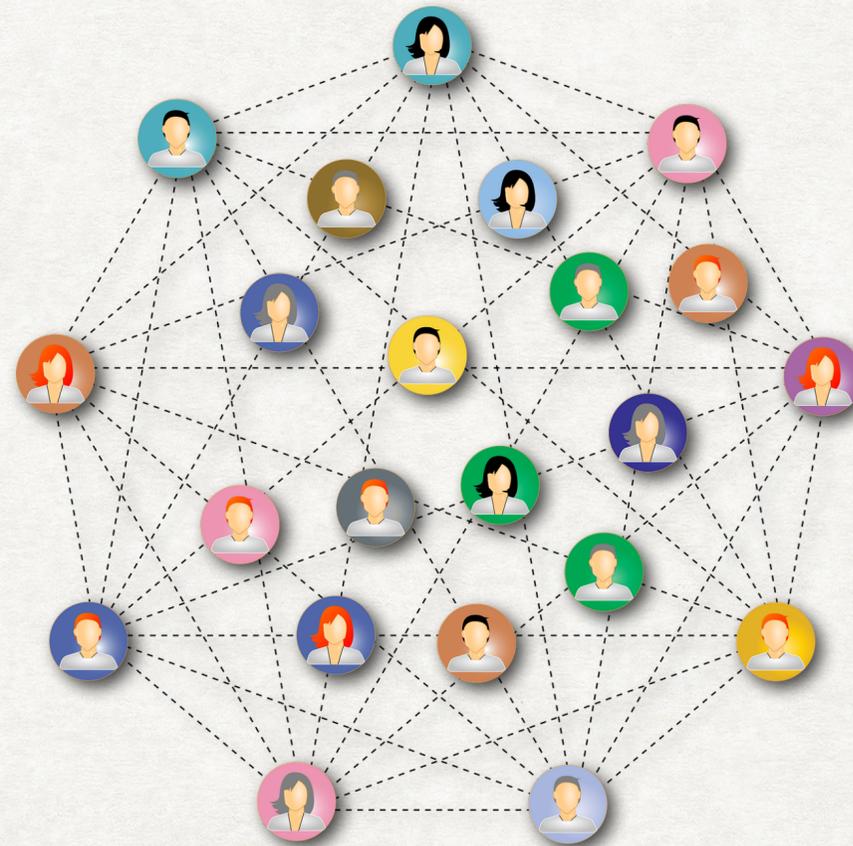


# USING SIMULATIONS TO STUDY ONLINE DISINFORMATION



MISINFOCON '22

**S. MEHTA, B. STATE, R. BONNEAU,  
J. NAGLER, P.H. TORR, A.G. BAYDIN**

# ABOUT ME



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

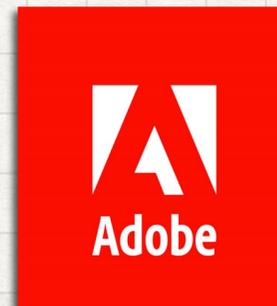
**2022** Ph.D. ML Engineering Intern at Twitter  
Civic Integrity, Misinformation

**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)



# 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

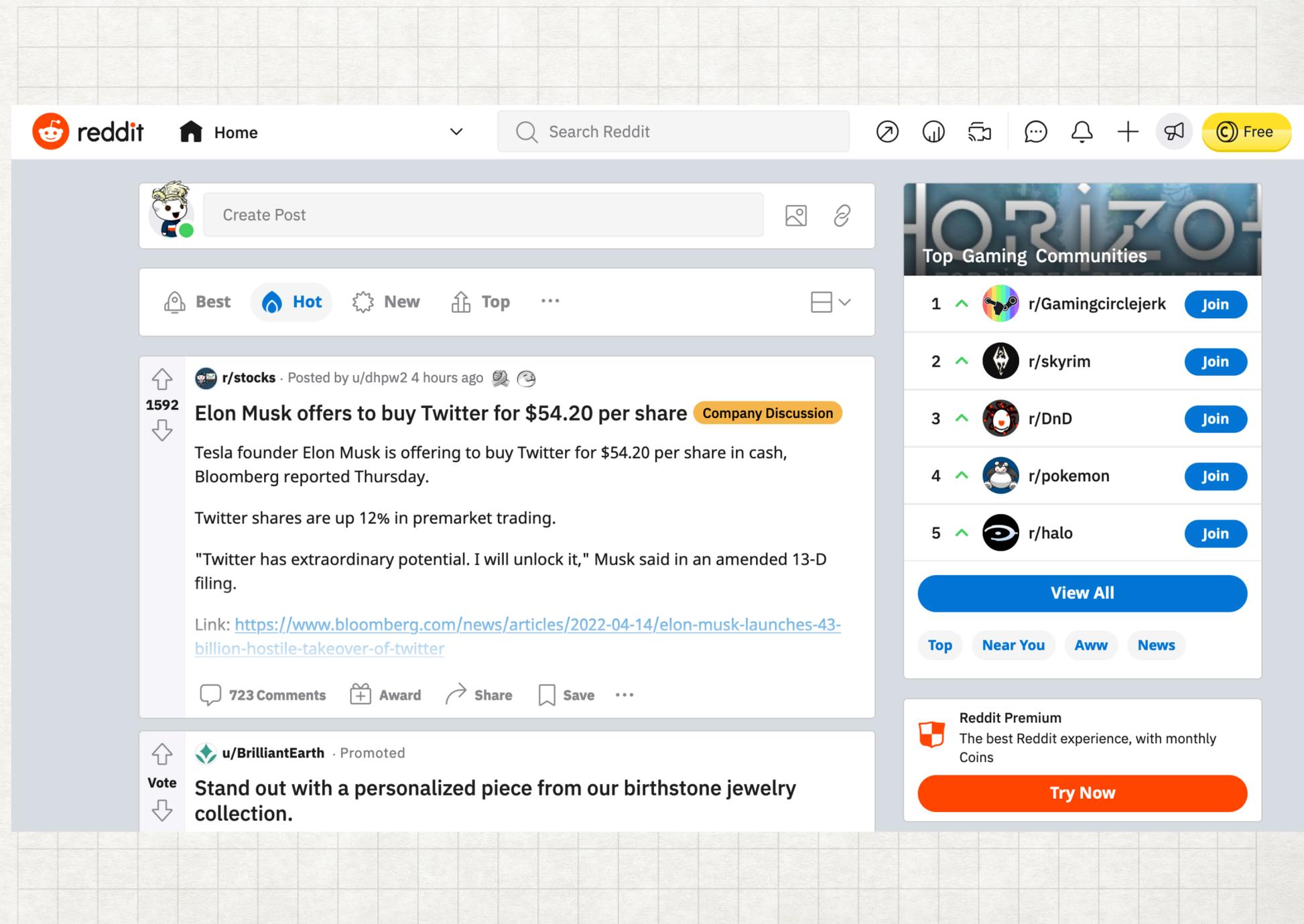
- RQ: Quantify the relative impact of different algorithmic choices that a platform makes, for instance:
  - To maximize engagement or not to maximize engagement?
  - Should we promote diverse content that isn't getting early views?
  - Dealing with "controversial" opinions?

# 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?

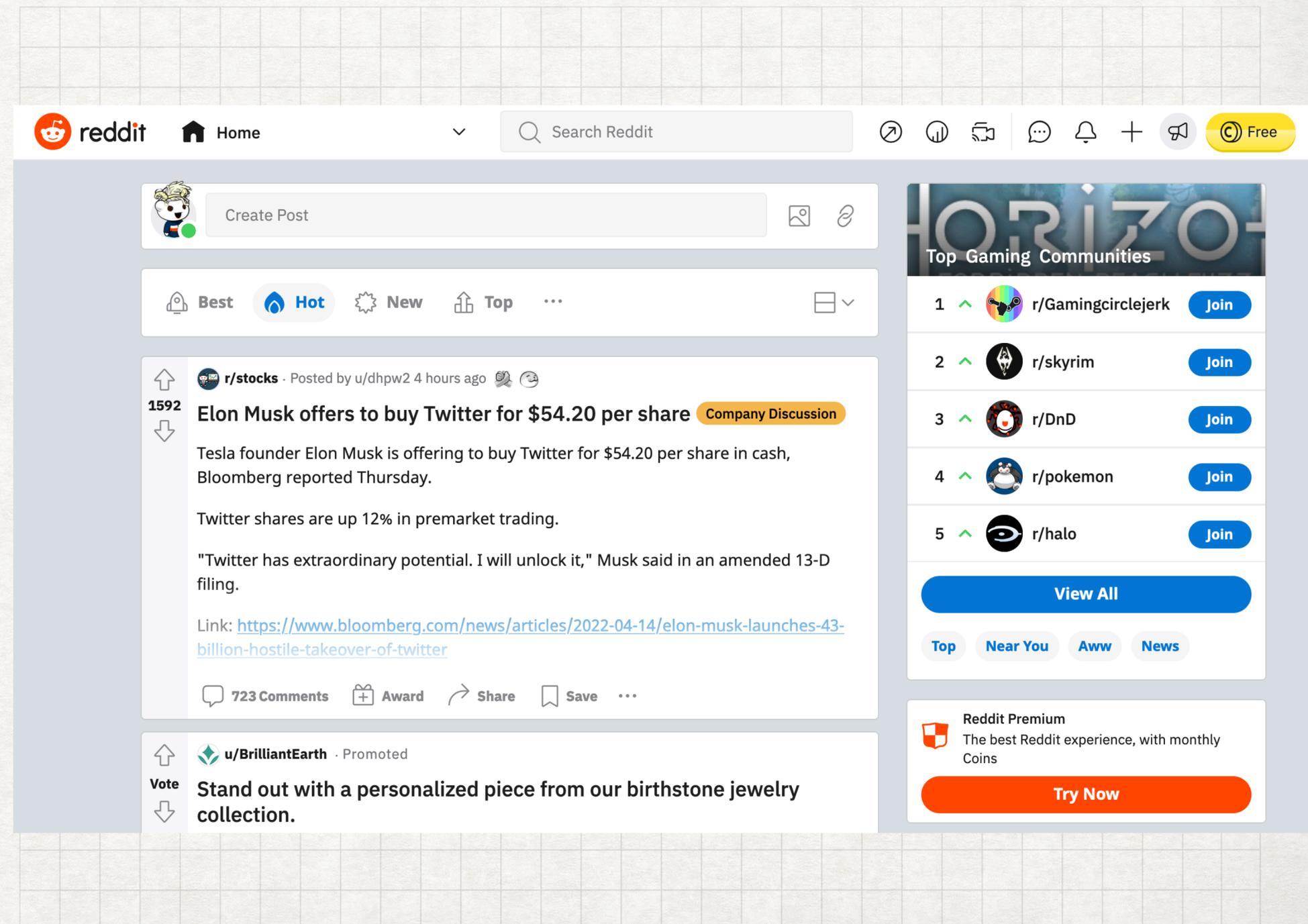
# SIMULATE A SOCIAL NETWORK

- 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

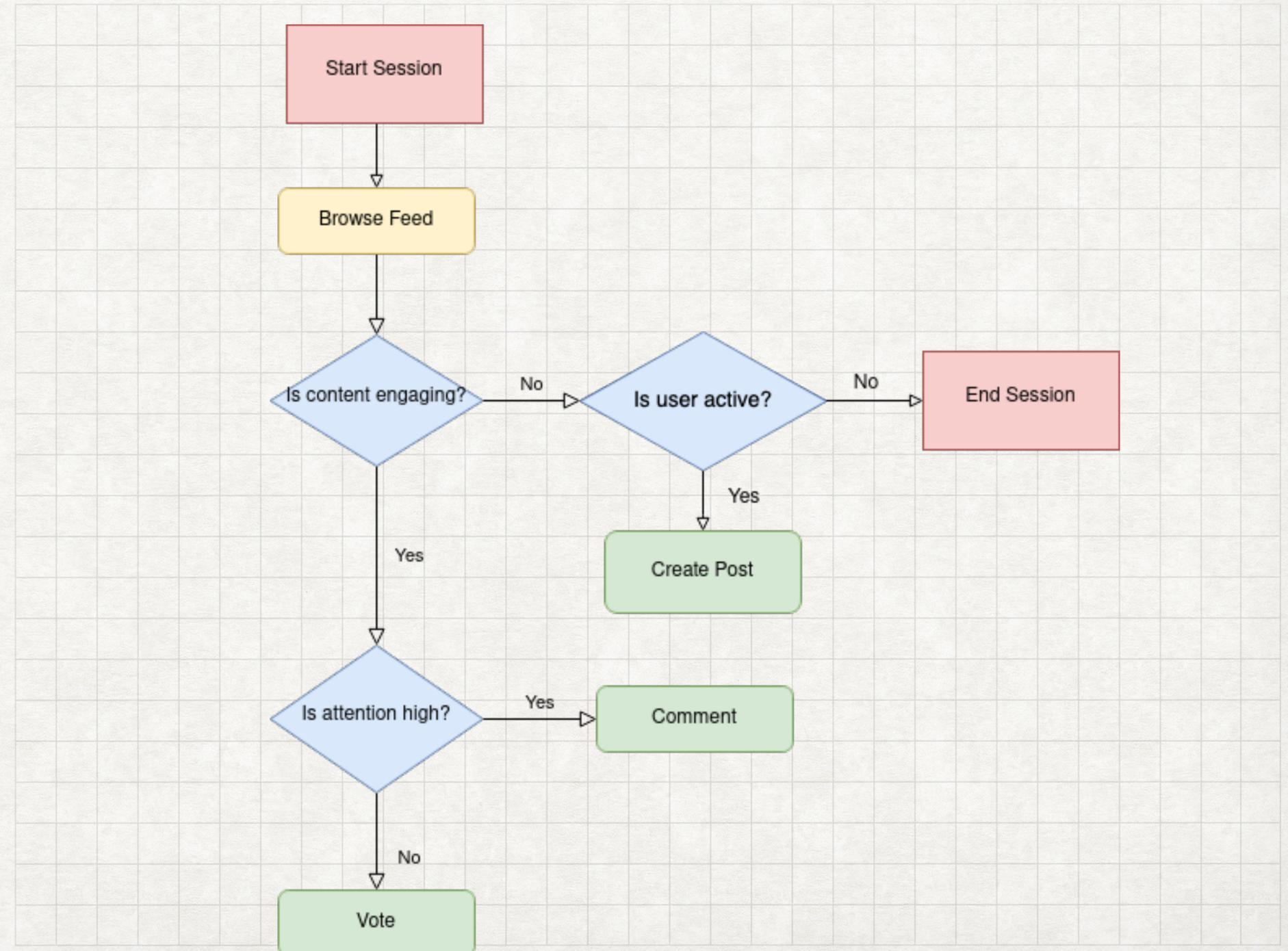
1: **procedure** REDDIT  
2: *Sample latents:*  
3:  $\mathbf{v} \leftarrow \{v_i \sim \text{Uniform}(0, 1)\}_{i=1, j=1}^{N, S}$  ▷ Interaction propensity over subreddit categories  
4: *Simulate:*  
5:  $\Phi_{1:N} \leftarrow \langle \rangle$  ▷ User Activity  
6: **for**  $t = 1 : T$  **do**  
7:     **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

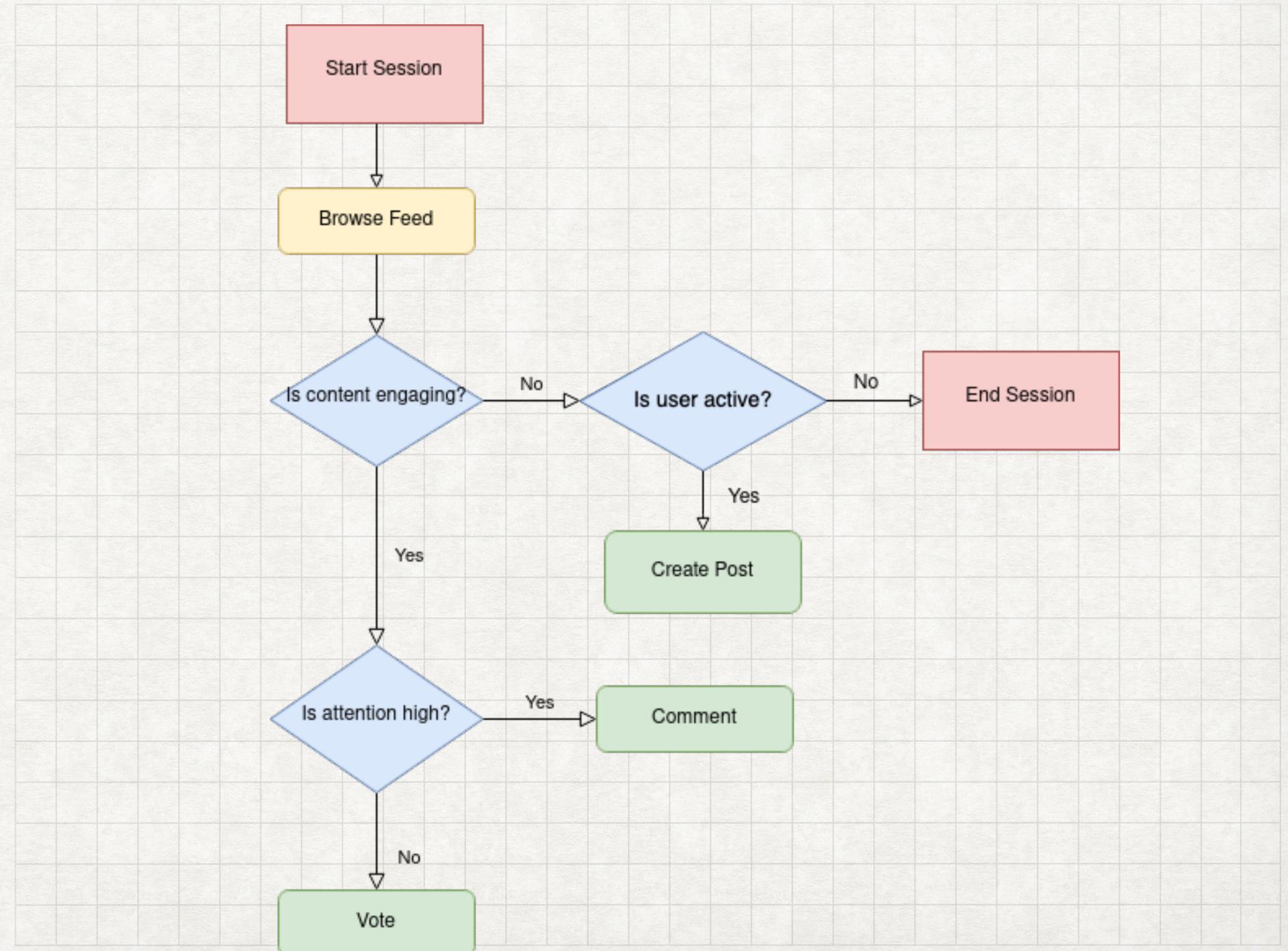
- Developing a model for a user's posting behavior
- Creating a story of how a user interacts with Reddit



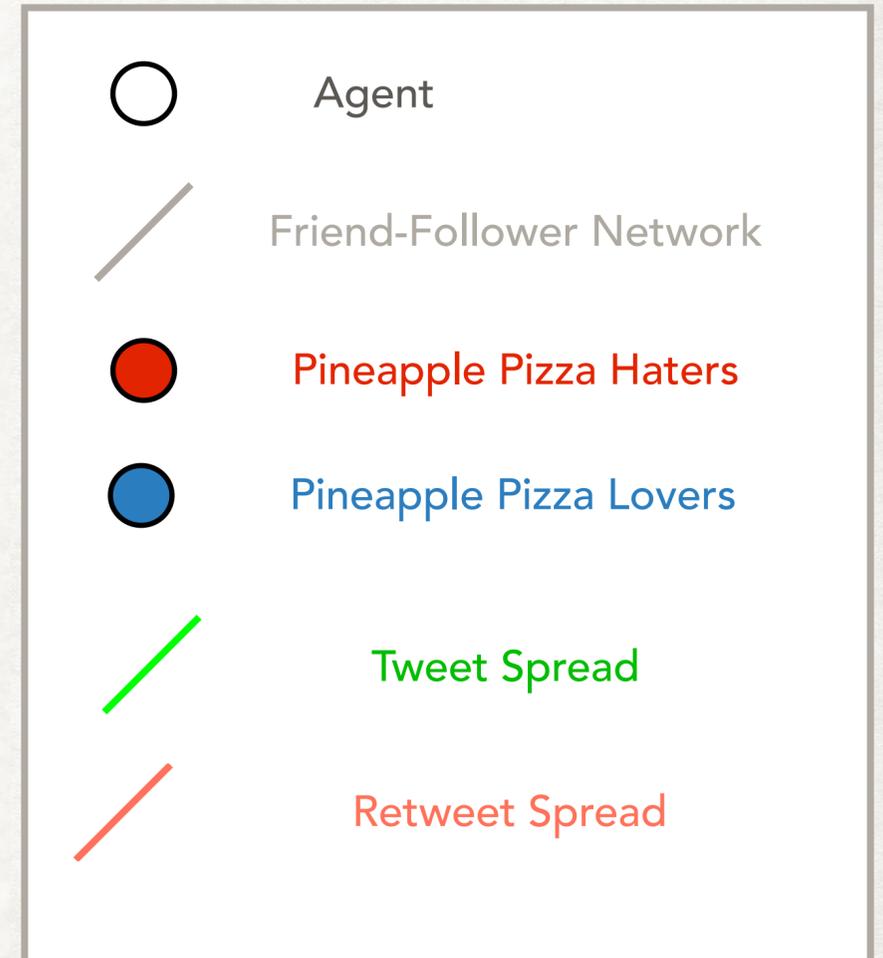
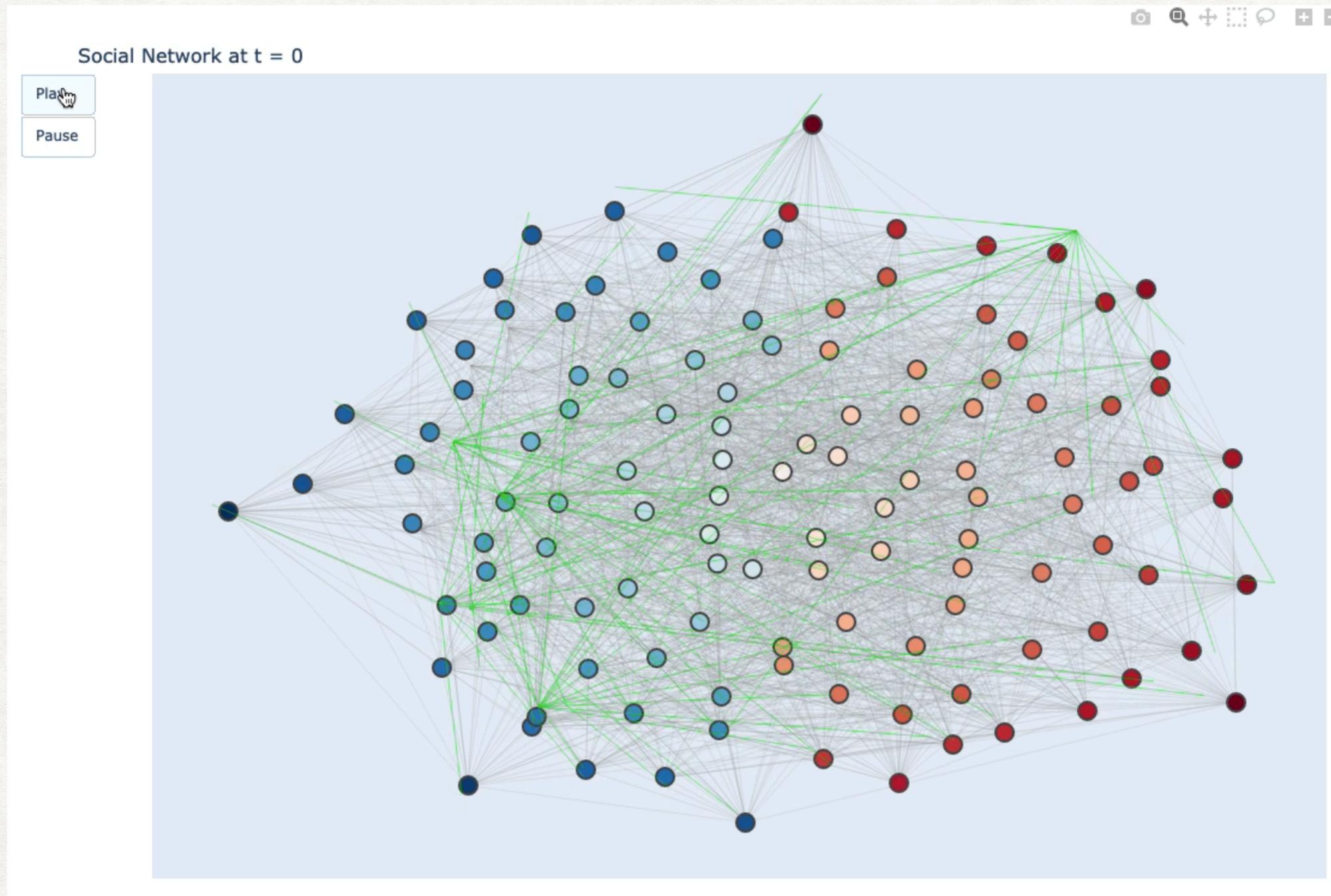
# SIMULATING SOCIAL NETWORKS

## 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!

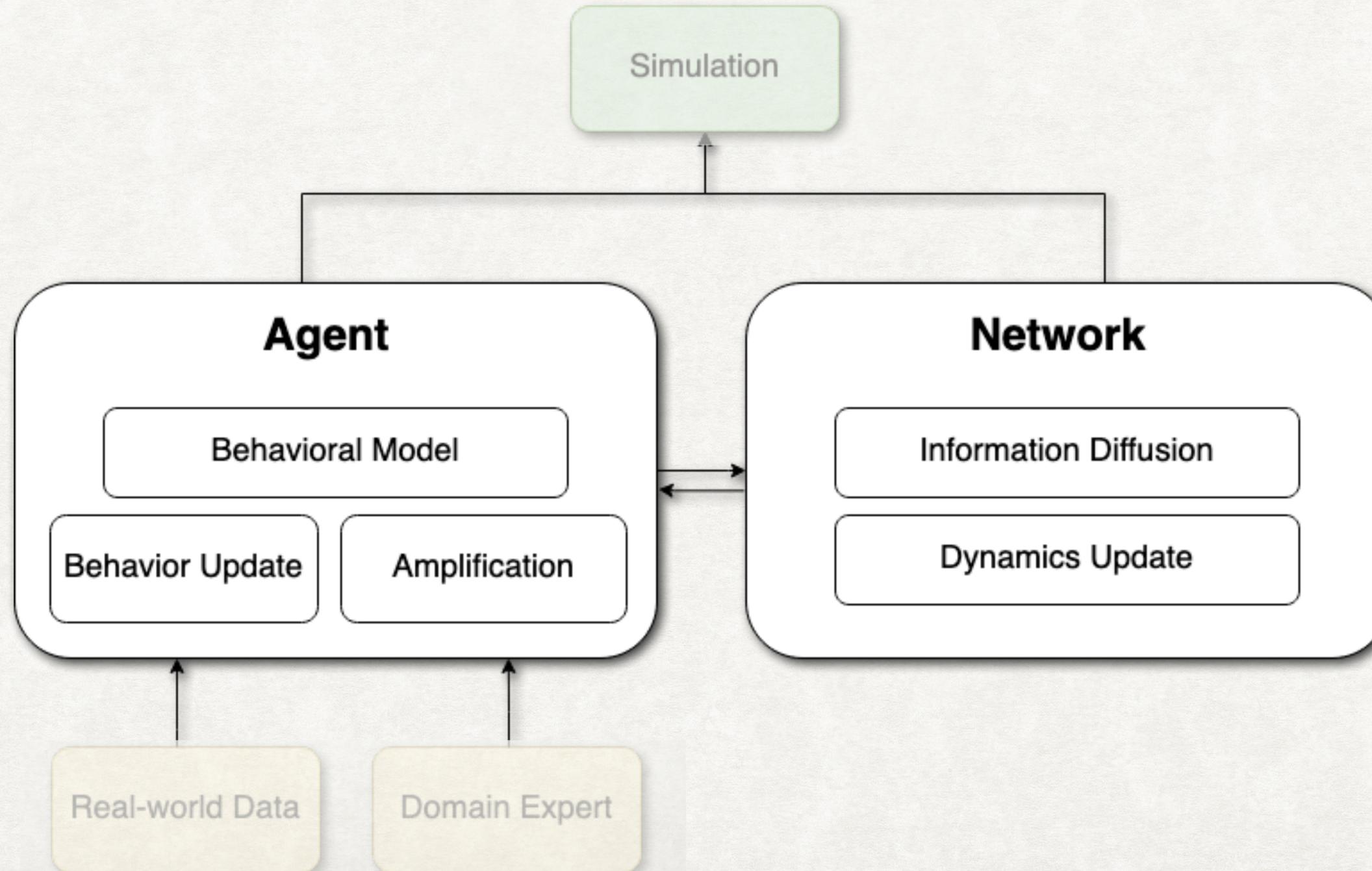


# SIMULATING SOCIAL NETWORKS

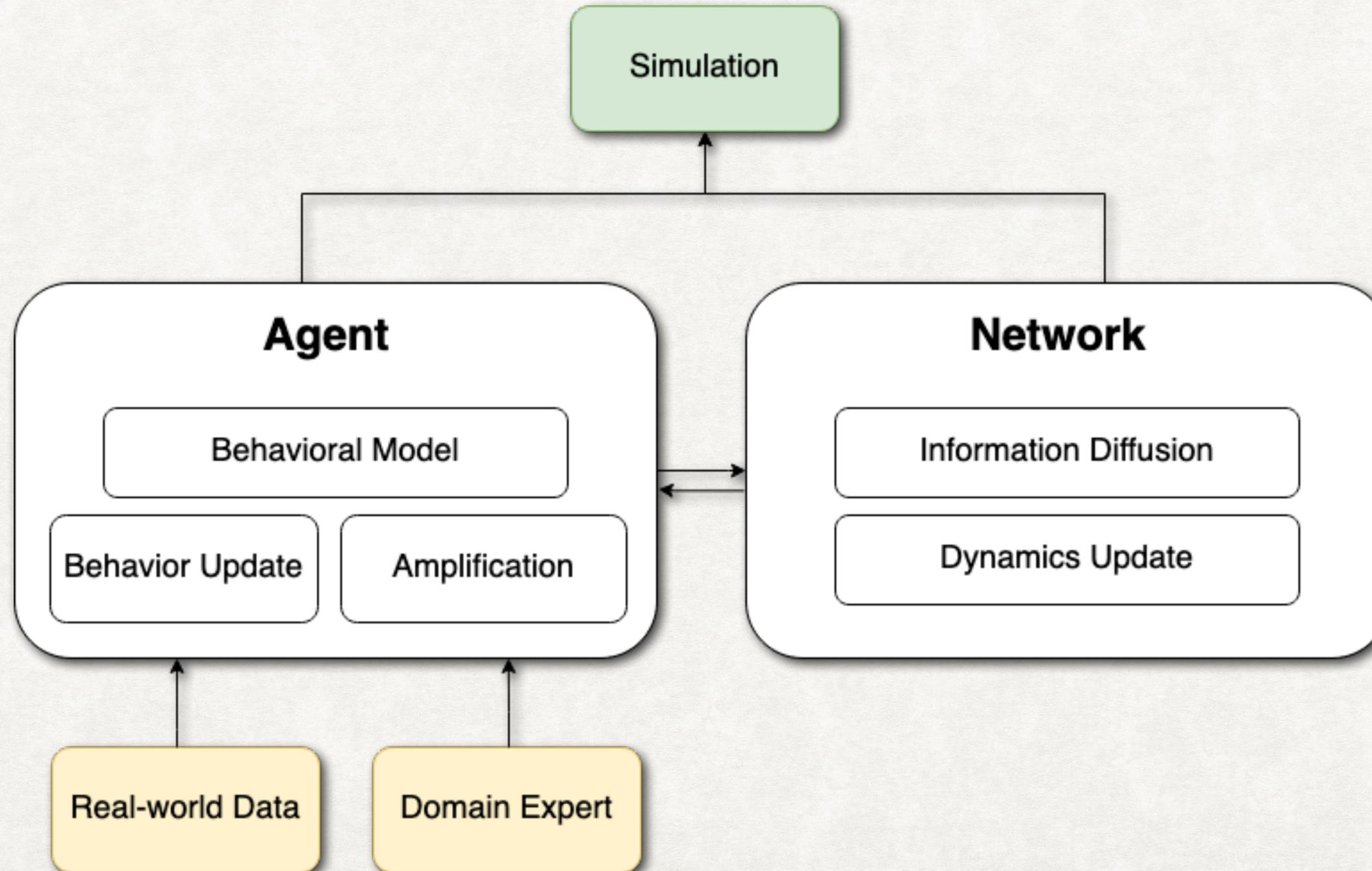


<https://youtu.be/GV5XuftiD7s>

# 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

# REDDIT RECOMMENDER SYSTEMS

The screenshot shows the Reddit homepage interface. 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 community, titled 'Elon Musk offers to buy Twitter for \$54.20 per share', with 1592 upvotes. The post text describes the offer and includes a link to a Bloomberg article. To the right, there is a sidebar titled 'Top Gaming Communities' 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

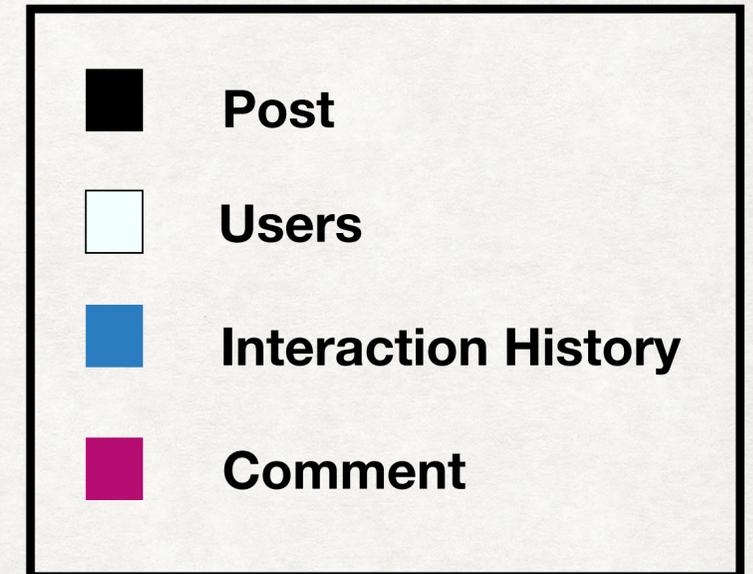
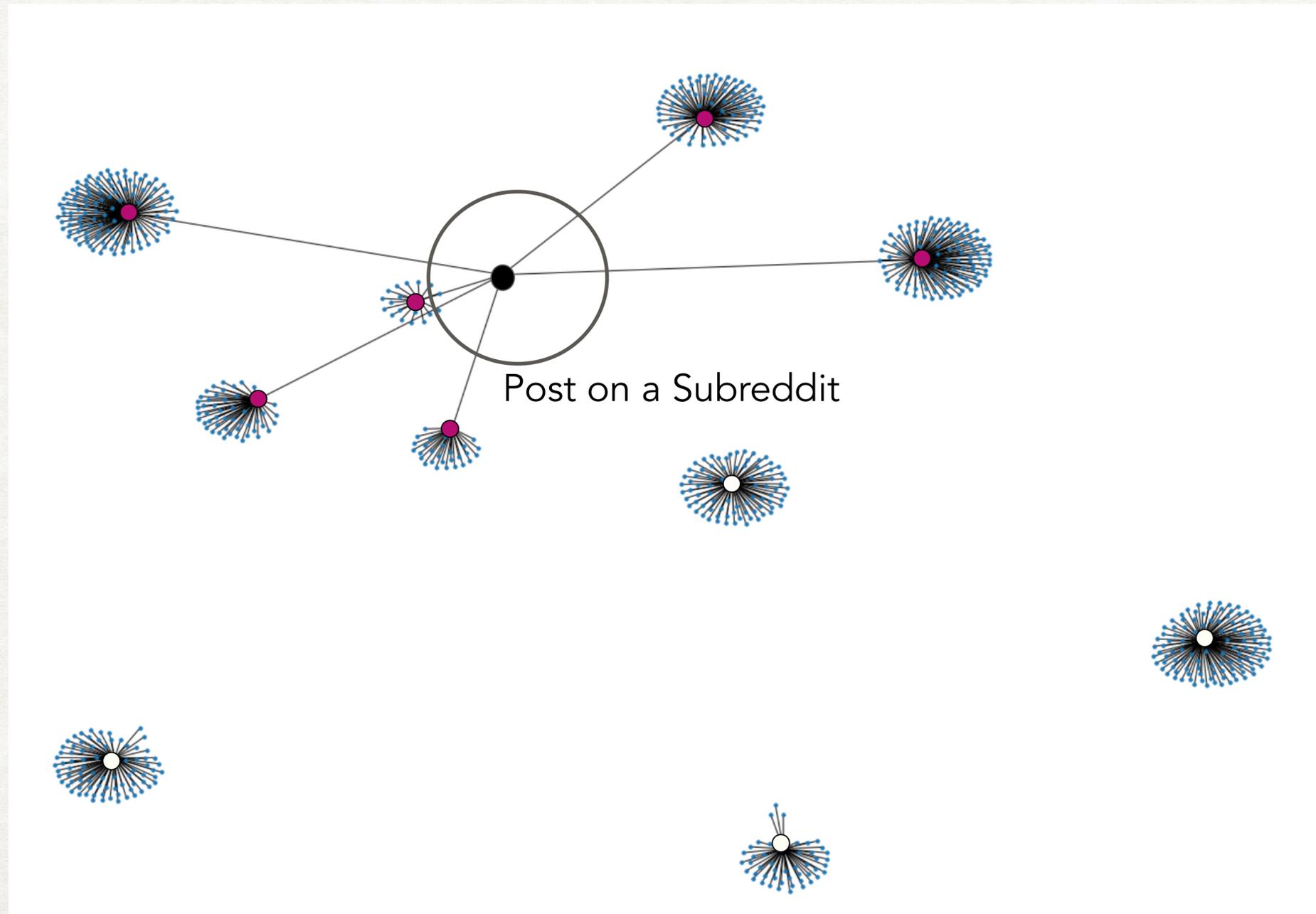
Reddit Premium

# RANKING AND RECOMMENDATION ALGORITHMS

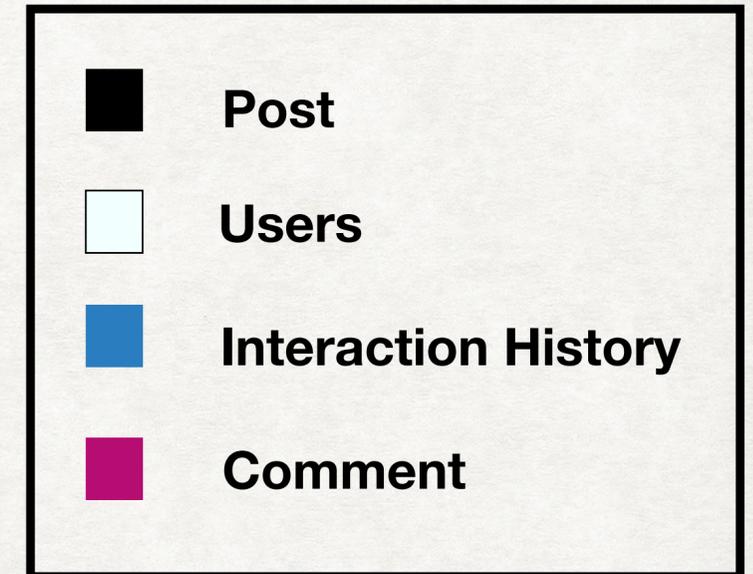
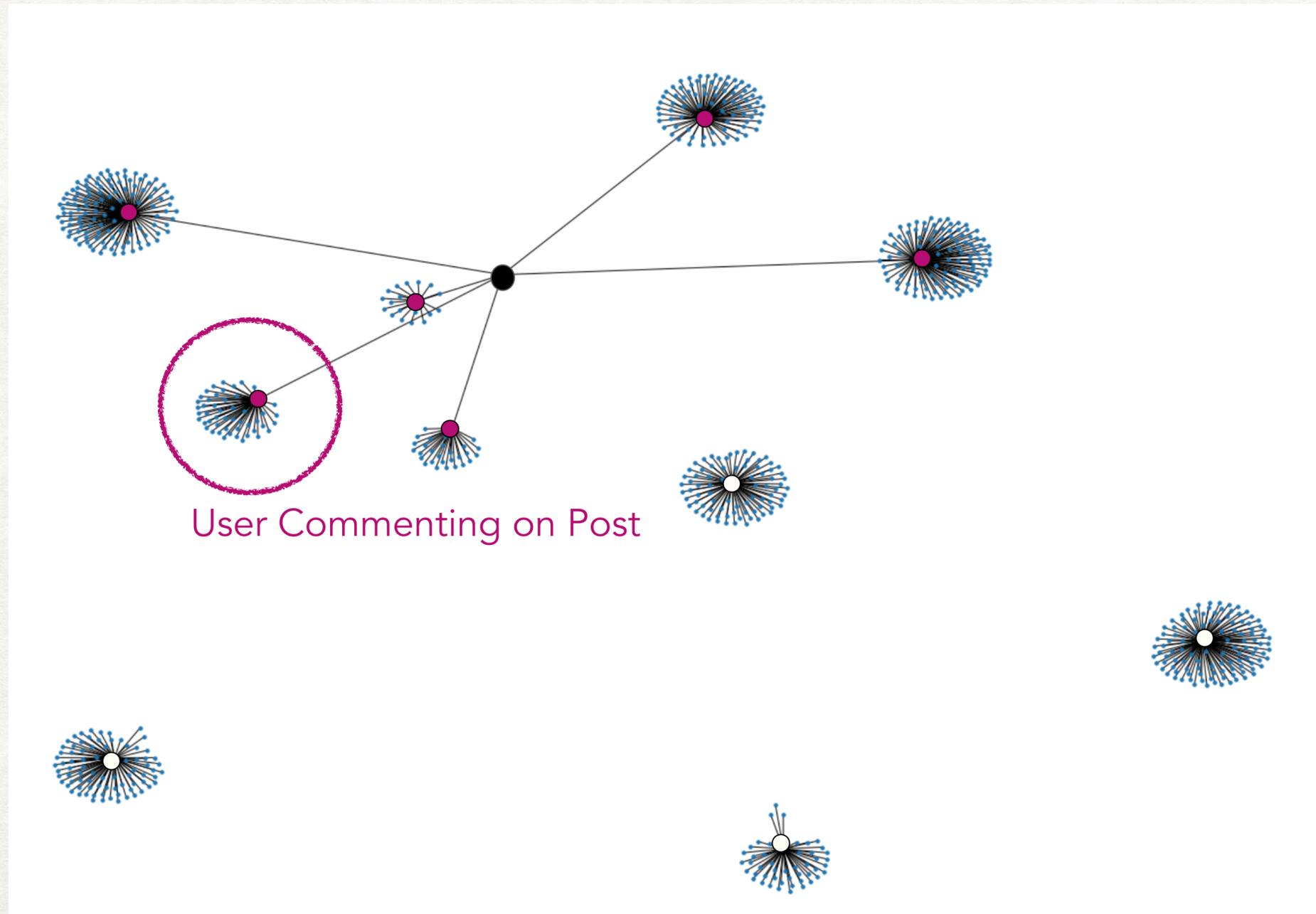
New	Top	Rising	Controversial	Best (Personalized)
Age of Post	Age of Post			Age of Post
	# of Upvotes	# of Upvotes	# of Upvotes	# of Upvotes
			# of Downvotes	
		Age of Votes		
		Age of Comments		
				Relevance to User
				Subreddit Membership

# USING REAL-WORLD DATA TO DRIVE THE SIMULATIONS

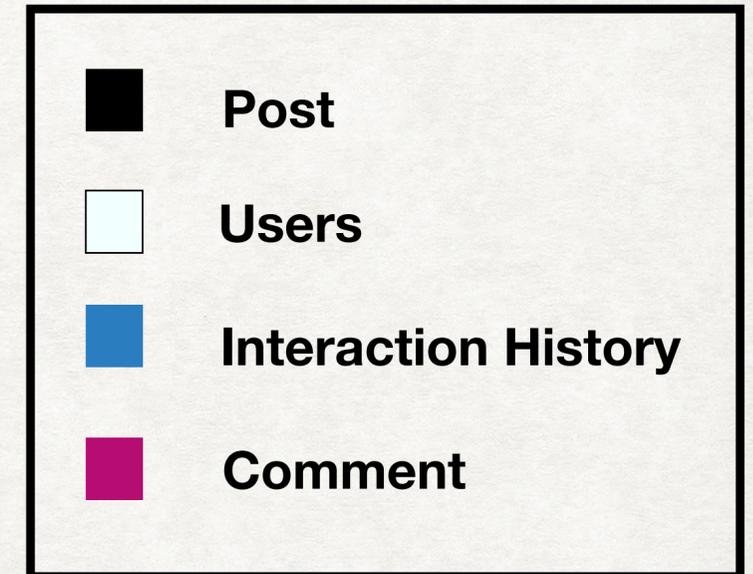
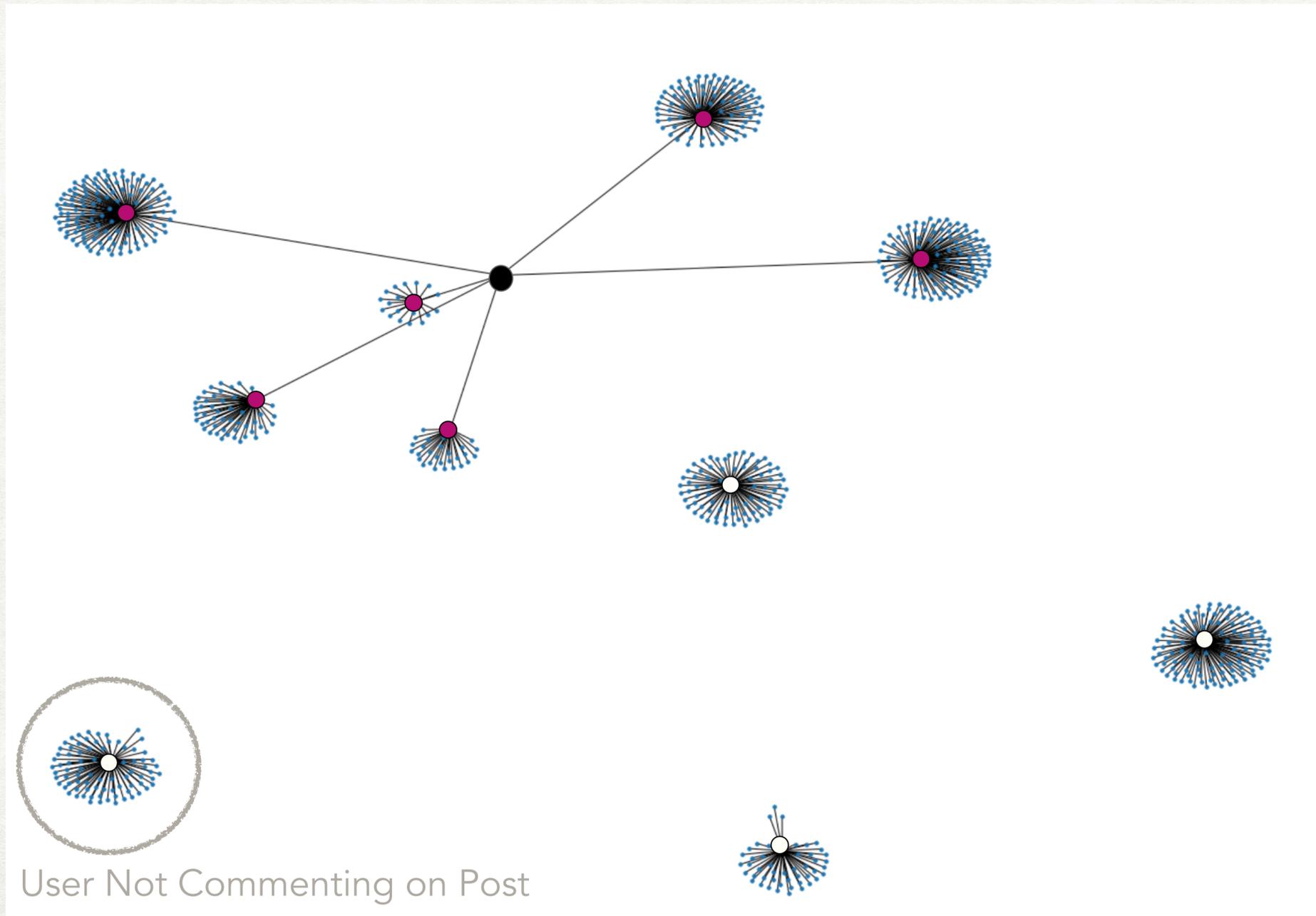
# REDDIT POST



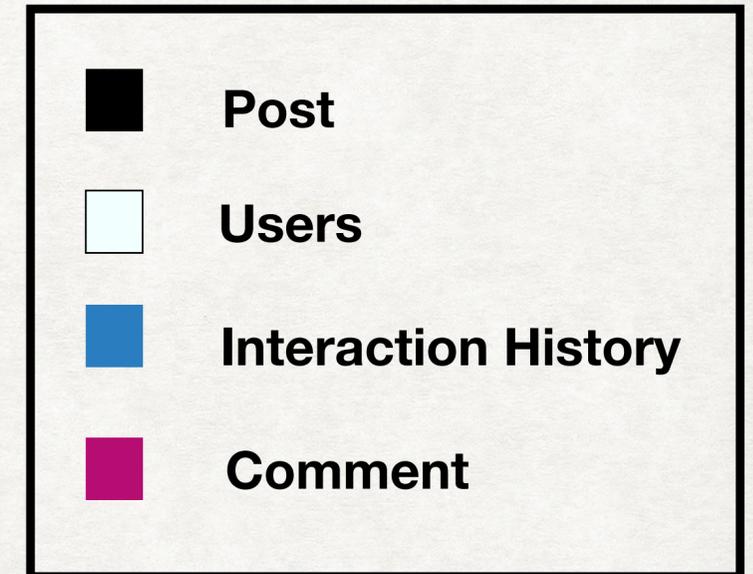
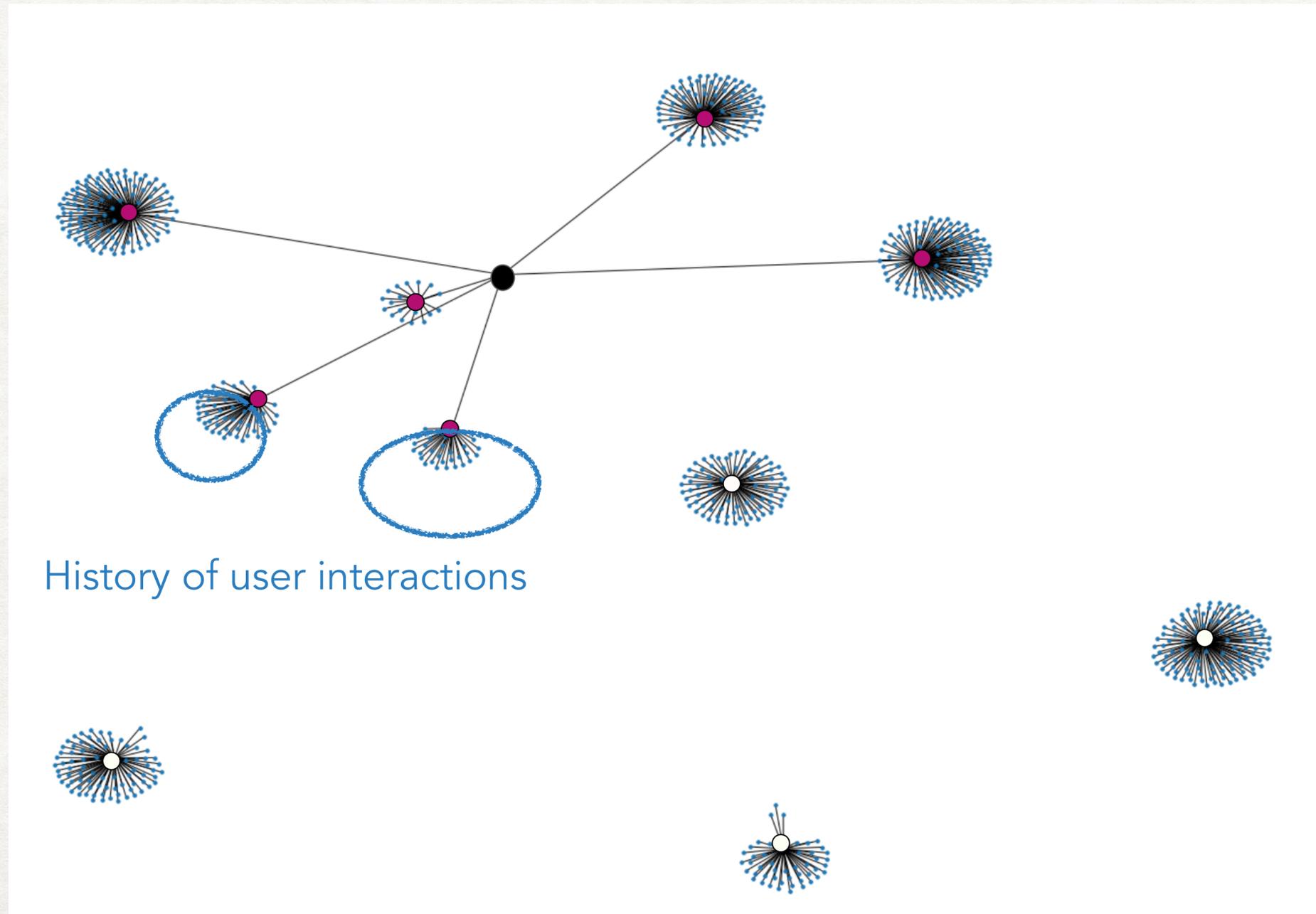
# REDDIT POST



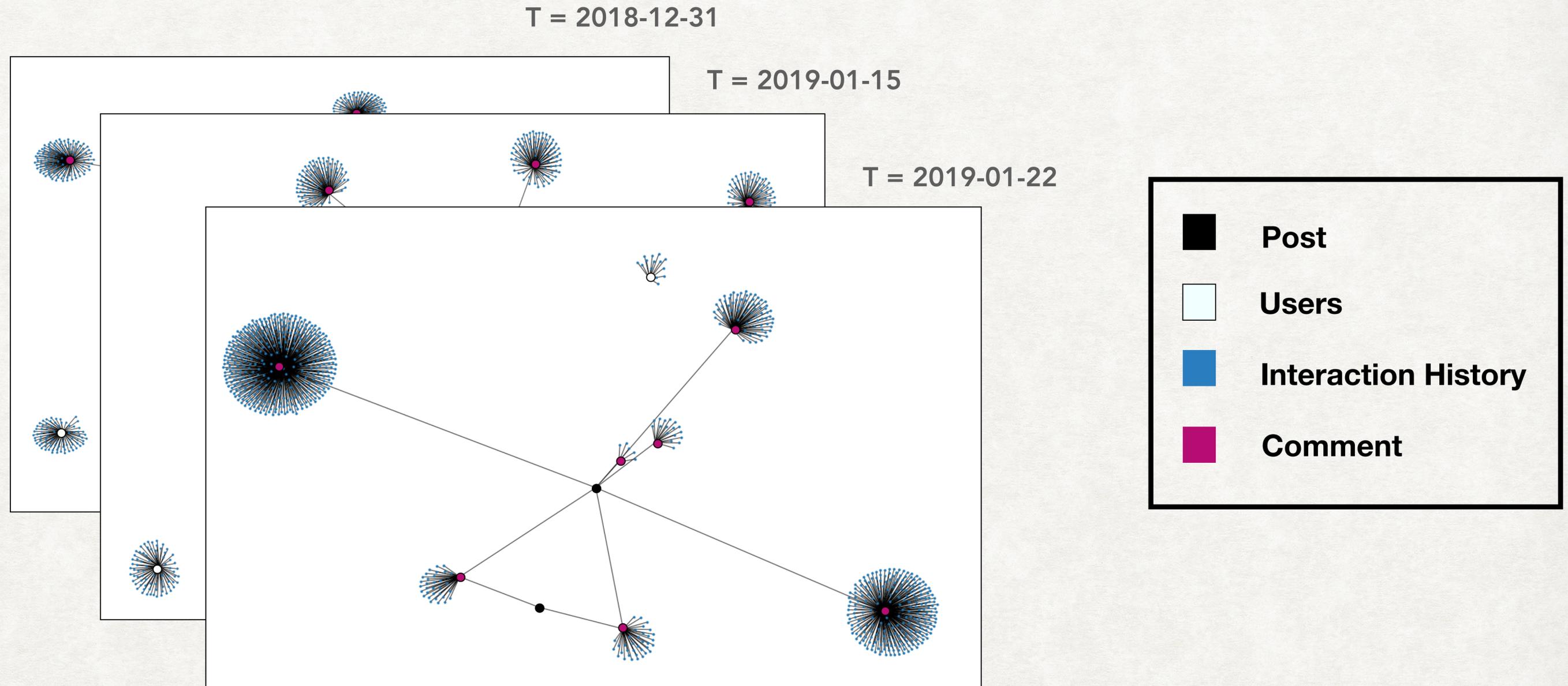
# REDDIT POST



# REDDIT POST

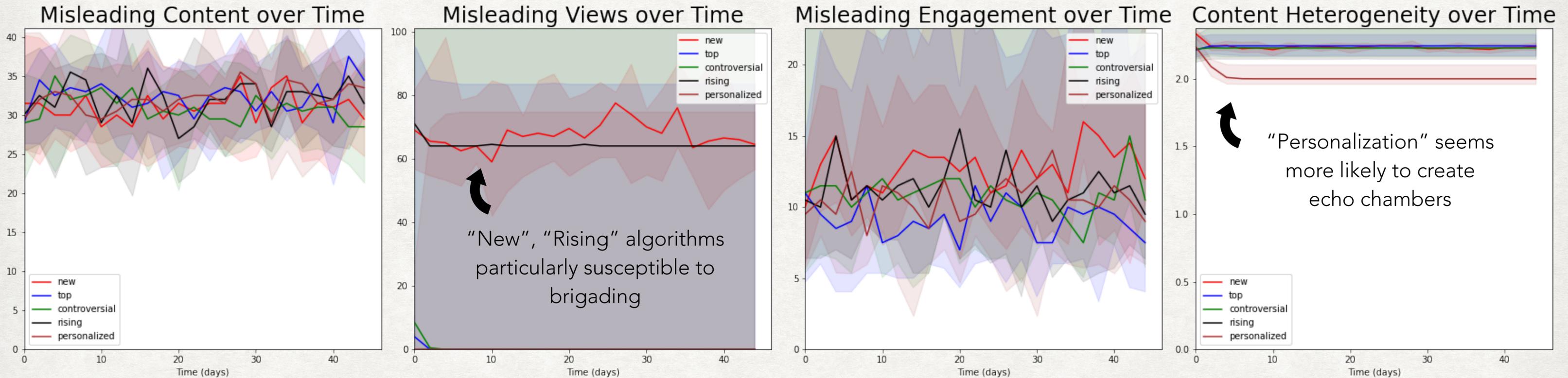


# REDDIT POSTS



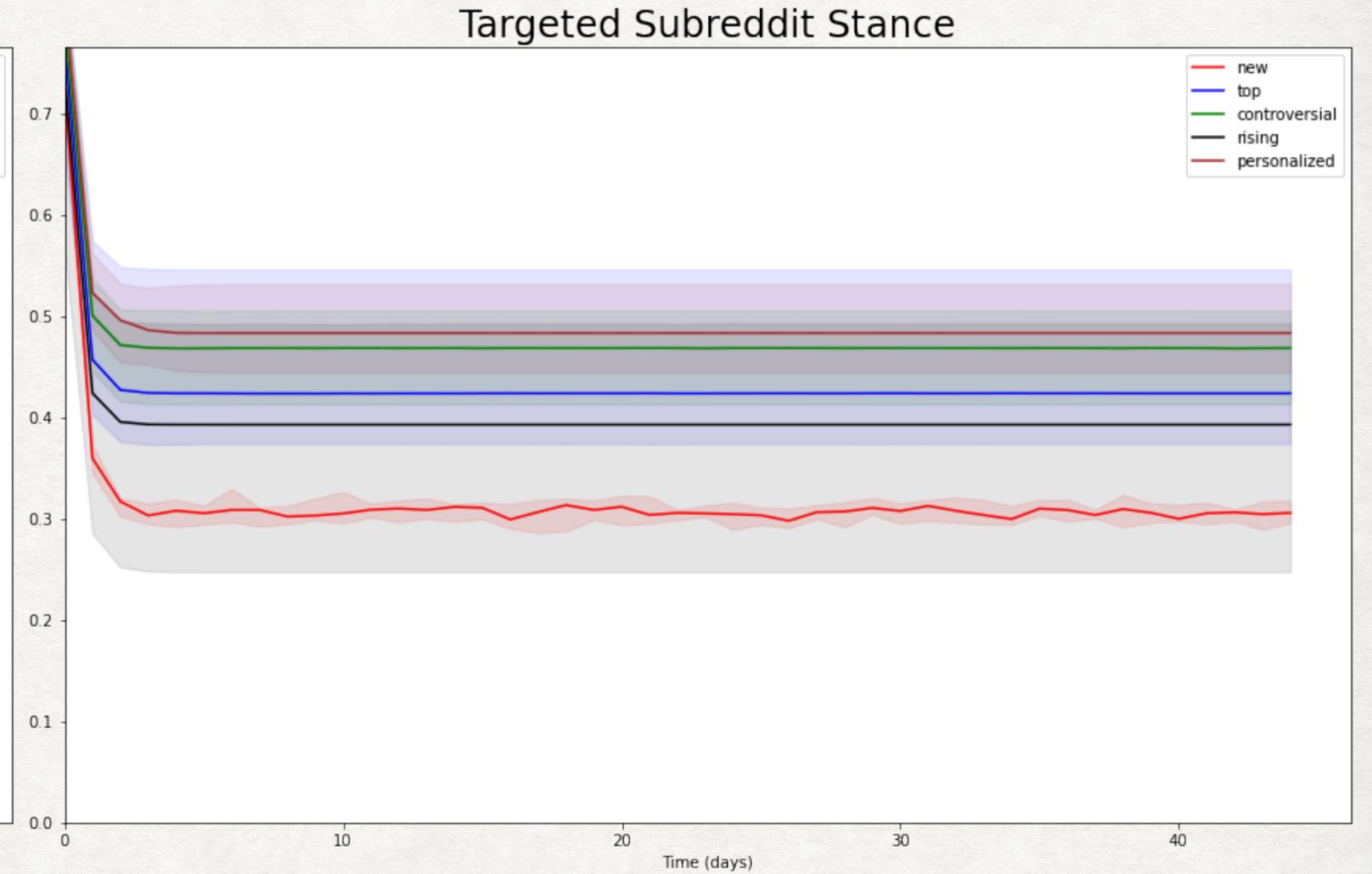
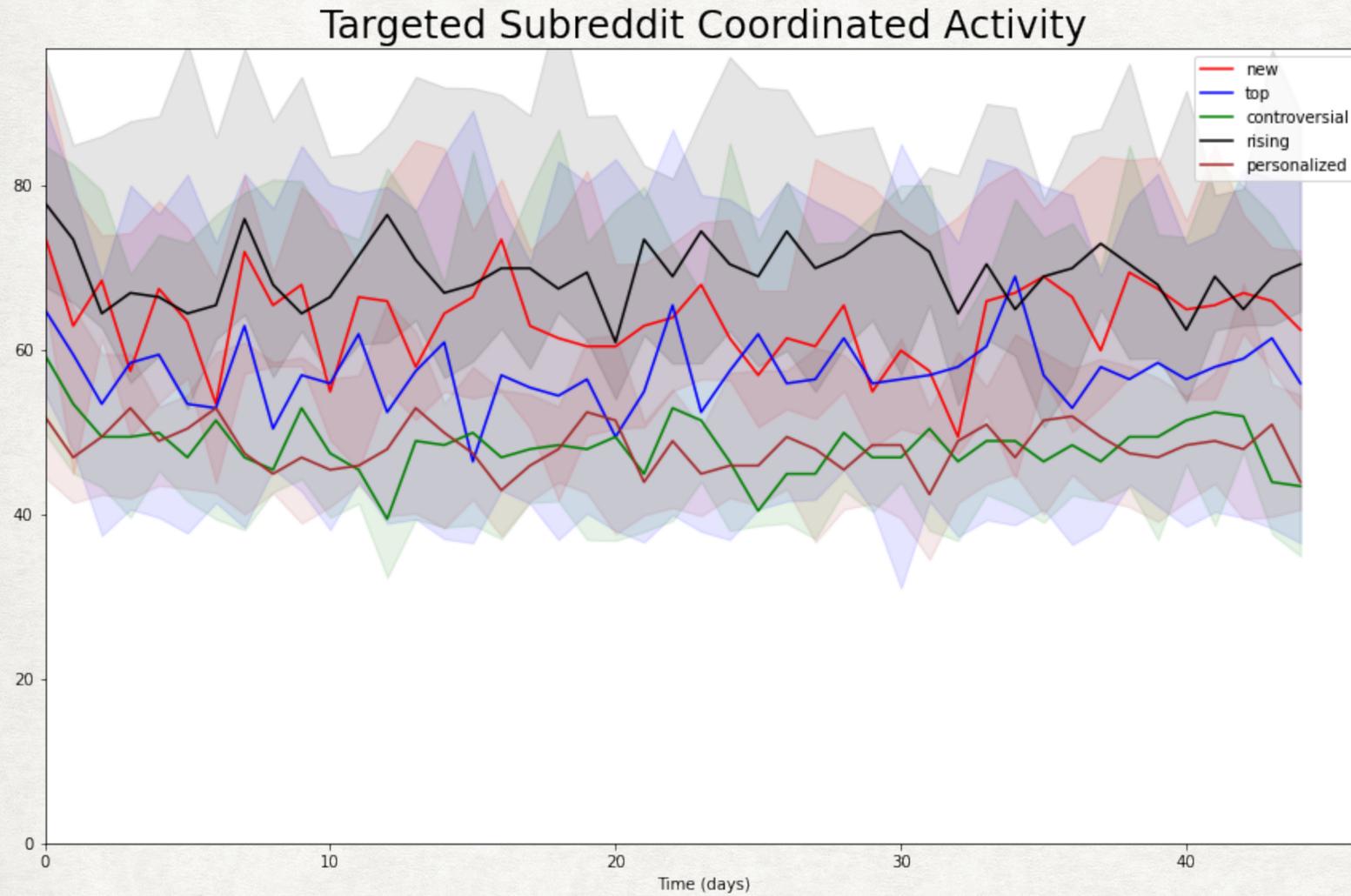
**HOW SUSCEPTIBLE ARE ALGORITHMS TO  
COORDINATED INAUTHENTIC BEHAVIOR?**

# 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

set to

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

Aisha Malik @

## 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

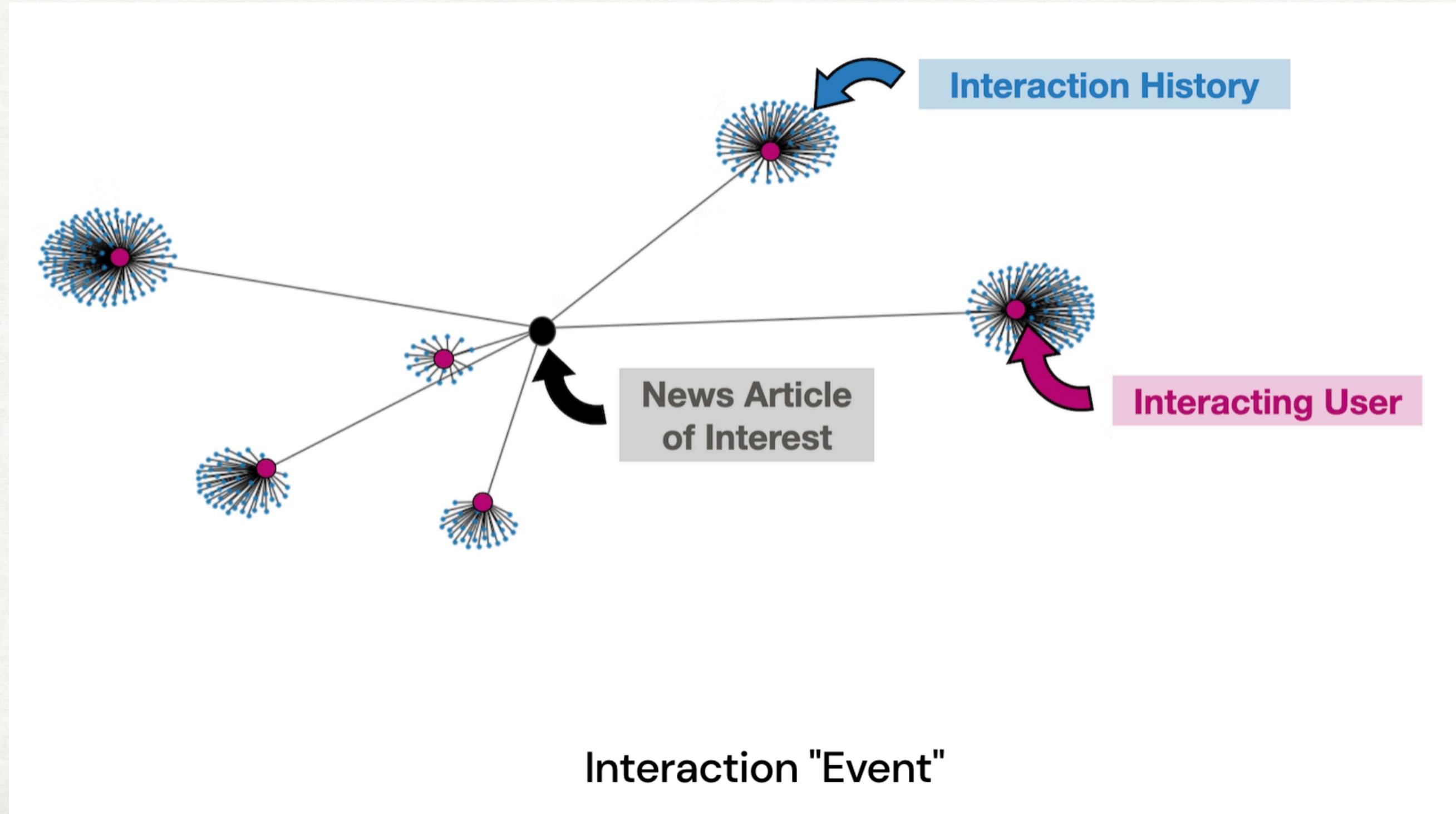
Switch to latest Tweets

show up as they happen.

preferences

cancel

# APPLYING SIMPPL'S TECH TO NEWS ARTICLES



# HELPING LOCAL NEWS UNDERSTAND AUDIENCES



## You're Getting Upvotes!

5

r/VTGuns

9

r/VoteDEM

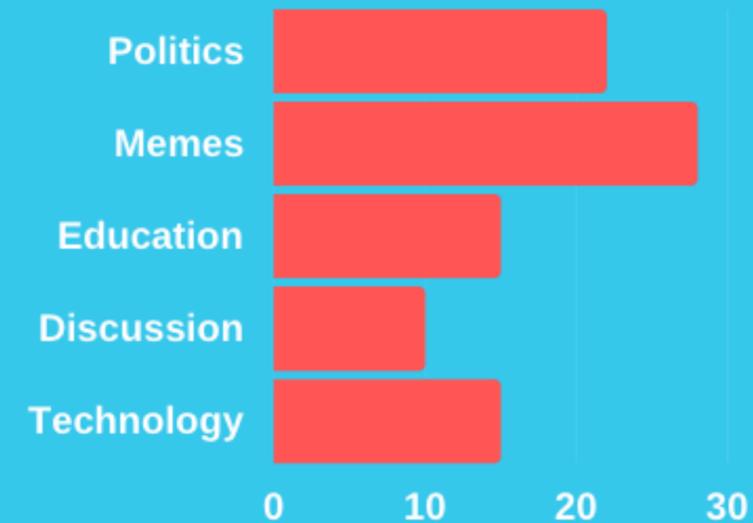
47<sup>↑</sup>

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

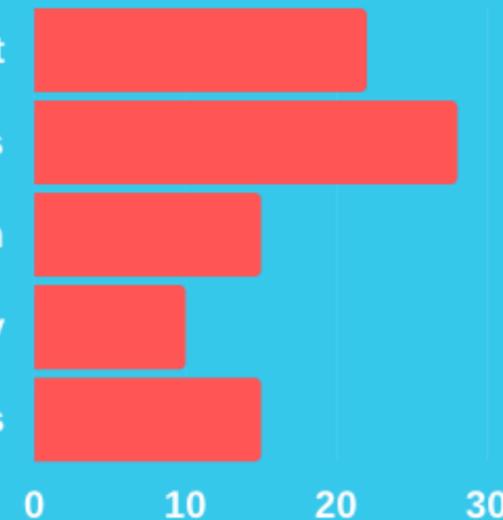
@sevendaysvt

@calmatters

Tech

Silicon Valley

Politics



## Related Engagement

14

Politics

23

Elections

86 ↑

Bernie Sanders

@sevendaysvt

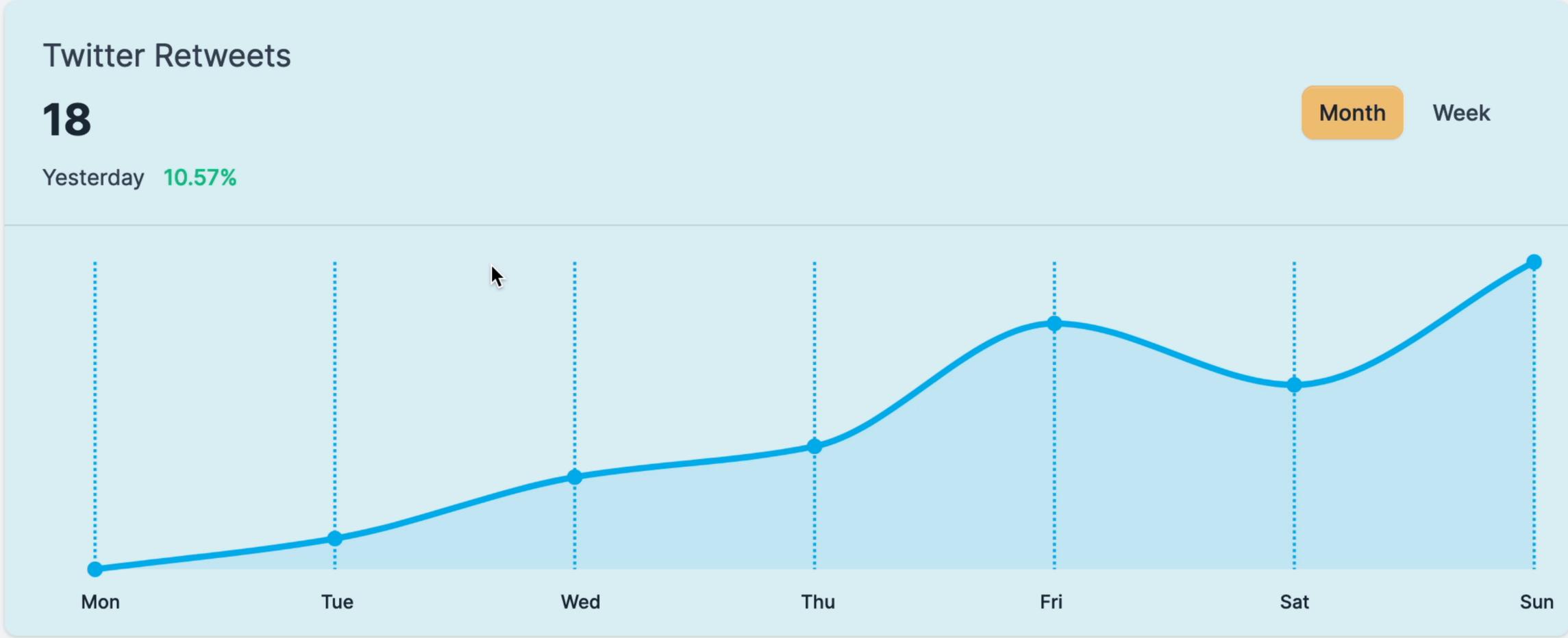


# DEMO

- ⚡ Projection Dash
  - 🕒 Audience Dashboard
  - 📅 Projections
  - 📊 Engagement >
  - 📧 Messages >
  - 📁 External Data >
  - ⚙️ Settings
  - 🔍 Product Page
  - ⚙️ Support
- 🔥 Update Projections

🔔  Current User: admin

+ Social Media Monitor



New Users  
**23**

Bounce Rate  
**12.88%**

Site Visits  
**82,7**

⚙️ Settings

# SUMMARY

- We simulated coordinated campaigns seeking to manipulate public debate using multiple authentic/inauthentic (fake) accounts to mislead people.
- Goal: Quantify the harms arising from CIB on Social Networks
  - Measure its effects on ranking and recommendation algorithms
  - Use real-world networks and behavior to simulate counterfactual outcomes
  - *Next Steps: Simulate Interventions*

# COLLABORATORS



Jonathan Nagler



Richard Bonneau



Philip Torr



Atılım Güneş Baydin



Bogdan State

# LET'S TALK!

[@swapneel\\_mehta](#)

[swapneelm.github.io](#)

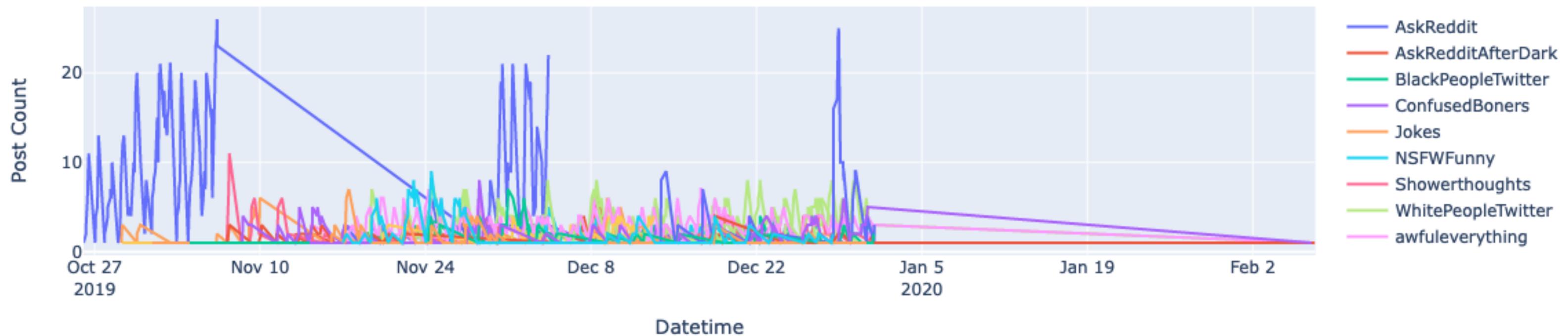
[swapneel.mehta@nyu.edu](mailto:swapneel.mehta@nyu.edu)

[ai4abm.org](#)

# REAL-WORLD MODELING

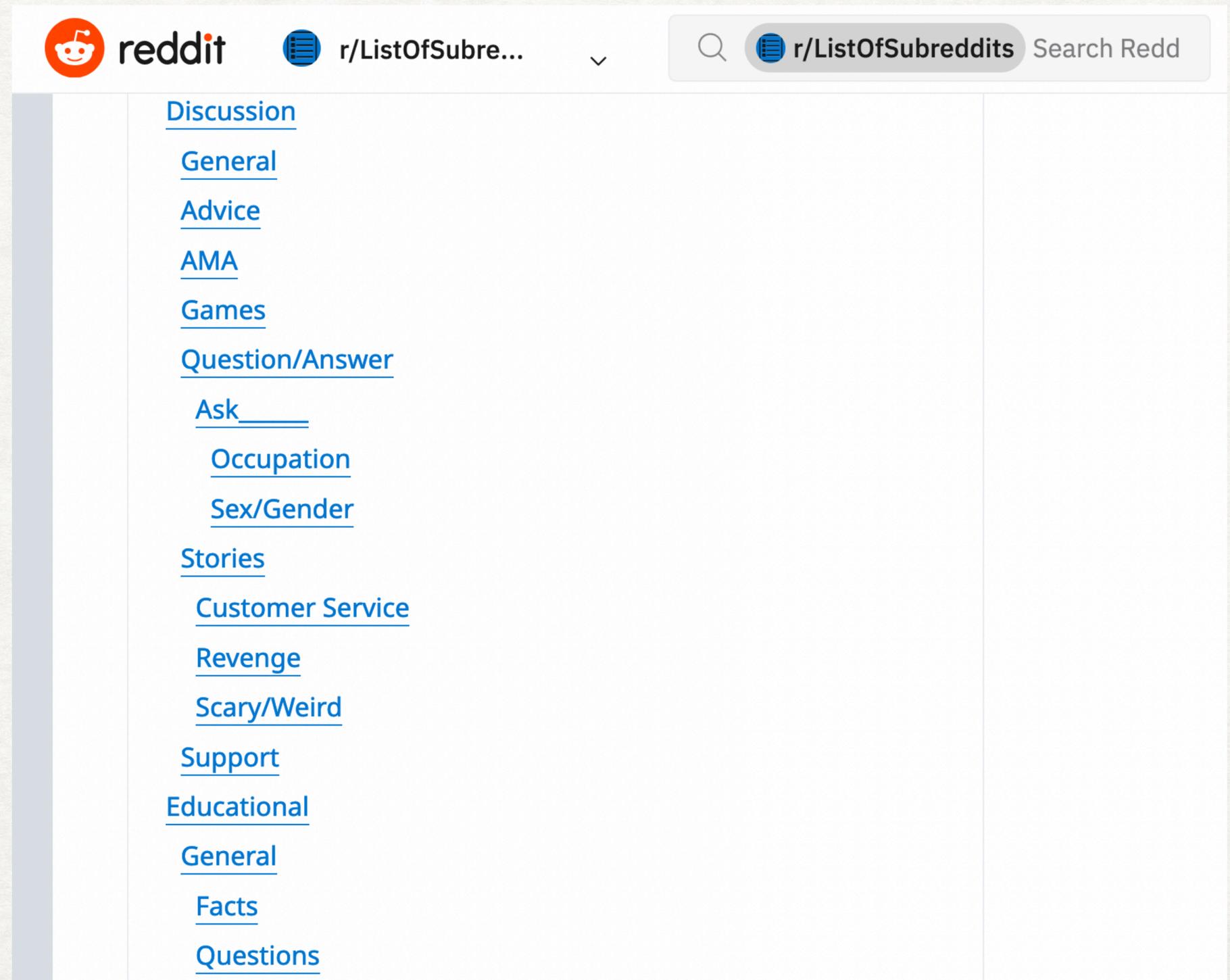
REDDIT USER 'JONNYCREEPYCREPES3'

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

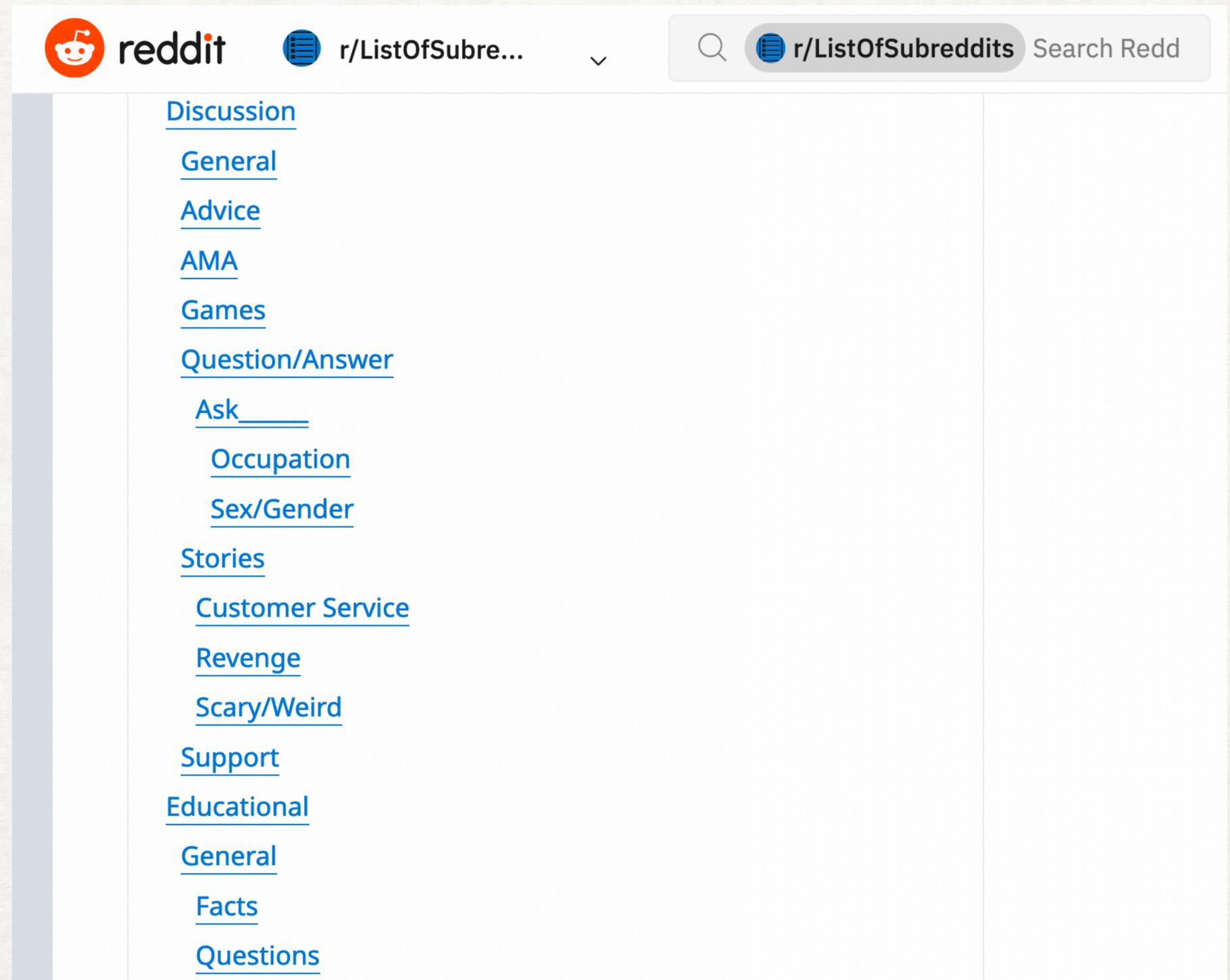


The screenshot shows the top navigation bar of the Reddit website. On the left is the Reddit logo and the word "reddit". In the center is a dropdown menu for the subreddit "r/ListOfSubre...". On the right is a search bar with the text "r/ListOfSubreddits" and "Search Redd". Below the navigation bar is 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](#)

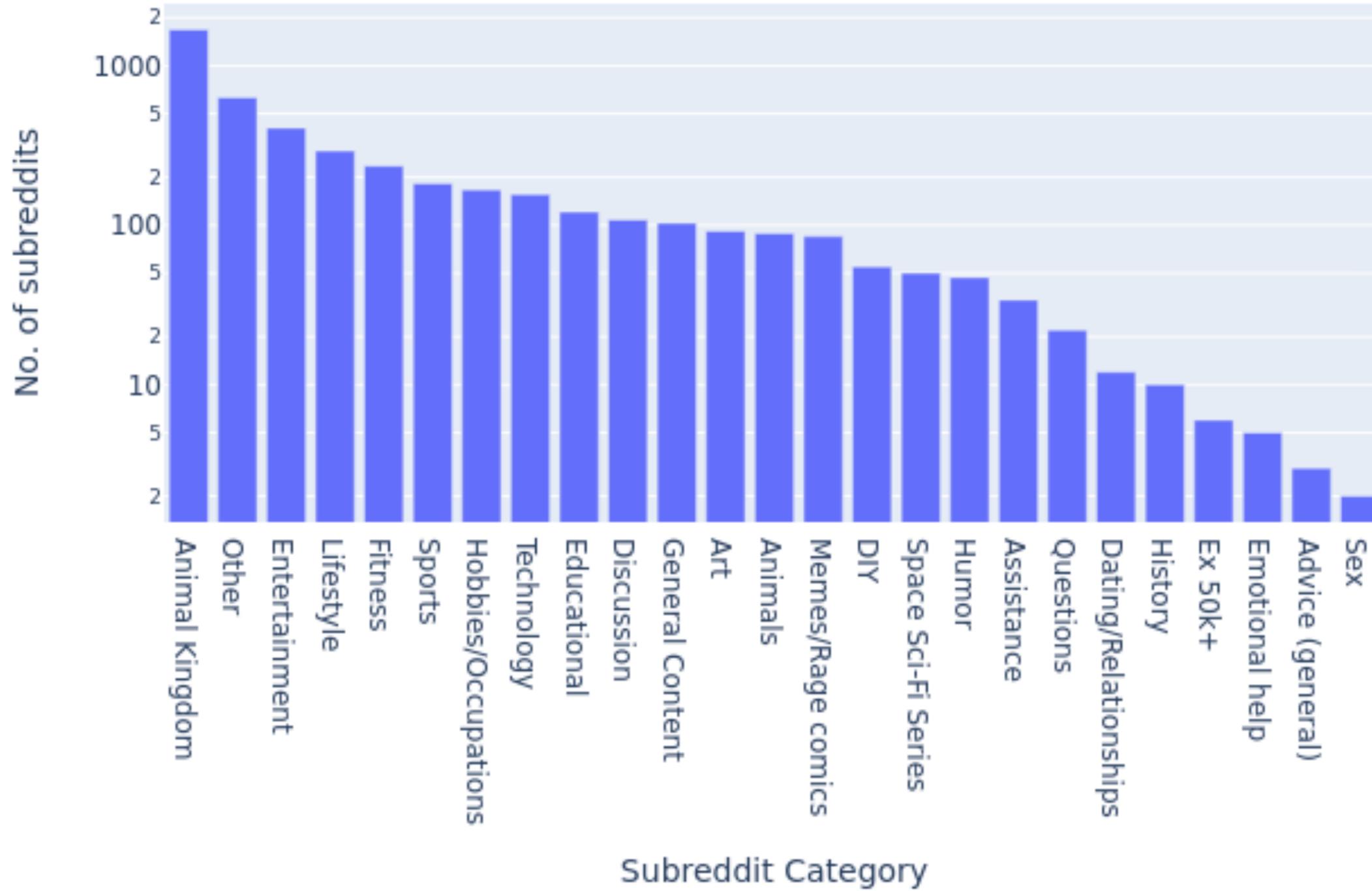
# REDDIT WIKI PAGES

- 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

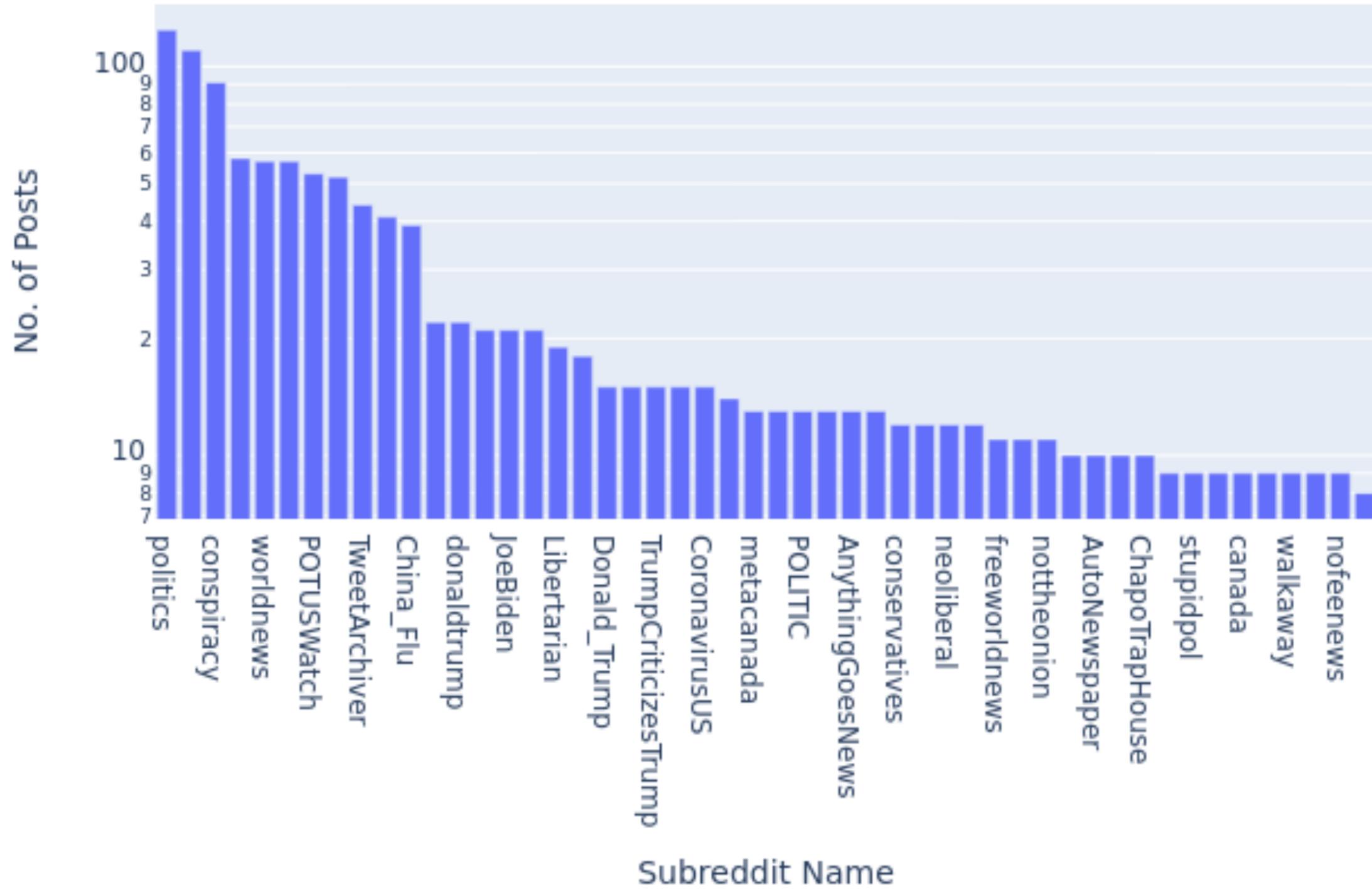


The screenshot shows the top navigation bar of the Reddit website. On the left, there is the Reddit logo and the word "reddit". To its right is a dropdown menu for the current subreddit, "r/ListOfSubre...". On the far right, there is a search bar with the text "r/ListOfSubreddits" and "Search Redd". Below the navigation bar, a vertical list of links is displayed, representing the hierarchy of categories for the subreddit. The links are: Discussion, General, Advice, AMA, Games, Question/Answer, Ask, Occupation, Sex/Gender, Stories, Customer Service, Revenge, Scary/Weird, Support, Educational, General, Facts, and Questions. Each link is underlined and blue.

# CATEGORIZE POSTS ACROSS SUBREDDITS

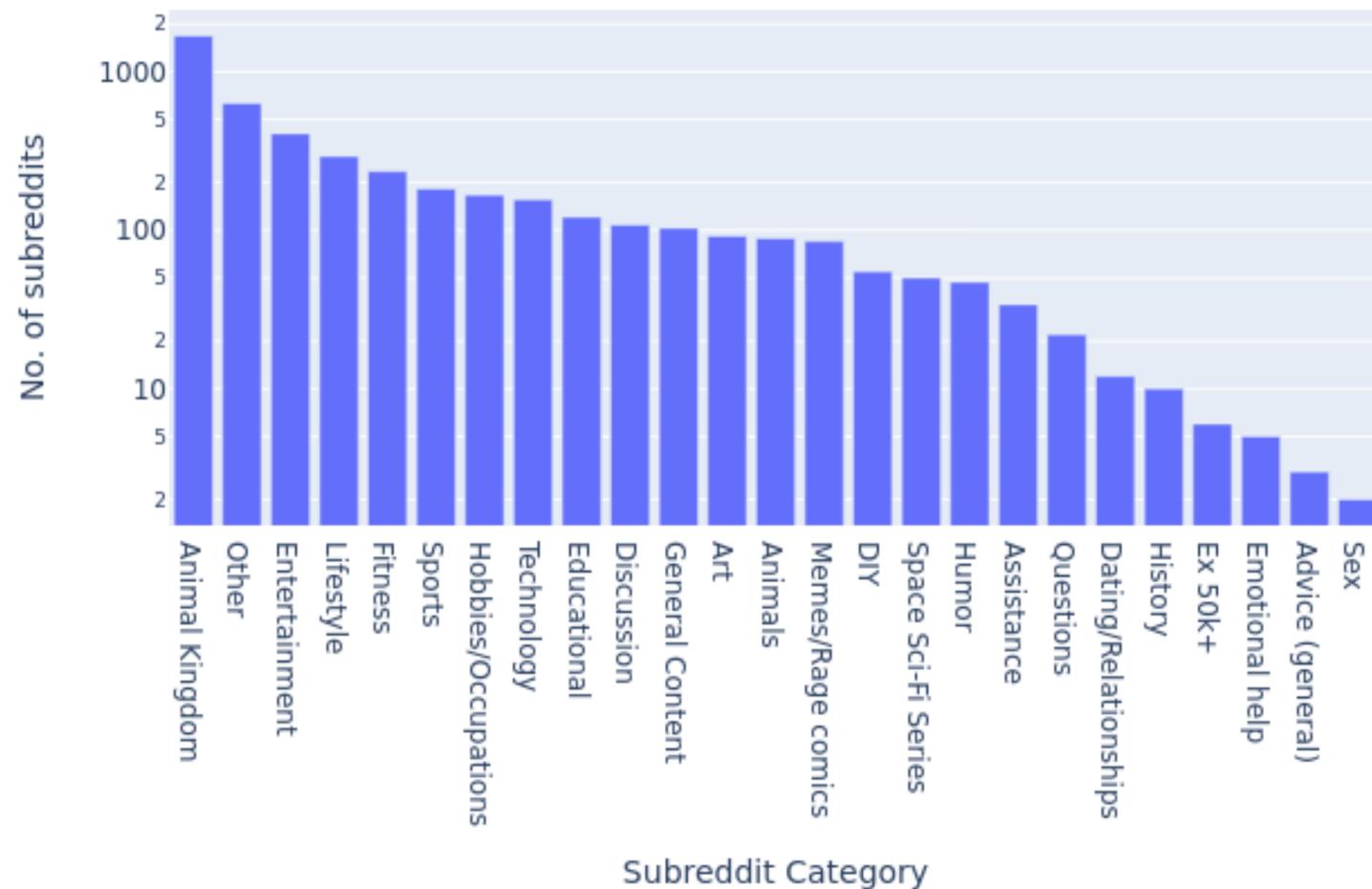


# CATEGORIZE POSTS ACROSS SUBREDDITS



# CATEGORIZE POSTS ACROSS SUBREDDITS

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



['mormon', 'politics']

## Predicted Subreddits:

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

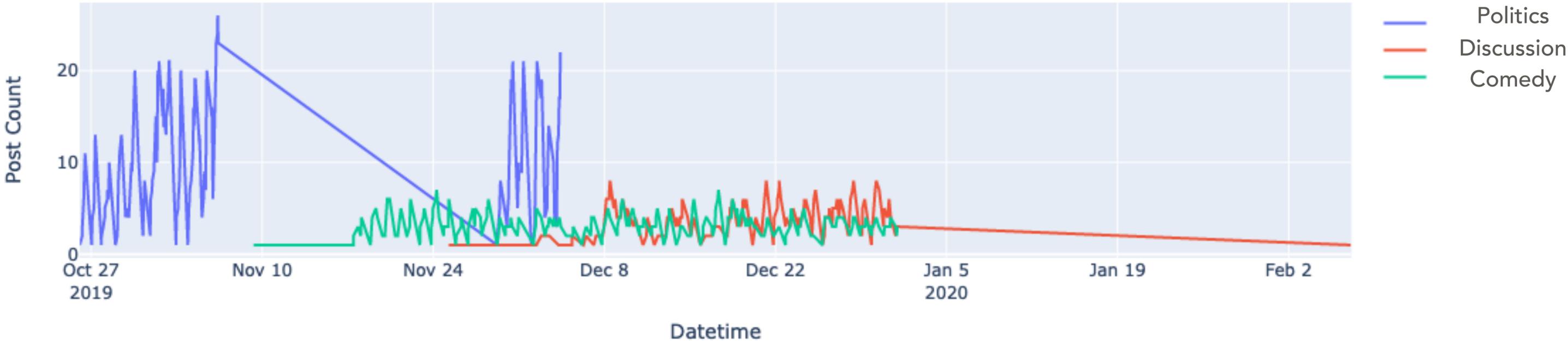
## Predicted Category:

[[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

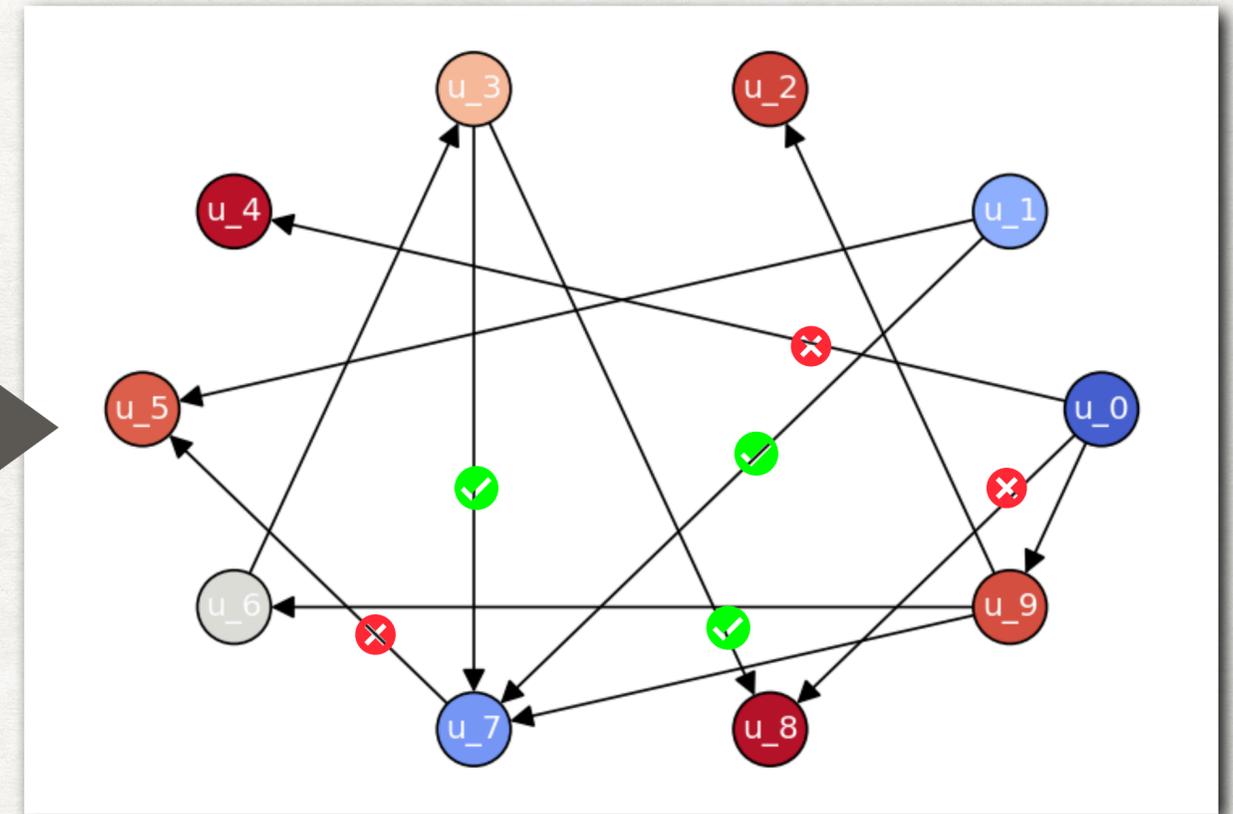
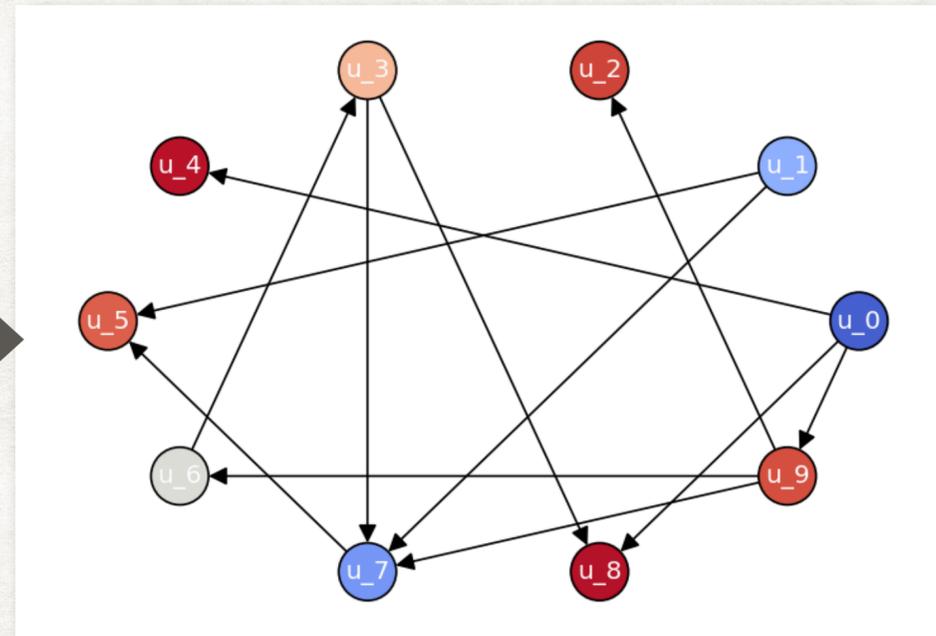
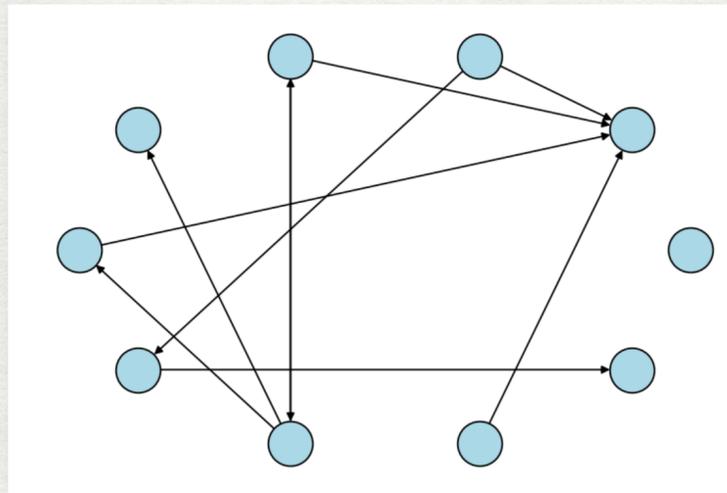
## 3. HYBRID

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