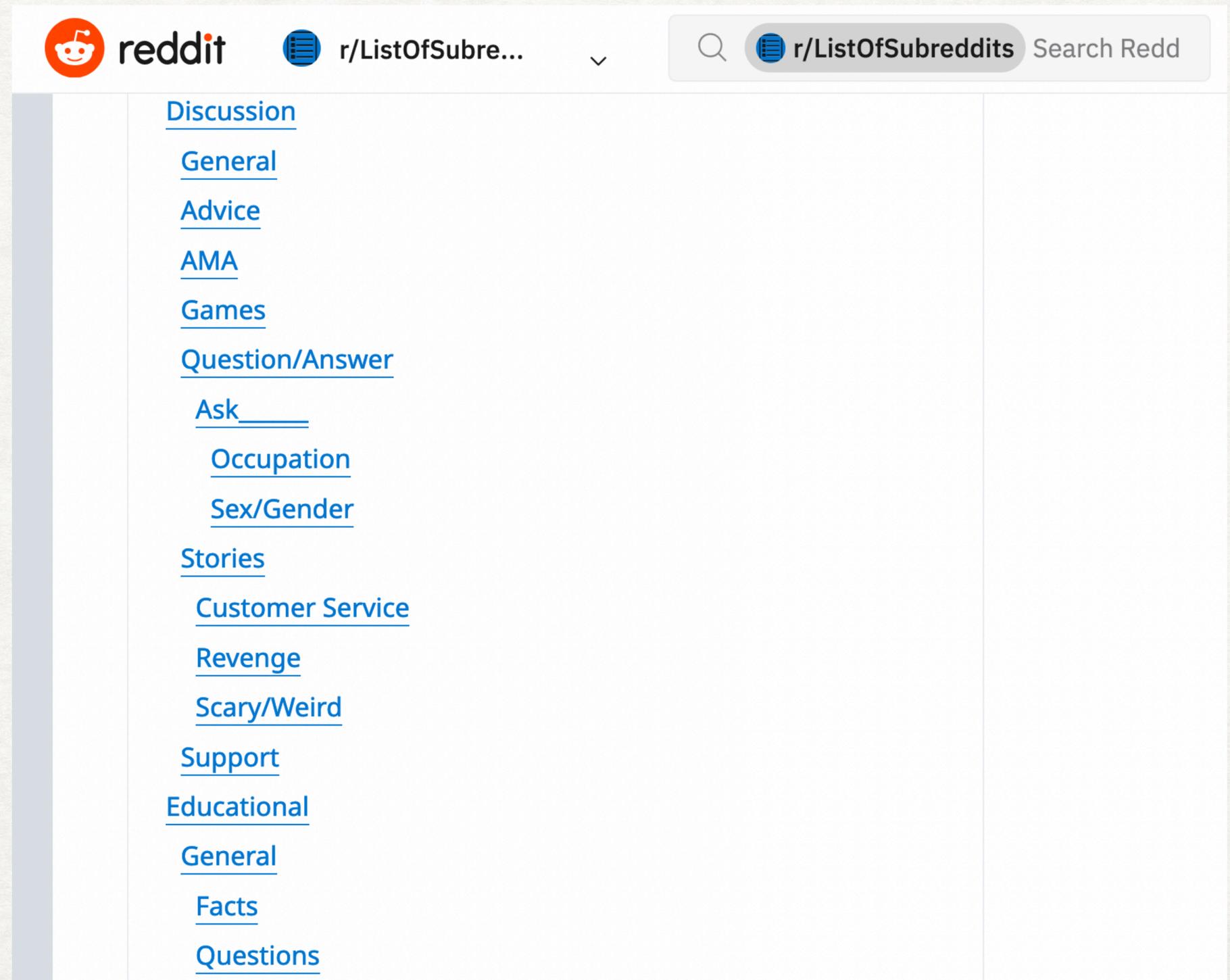


The screenshot shows the top navigation bar of the Reddit website. On the left is the Reddit logo and the word "reddit". In the center is a dropdown menu for the subreddit "r/ListOfSubre...". On the right is a search bar with the text "r/ListOfSubreddits" and "Search Redd". Below the navigation bar is a list of categories, each with a blue underline:

- [Discussion](#)
- [General](#)
- [Advice](#)
- [AMA](#)
- [Games](#)
- [Question/Answer](#)
- [Ask\\_\\_\\_\\_\\_](#)
- [Occupation](#)
- [Sex/Gender](#)
- [Stories](#)
- [Customer Service](#)
- [Revenge](#)
- [Scary/Weird](#)
- [Support](#)
- [Educational](#)
- [General](#)
- [Facts](#)
- [Questions](#)

# REDDIT WIKI PAGES

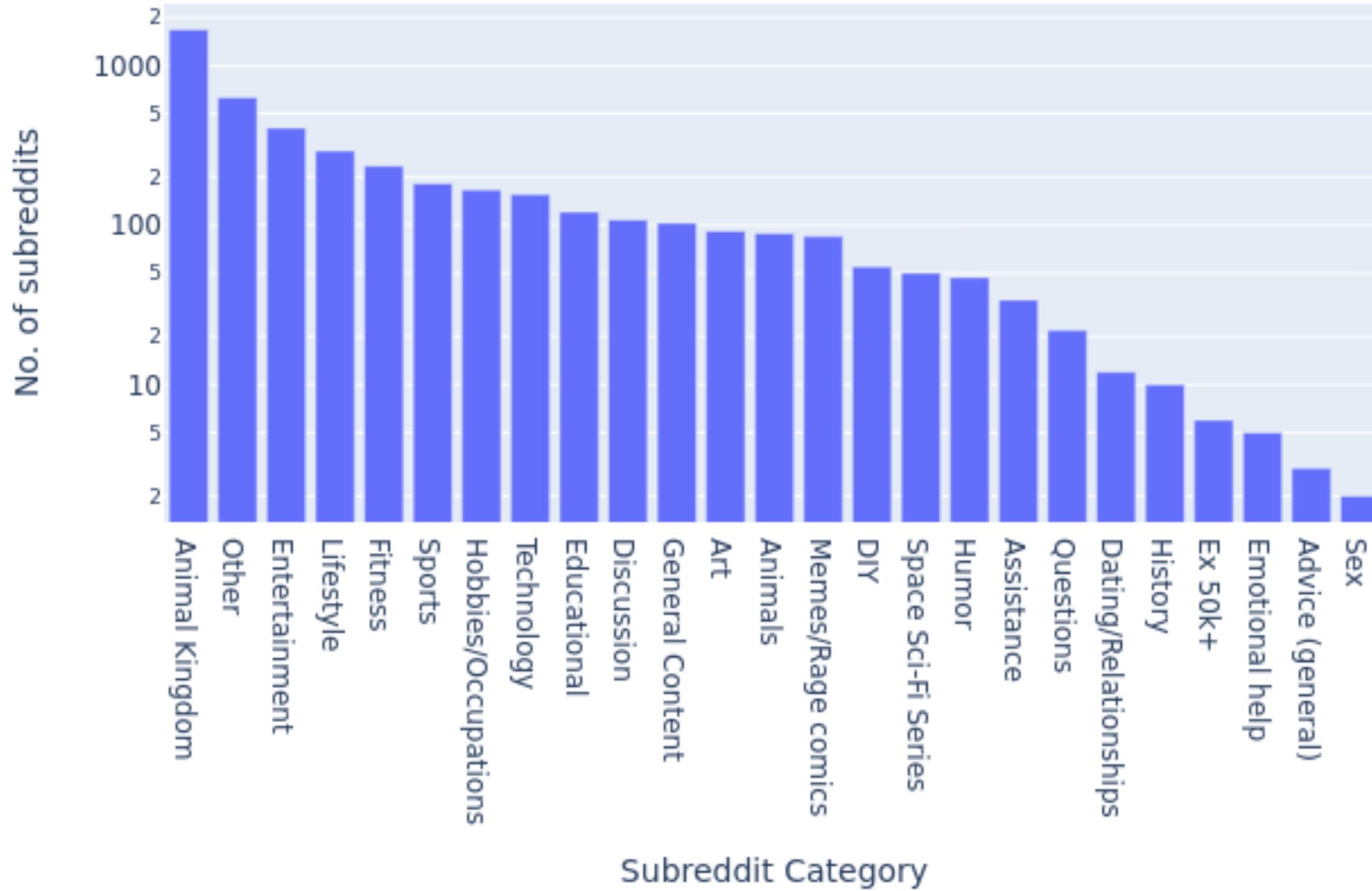
- User-driven hierarchical categorization of subreddits
- Discussion > Stories > Customer Service
- 4998 subreddits
- 5-level hierarchy of categories
- For each new subreddit
  - Split words > match embeddings > associate category



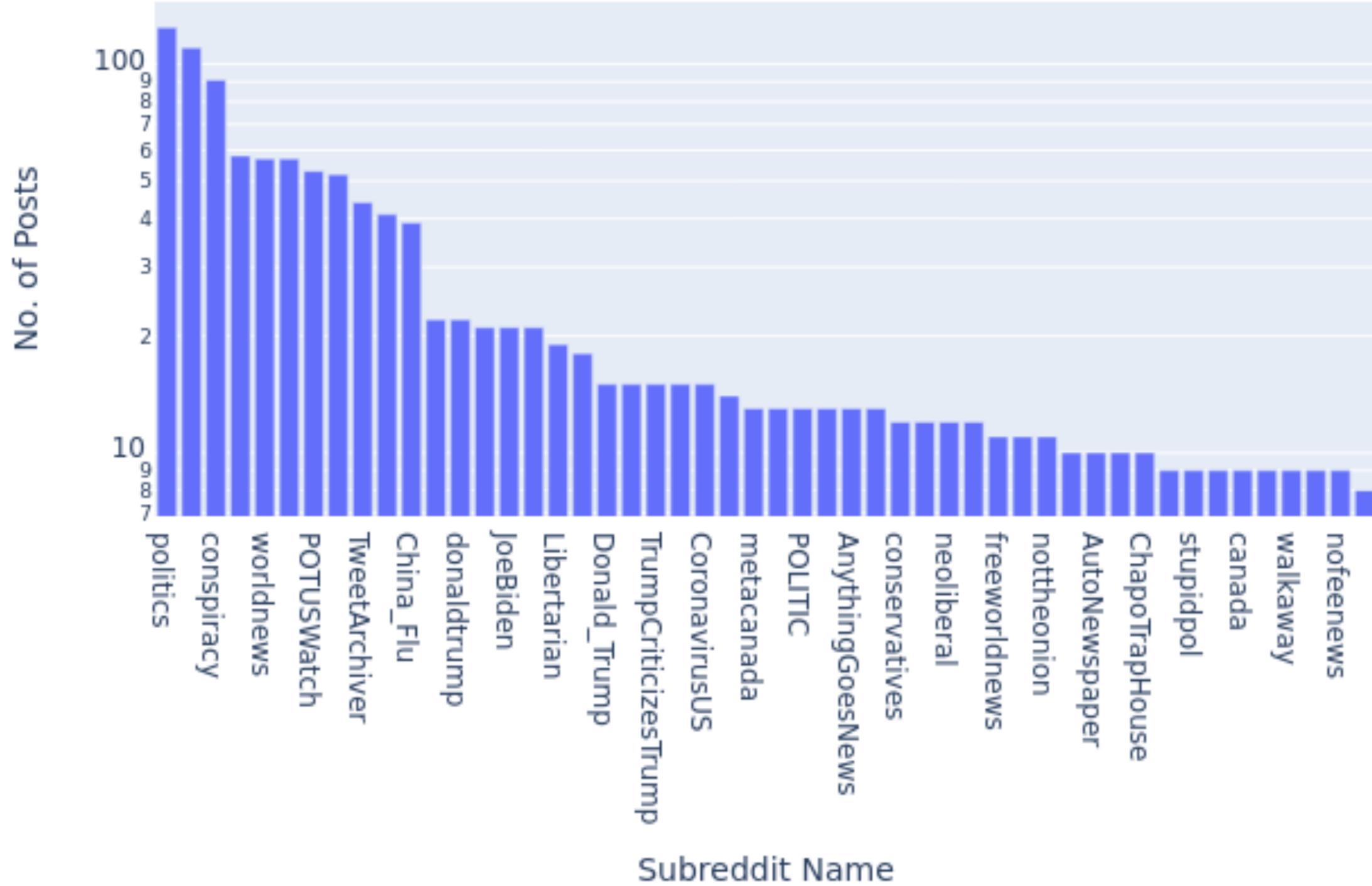
The screenshot shows the top navigation bar of the Reddit website. On the left, there is the Reddit logo and the word "reddit". In the center, there is a dropdown menu for the current subreddit, "r/ListOfSubre...". On the right, there is a search bar with the text "r/ListOfSubreddits" and "Search Redd". Below the navigation bar, a list of categories is displayed, each with a blue underline:

- [Discussion](#)
- [General](#)
- [Advice](#)
- [AMA](#)
- [Games](#)
- [Question/Answer](#)
- [Ask\\_\\_\\_\\_\\_](#)
- [Occupation](#)
- [Sex/Gender](#)
- [Stories](#)
- [Customer Service](#)
- [Revenge](#)
- [Scary/Weird](#)
- [Support](#)
- [Educational](#)
- [General](#)
- [Facts](#)
- [Questions](#)

# CATEGORIZE POSTS ACROSS SUBREDDITS

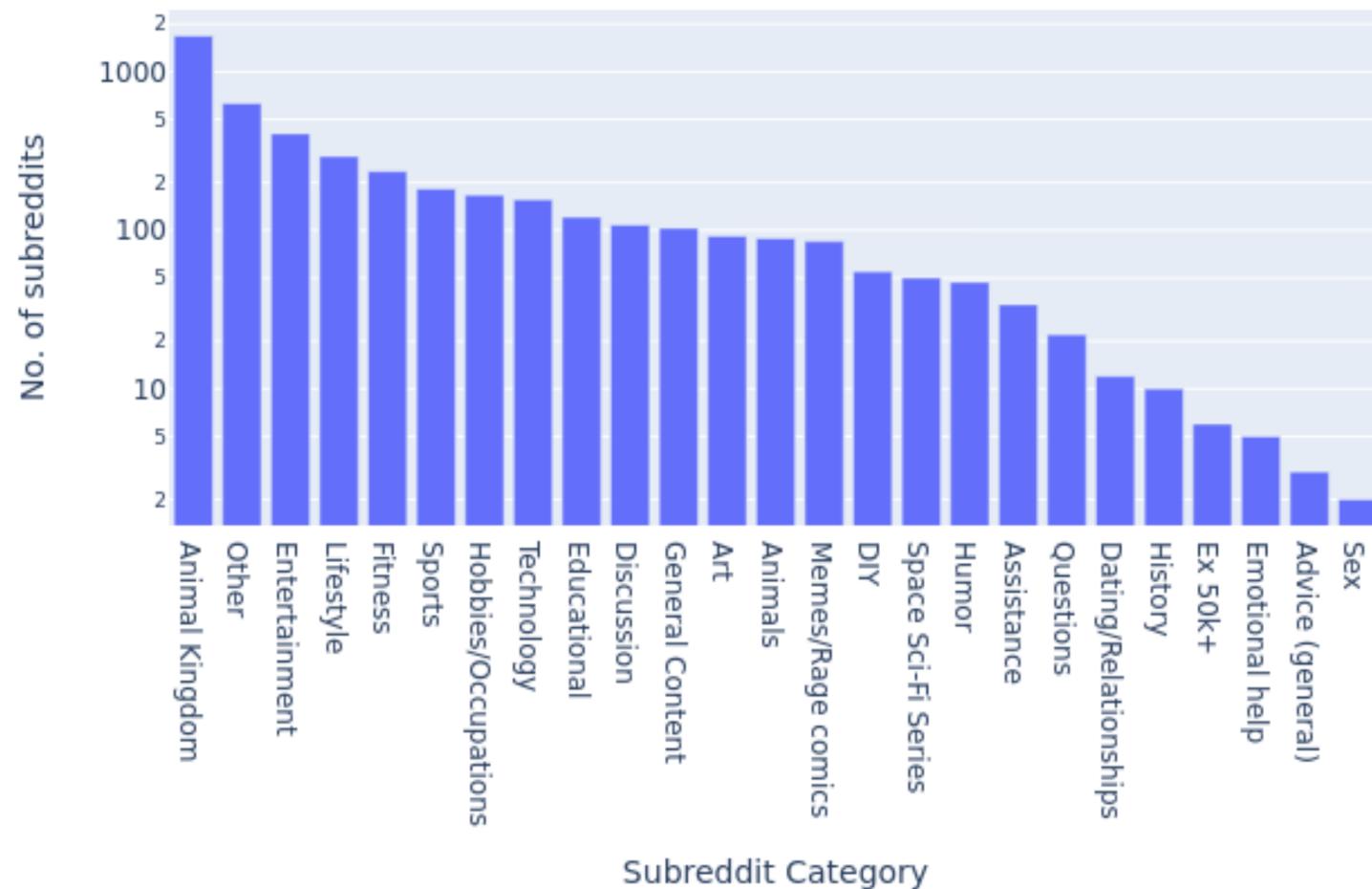


# CATEGORIZE POSTS ACROSS SUBREDDITS



# CATEGORIZE POSTS ACROSS SUBREDDITS

Sample of unique subreddit categories ordered by no. of subreddits belonging



['mormon', 'politics']

## Predicted Subreddits:

['mormonhistory', 'ldshistory', 'christianhistory', 'jewishhistory', 'historicalreligion']

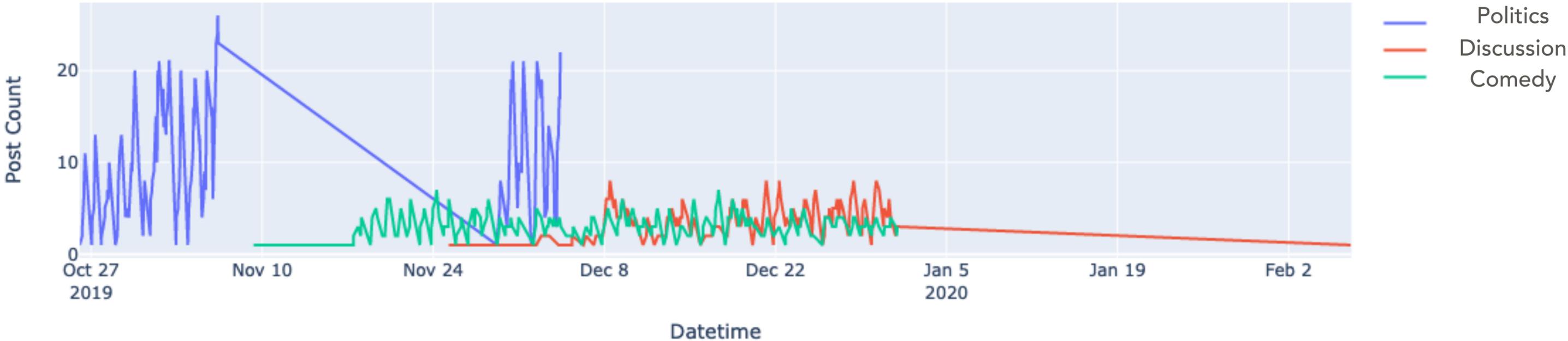
## Predicted Category:

[[None, 'History of People', None, None, None],

# MODELING CATEGORIES

## REDUCING SUBREDDIT DIMENSIONALITY

Subreddit-wise Post Count in 12H bins by jonnycreepycrepes3



# REAL-WORLD INTERVENTIONS

## 1. AGENT - LEVEL

Awareness campaigns, training, ideological change

## 2. NETWORK - LEVEL

Reduced sharing, visibility, confirmation of retweets

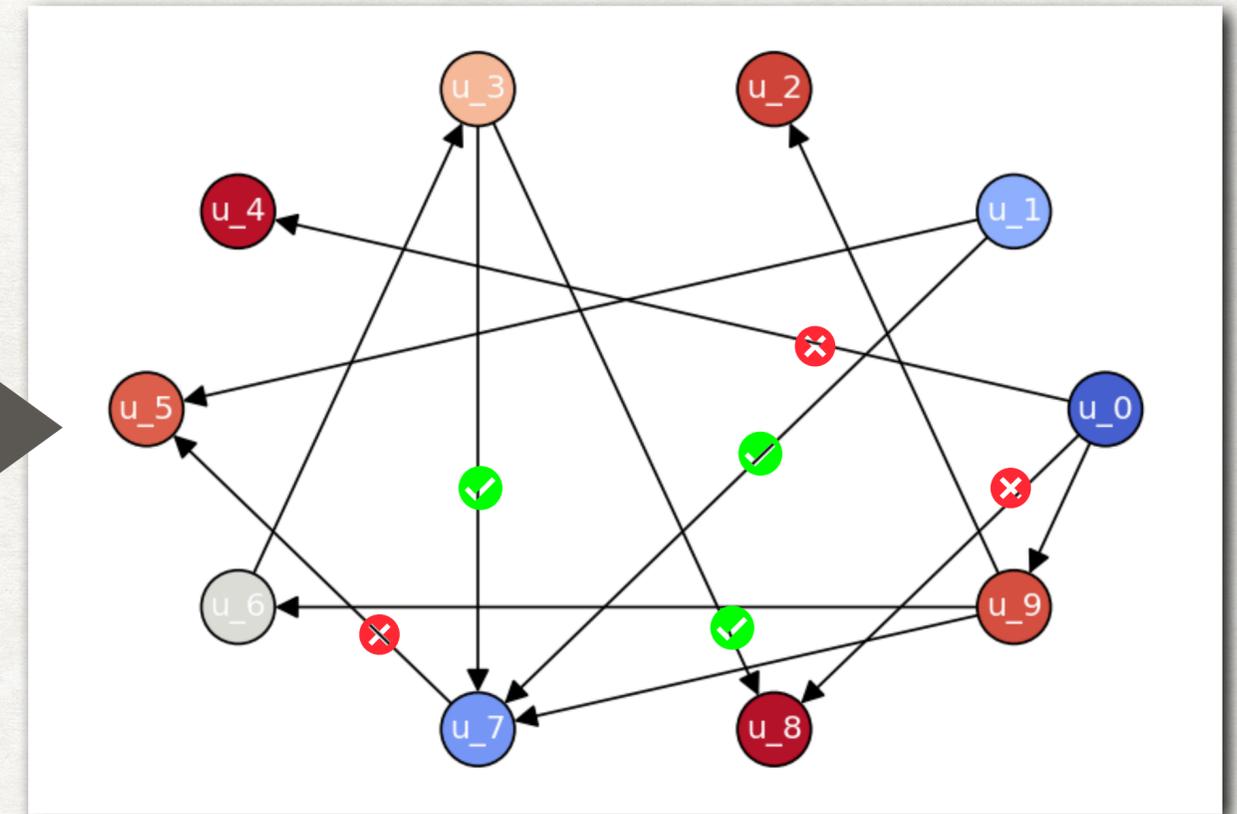
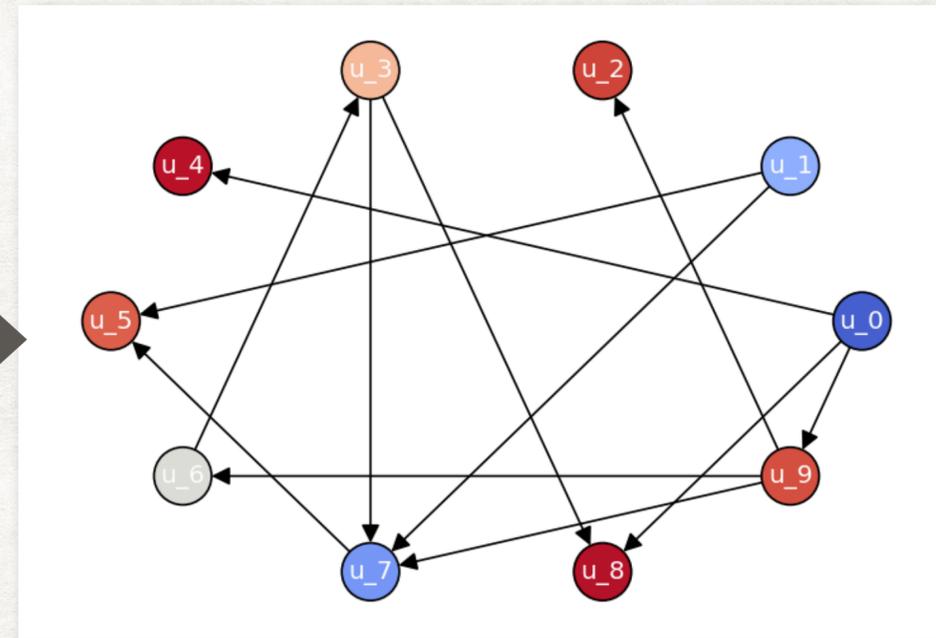
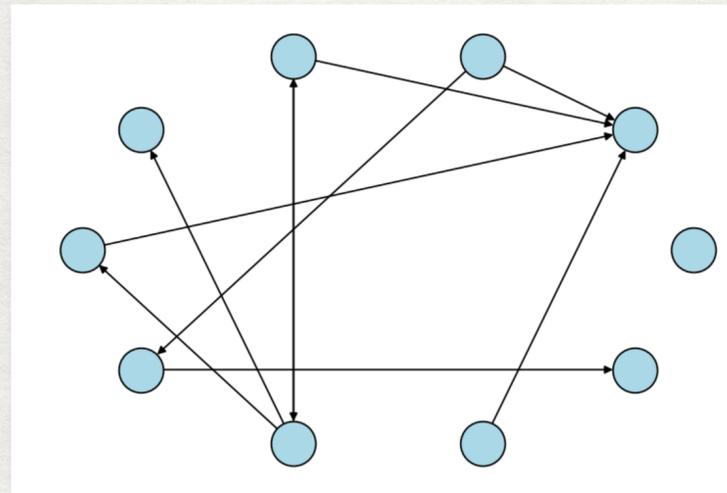
## 3. HYBRID

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