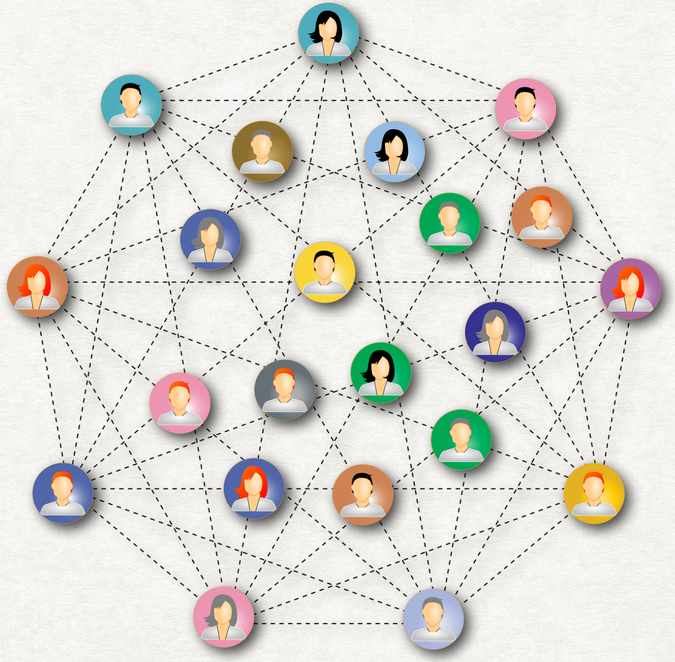
USING SIMULATIONS TO STUDY ONLINE DISINFORMATION



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

MISINFOCON ‘22

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

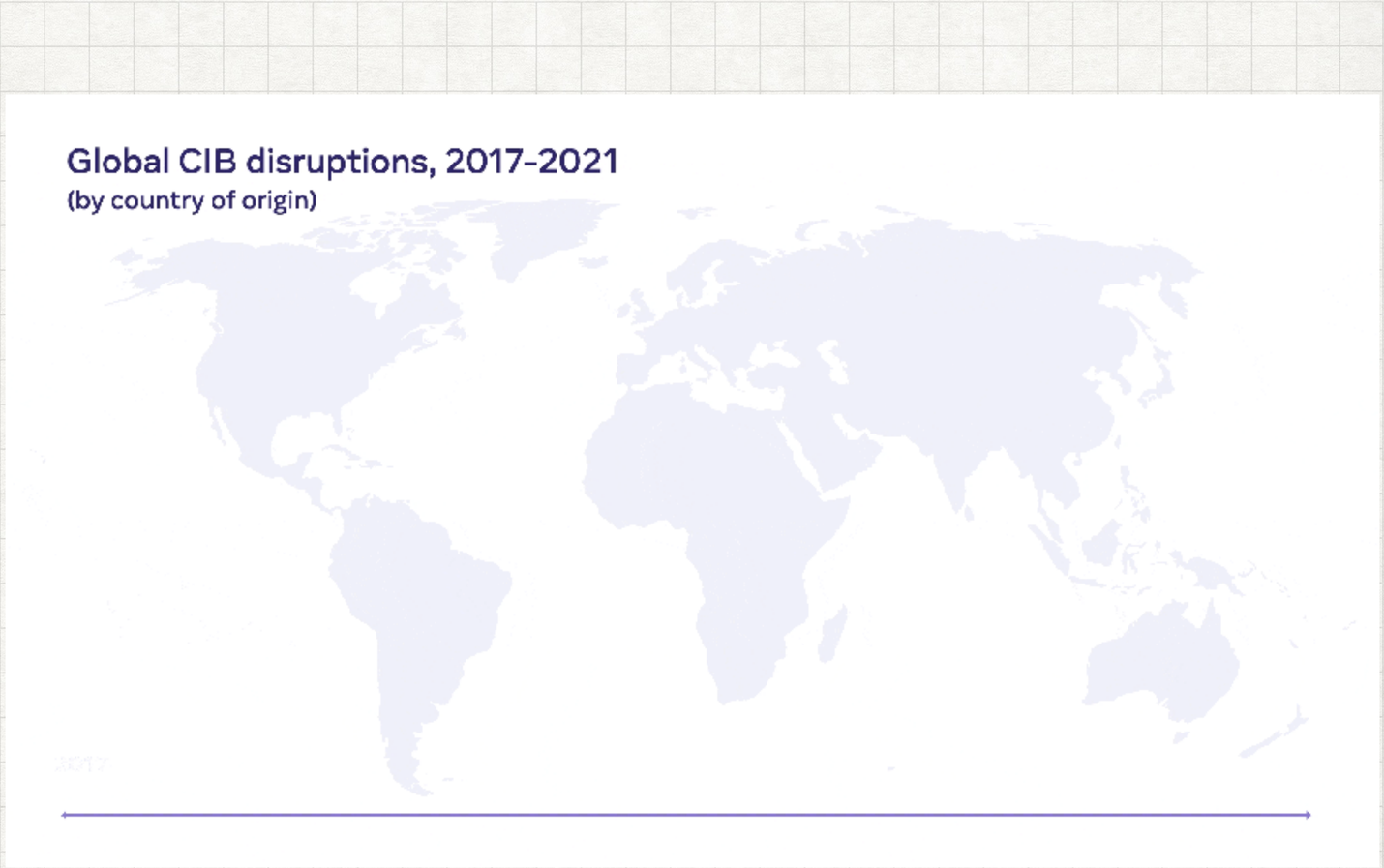
2018 - 19 ML + Physics at the European Org. for Nuclear Research

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.

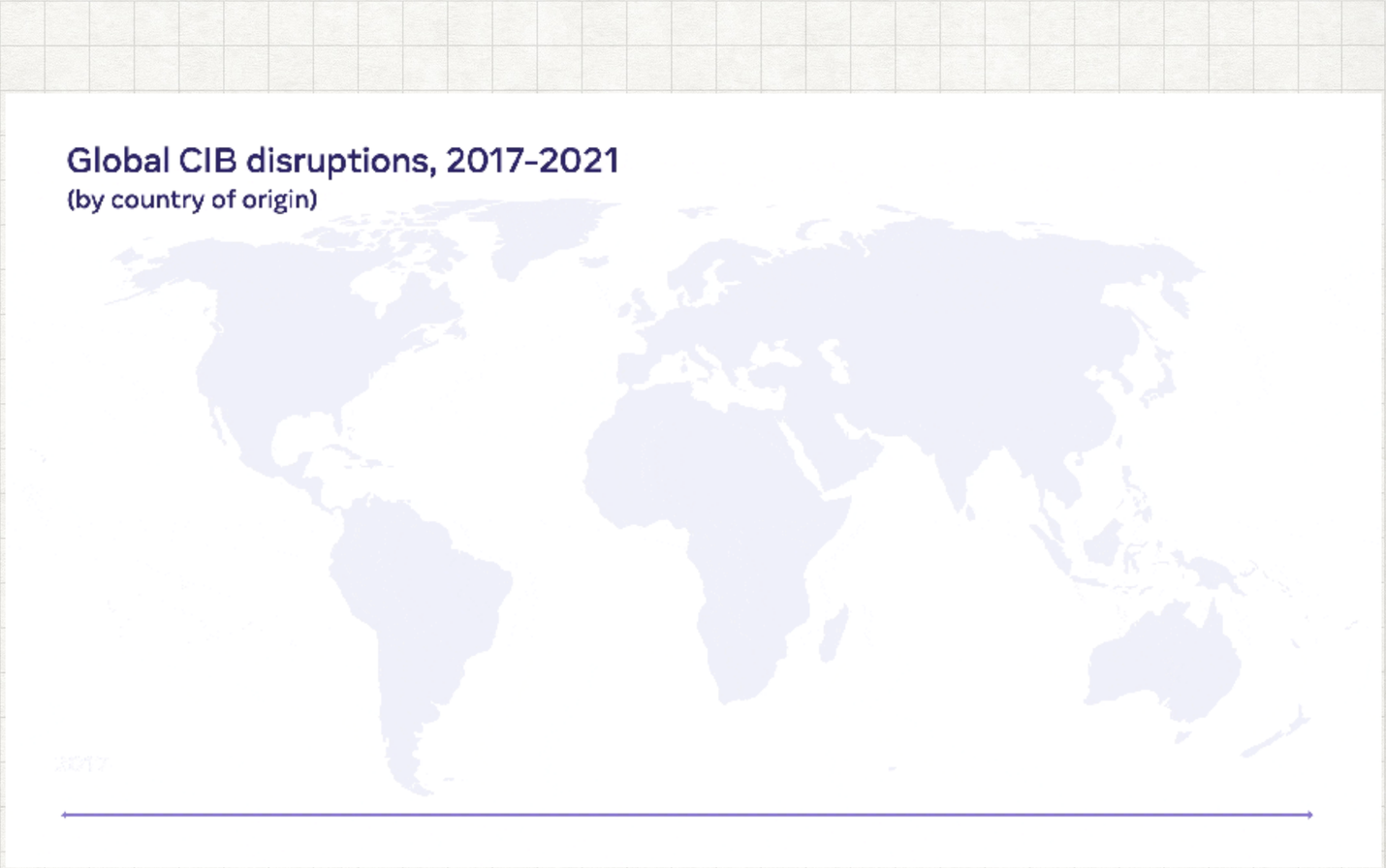
COORDINATED INAUTHENTIC BEHAVIOR

- 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.



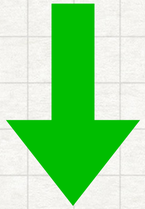
COORDINATED INAUTHENTIC BEHAVIOR

- 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

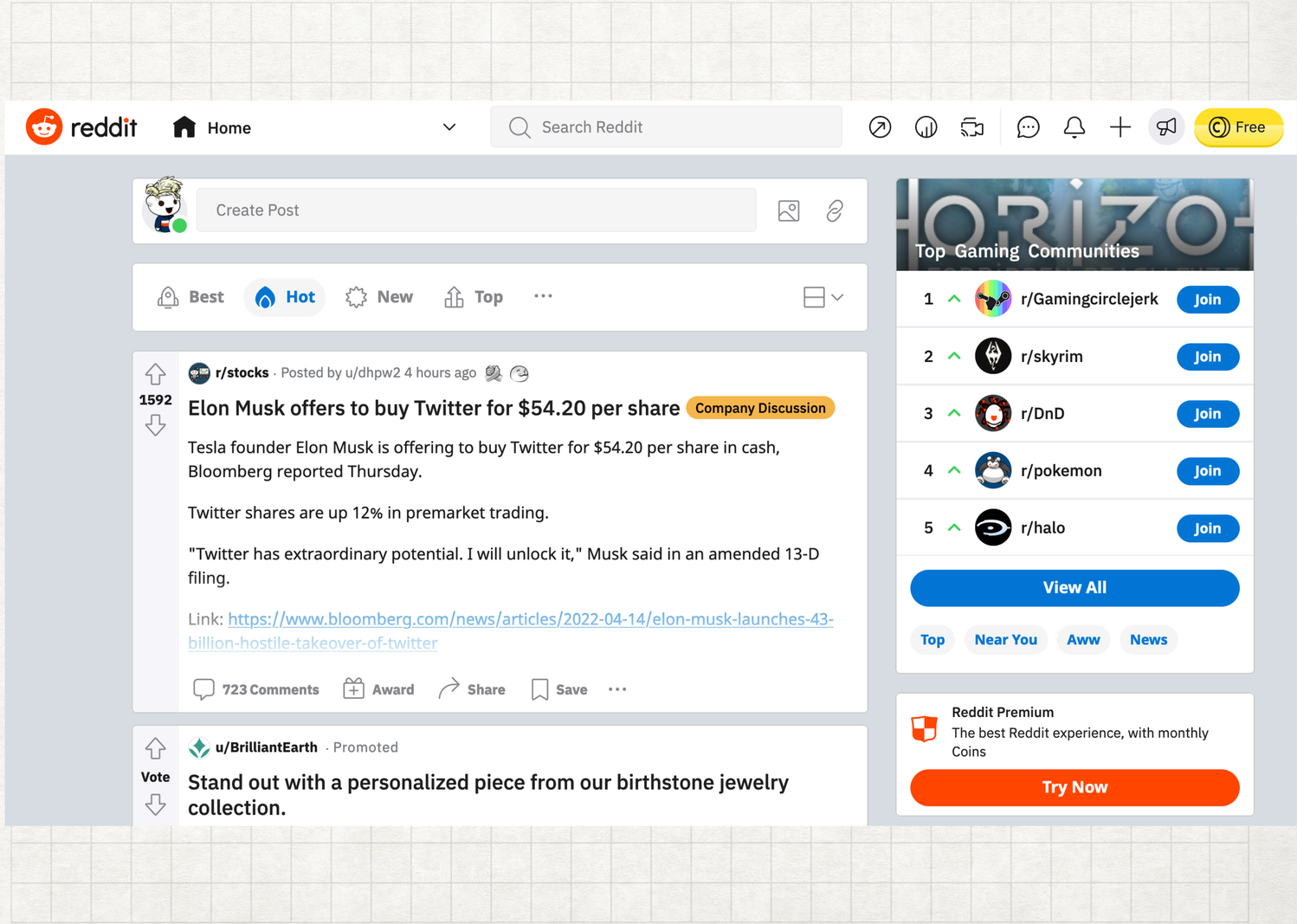
- RQ: Quantify the relative impact of different algorithmic choices that a platform makes, for instance:
  - To maximize engagement or not to maximize engagement?
  - Should we promote diverse content that isn't getting early views?
  - Dealing with "controversial" opinions?

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?

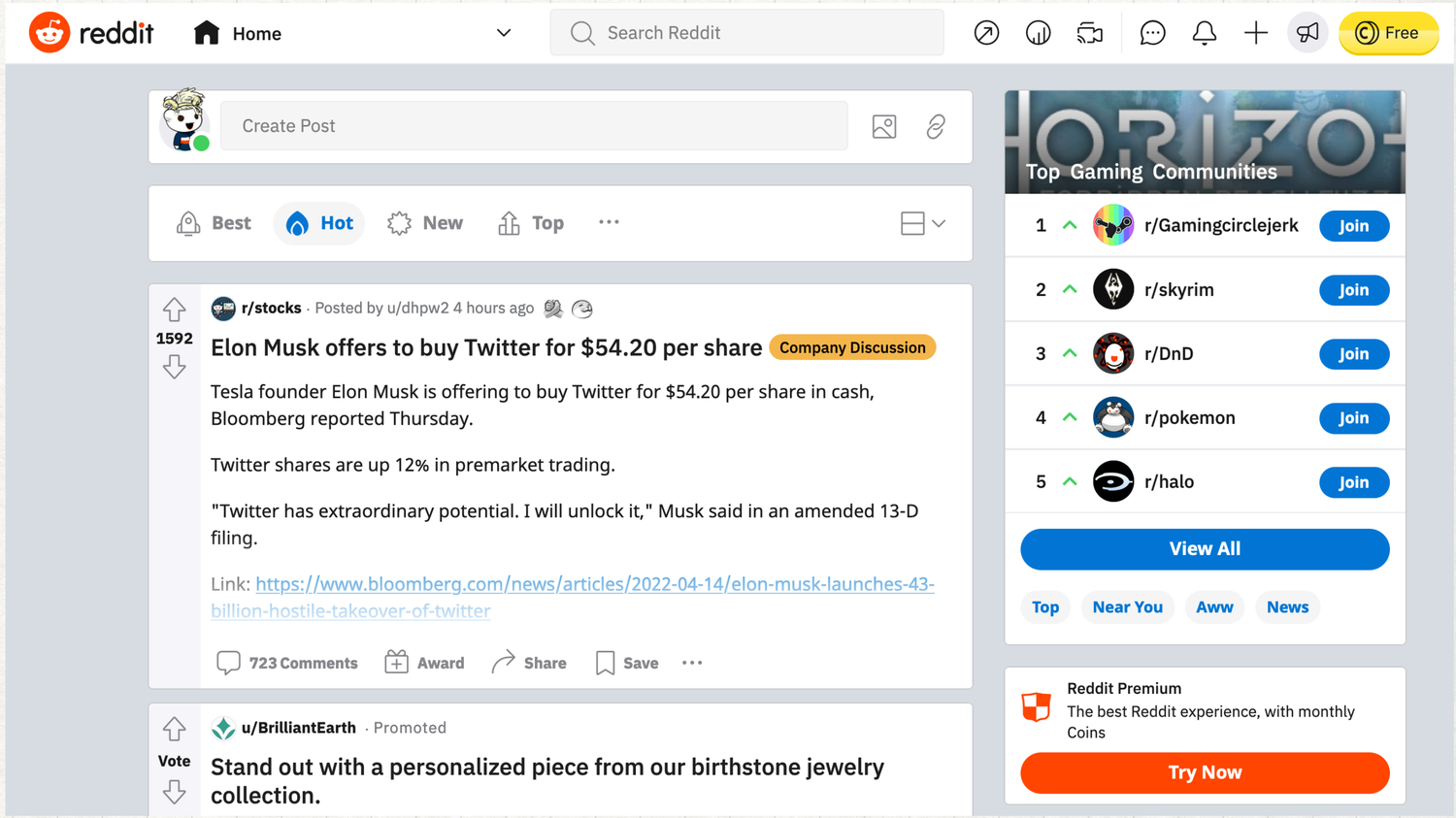
SIMULATE A SOCIAL NETWORK

- 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

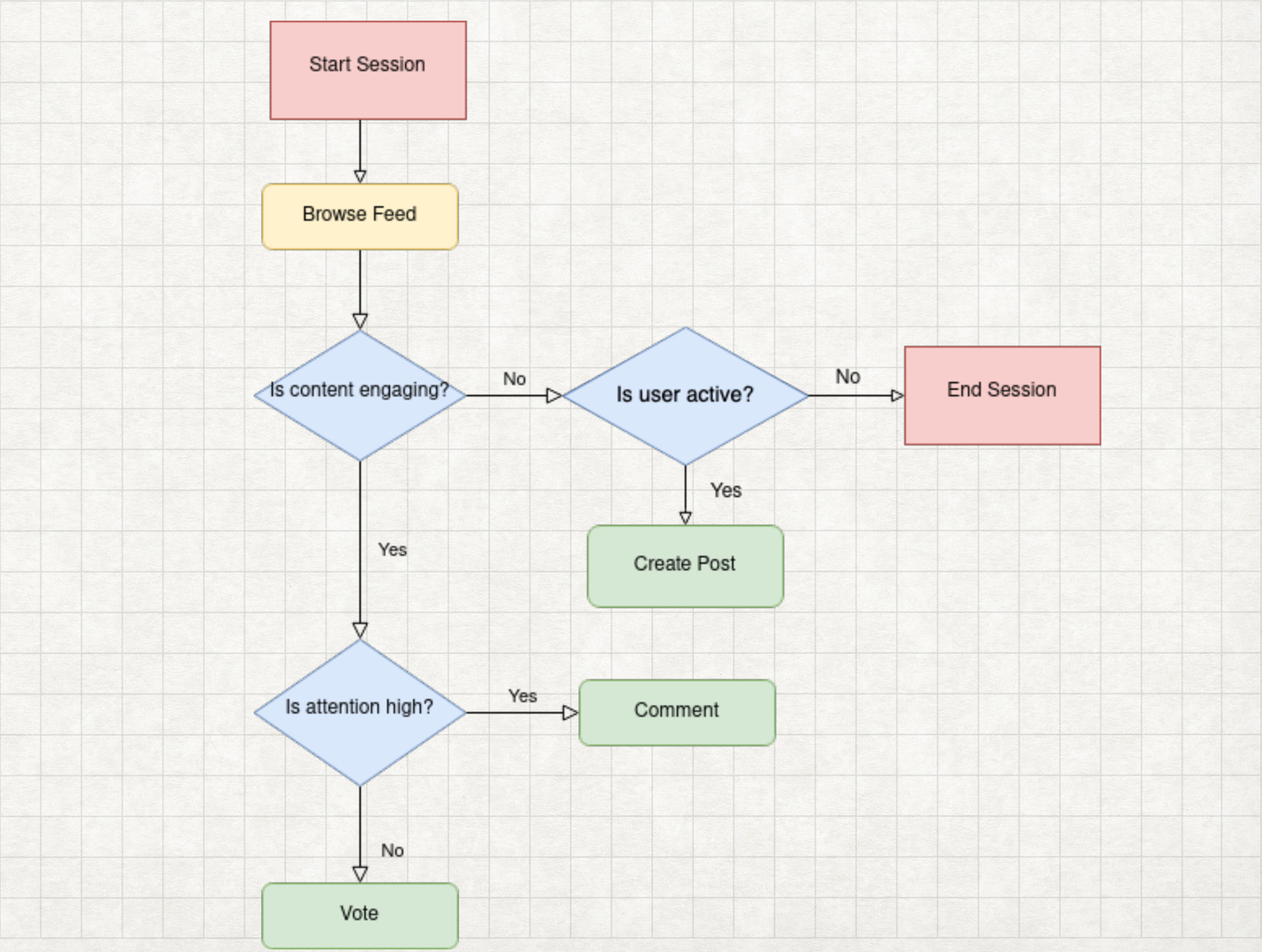
$N \in \mathbb{N}$: Number of users
$T \in \mathbb{N}$: Time steps
$S \in \mathbb{N}$: Number of sub-reddit categories
$\{\pi_i \in \text{Uniform}(0,1)\}_{i=1,j=1}^{N,S}$ ▷ Interaction frequency

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

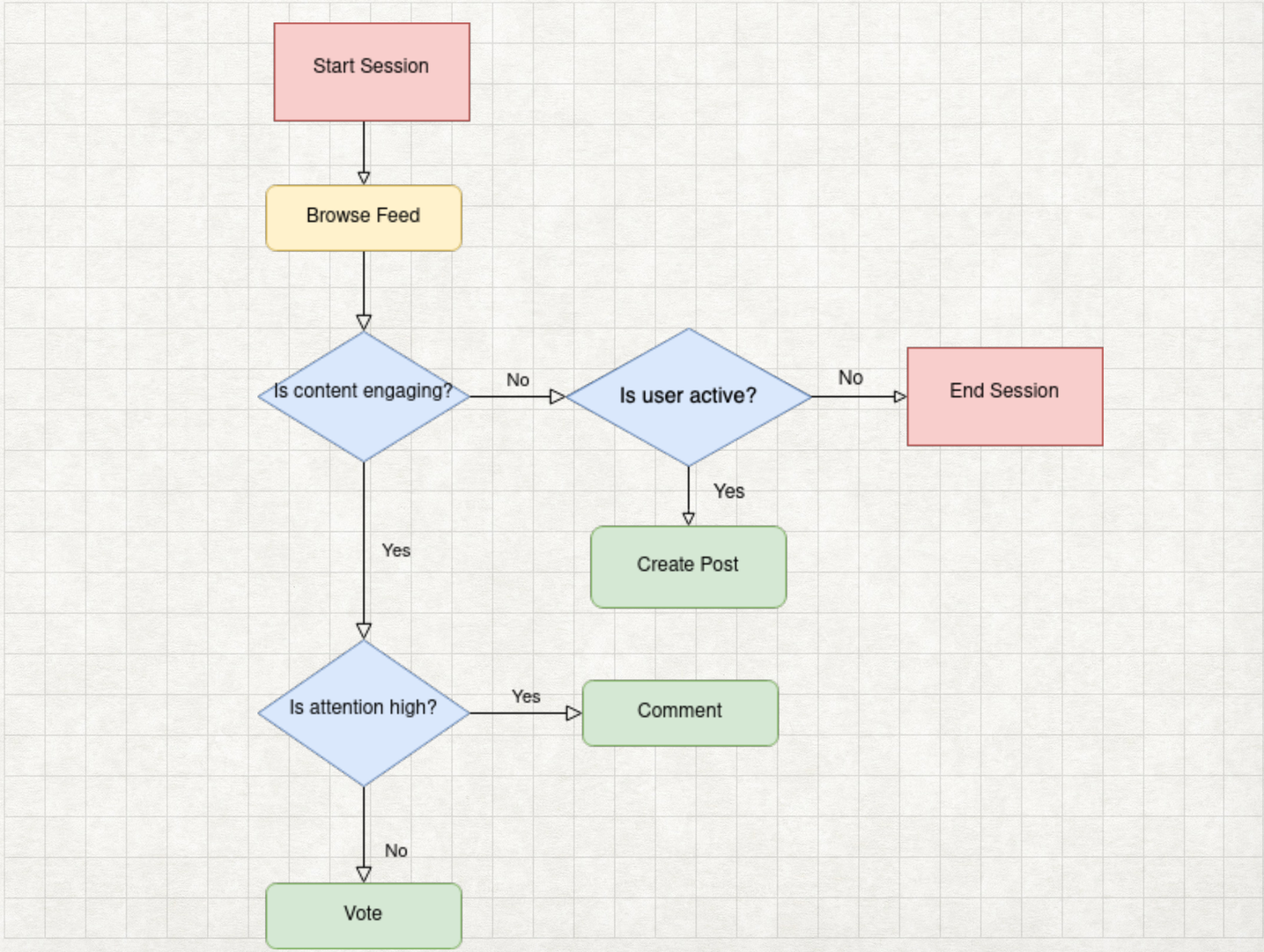
- Developing a model for a user's posting behavior
- Creating a story of how a user interacts with Reddit



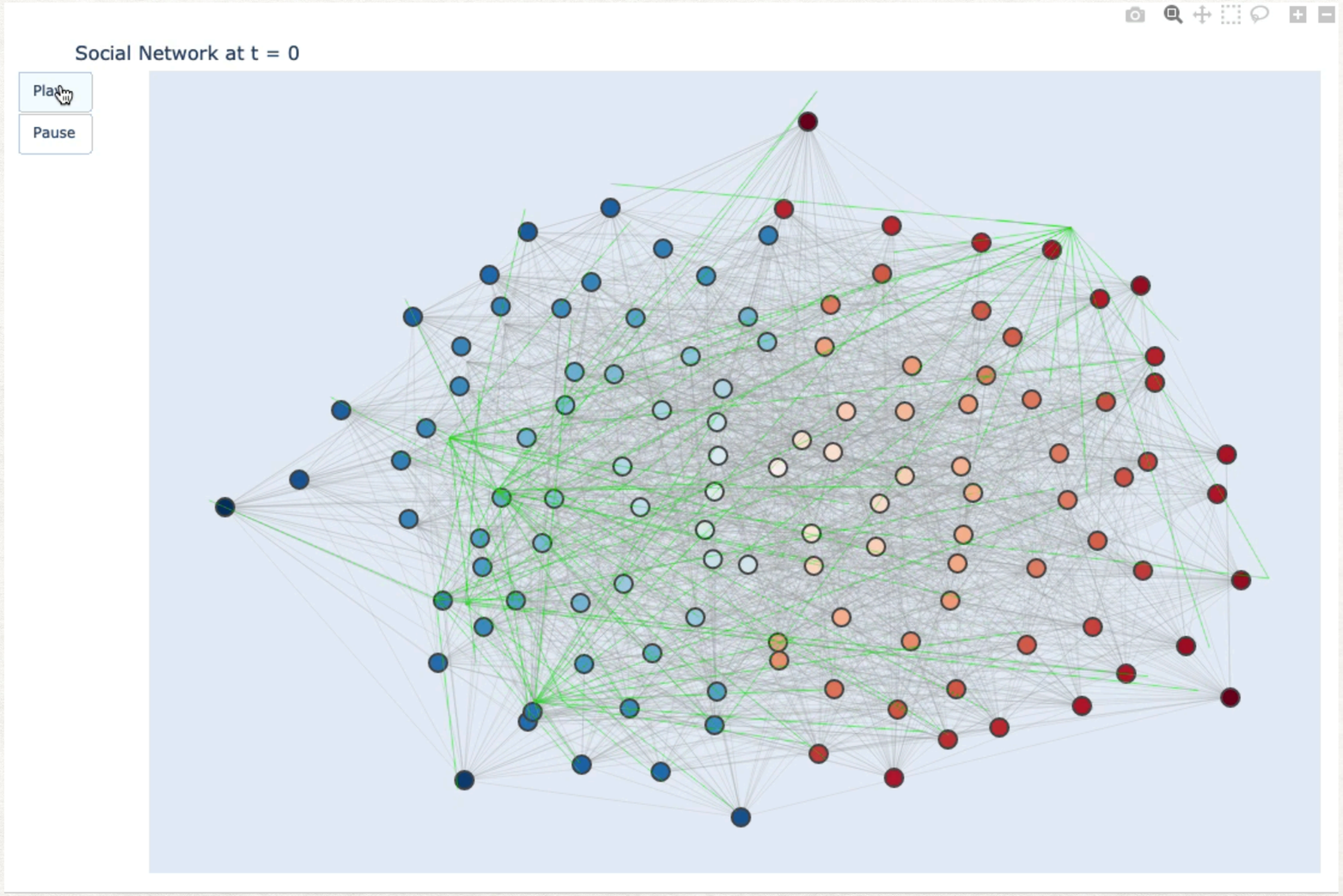
SIMULATING SOCIAL NETWORKS

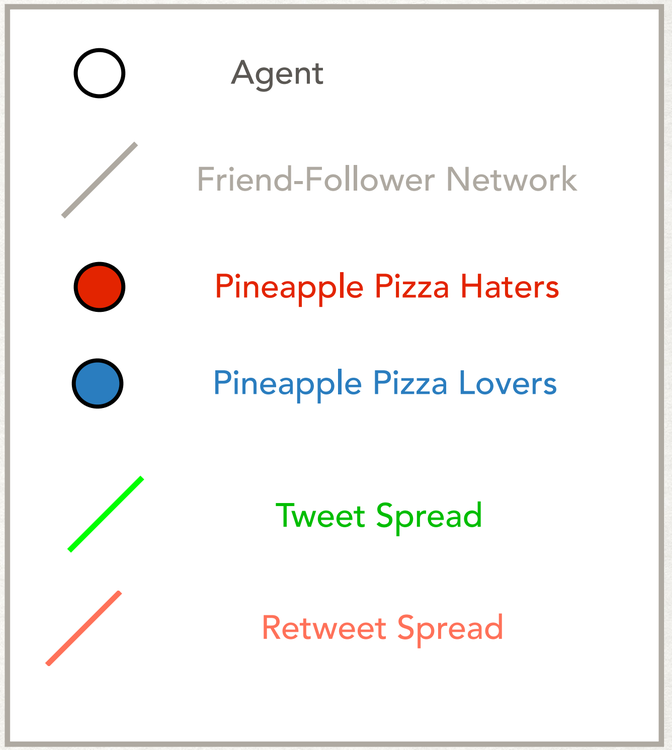
REDDIT

- 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!



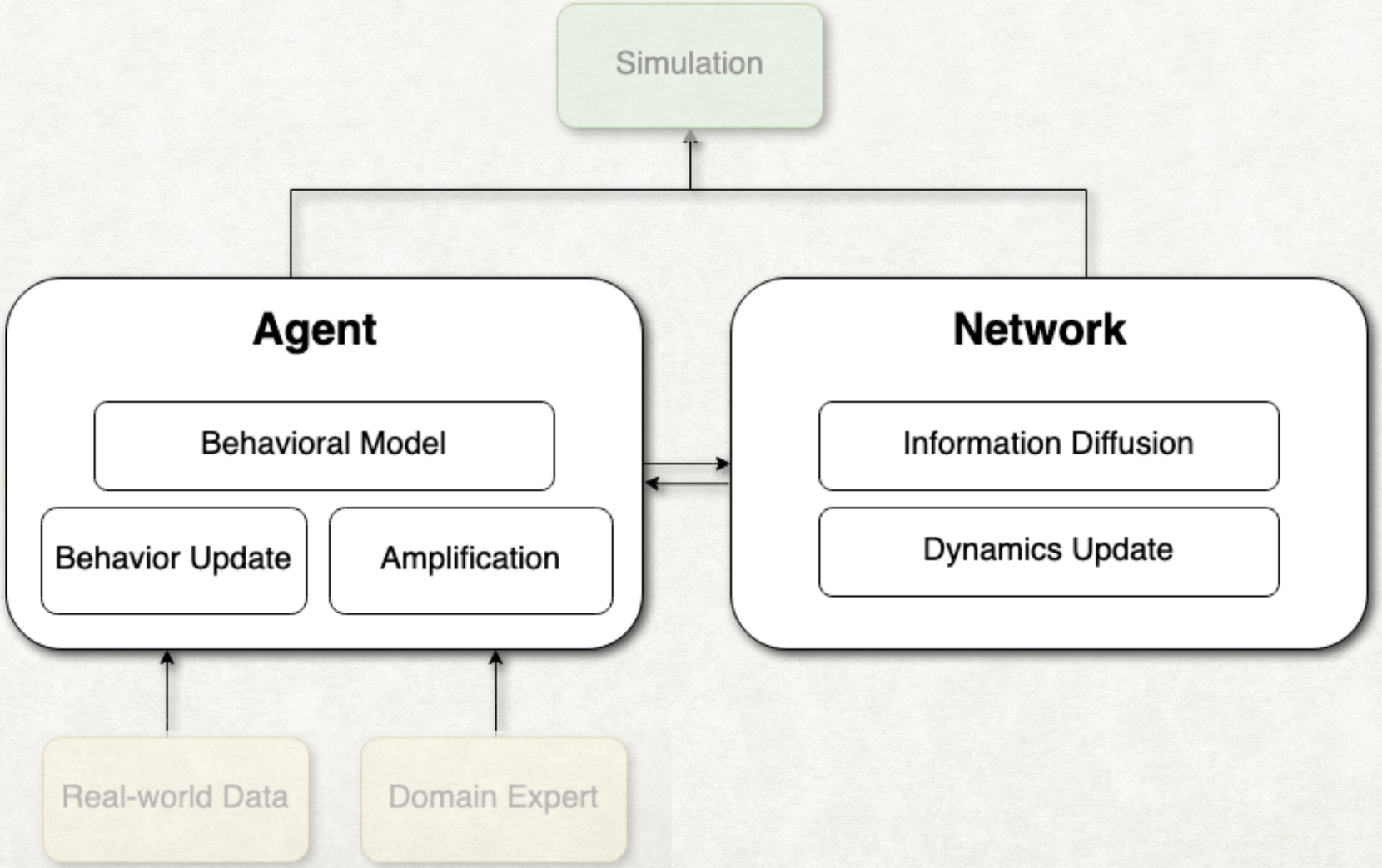
SIMULATING SOCIAL NETWORKS



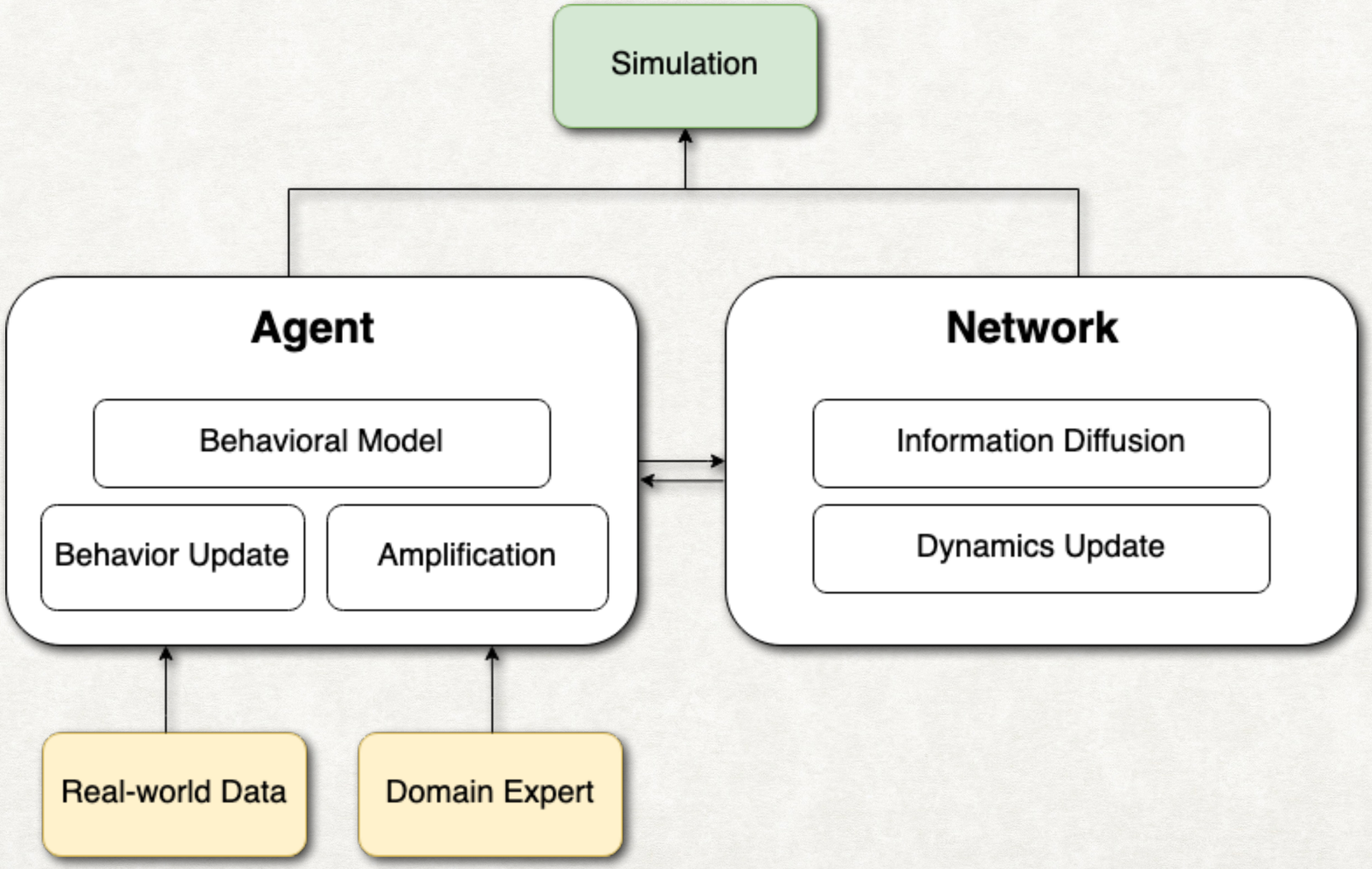


https://youtu.be/GV5XuftiD7s

SIMPPL


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

SIMPPL

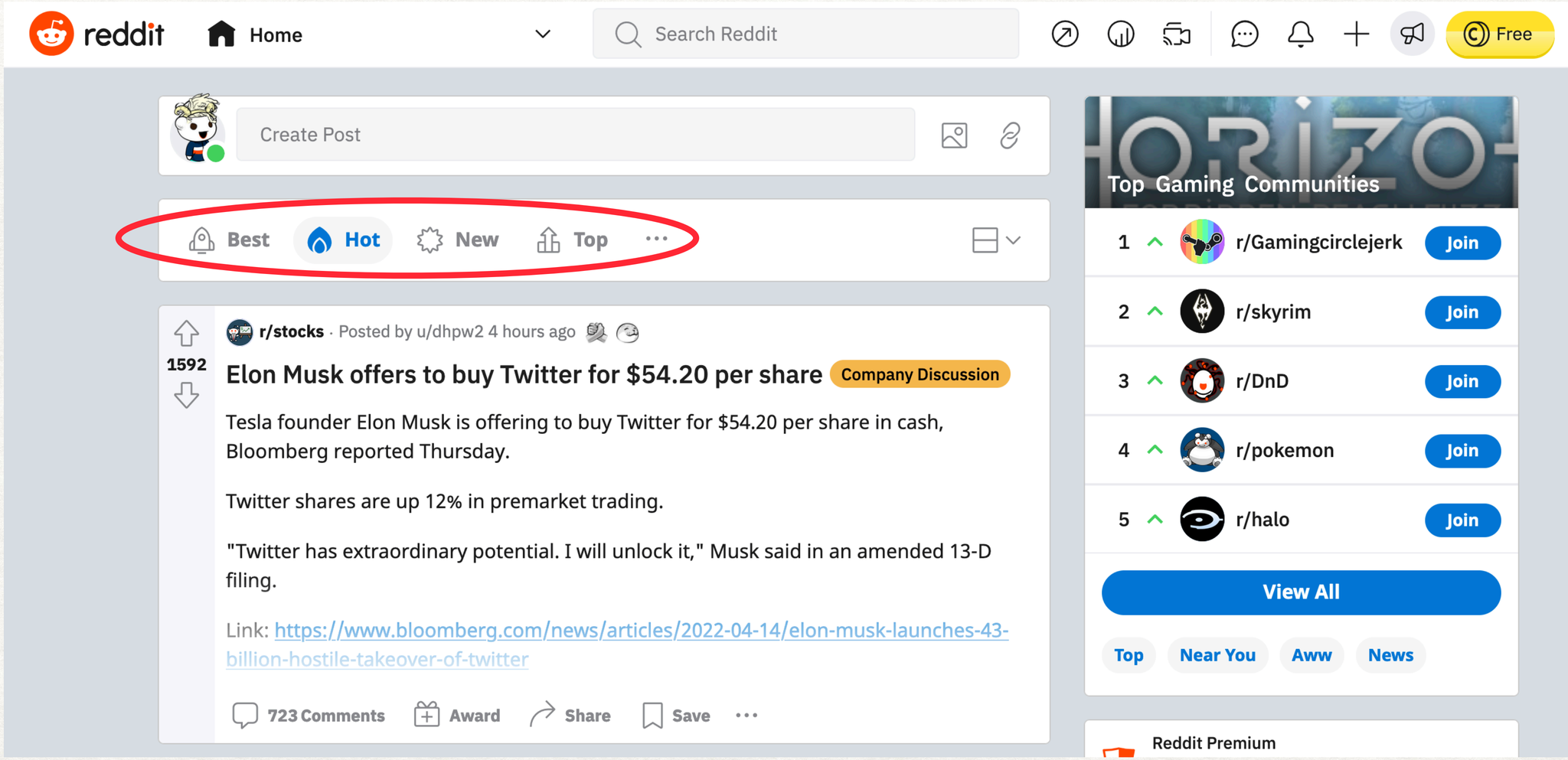

Simulation
Agent
Behavioral Model
Behavior Update
Amplification
Network
Information Diffusion
Dynamics Update
Real-world Data
Domain Expert

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

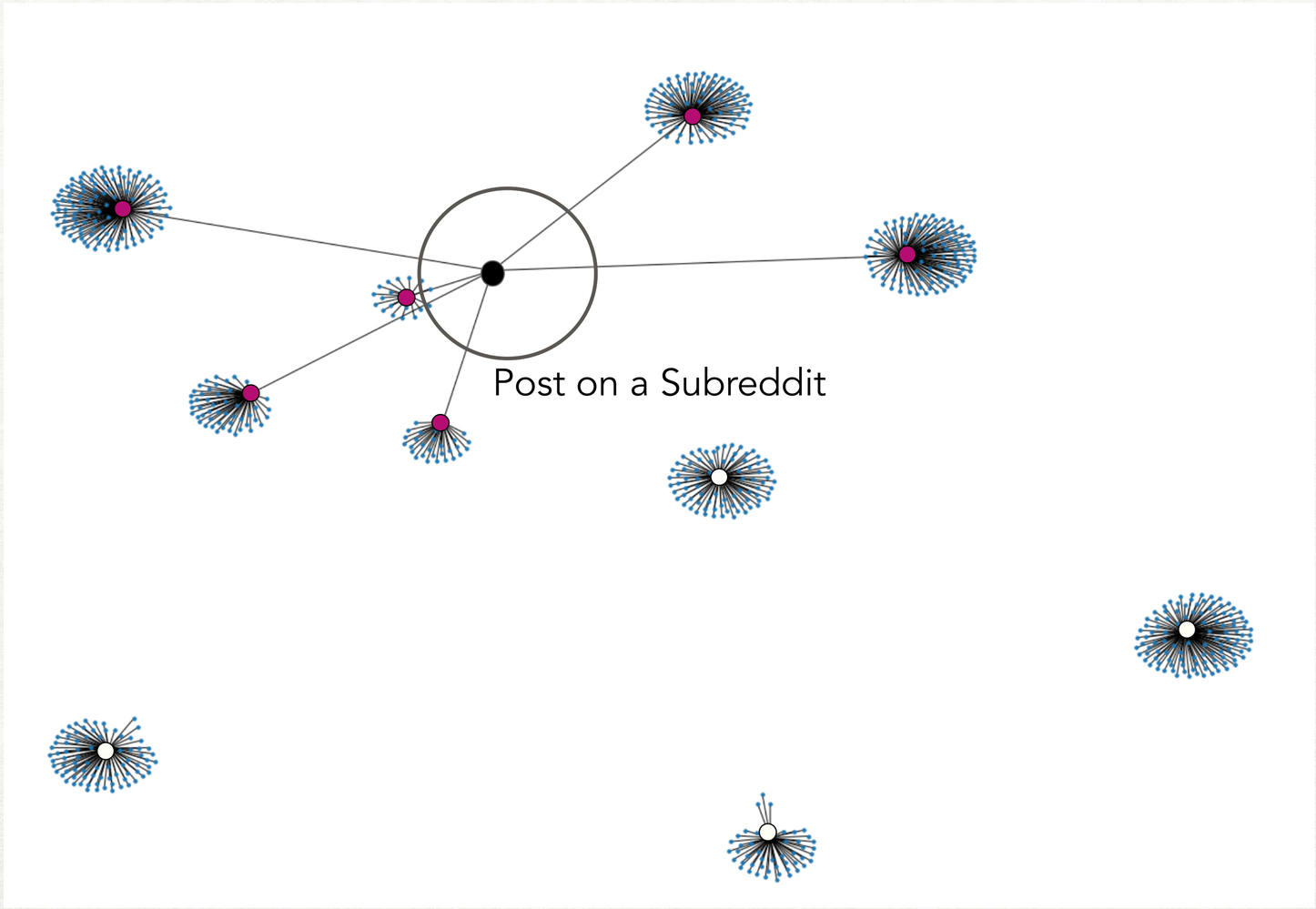


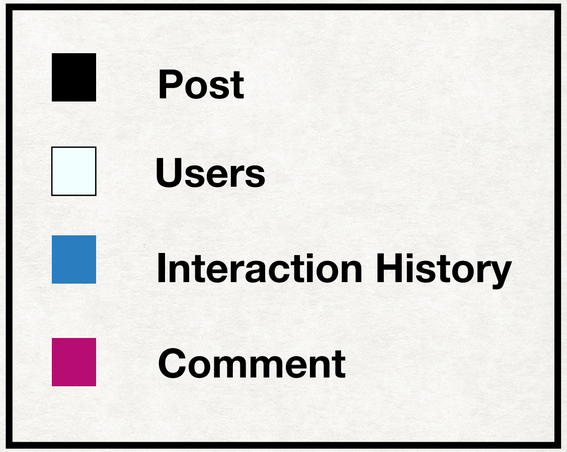
RANKING AND RECOMMENDATION ALGORITHMS

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

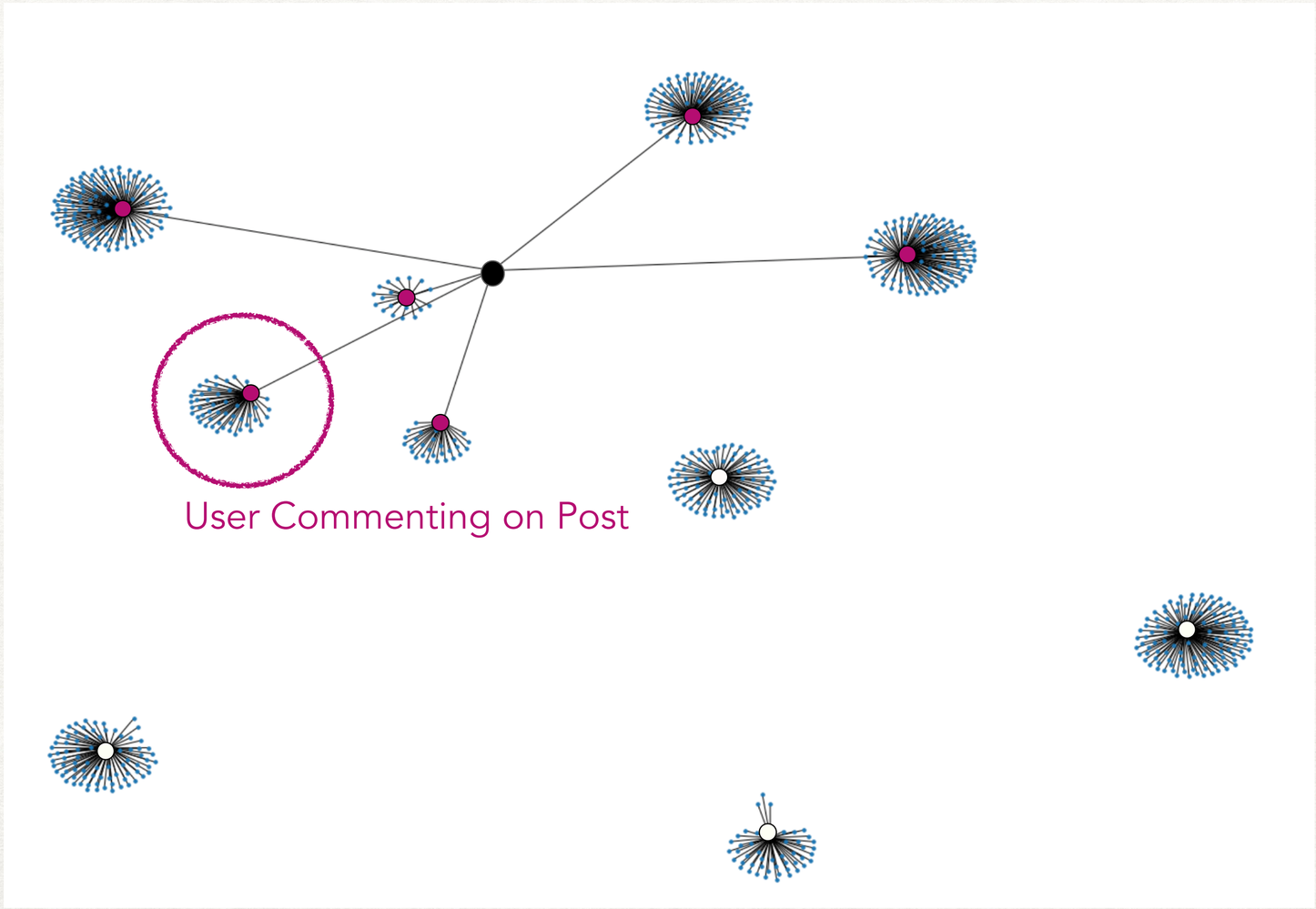
USING REAL-WORLD DATA TO DRIVE THE SIMULATIONS

REDDIT POST


Post on a Subreddit


Post
Users
Interaction History
Comment

REDDIT POST



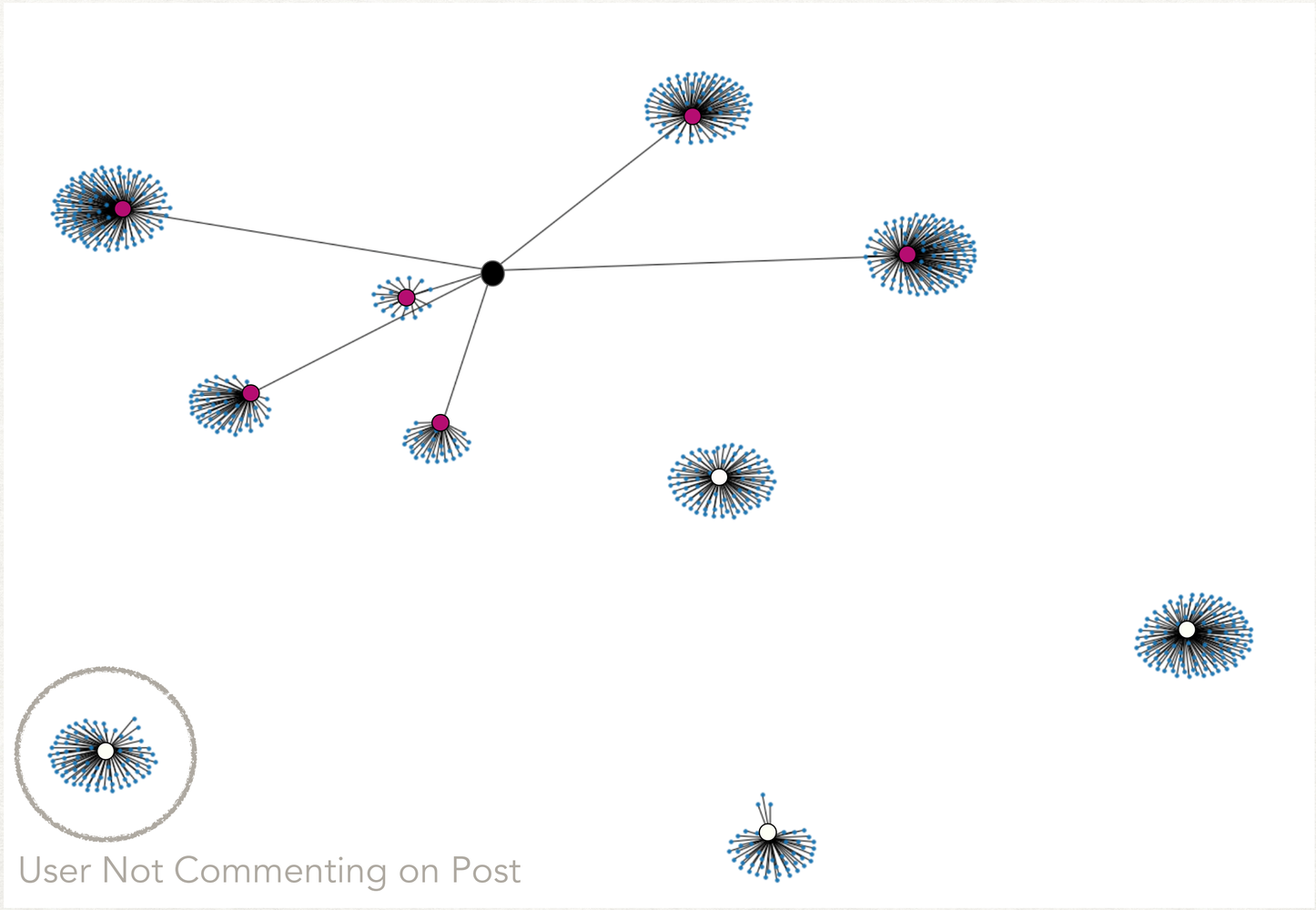
Post

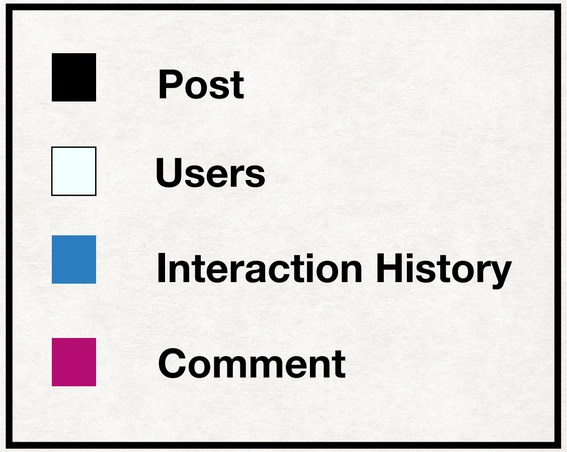
Users

Interaction History

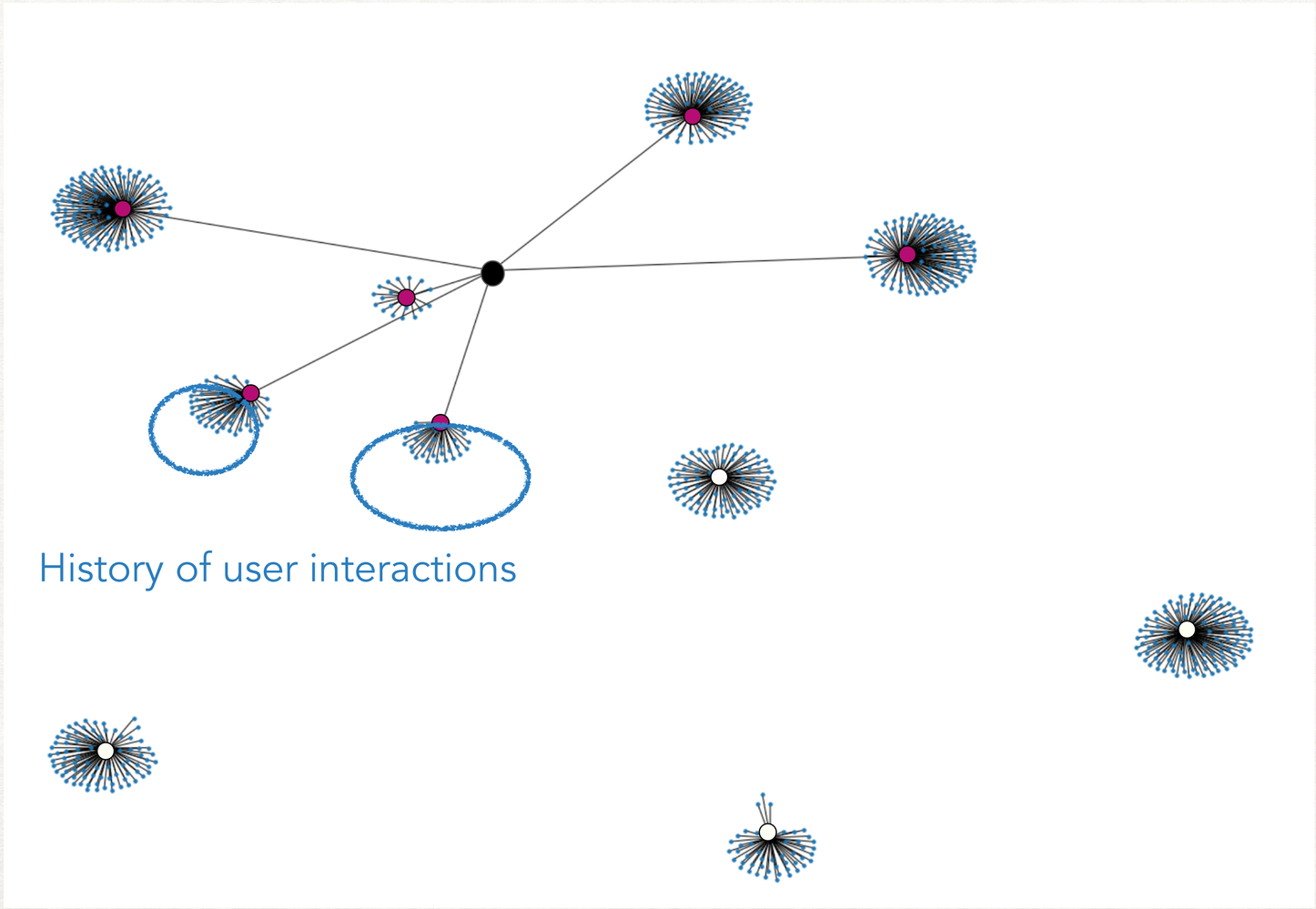
Comment

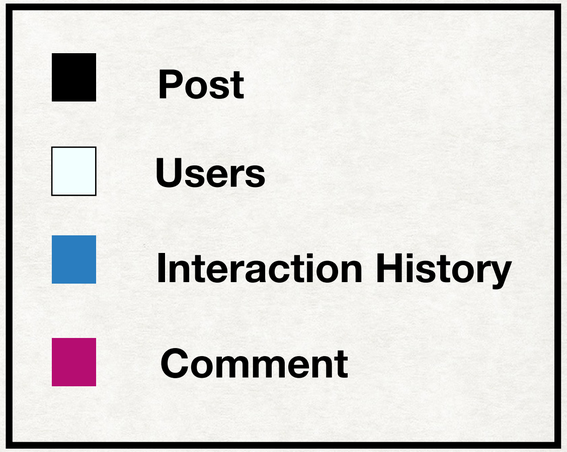
REDDIT POST


User Not Commenting on Post

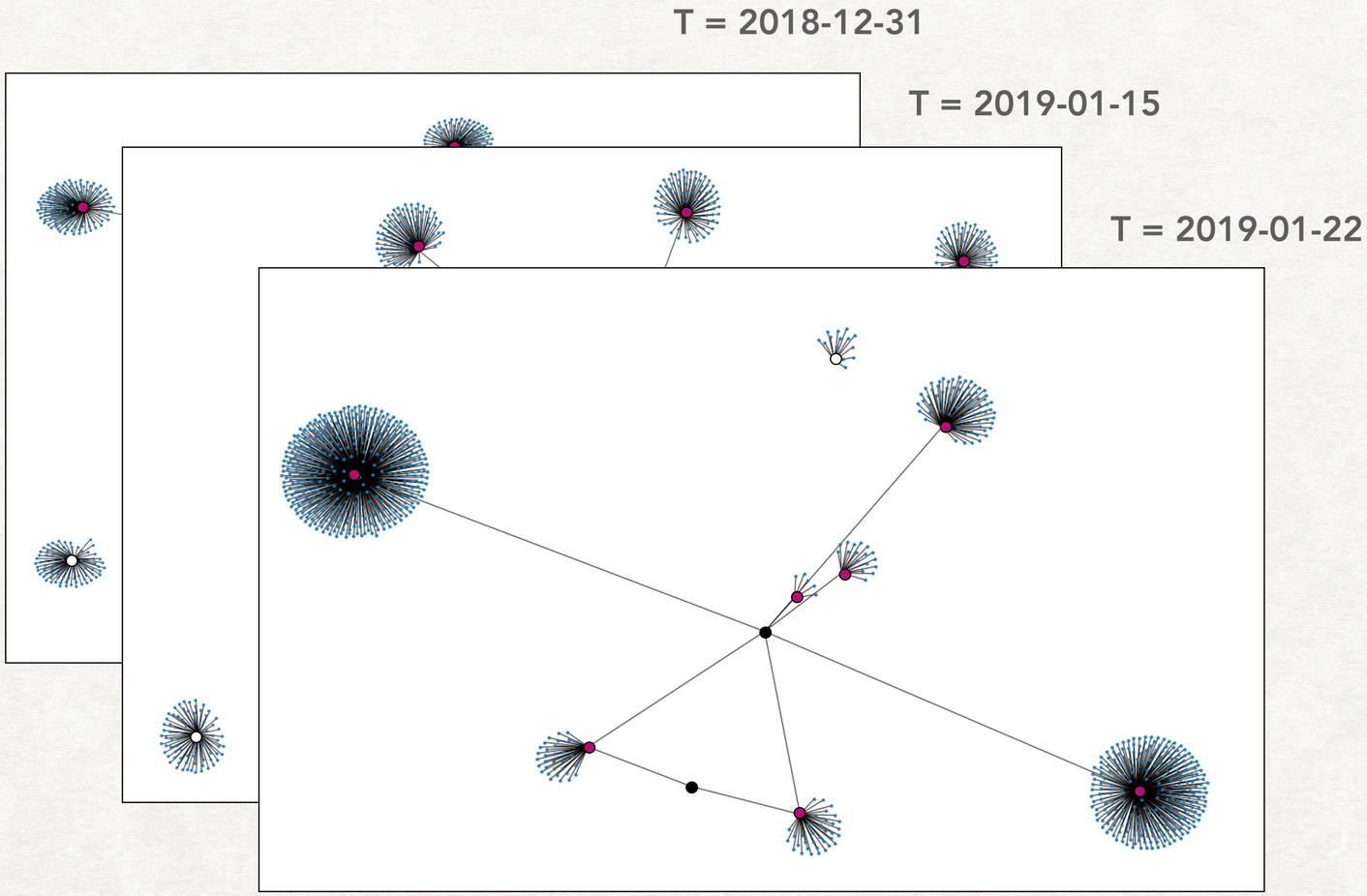

Post
Users
Interaction History
Comment

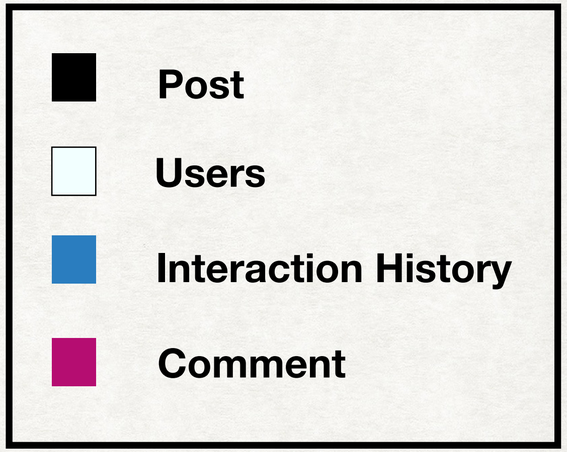
REDDIT POST


History of user interactions


Post
Users
Interaction History
Comment

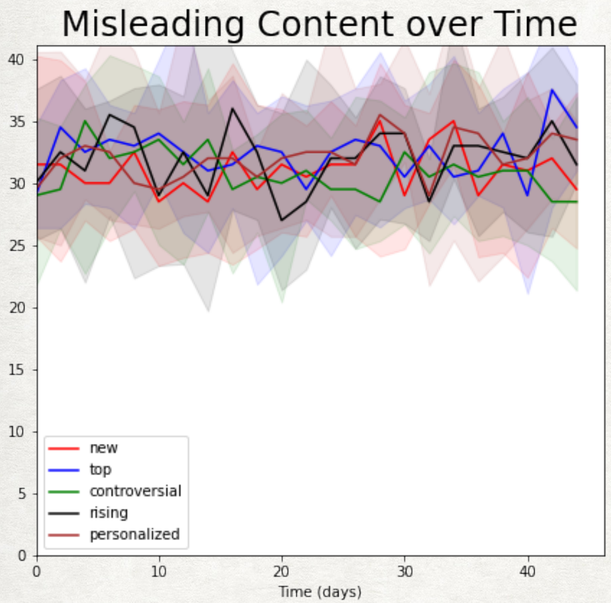
REDDIT POSTS

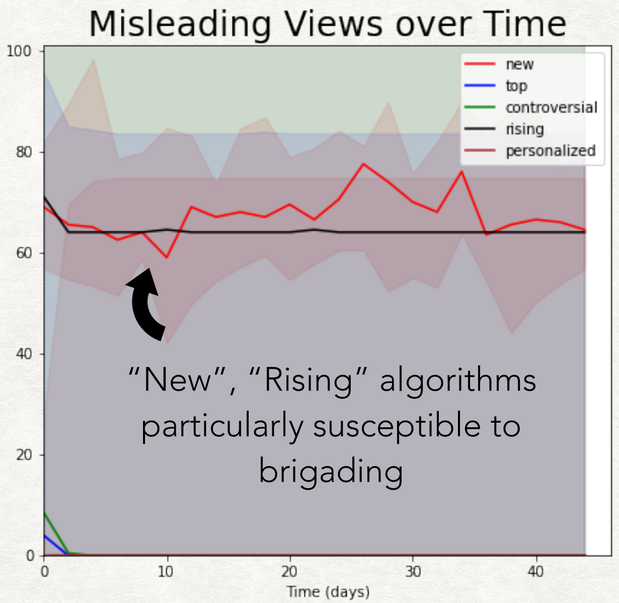


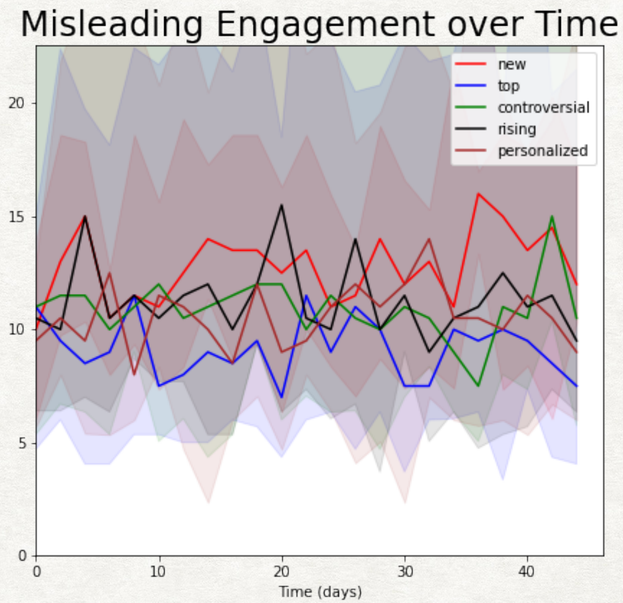


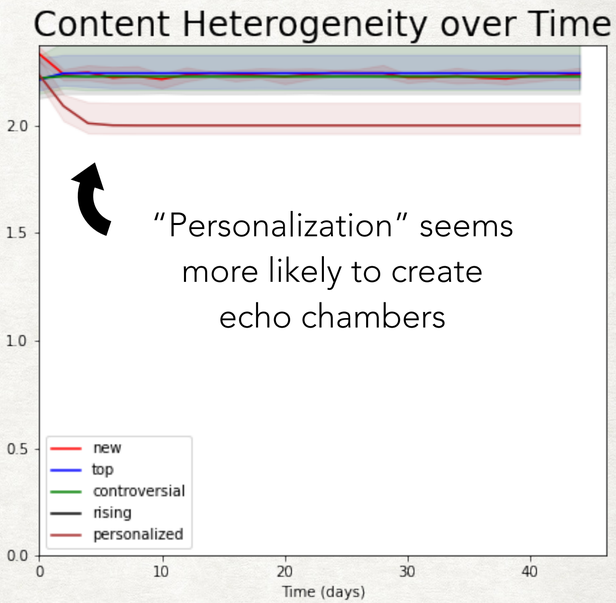
HOW SUSCEPTIBLE ARE ALGORITHMS TO COORDINATED INAUTHENTIC BEHAVIOR?

WHAT ARE THE EFFECTS OF BRIGADING?



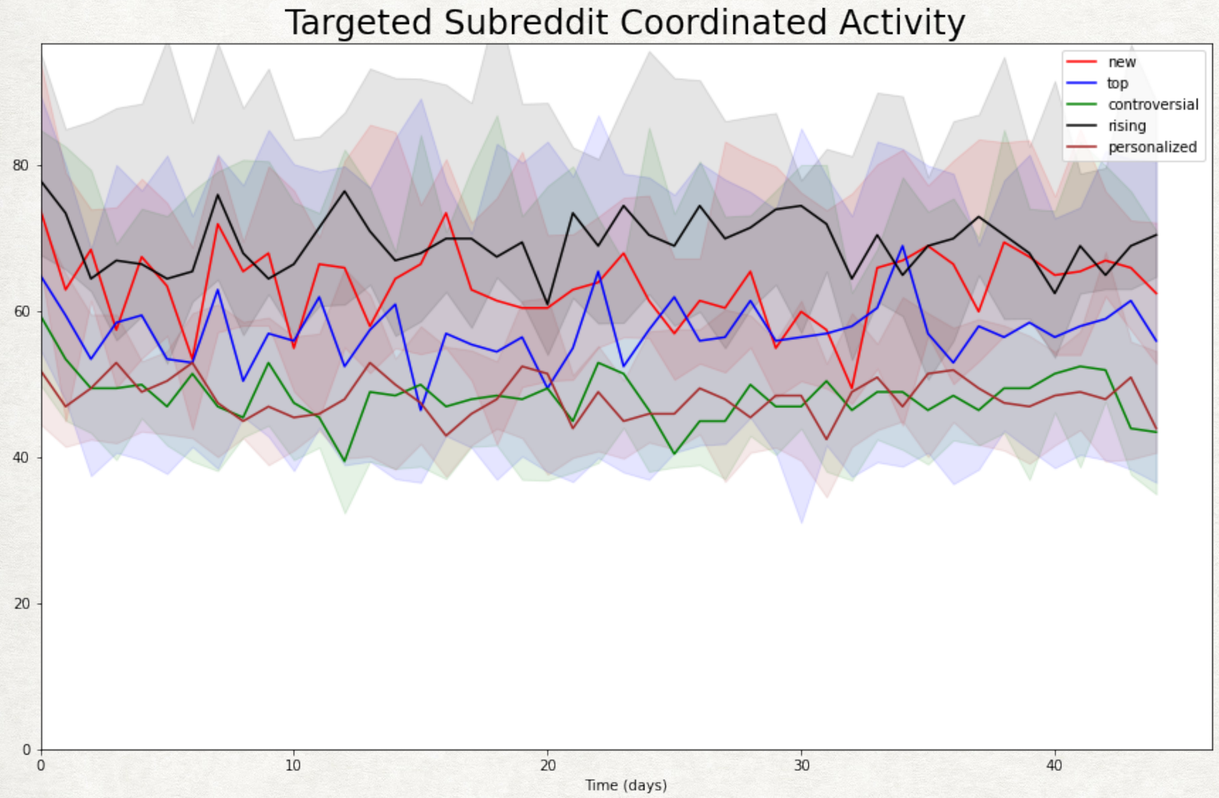


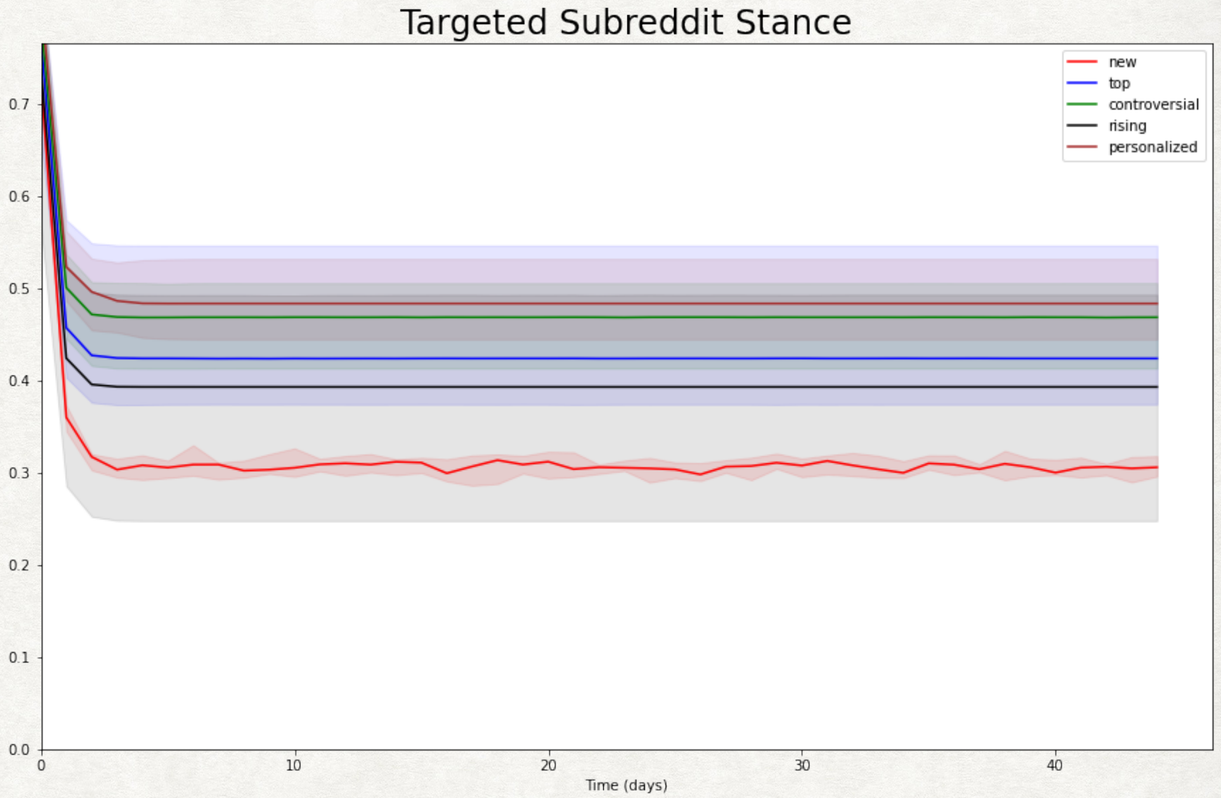




Similar levels of misleading content leads to different emergent dynamics

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

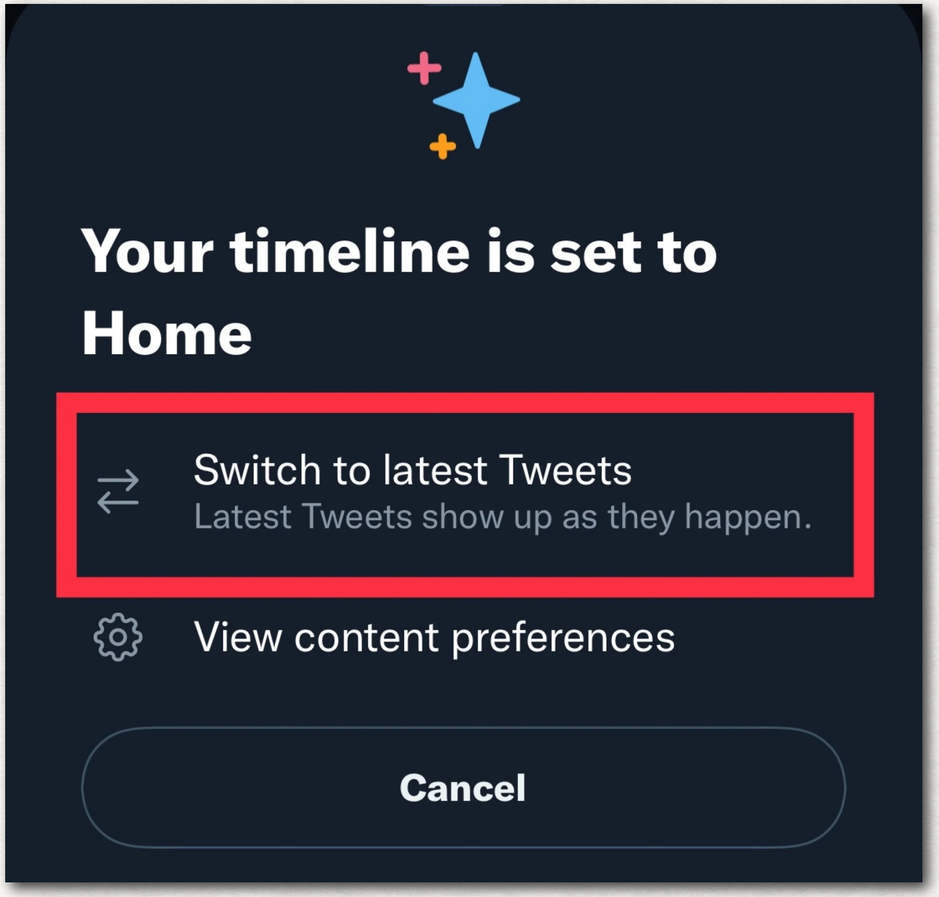
TECH • BIG TECH

Facebook Is Finally Giving People A Non-Algorithmic News Feed

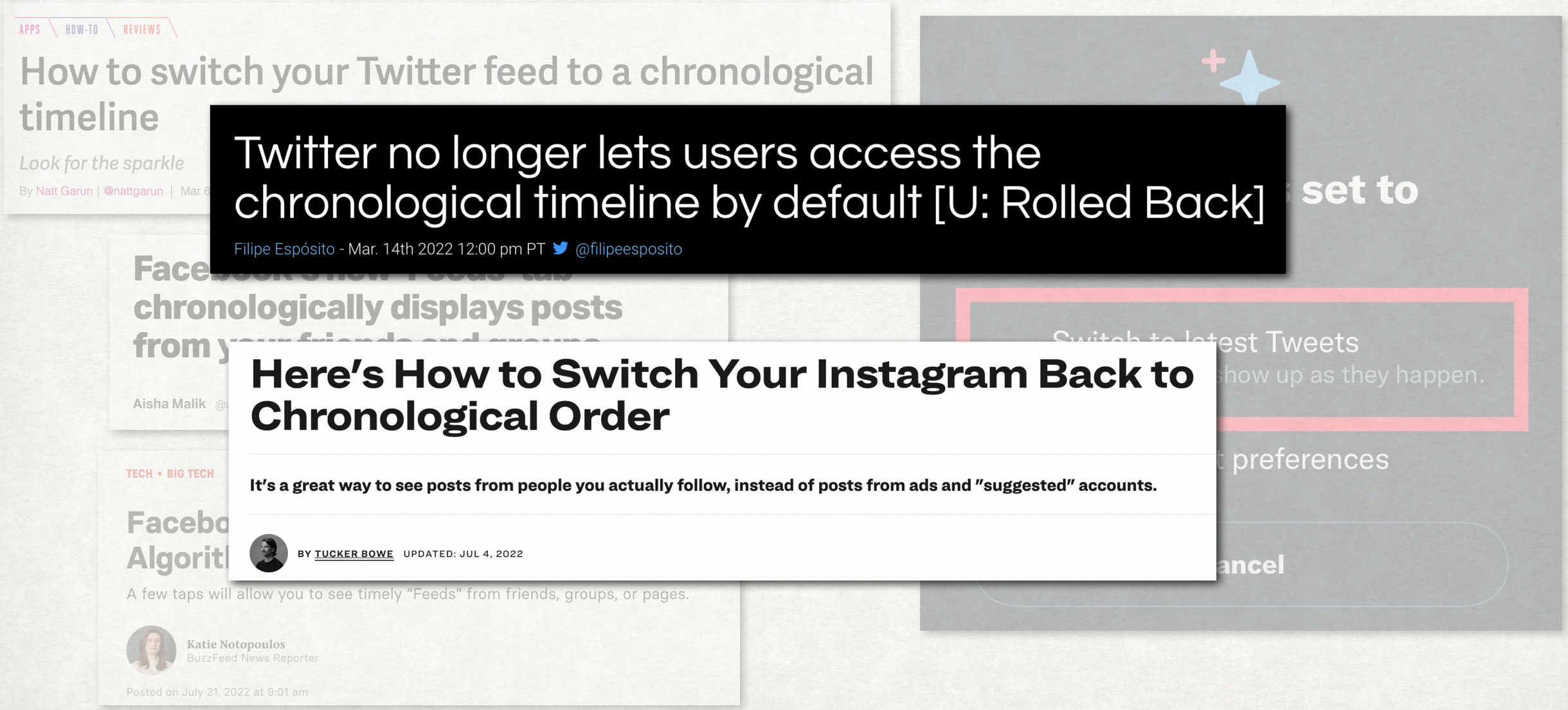
A few taps will allow you to see timely "Feeds" from friends, groups, or pages.

Katie Notopoulos
BuzzFeed News Reporter

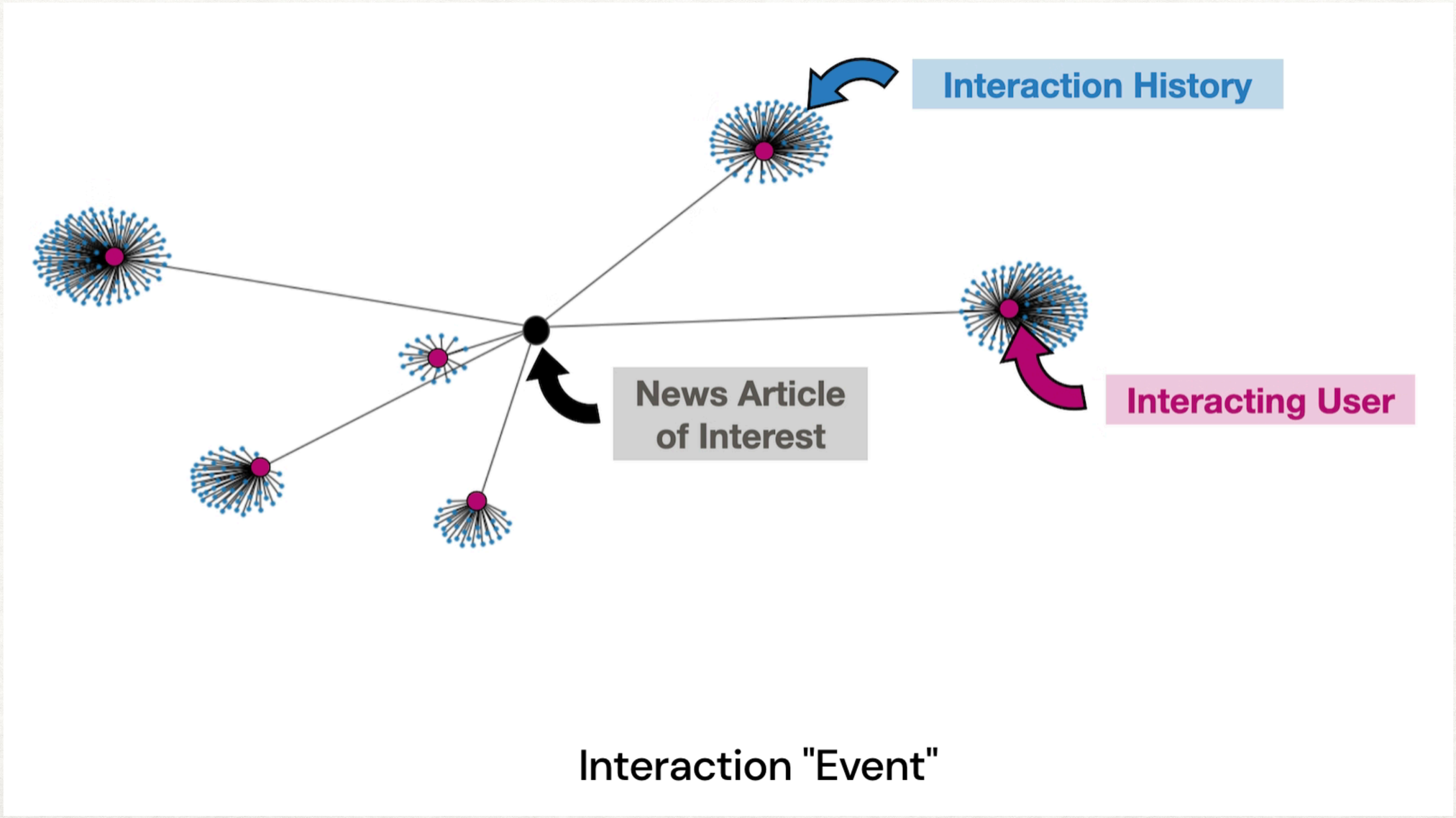
Posted on July 21, 2022 at 9:01 am



CONTENT DISTRIBUTION CHOICES



APPLYING SIMPPL'S TECH TO NEWS ARTICLES



HELPING LOCAL NEWS UNDERSTAND AUDIENCES

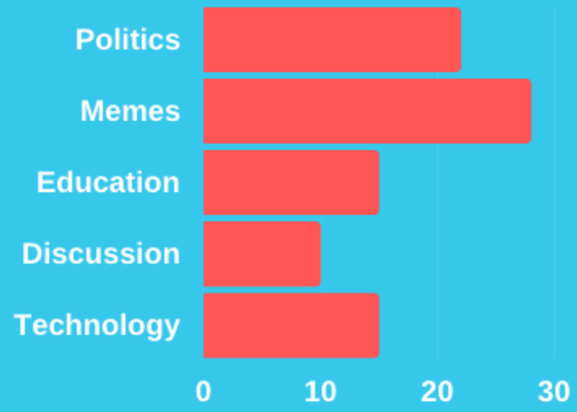
You're Getting Upvotes!

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

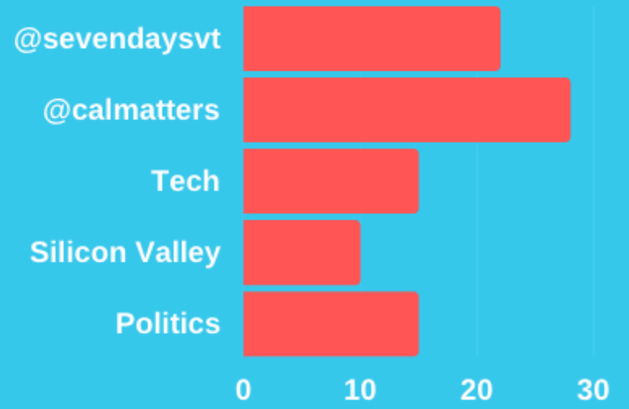
HELPING LOCAL NEWS UNDERSTAND AUDIENCES

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

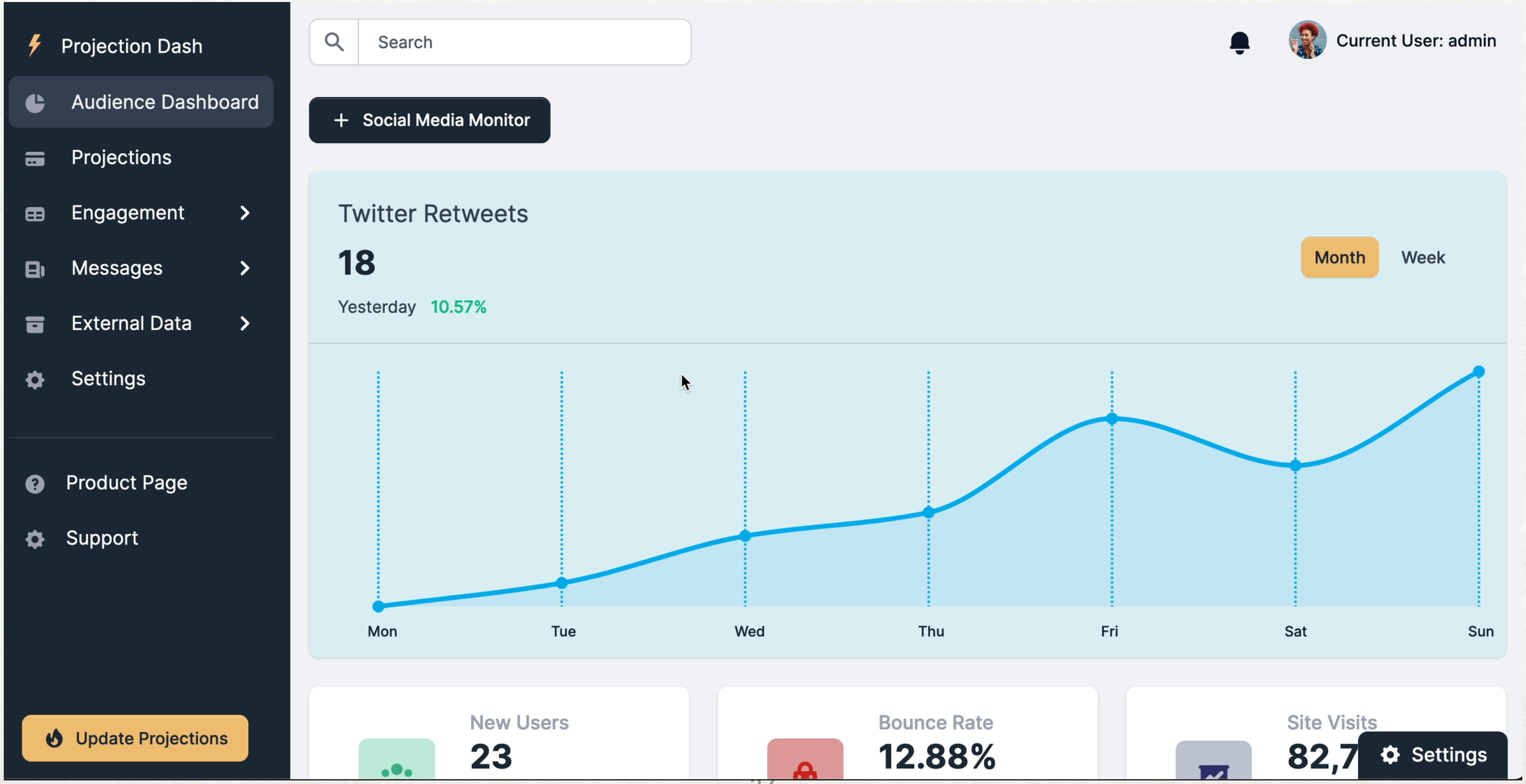


Related Engagement

14	23	86 ↑
Politics	Elections	Bernie Sanders

@sevendaysvt

DEMO


Projection Dash
Audience Dashboard
Projections
Engagement
Messages
External Data
Settings
Product Page
Support
Update Projections
Search
Current User: admin
Social Media Monitor
Twitter Retweets
18
Yesterday 10.57%
Month
Week
Mon
Tue
Wed
Thu
Fri
Sat
Sun
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

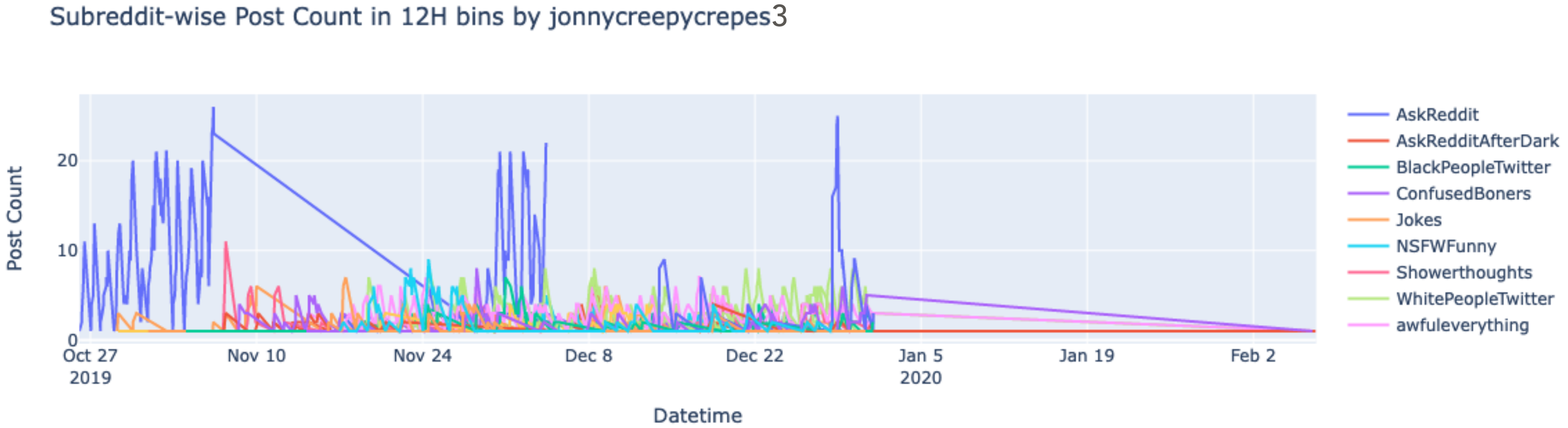
LET'S TALK!

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

ai4abm.org

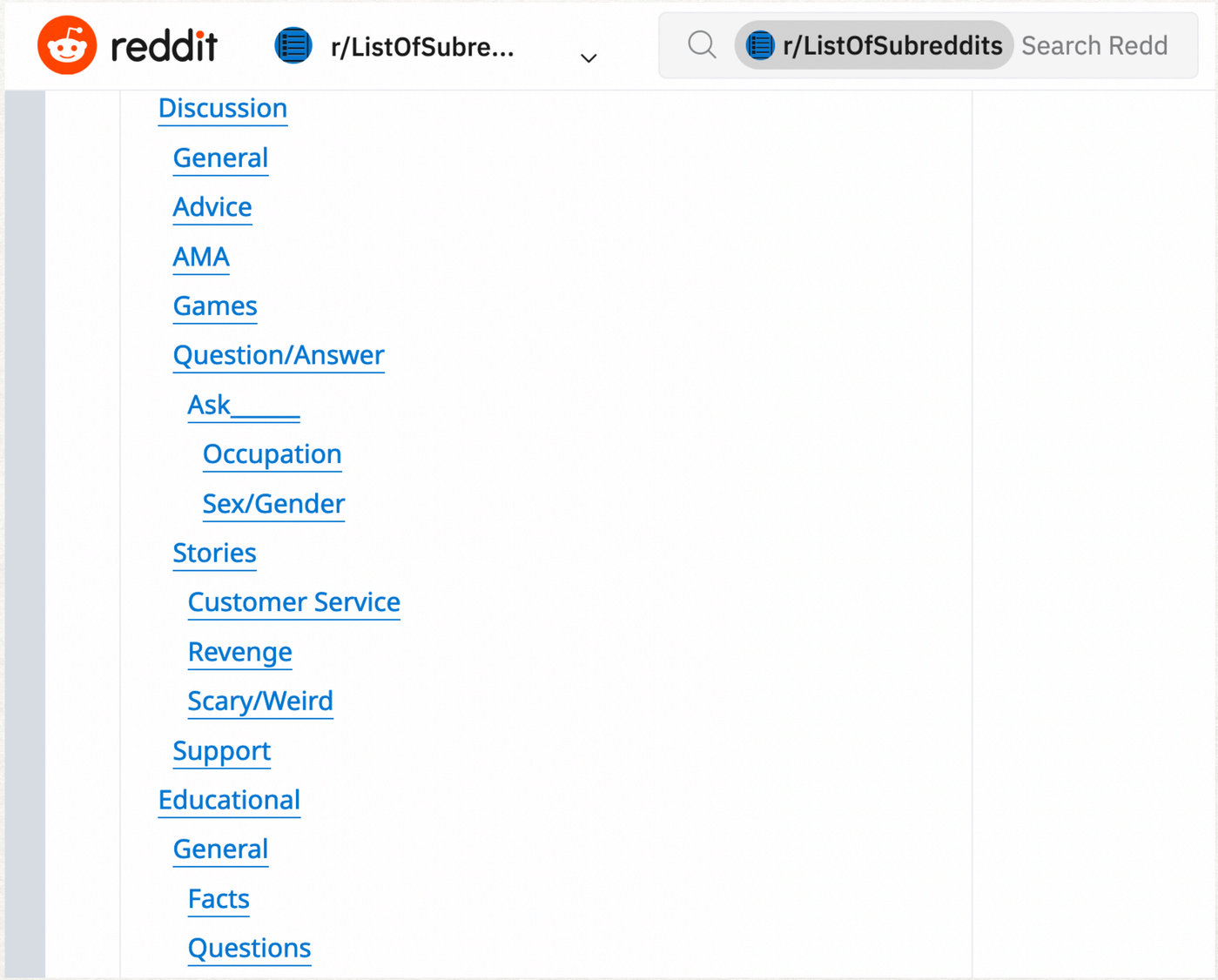
REAL-WORLD MODELING

REDDIT USER 'JONNYCREEPYCREPES3'



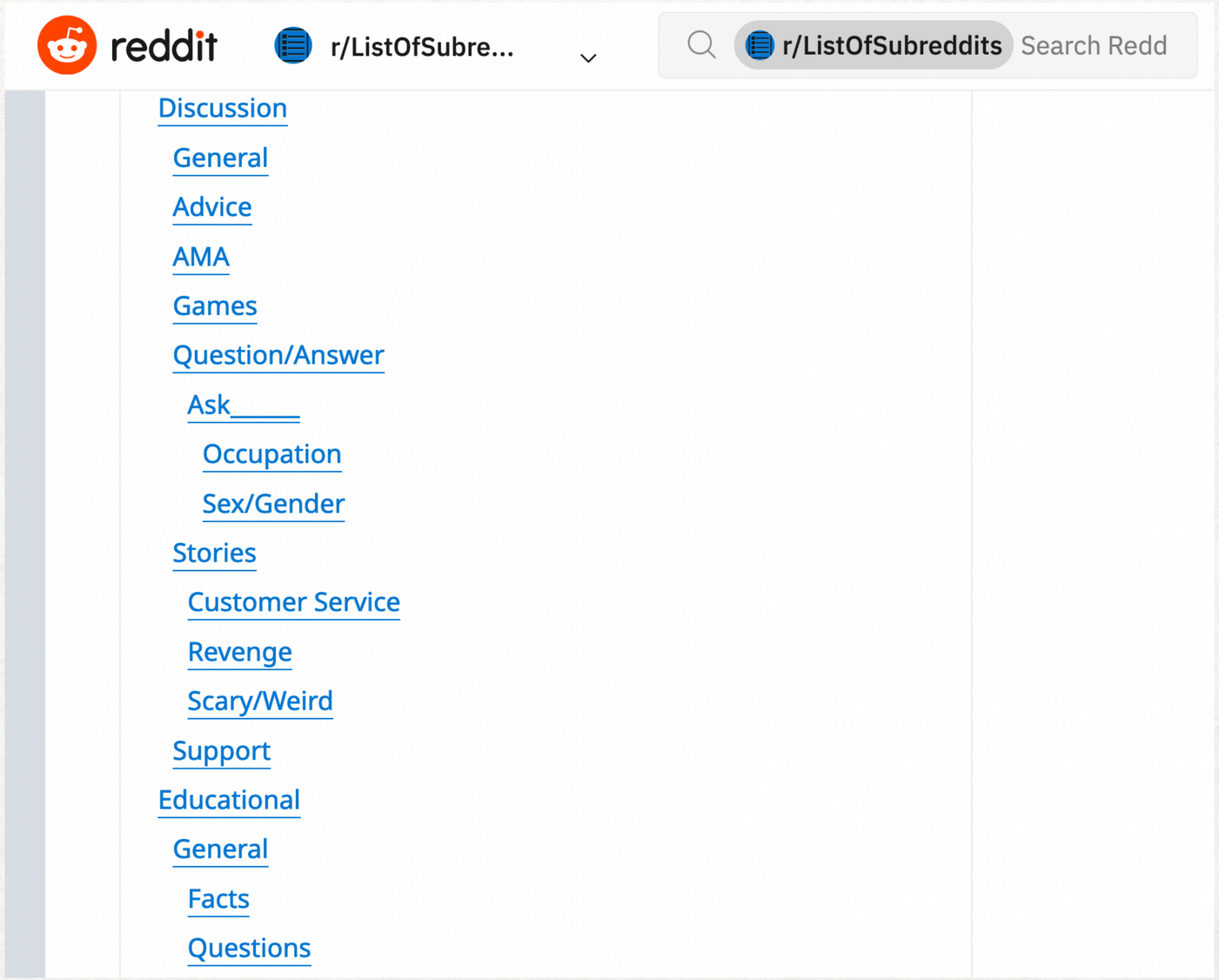
REDDIT WIKI PAGES

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

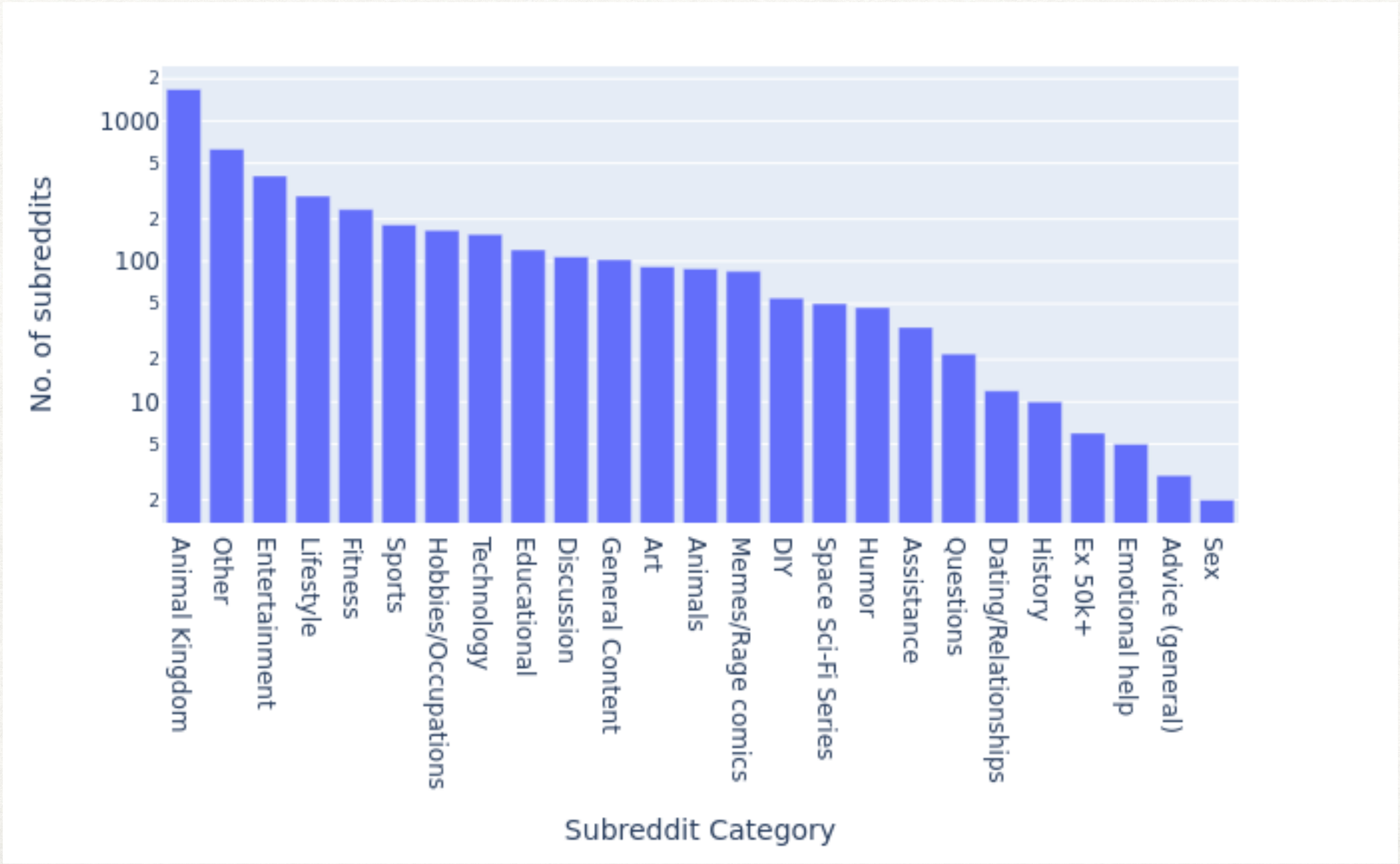


REDDIT WIKI PAGES

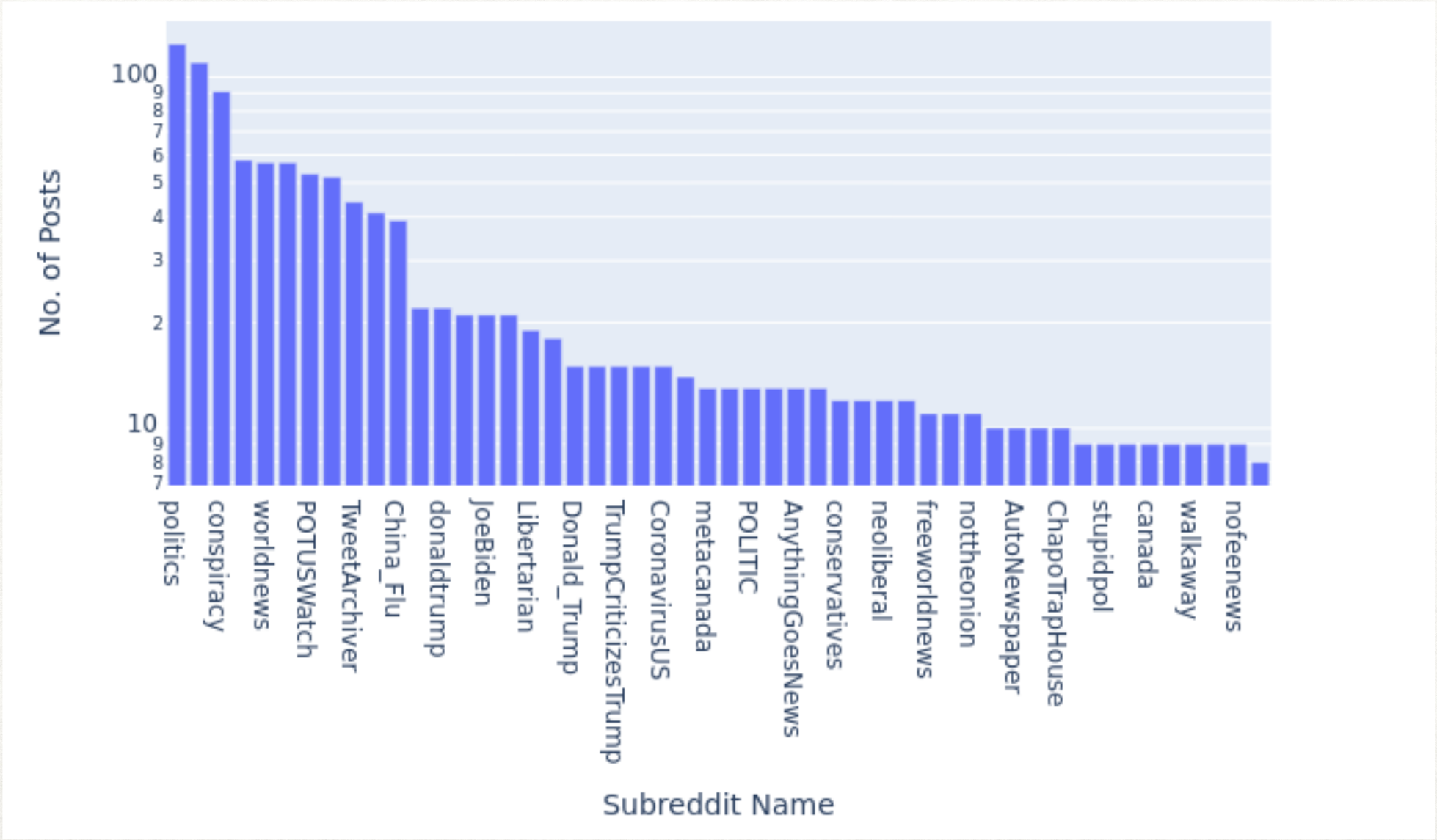
- User-driven hierarchical categorization of subreddits
- Discussion > Stories > Customer Service
- 4998 subreddits
- 5-level hierarchy of categories
- For each new subreddit
  - Split words > match embeddings > associate category



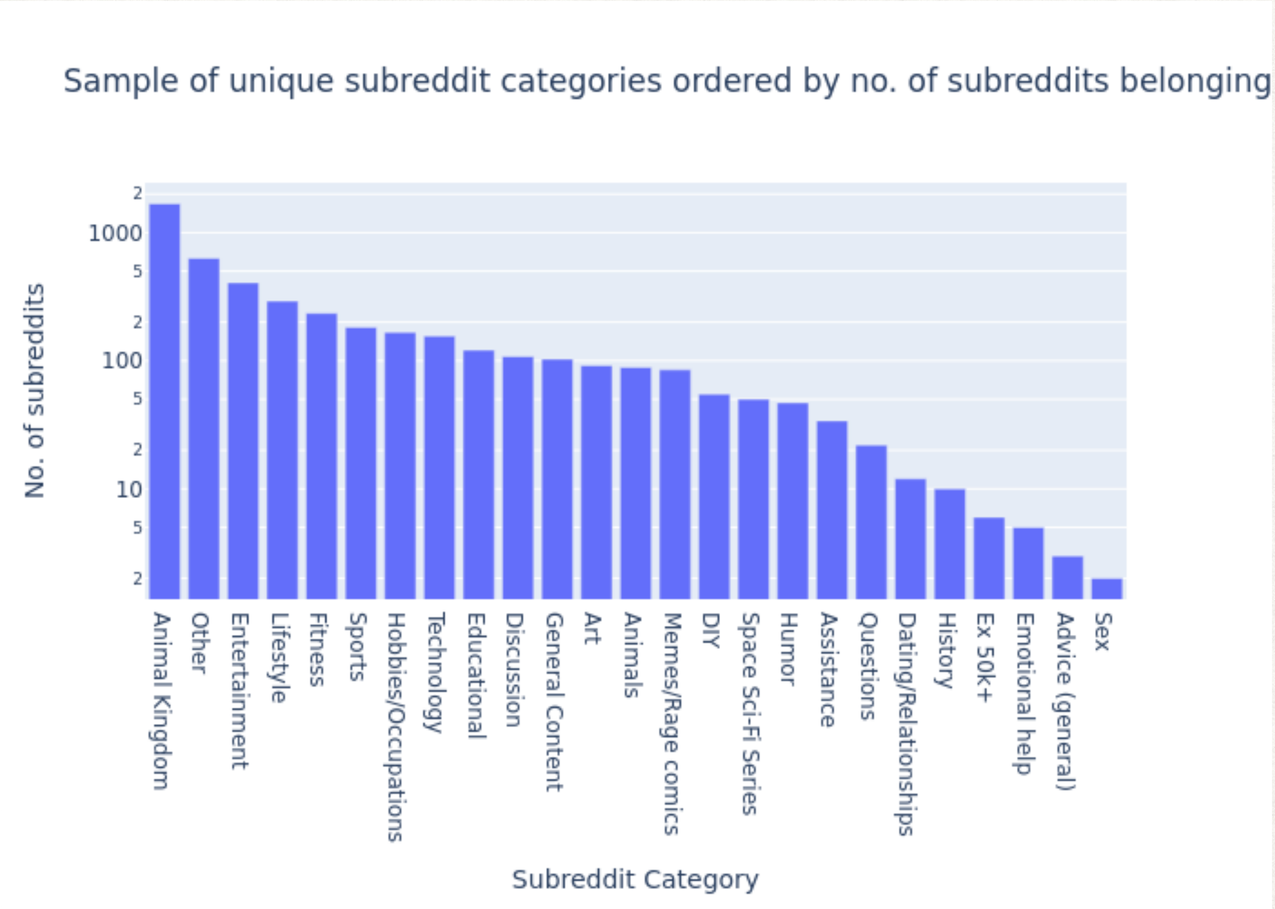
CATEGORIZE POSTS ACROSS SUBREDDITS



CATEGORIZE POSTS ACROSS SUBREDDITS



CATEGORIZE POSTS ACROSS SUBREDDITS



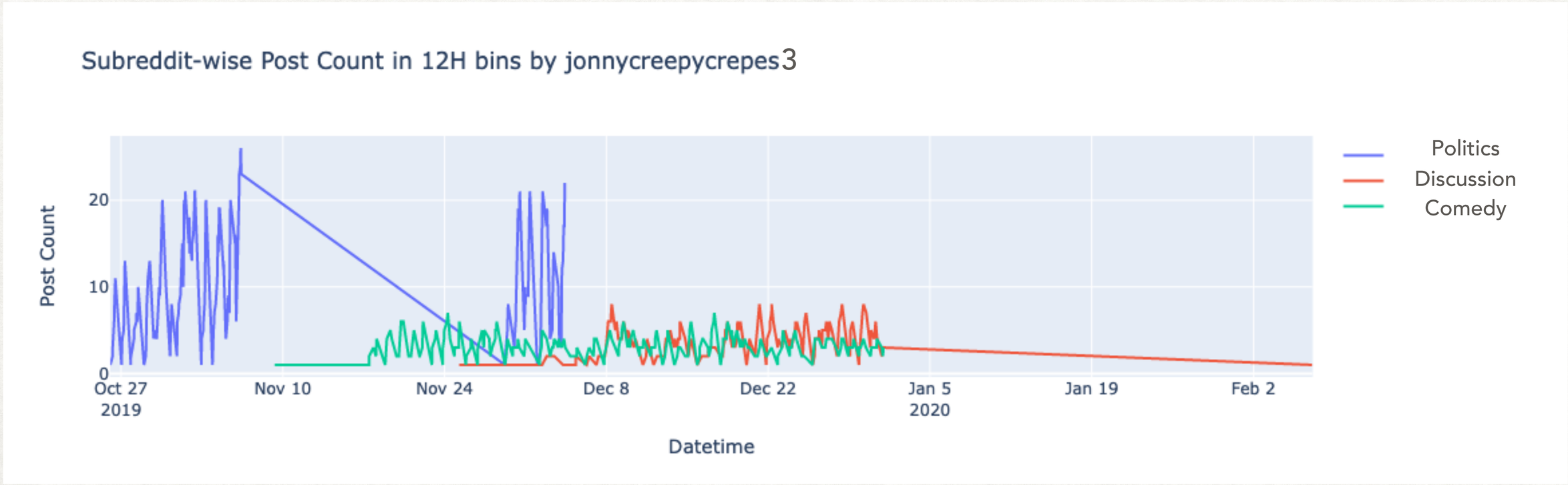
['mormon', 'politics']

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

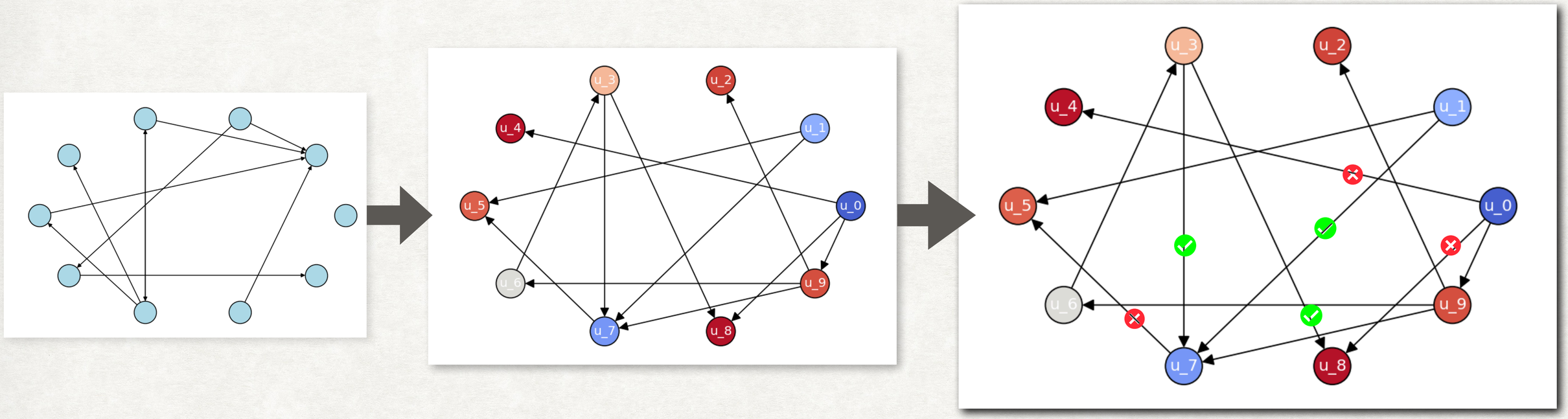
3. **HYBRID**

   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