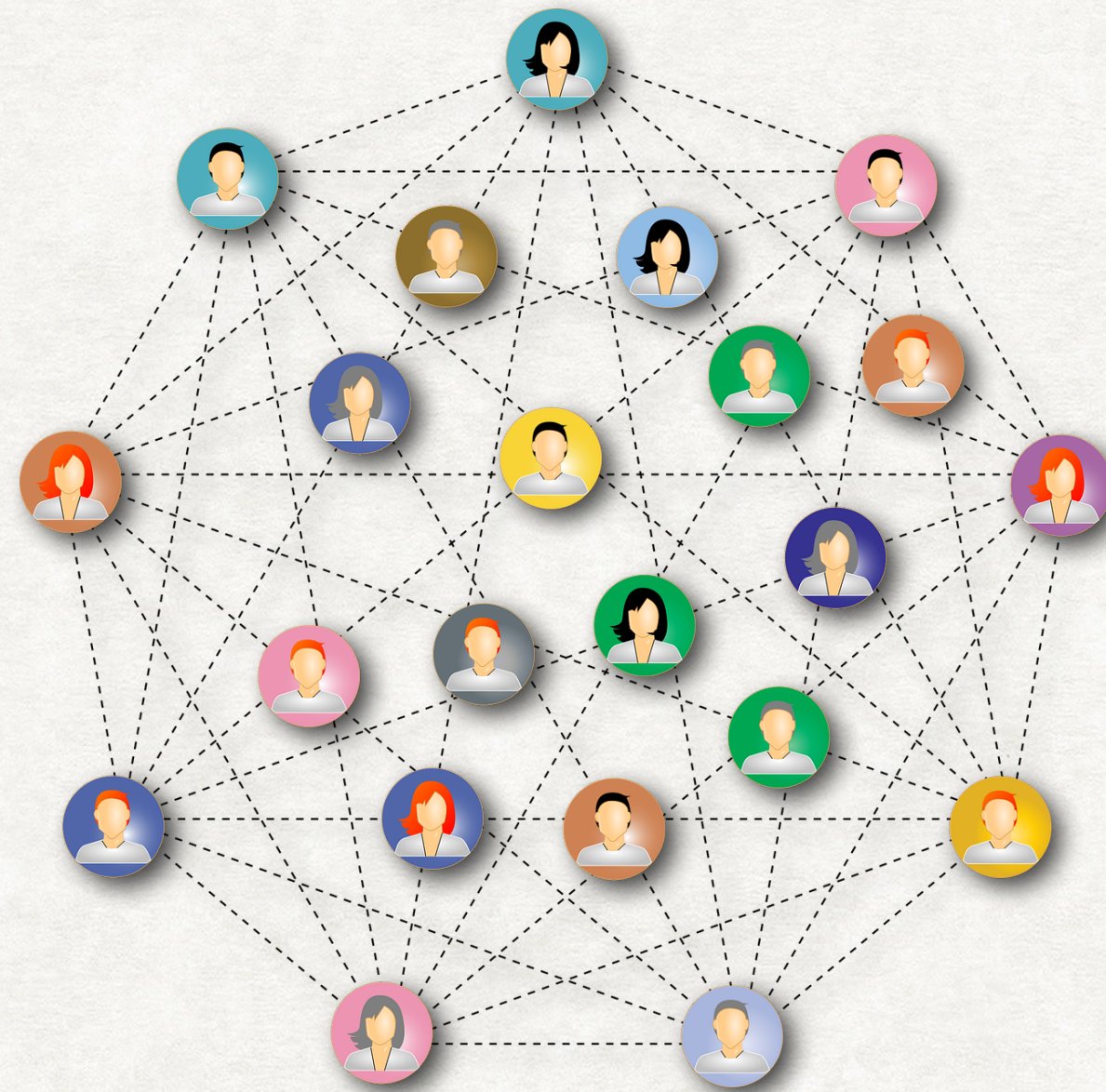


# ESTIMATING THE IMPACT OF COORDINATED INAUTHENTIC BEHAVIOR ON CONTENT RECOMMENDATIONS IN SOCIAL NETWORKS



AI4ABM WORKSHOP,  
ICML '22

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



# 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!

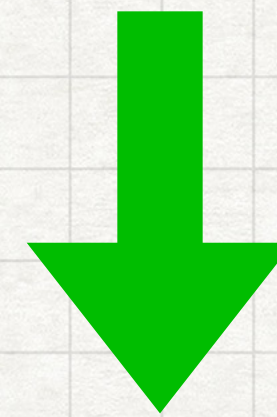




**(COST TO)  
MITIGATE THESE INFLUENCE OPS!**



**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?
  - What about "controversial" opinions?



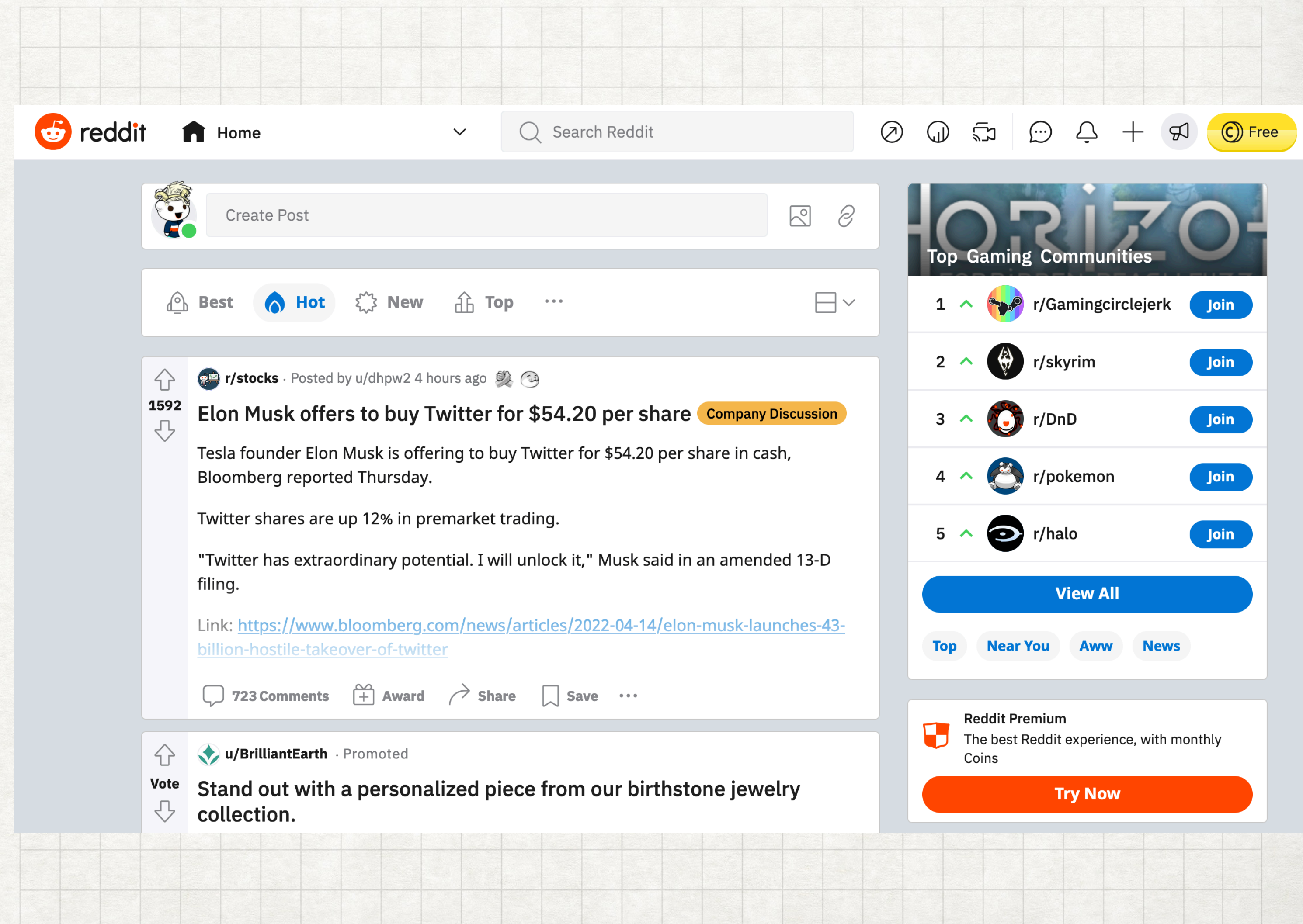
# RESEARCH GOALS

- 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?
  - What about "controversial" opinions?
  - Susceptibility varies by community. Will penalties on engagement with disinformation be fair to apply?



# SIMULATE A SOCIAL NETWORK

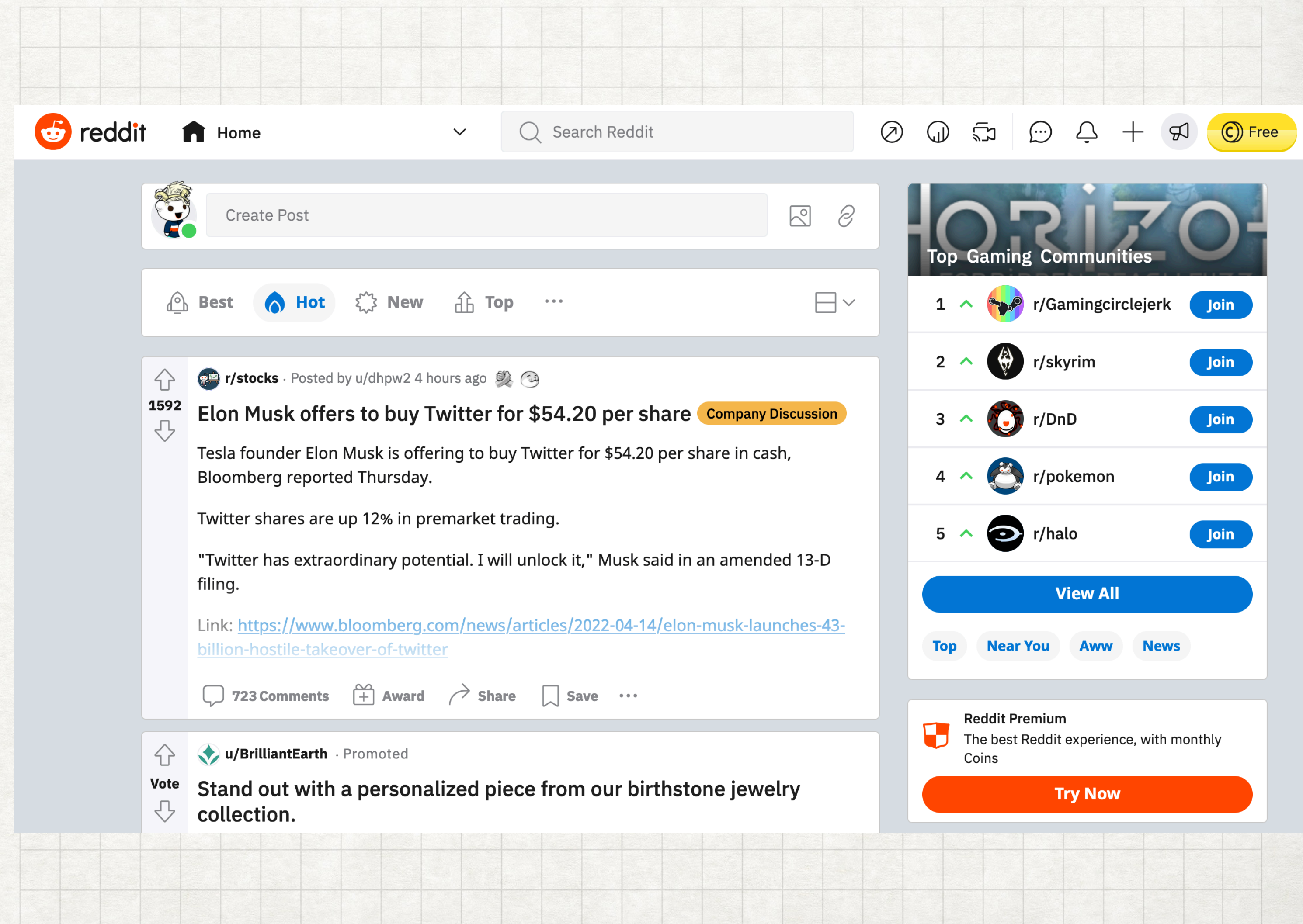
- Reddit is a pseudonymous social network comprising users who are part of like-minded groups or subreddits
- It has a community-based structure
- The state-action space for a user includes:
  - Create a post/comment
  - Upvote a post/comment
  - Downvote a post/comment
  - Cross-post an existing post





# 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  
 $\{\pi_i \in \text{Uniform}(0, 1)\}_{i=1, j=1}^{N, S}$

1: **procedure** REDDIT

2: *Sample latents:*

3:  $\mathbf{v} \leftarrow \{v_i \sim \text{Uniform}(0, 1)\}_{i=1, j=1}^{N, S}$

4: *Simulate:*

5:    $\Phi_{1:N} \leftarrow \langle \rangle$

6:   **for**  $t = 1 : T$  **do**

7:     **for**  $i = 1 : N$  **do**

8:        $\gamma \sim \text{Categorical}(\pi_i)$

9:        $\tau \sim \text{Bernoulli}(v_{i, \gamma})$

10:        $\Phi_i \leftarrow \Phi_i + \langle \tau \rangle$

11: **return**  $\Phi_{1:N}$

▷ Interaction frequency

▷ Interaction propensity over subreddit categories

▷ User Activity

▷ Choose Subreddit (category)

▷ Interact with Subreddit (category)

▷ Append to user's activity

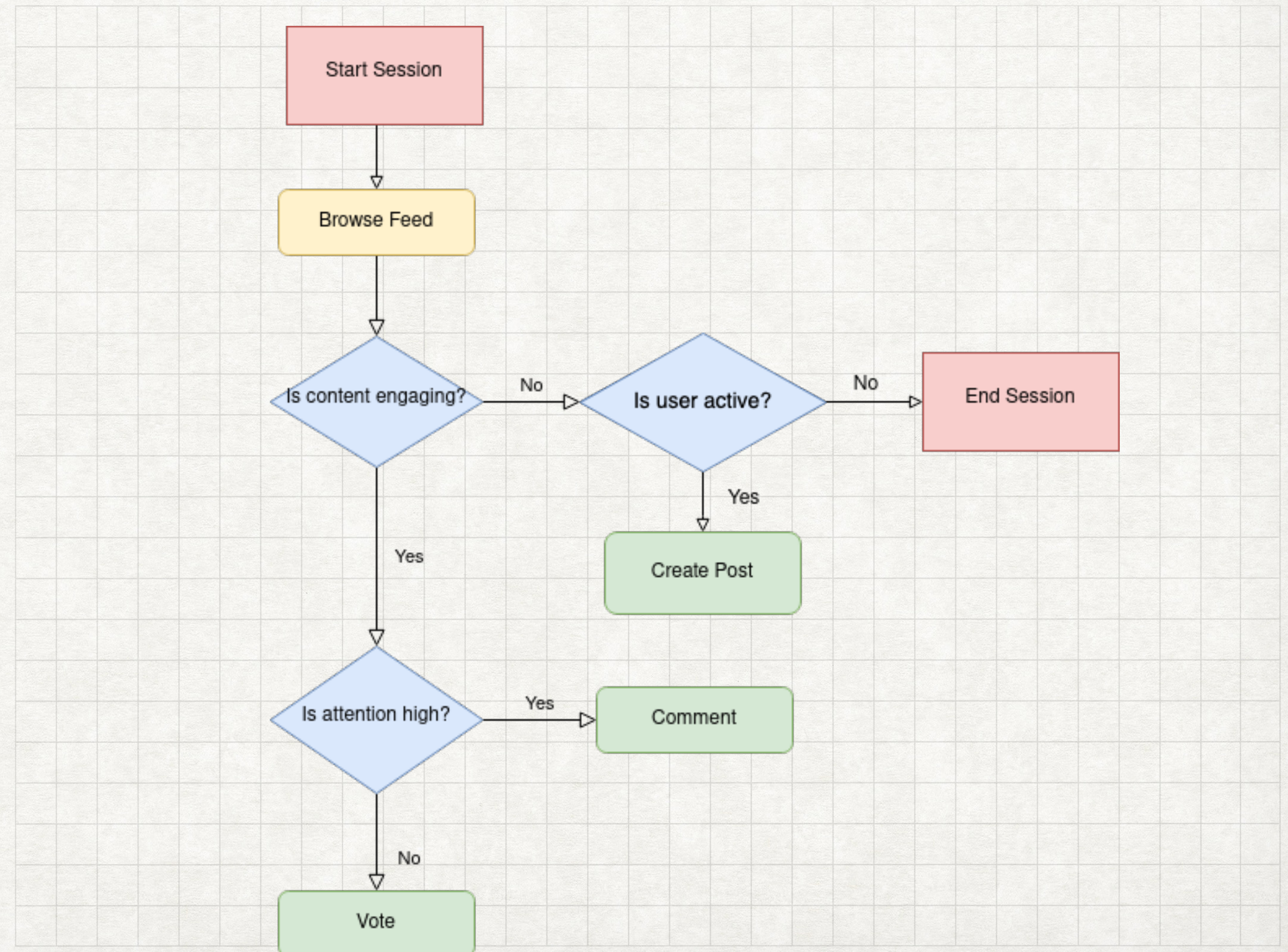
▷ All user activity



# SIMULATING SOCIAL NETWORKS

## REDDIT

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

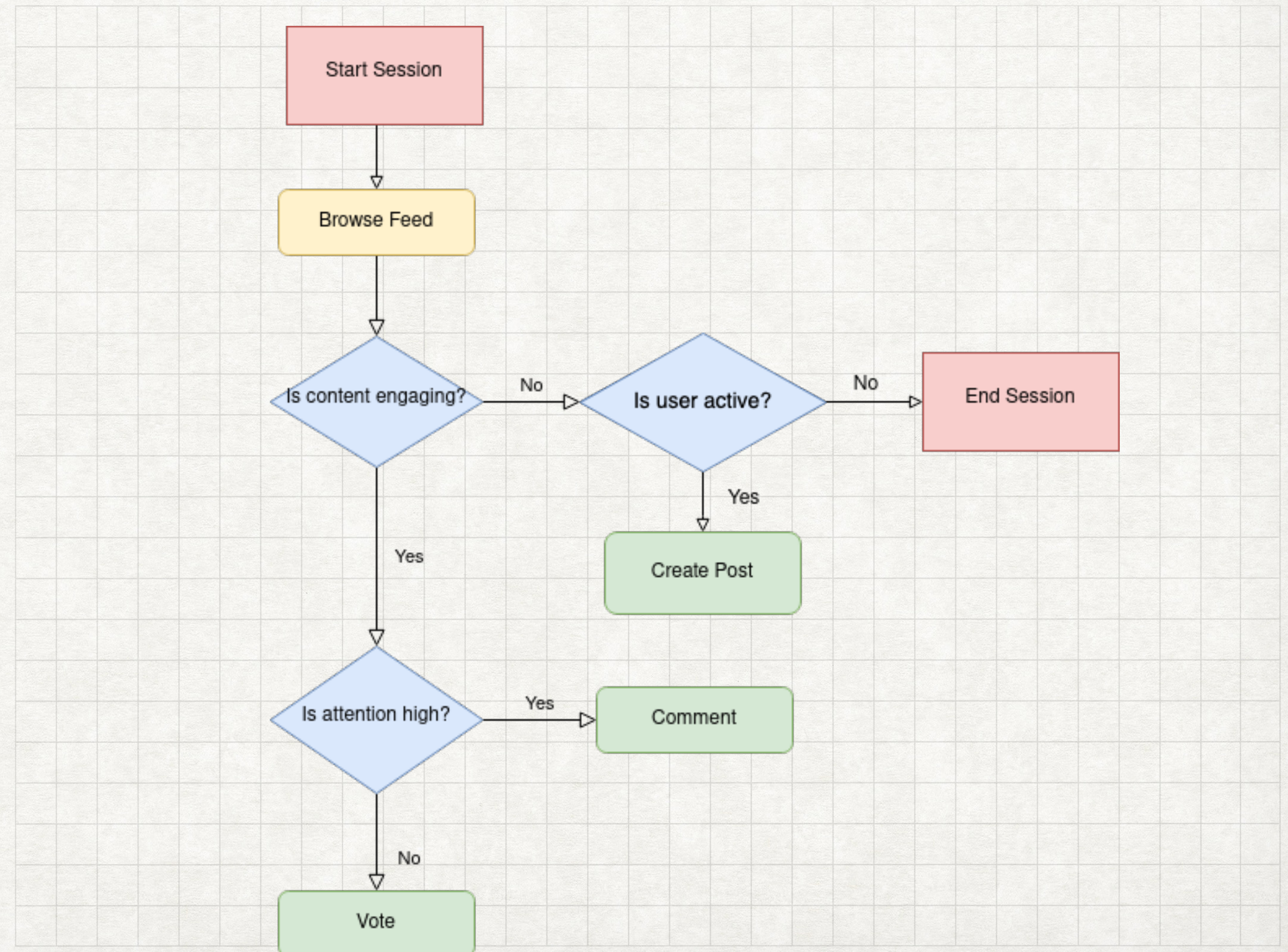




# SIMULATING SOCIAL NETWORKS

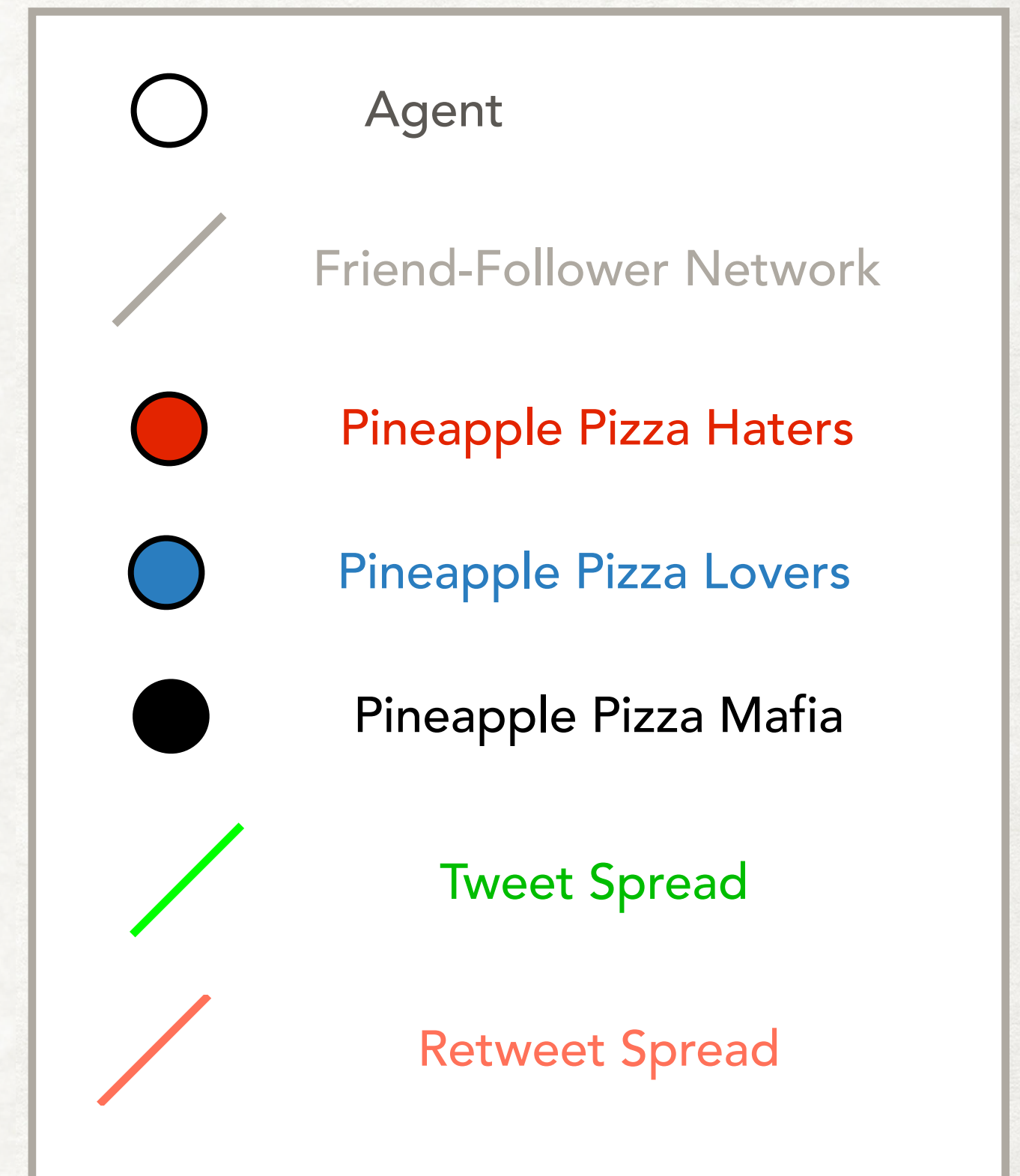
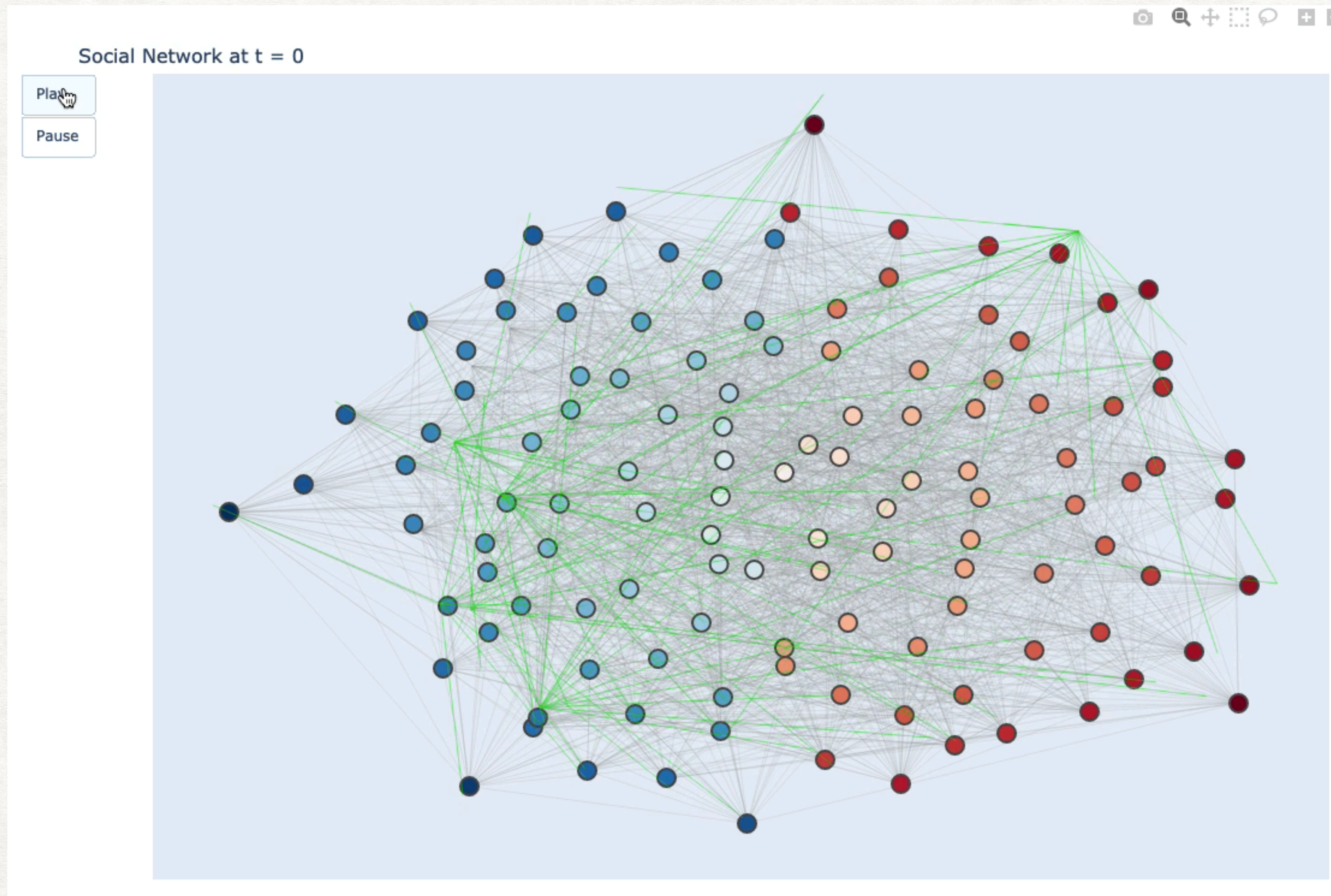
## REDDIT

- 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!





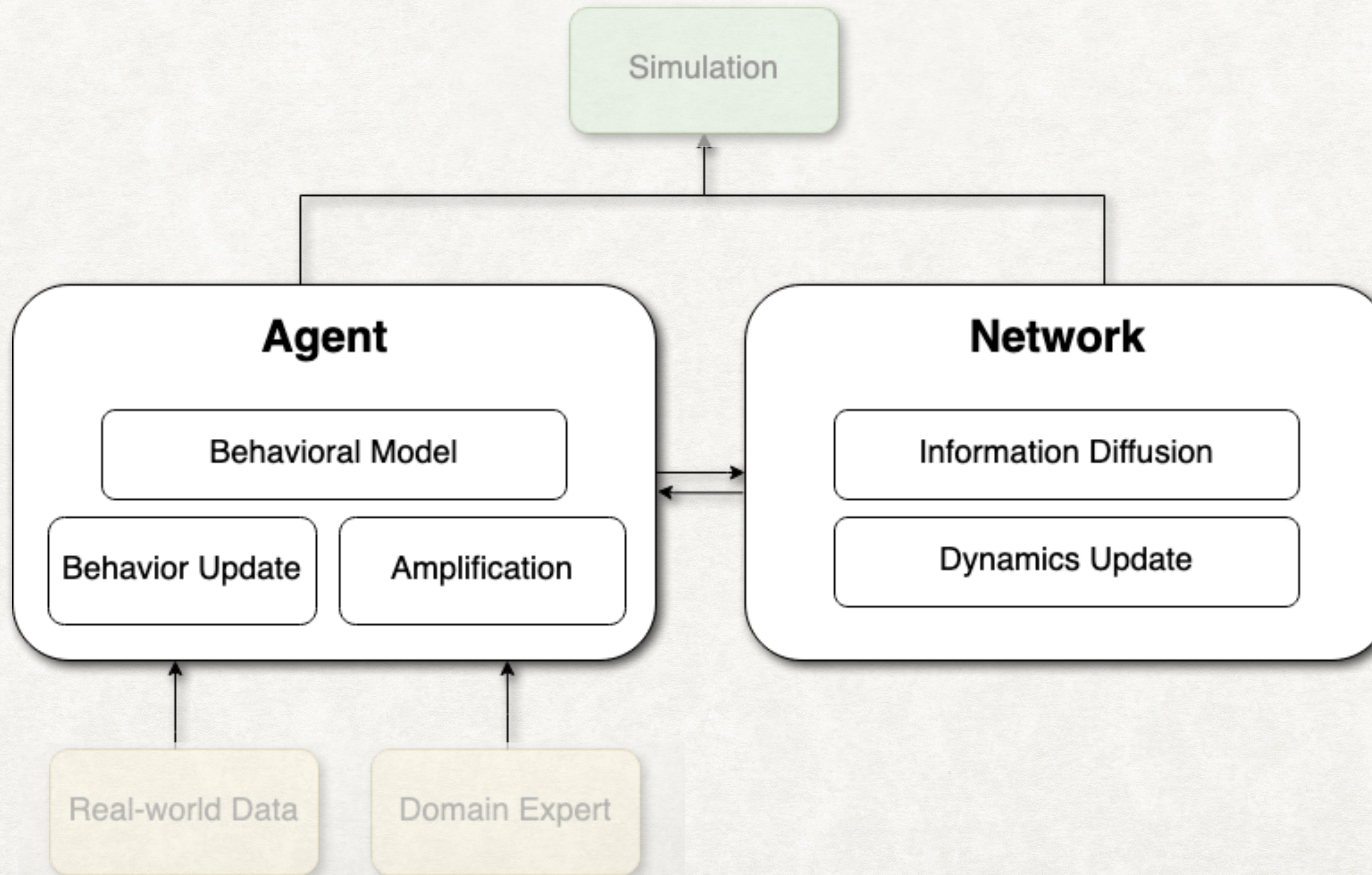
# SIMULATING SOCIAL NETWORKS



<https://youtu.be/GV5XuftiD7s>

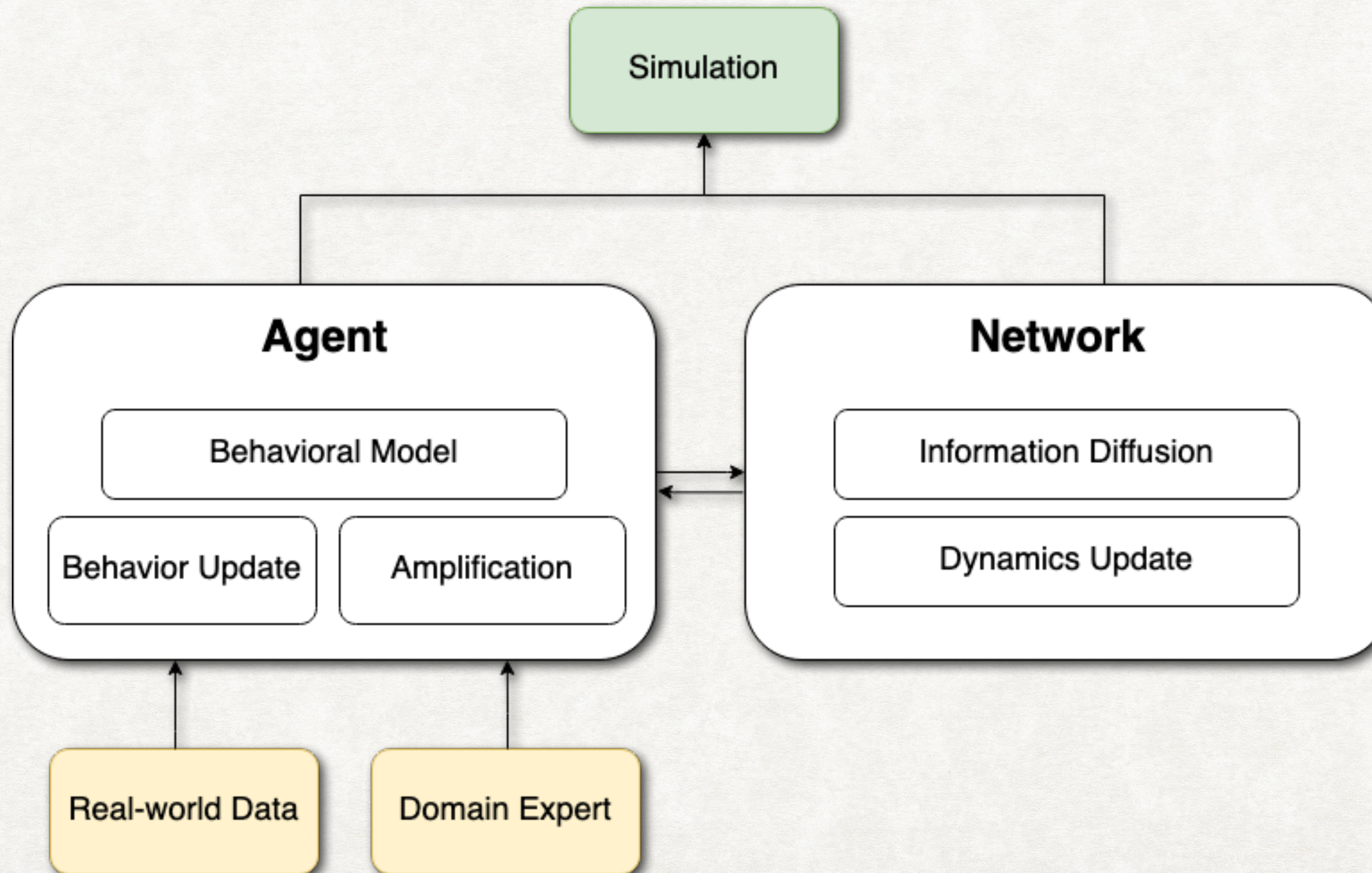


# SIMPPL





# 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


 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 DATA COLLECTION

r/politics

**r/politics  
Posts**

**+**

**r/politics  
Comments**

**2.7M posts**

**23M comments**



# REDDIT DATA COLLECTION

r/politics

**r/politics  
Posts**

2.7M posts

+

**r/politics  
Comments**

23M comments



**Author**

255K users

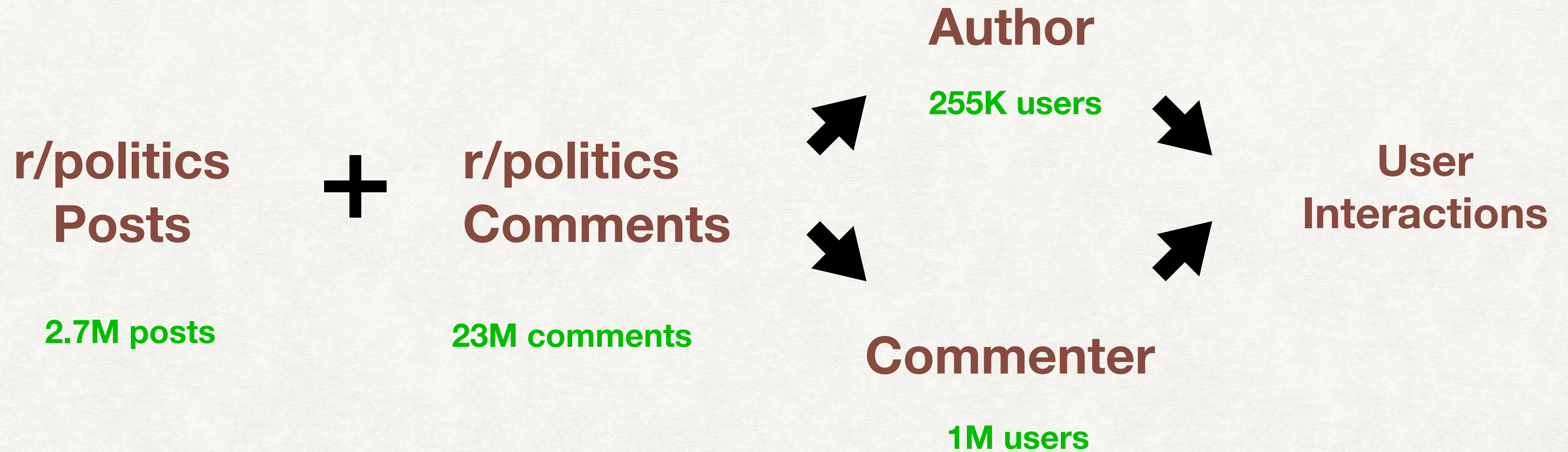
**Commenter**

1M users



# REDDIT DATA COLLECTION

r/politics





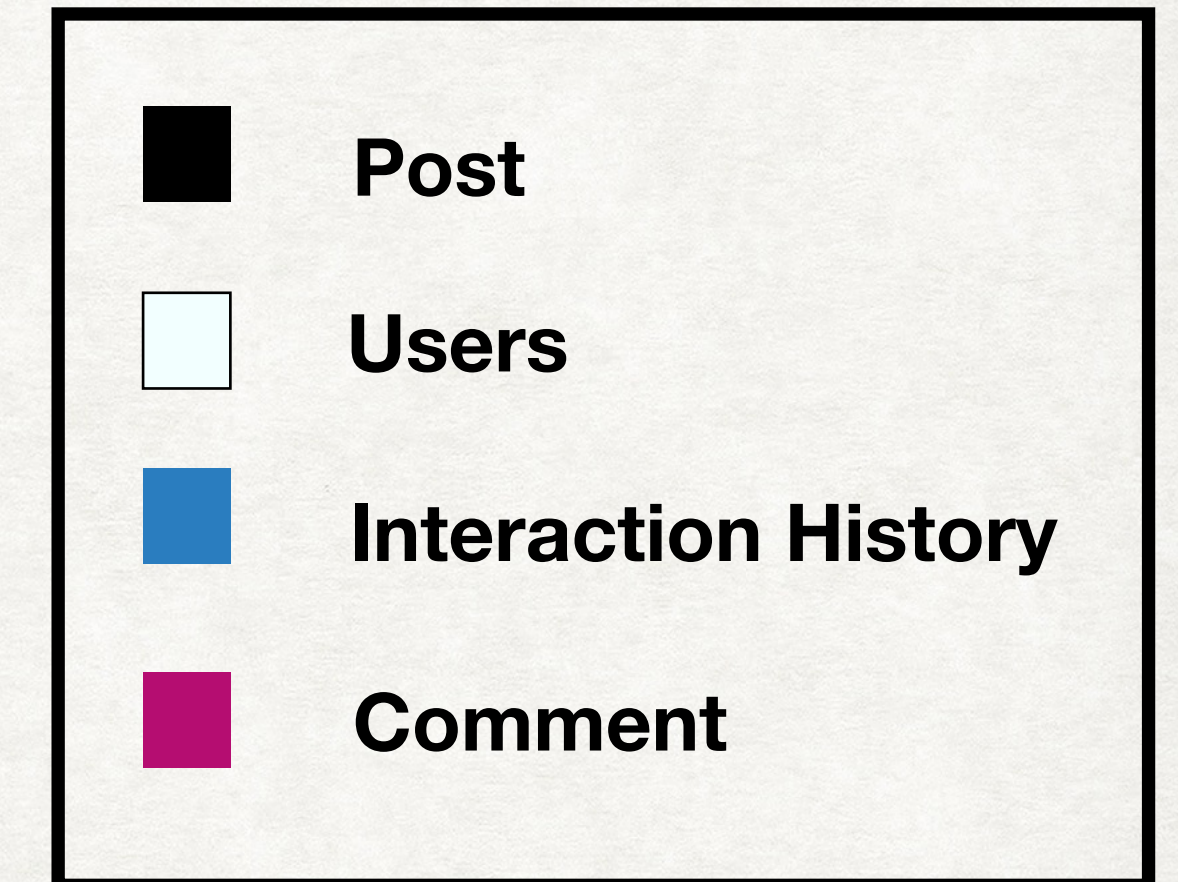
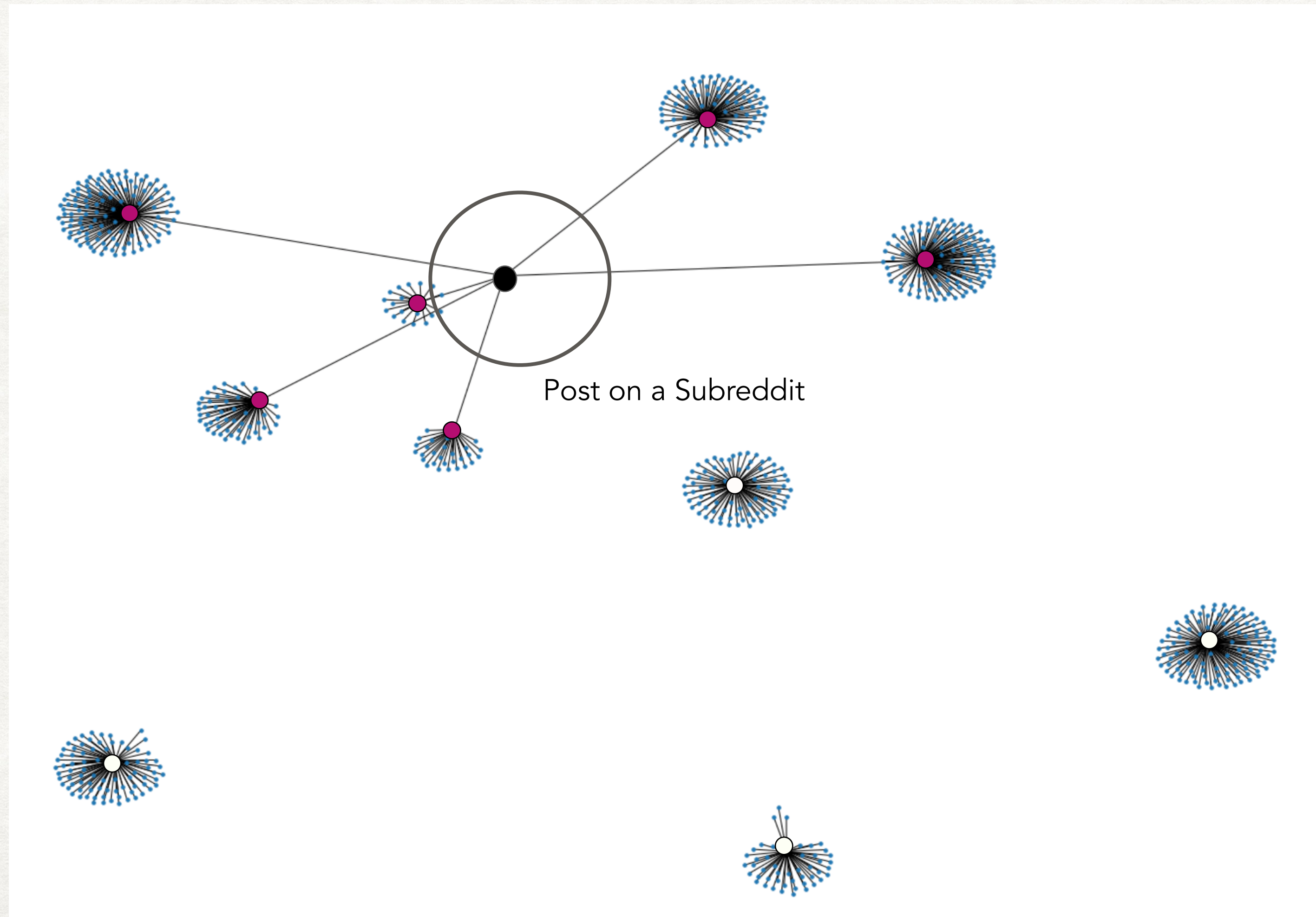
# REDDIT DATA COLLECTION

r/politics



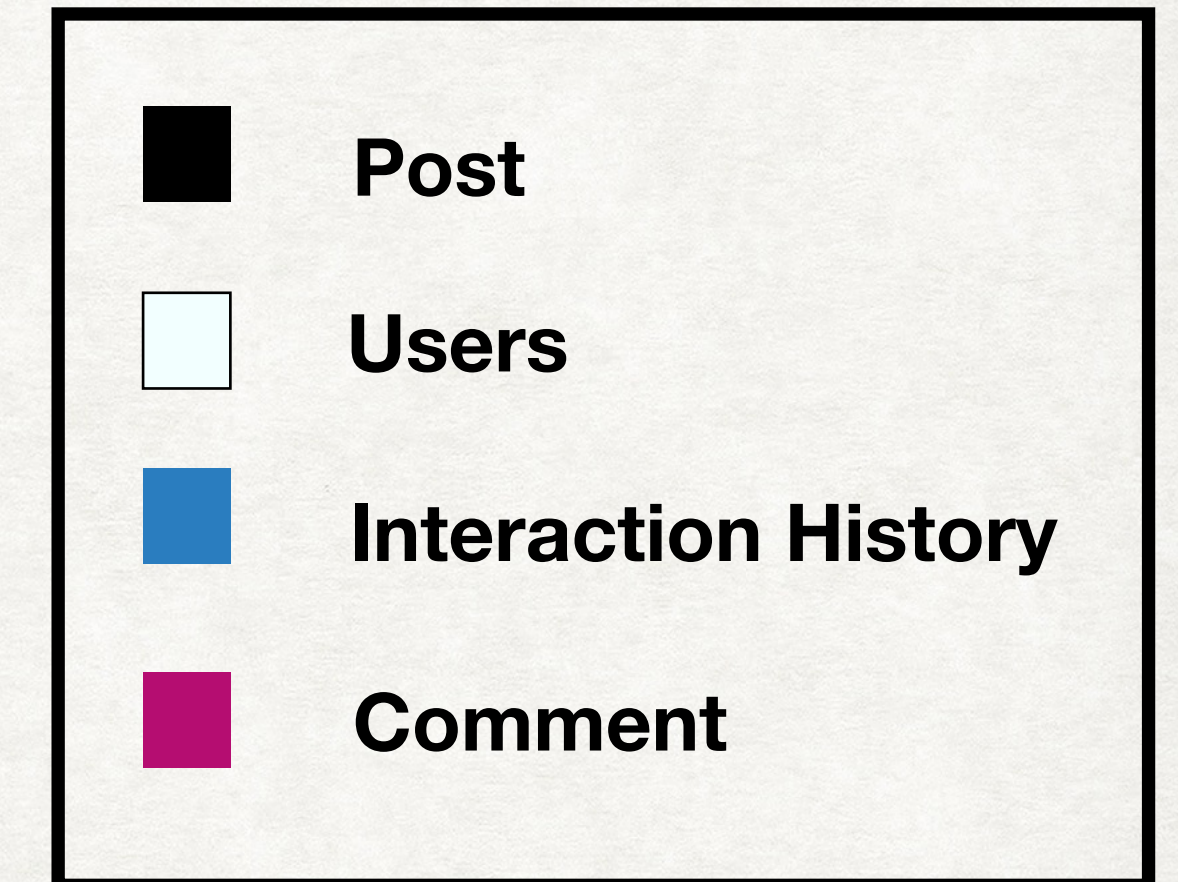
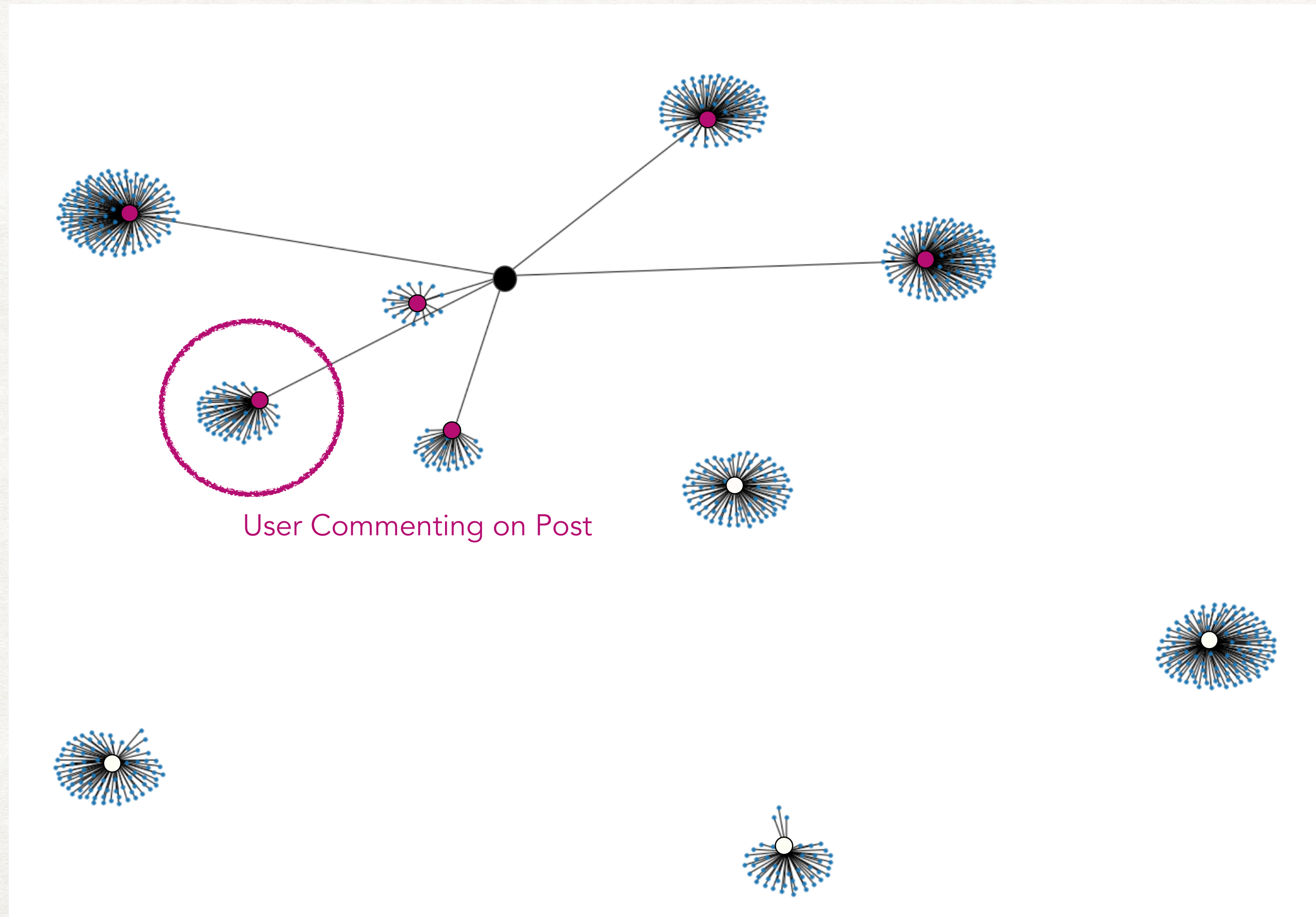


# REDDIT POST



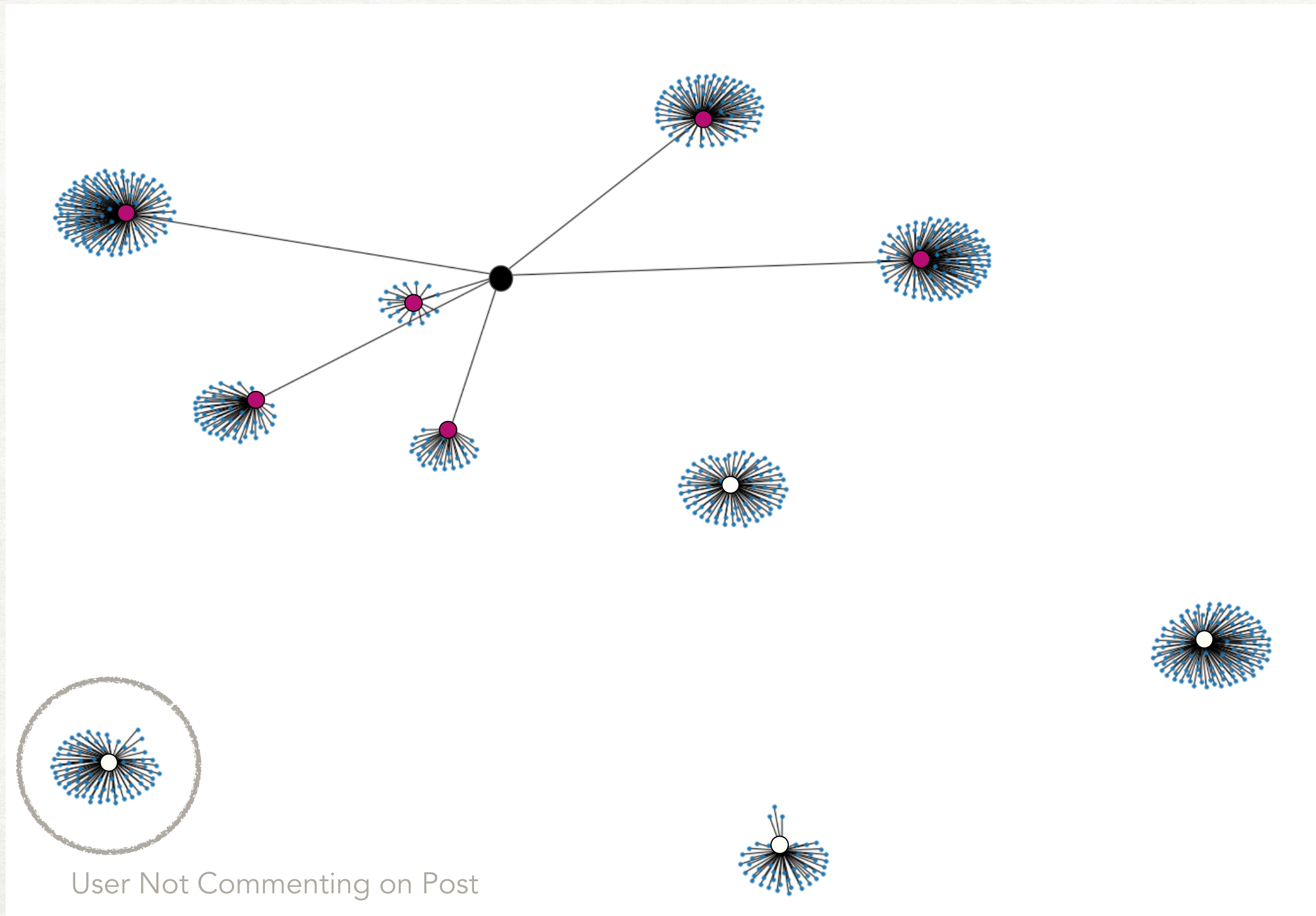


# REDDIT POST



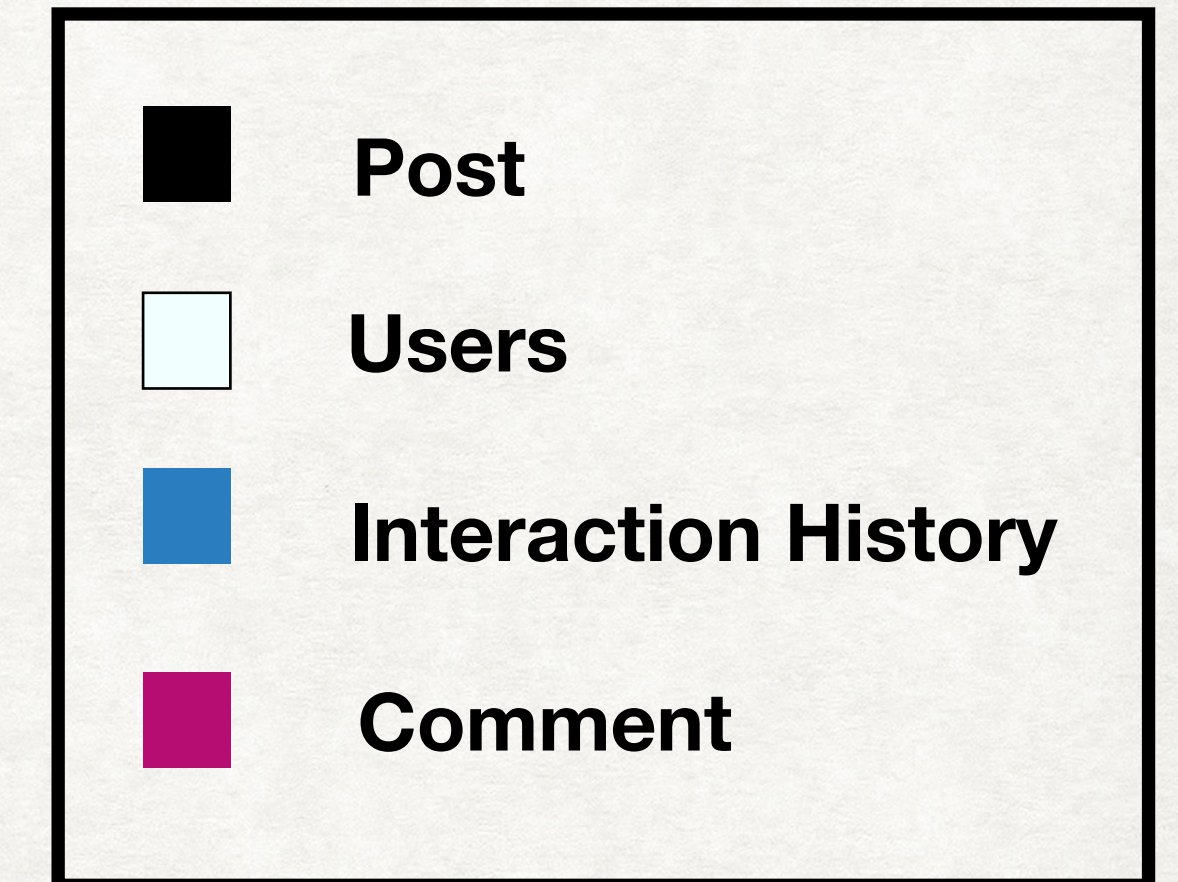
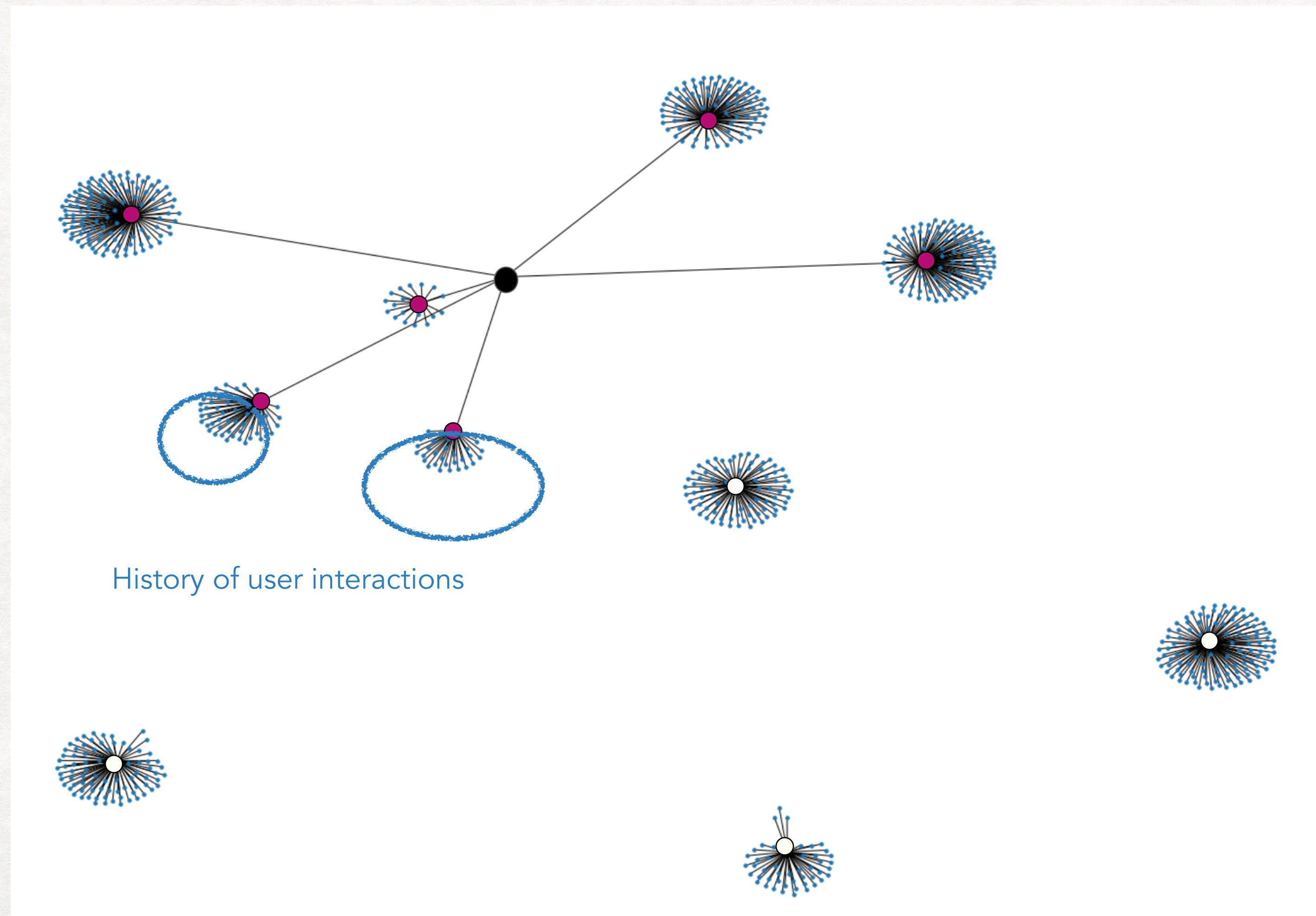


# REDDIT POST



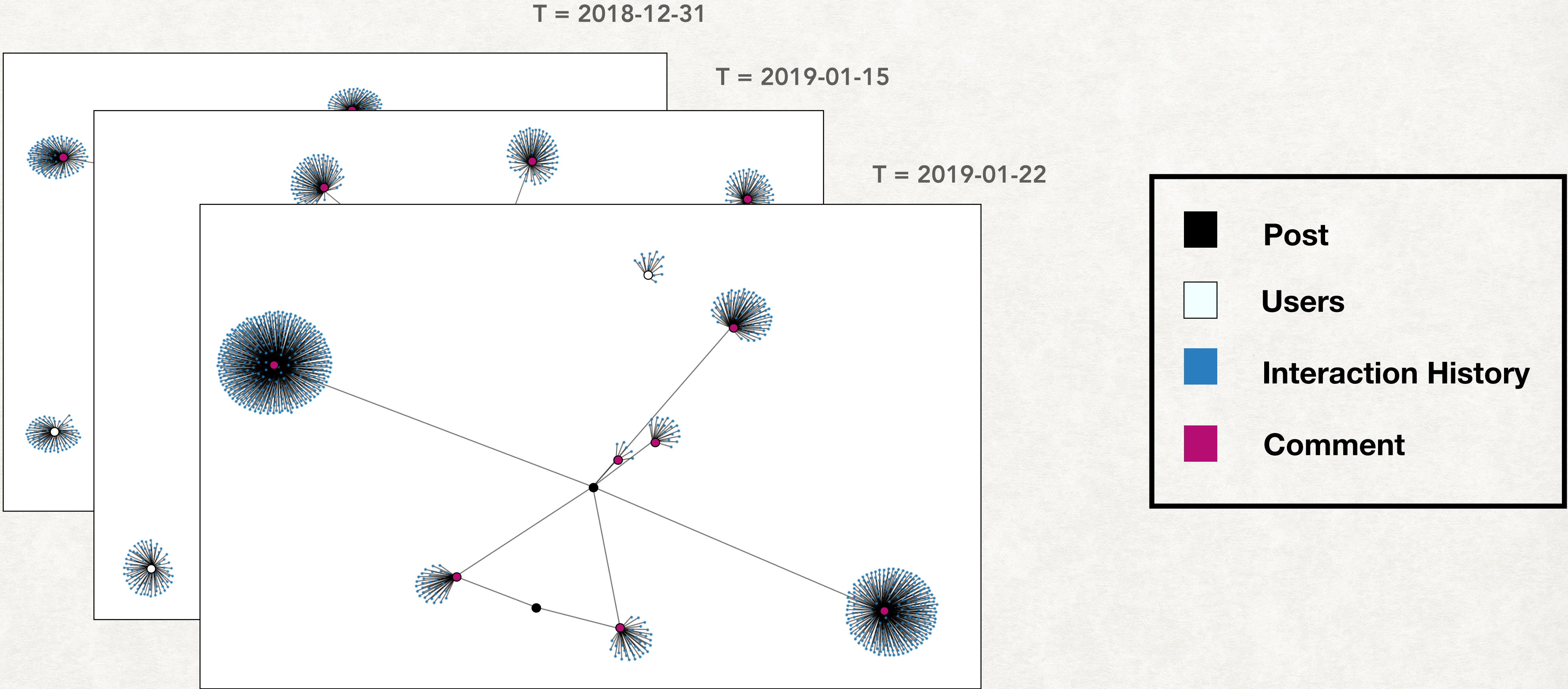


# REDDIT POST





# POLITICS POSTS

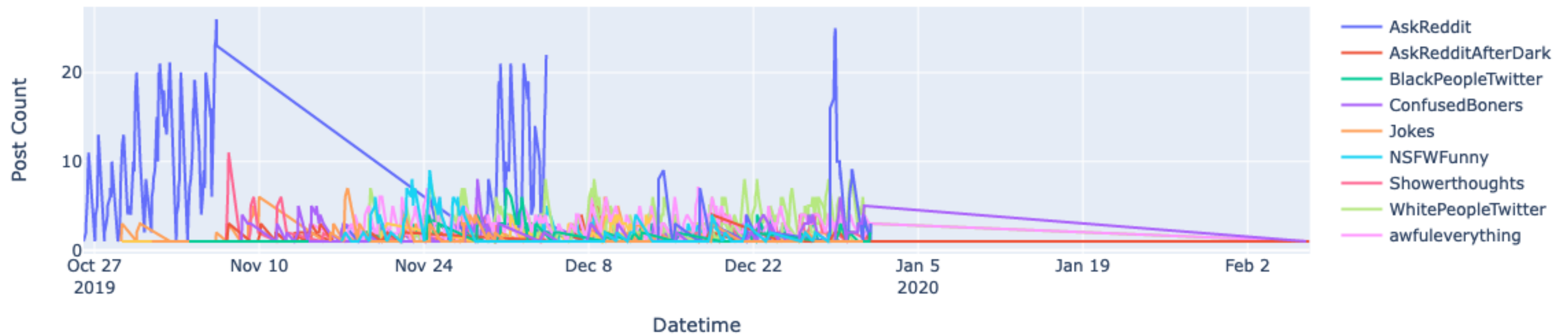




# REAL-WORLD MODELING

## REDDIT USER 'JONNYCREEPYCREPES3'

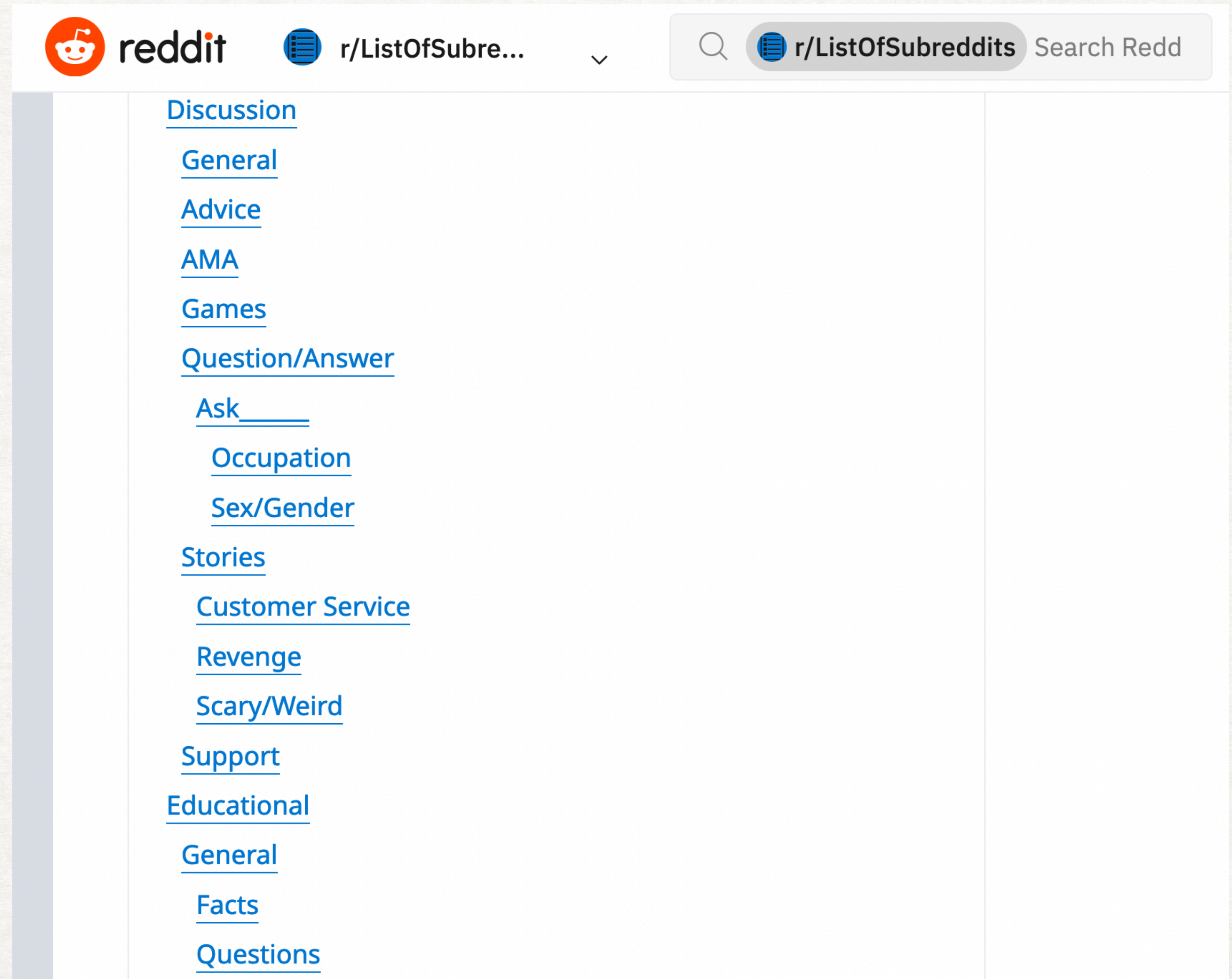
Subreddit-wise Post Count in 12H bins by jonnycreepycrepes3





# REDDIT WIKI PAGES

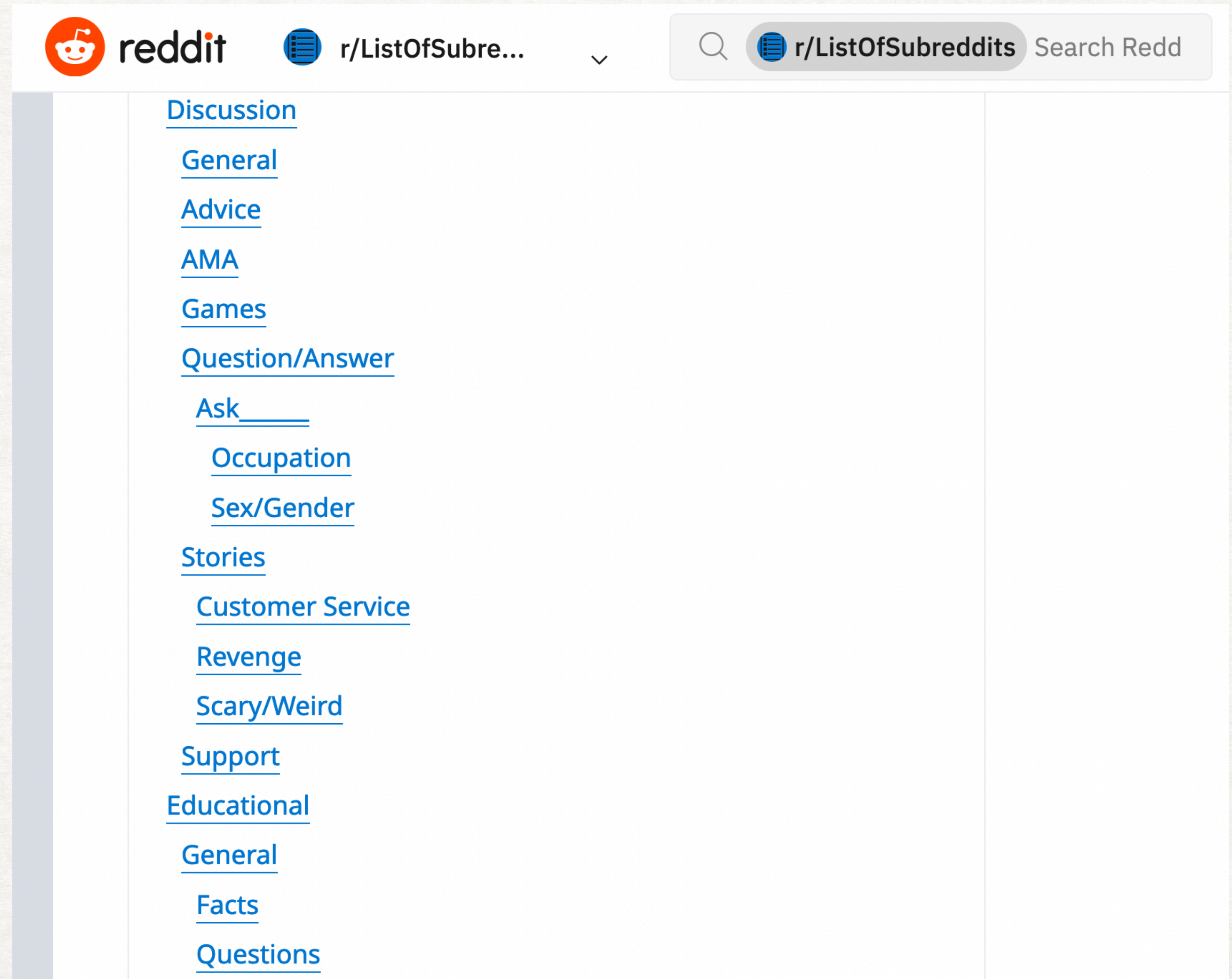
- User-driven hierarchical categorization of subreddits
- Discussion > Stories > Customer Service
- 4998 subreddits
- 5-level hierarchy of categories





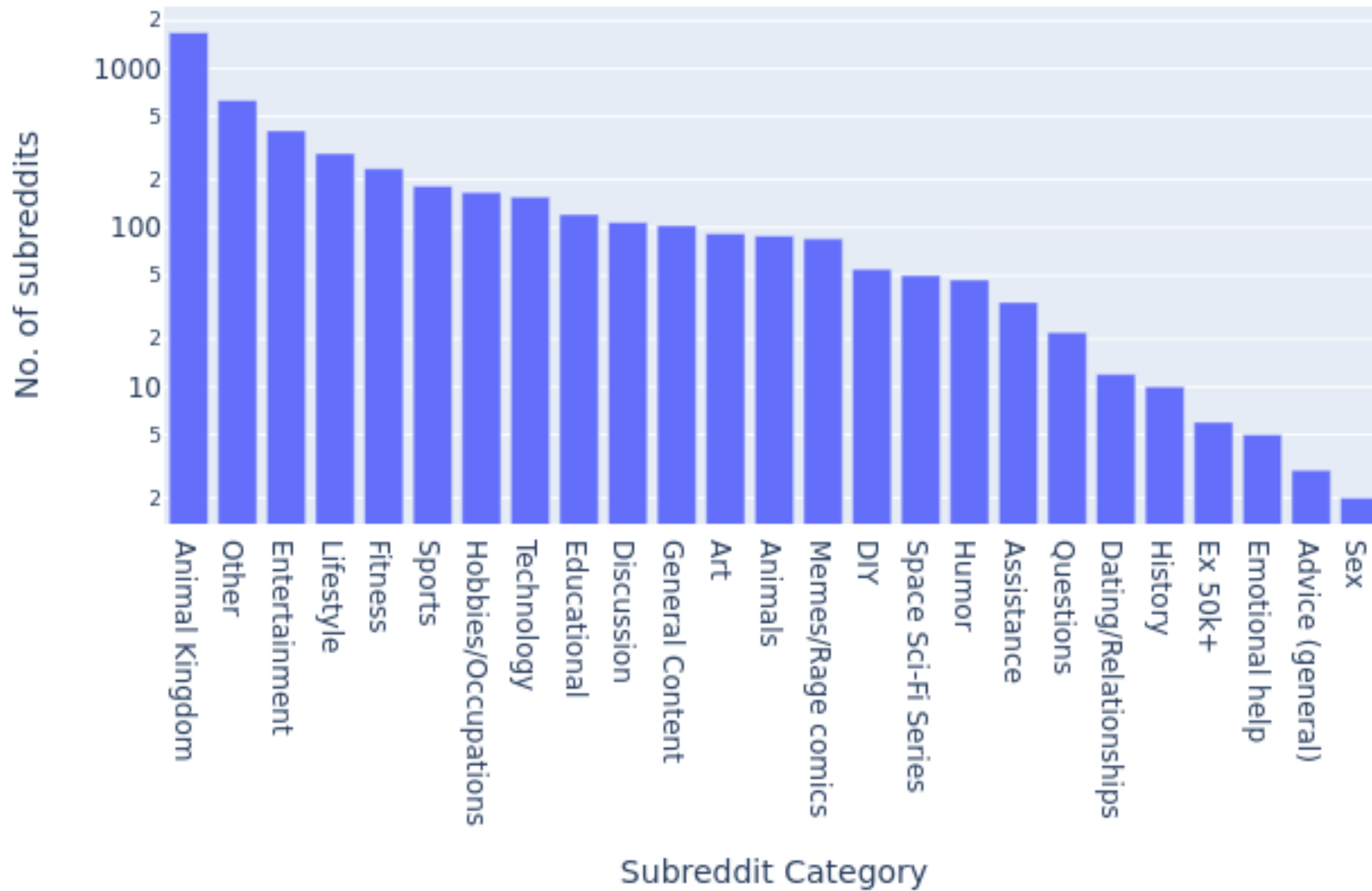
# REDDIT WIKI PAGES

- 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



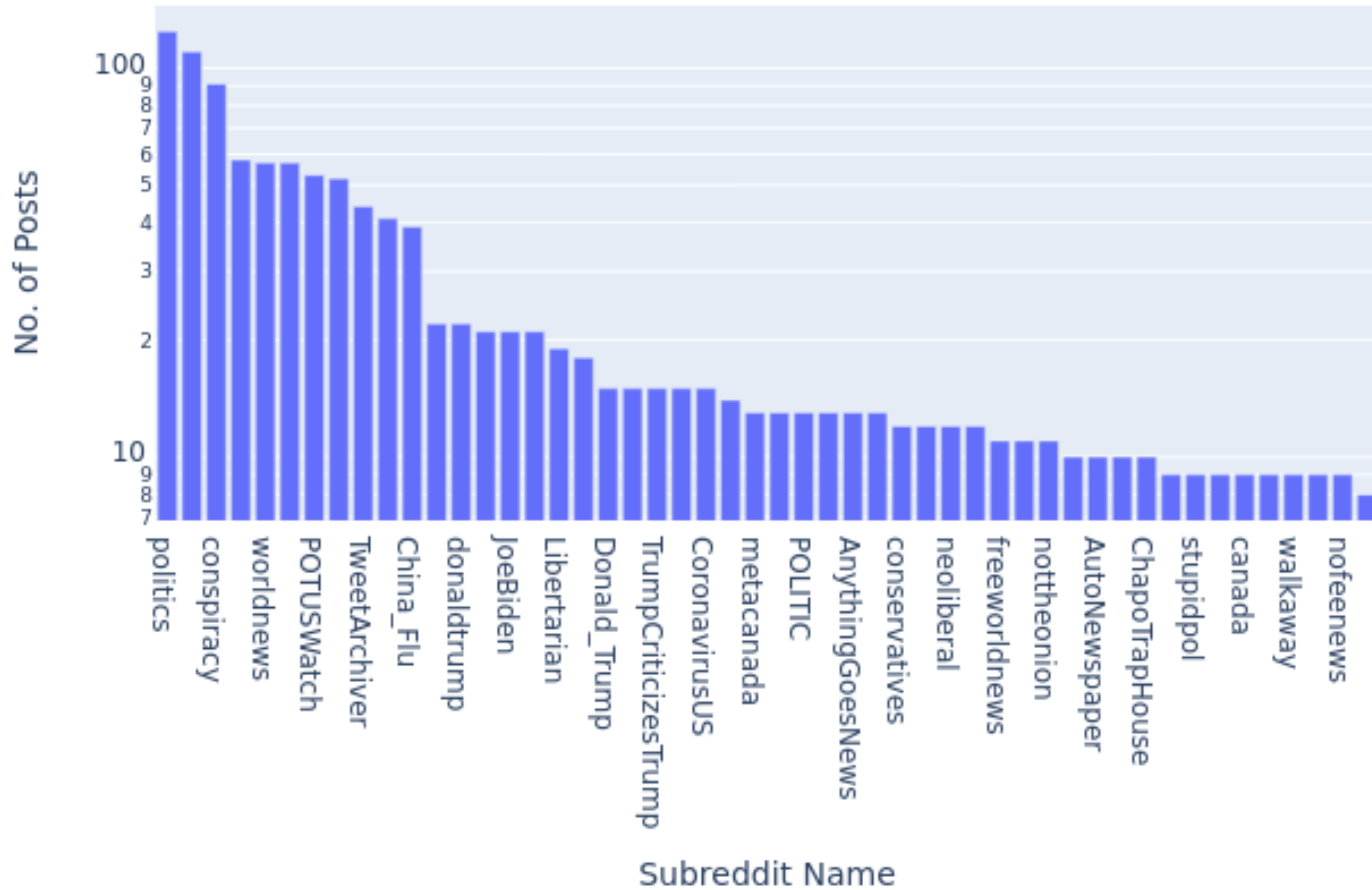


# CATEGORIZE POSTS ACROSS SUBREDDITS





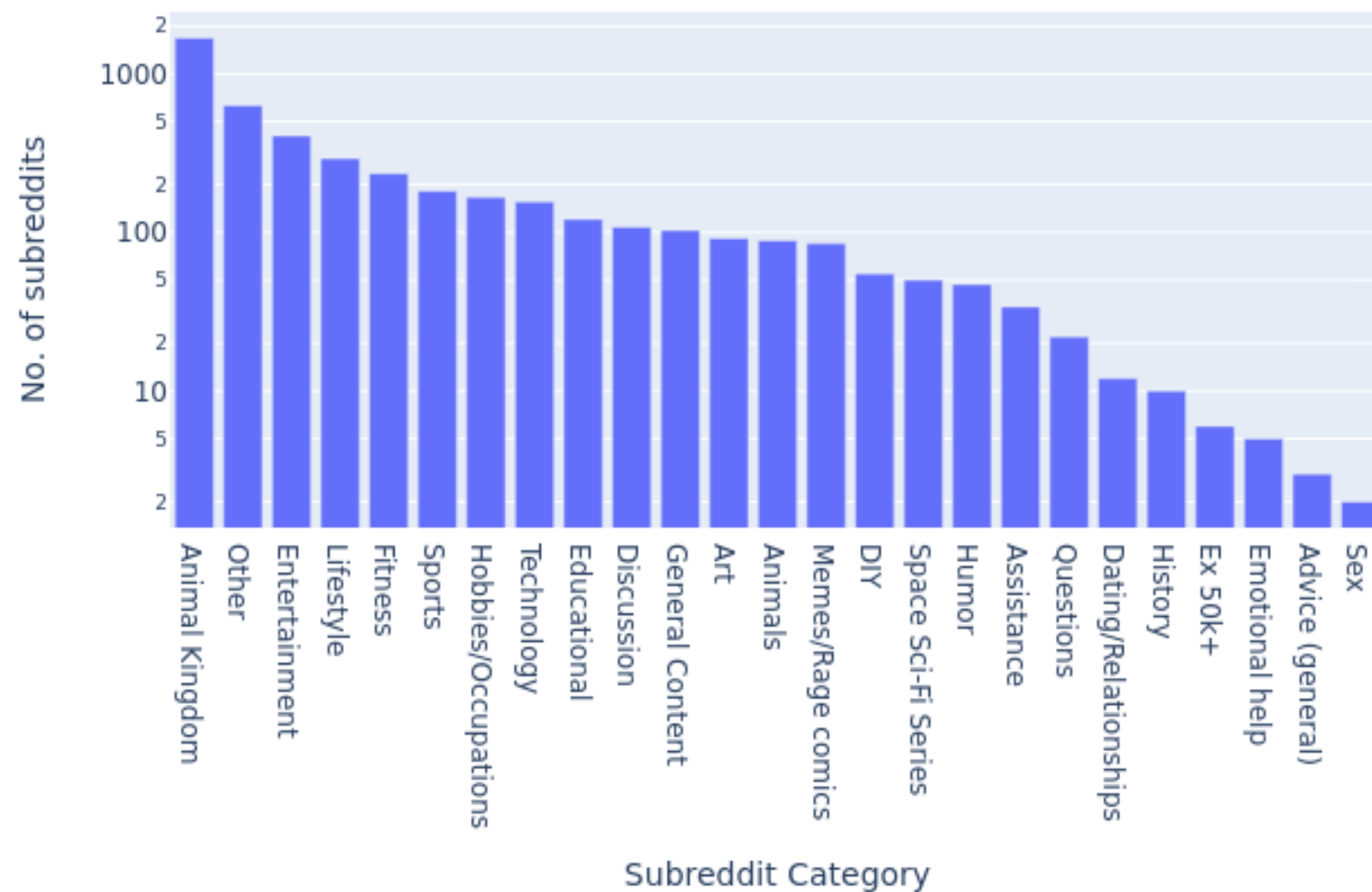
# CATEGORIZE POSTS ACROSS SUBREDDITS





# CATEGORIZE POSTS ACROSS SUBREDDITS

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



['mormon', 'politics']

## Predicted Subreddits:

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

## Predicted Category:

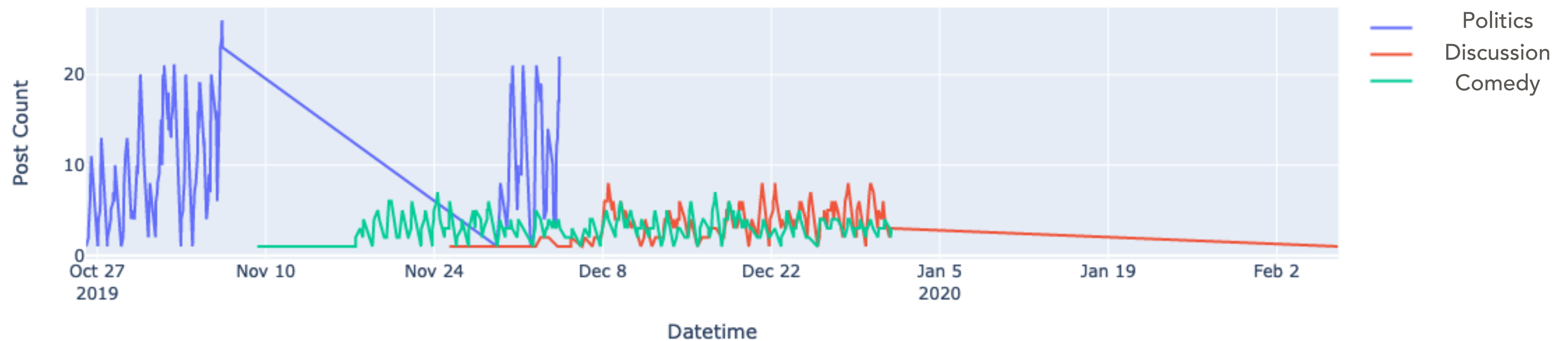
[[None, 'History of People', None, None,  
None],



# MODELING CATEGORIES

## REDUCING SUBREDDIT DIMENSIONALITY

Subreddit-wise Post Count in 12H bins by jonnycreepycrepes3

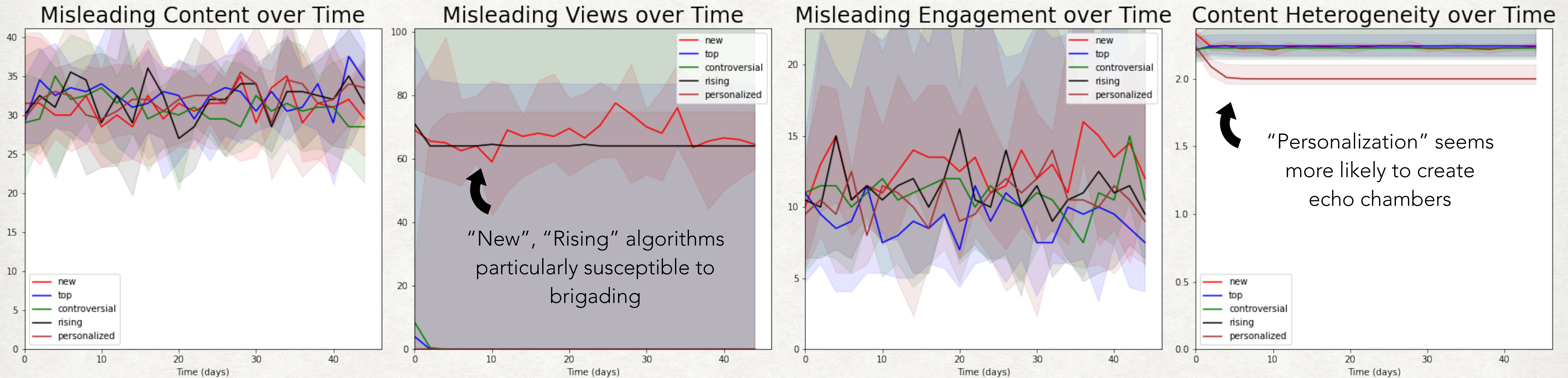




# ESTIMATE ALGORITHMIC SUSCEPTIBILITY TO COORDINATED ATTACKS ON REDDIT



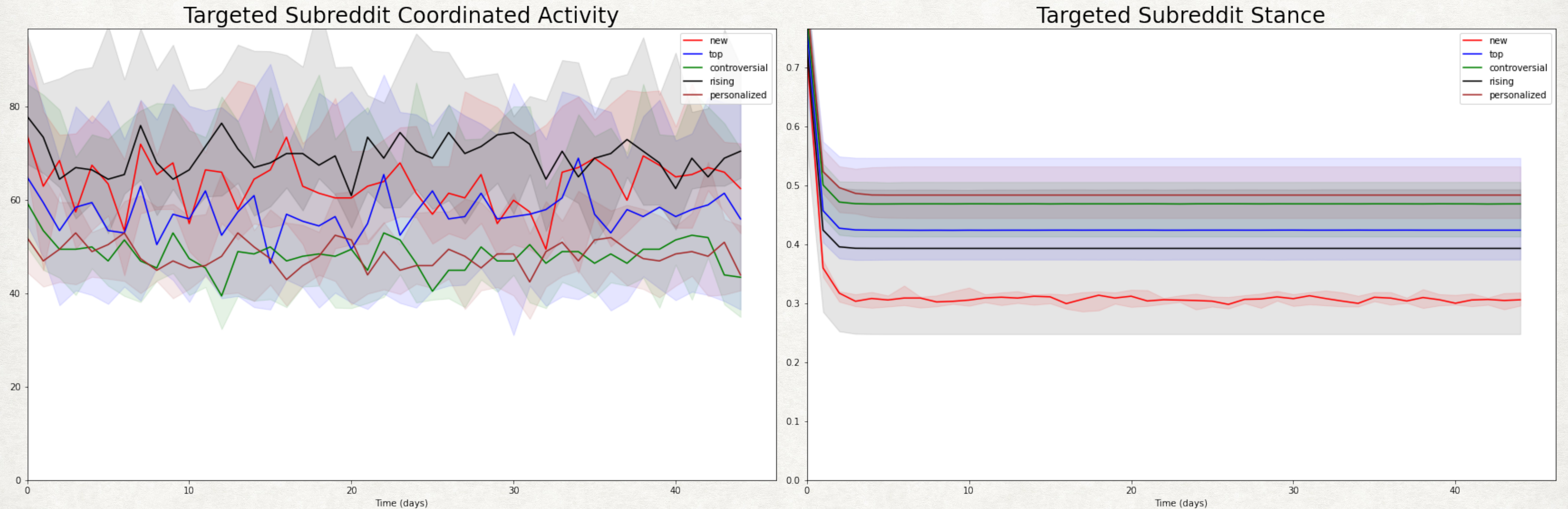
# WHAT ARE THE EFFECTS OF BRIGADING?



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

Facebook  
Algorithm

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

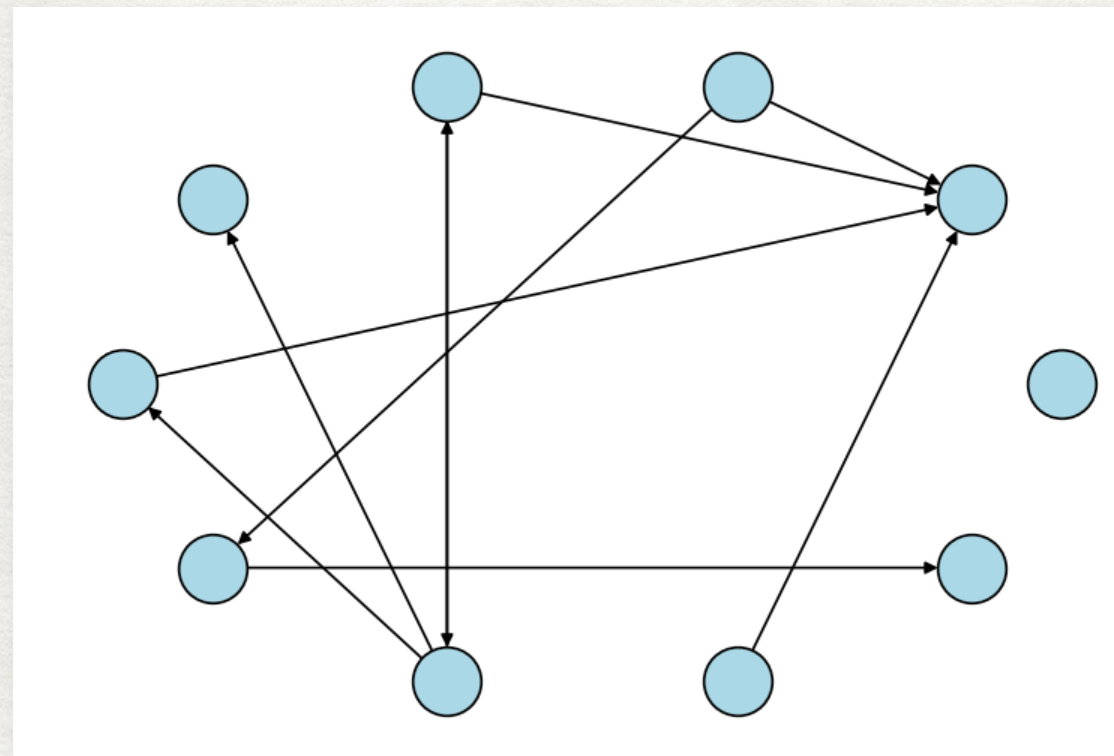


# SUMMARY

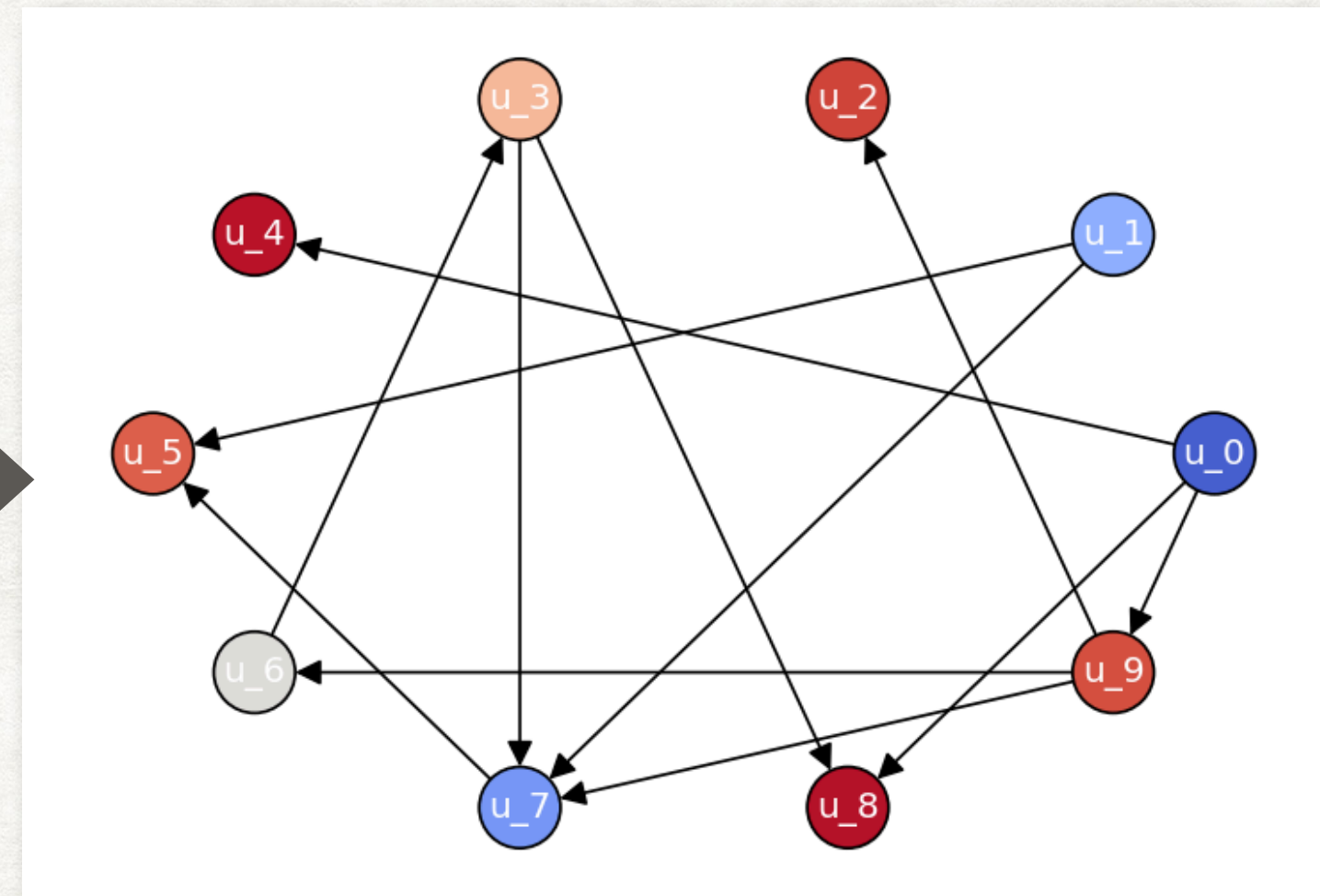
- We simulate coordinated campaigns seeking to manipulate public debate using multiple authentic/inauthentic (fake) accounts to mislead people.
- Goal: Quantify the harms arising from CLB on Social Networks
  - Measure its effects on ranking and recommendation algorithms
  - Use real-world networks and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



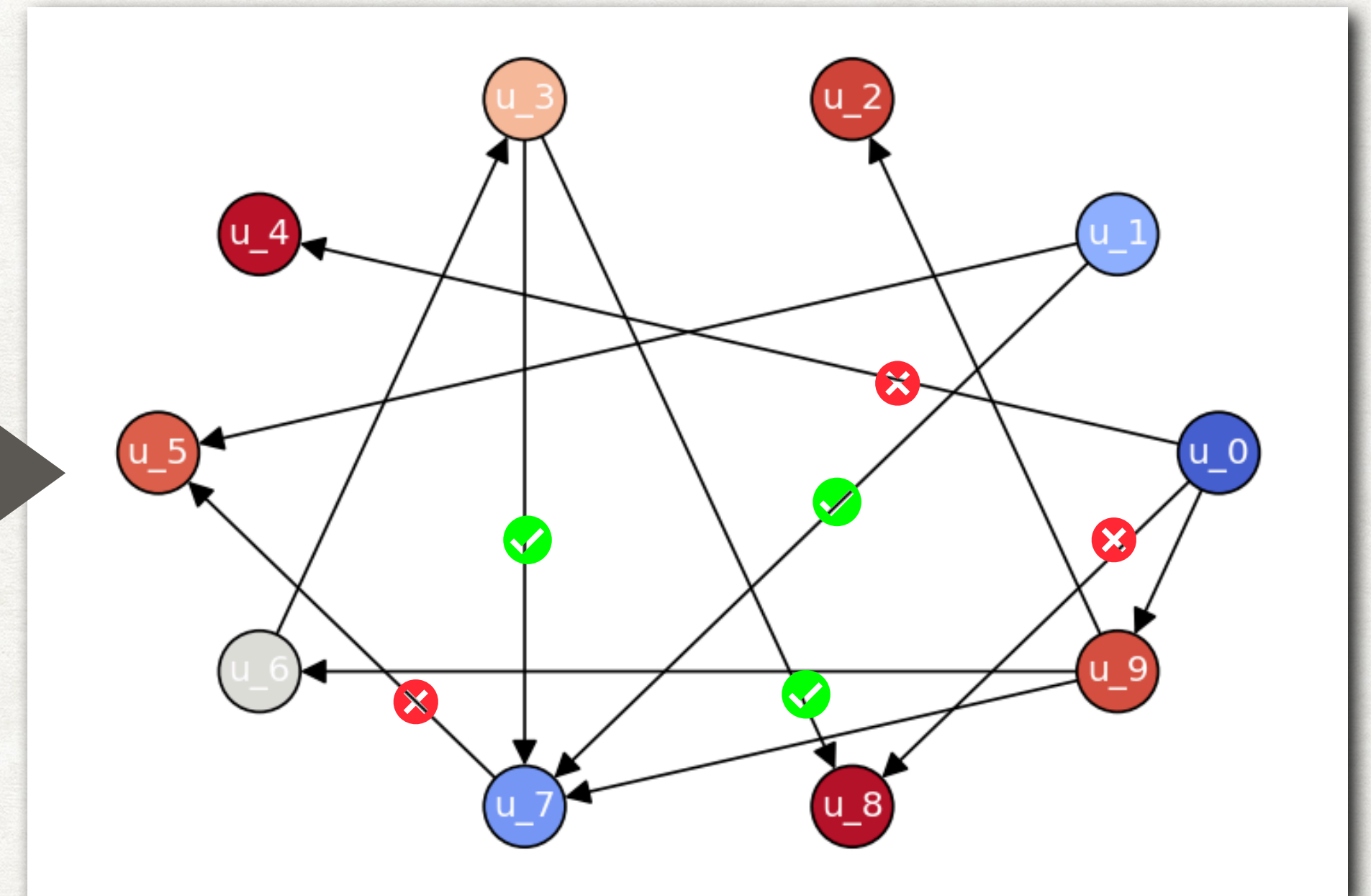
# INTERVENTIONS TO LIMIT DISINFORMATION



Agents + Networks



Agents + Networks + Behaviors



Agents + Networks + Behaviors + Interventions



# 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



# 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](#)