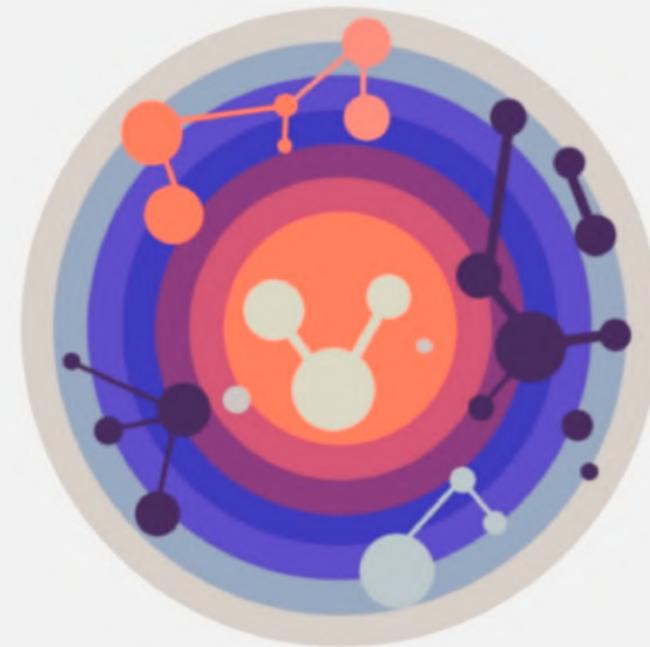SimPPL

Network Science Reimagined.

**Trigger Warning: Discussion of Online Abuse**

**You can skip 1 slide.**

# Online Platforms have governance problems.

## 01 Disinformation

Spread of misleading and malicious content by networks of coordinated actors who game algorithms and policies to manipulate the public.

## 02 Online Abuse

Racist, identity, and sexual attacks targeting minorities are at historic highs. Child sexual abuse material is rising astronomically despite safeguards.

# Disinformation is undermining democracy.

## 72%

Americans report
exposure to perceived
misleading info. online

## 60%

US adults report
confusion in being able
to tell apart fake news

# 01 Disinformation is undermining democracy.

**72%**

Americans report exposure to perceived misleading info. online

**60%**

US adults report confusion in being able to tell apart fake news

**~30%**

lower trust in news for US adults since 2021

**18%**

Reduction in knowledge for adults who get all their news from social media

## 01 Disinformation is undermining democracy.

**72%**

Americans report exposure to perceived misleading info. online

**60%**

US adults report confusion in being able to tell apart fake news

**~30%**

lower trust in news for US adults since 2021

**18%**

Reduction in knowledge for adults who get all their news from social media

Black And Latino Voters Flooded With Disinformation In Election's Final Days

October 30, 2020 · 7:49 AM ET

Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power

Misinformation on Facebook got six times more clicks than factual news during the 2020 election, study says

**Fake news spreads faster than true news on Twitter—thanks to people, not bots**

Tweets containing falsehoods were 70% more likely to be retweeted than truthful tweets

Gallup, Pew, Statista, Guardian

# Dealing with Misleading Content and Activities



## Identification

What information is being shared online?

# Dealing with Misleading Content and Activities

## Identification

What information is being shared online?

## Assessment

Is this misleading?
Is this coordinated?

# Dealing with Misleading Content and Activities



## Identification

What information is being shared online?

## Assessment

Is this misleading?
Is this coordinated?

## Actioning

What can we do to limit its harms?

# We make Identification and Verification Easier

## Identification

What information is being shared online?

## Assessment

Is this misleading?
Is this coordinated?

## Actioning

What can we do to limit its harms?

# Operationalize Metrics to Augment Disinformation Identification and Verification by <u>Anyone</u>

# Case Study: Russian state-backed media is a major part of its "pillars of disinformation"

01 **Official Govt. Communications**

02 **State-funded Global Messaging**

03 **Proxy Sources**

04 **Social Media**

05 **Cyber-enabled Disinformation**

U.S. DEPARTMENT of STATE

**GEC Special Report:**

**Pillars of Russia's Disinformation and Propaganda Ecosystem**

GEC Report

# RT.com and Sputnik News

Known sources of disinformation and influence operations.

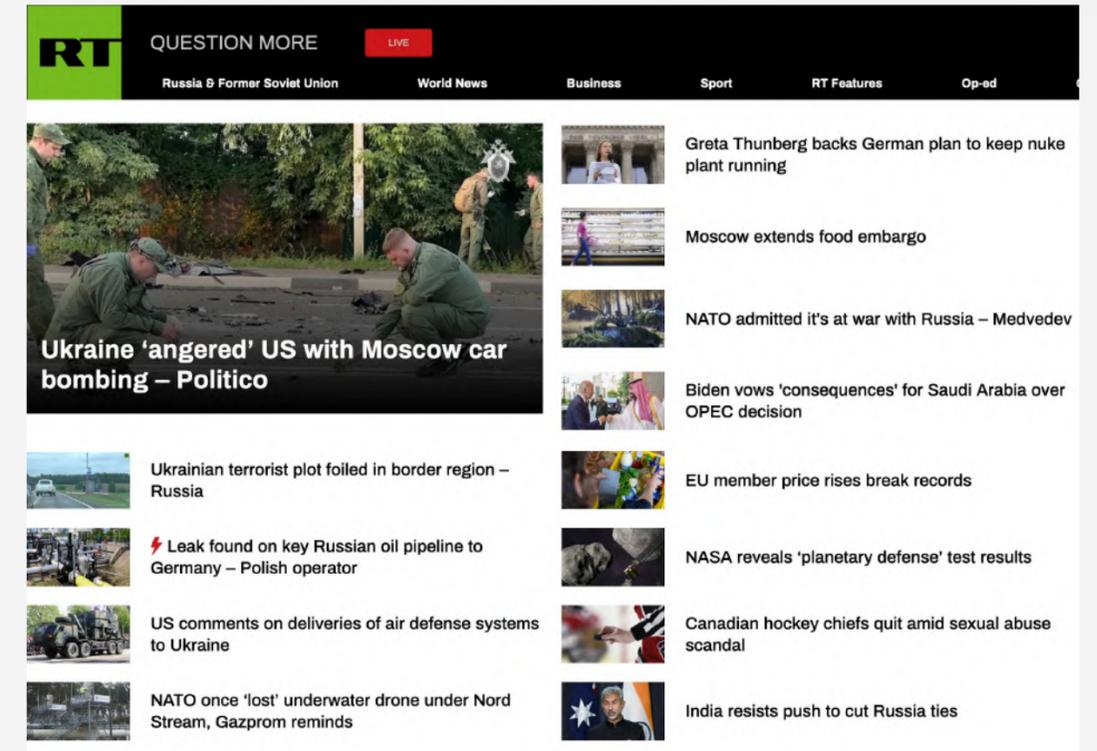We track 8000 articles posted since 2021.

Some believed to be manipulated.



GEC Special Report
KREMLIN-FUNDED MEDIA:
**RT AND SPUTNIK'S**
ROLE IN RUSSIA'S DISINFORMATION
AND PROPAGANDA ECOSYSTEM

EU bans RT, Sputnik over Ukraine disinformation
By Foo Yun Chee

**Weaponising news: RT, Sputnik and targeted disinformation**
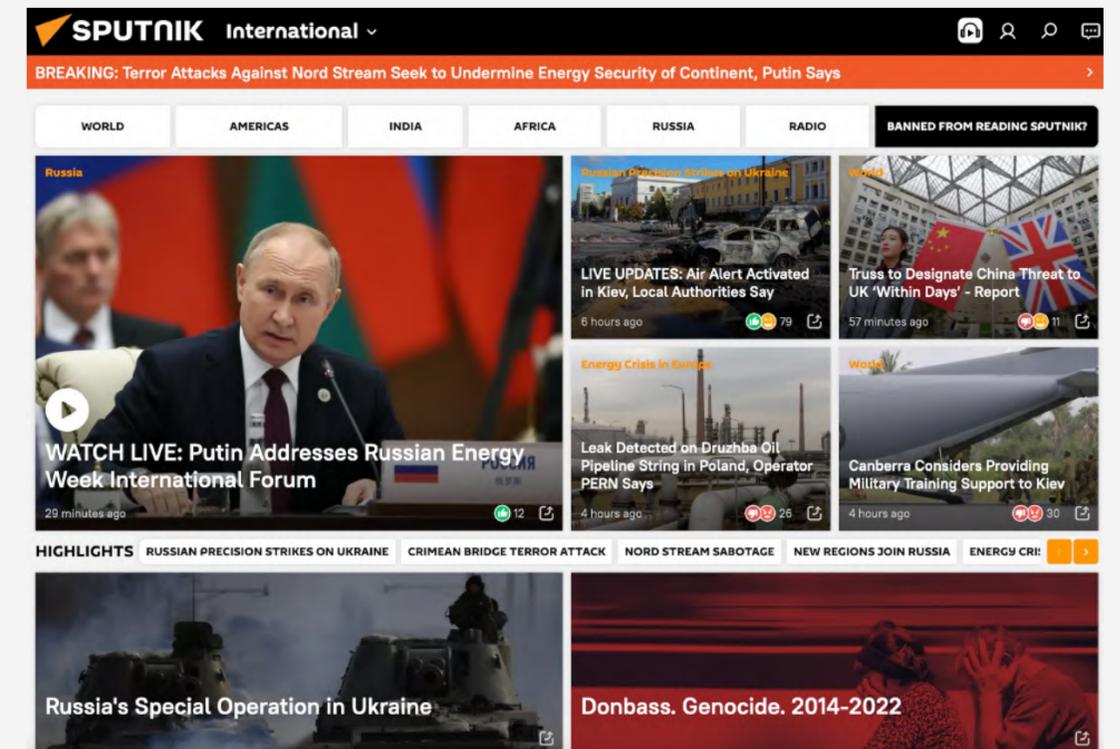
# How to help Identify Manipulative Narratives?

| headline | publication_date | article_content | article_url | source |
|---|---|---|---|---|
| Great Britain Imposes Sanctions Against Sputnik's Head Anton Anisimov | 2022-03-31 | London has expanded its sanctions list by making 14 additions, the web site of the British government says. … (text truncated) | https://sputniknews.com/20220331/great-britain-imposes-sanctions-against-sputniks-head-anton-anisimov-1094354237.html | sputnik |
| Hindu Group Marches to Support Russian Military Op, Slams Zelensky as Ukraine's No1 Enemy - Video | 2022-03-31 | Although the US and its allies imposed sanctions on Russia in response to the military operation in Ukraine, India's government has refused to give in to pressure… | https://sputniknews.com/20220331/hindu-group-marches-to-support-russian-military-op-slams-zelensky-as-ukraines-no1-enemy---video-1094367263.html | sputnik |
| US tells citizens in Russia and Ukraine to leave immediately | 2022-03-31 | State Department says Russia is targeting Americans after basketball star is jailed on drug charges… | https://www.rt.com/russia/553074-state-department-americans-leave-immediately/ | rt |



**RT and Sputnik News Articles**
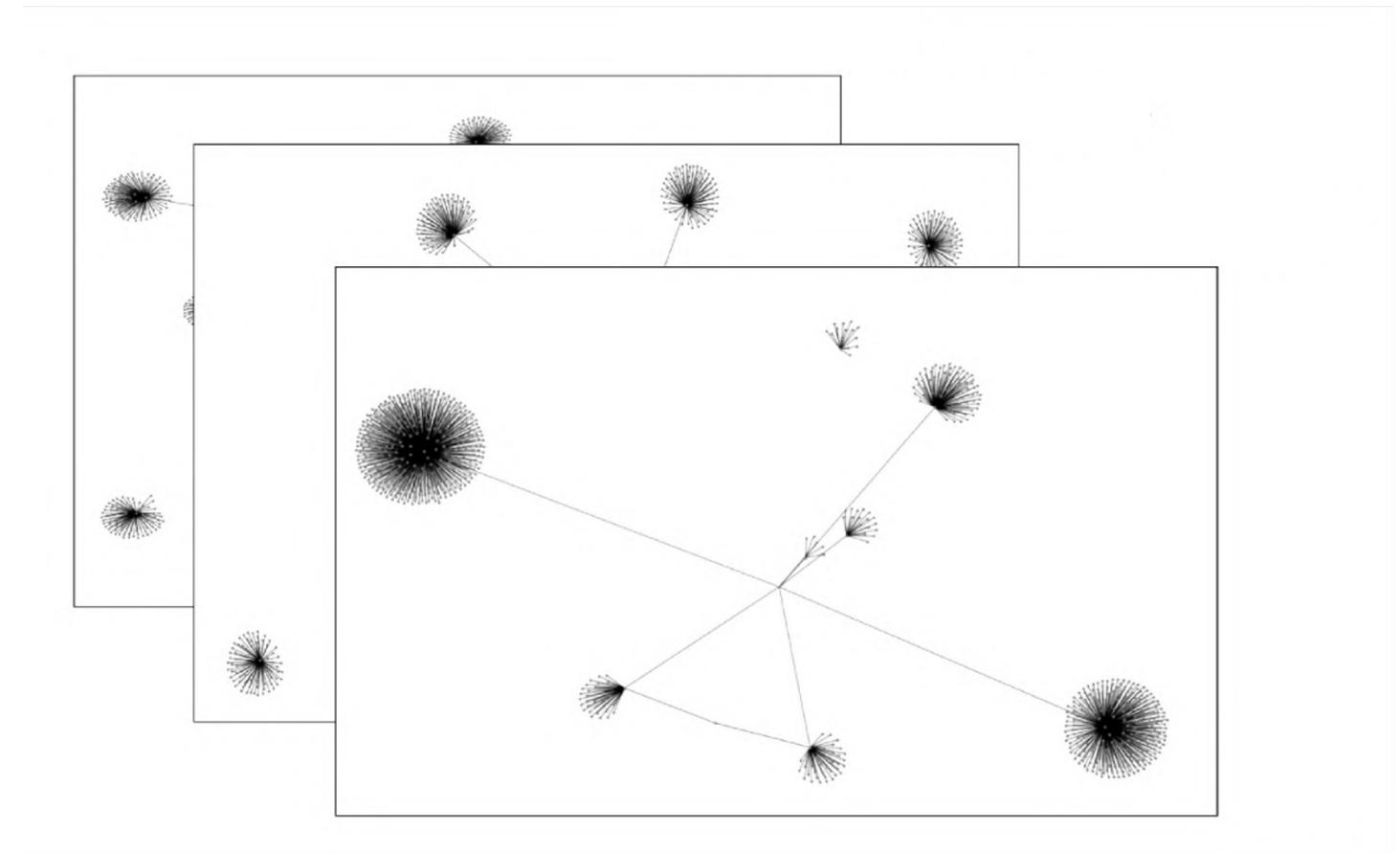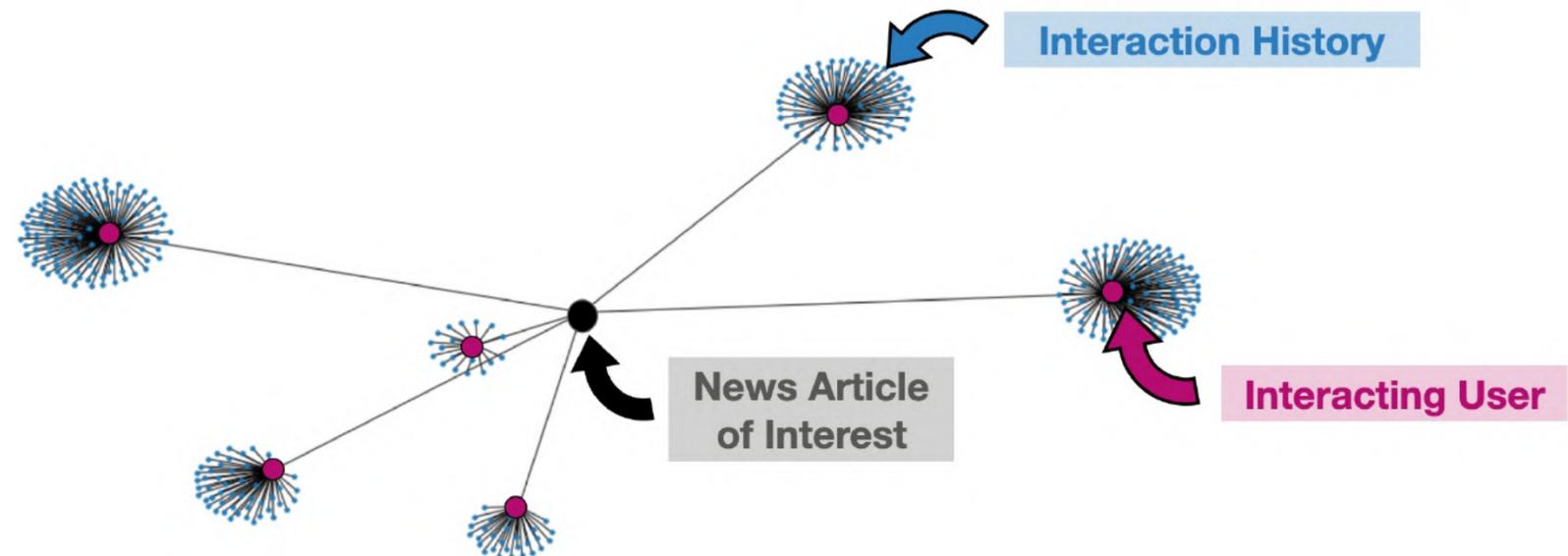
**Understand the spread of the post**

# Collect Real-time Data

**Tweets about the Article**

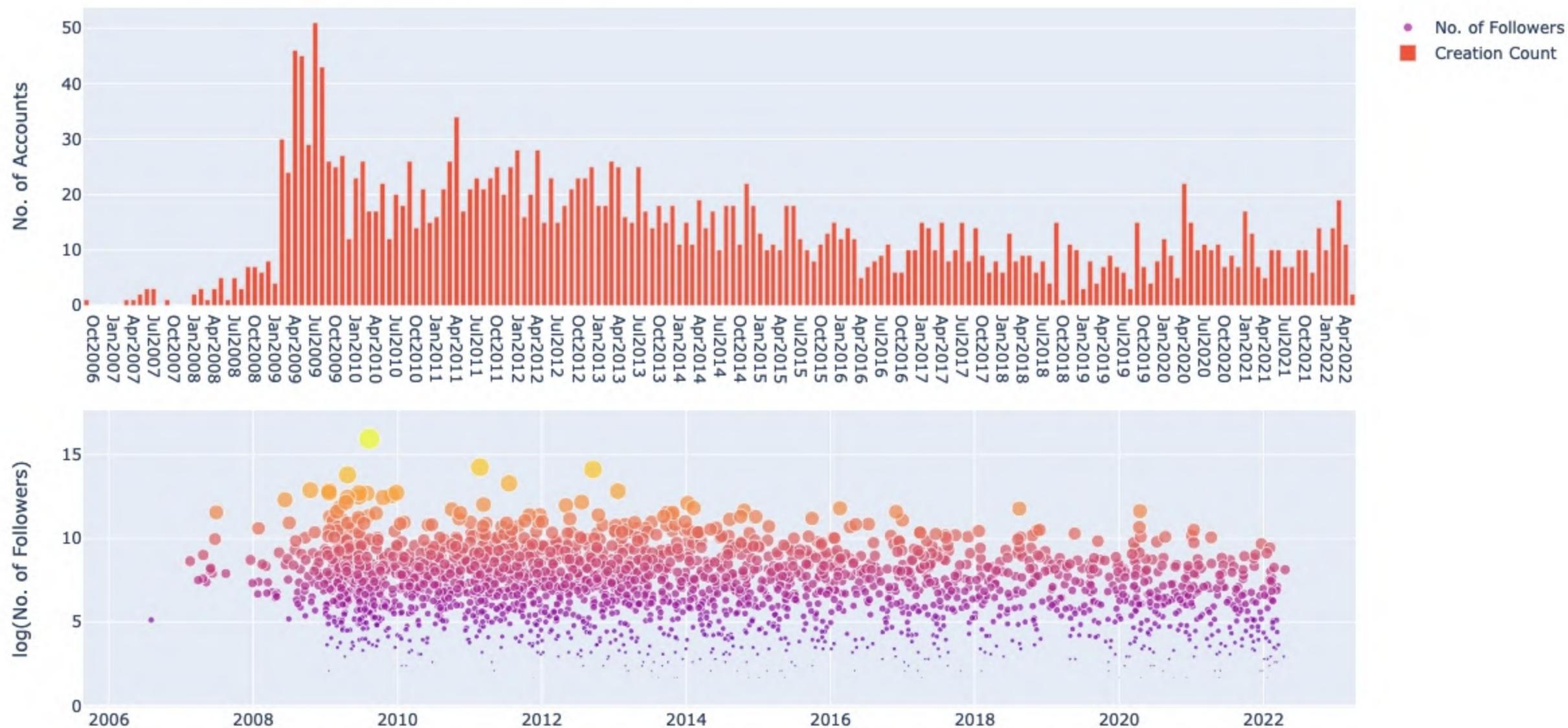**Accounts Posting the Tweets,**
**Account Tweet History, Metrics**

**Account  Followers,**
**Account Followers Tweet History**

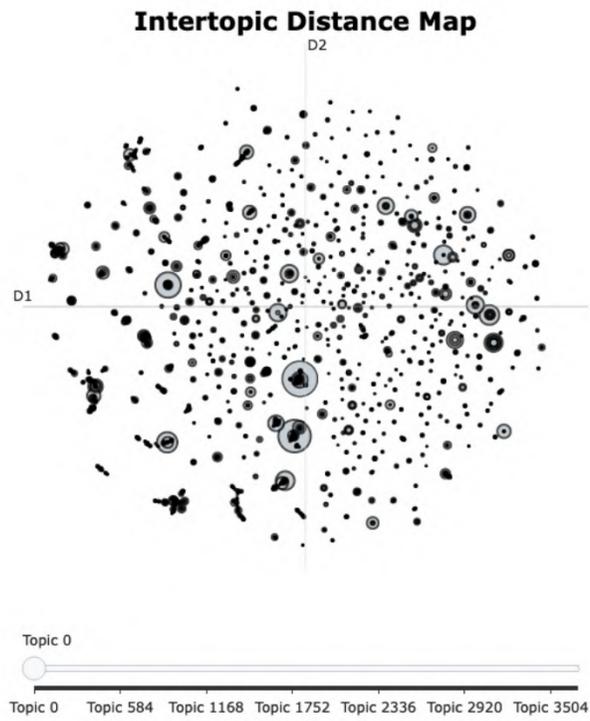**Favoriting, Retweeting,**
**  and Quote Tweeting Accounts**



Interaction History

News Article
of Interest

Interacting User

# Influential Accounts Created Years Ago are still Active!



Account Creation Month and Follower Count for Users

# Content-level analysis isn't enough to tell apart potentially coordinating accounts

# Find Interesting Groups: Signal from Noise

**Accounts that share potentially manipulated narratives (RT, Sputnik)**

**Accounts that follow 1 – 3 blue accounts**

**Accounts that follow 4+ blue accounts**

# Identify Narratives Among these Groups
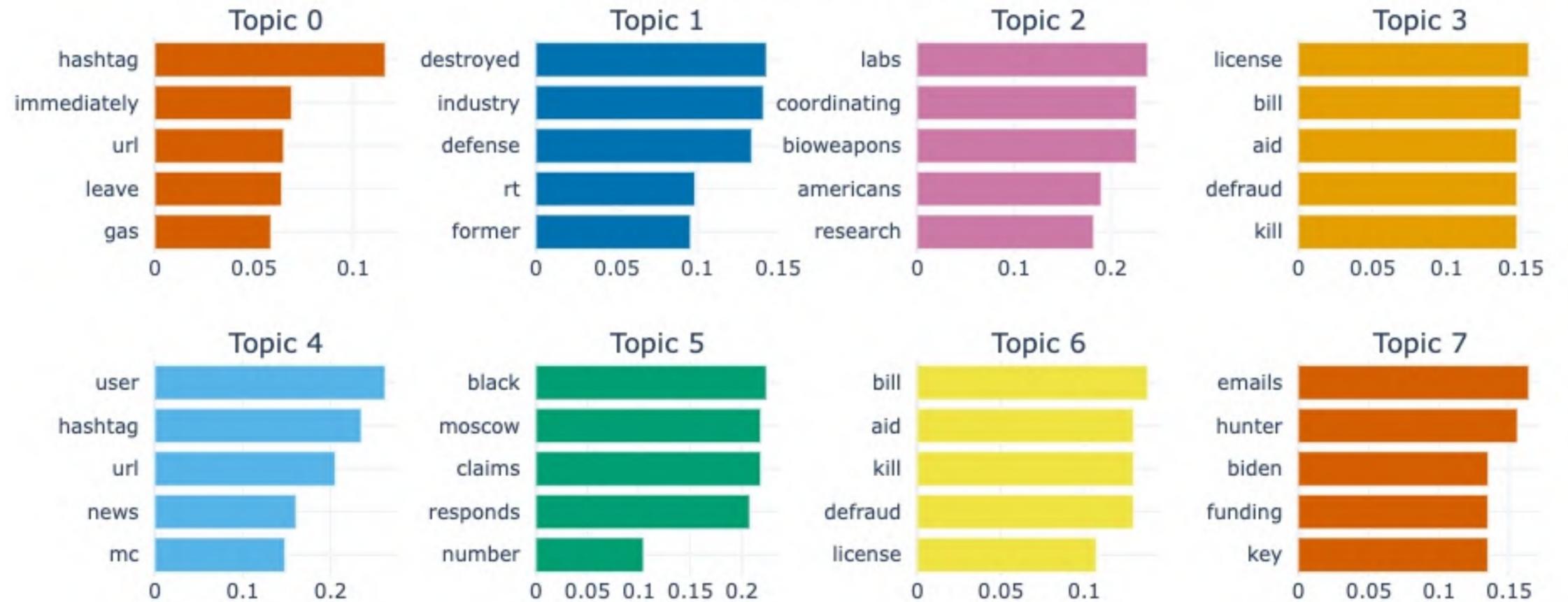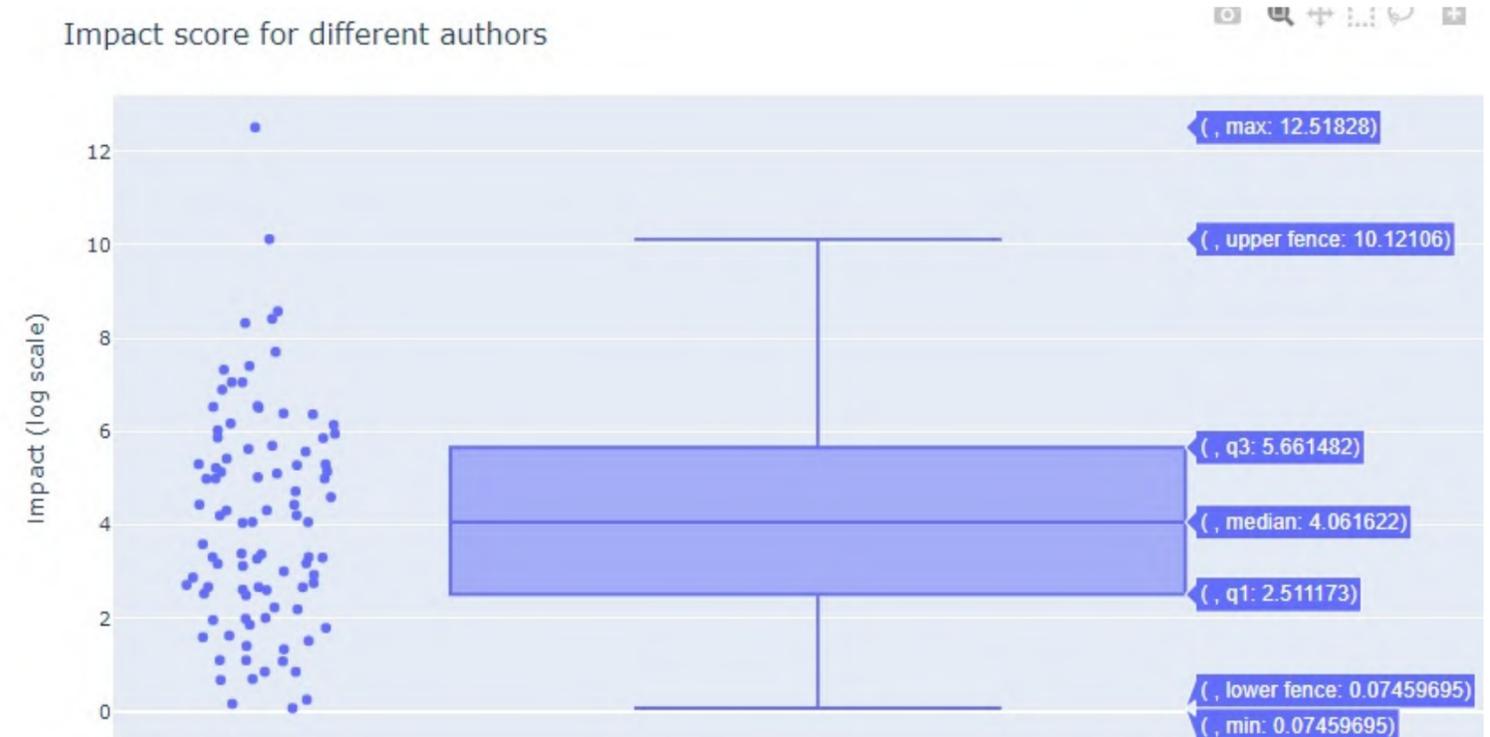
# Find Connections between Narratives

**List of Topics**

| topic_nr | top_10_words | topic_size |
|---|---|---|
| 0 | impeachment, democrats, impeached, impeach, republicans, bidens, inquiry, obstruction, quid, impeaching | 662 |
| 1 | normandy, agreements, minsk, format, contact, lugansk, ceasefire, donbass, agreed, subgroups | 520 |
| 2 | imposed, sanctions, embargo, agricultural, imports, products, introduced, restrictions, goods, exports | 499 |
| 3 | think, really, do, don, sort, necessarily, certainly, perhaps, kind, going | 485 |
| 4 | civilians, militants, shelling, joins, jihadists, hama, injured, killing, residential, killed | 480 |
| 5 | it, but, is, understand, this, rather, that, serious, confrontation, not | 462 |
| 6 | nazi, nazis, wwii, collaborators, nationalists, upa, galicia, heroes, bandera, nazism | 444 |
| 7 | nord, stream, pipeline, uniper, project, engie, wintershall, allseas, omv, liquefied | 425 |
| 8 | proposals, guarantees, eastward, eastwards, expansion, written, nato, ryabkov, expand, alliance | 419 |
| 9 | exercises, drills, troops, military, presence, exercise, rotational, zapad, deployment, deployed | 415 |
| 10 | votes, timoshenko, poroshenko, elections, vote, election, rada, petro, party, parliament | 401 |
| 11 | prepayment, cubic_x, naftogaz, cubic, gazprom, gas, gas_x, oettinger, prodan, transit | 398 |
| 12 | jit, mh, buk, crash, downed, malaysian, dutch, dubinsky, dsb, malaysia | 383 |
| 13 | issg, sar, syrian, jabhat, nusra, settlement, damascus, syria, al, astana | 355 |
| 14 | invasion, imminent, invade, hysteria, amassing, reports, incursion, kremlin, massing, troops | 338 |
| 15 | lethal, assistance, stinger, equipment, javelin, provide, aid, providing, portable, systems | 329 |
| 16 | precision, konashenkov, artillery, destroyed, civilian, residential, donetsk, dpr, forces, civilians | 285 |
| 17 | match, midfielder, finals, tournament, football, lewandowski, game, fans, league, cristiano | 272 |
| 18 | membership, alliance, nato, joining, bloc, join, hoop, jaap, accession, scheffer | 269 |
| 19 | crosstalking, soundcloud, rttv, crosstalk, clear, sleboda, talk, political, babich, podcast | 269 |

## Hierarchical Clustering
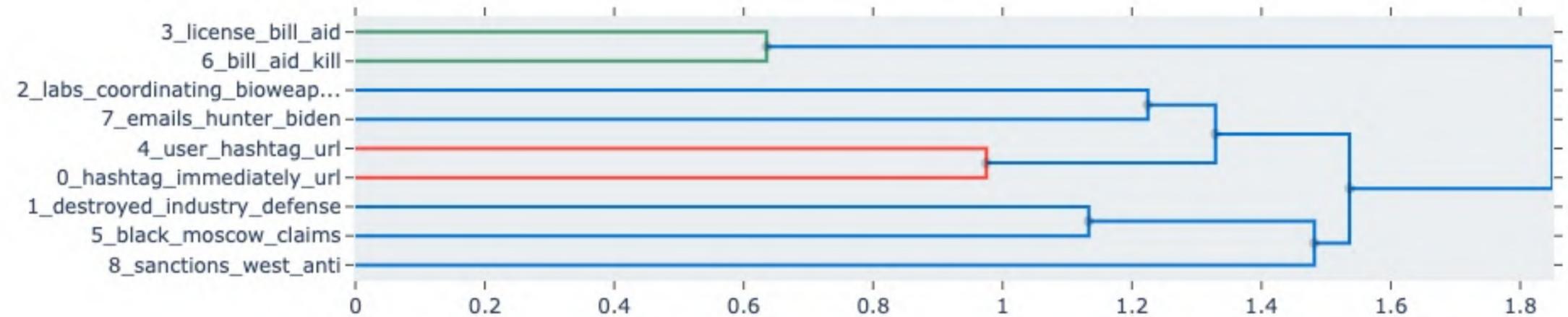
# Identify the most "Impactful" Accounts



| | username | followers_count |
|---|---|---|
| 44 | RT_com | 3087851 |
| 49 | SputnikInt | 356111 |
| 47 | ZANewsFlash | 29907 |
| 23 | AdsumP | 21122 |
| 80 | MendesOnca | 14952 |



Impact score for different authors

( , max: 12.51828)
( , upper fence: 10.12106)
( , q3: 5.661482)
( , median: 4.061622)
( , q1: 2.511173)
( , lower fence: 0.07459695)
( , min: 0.07459695)

Impact (log scale)

# Impact = Mean Toxicity x log(follower count)

# Some User Groups Share a LOT of Articles
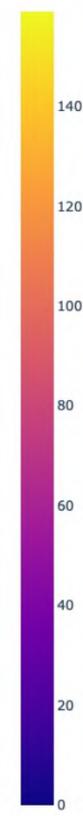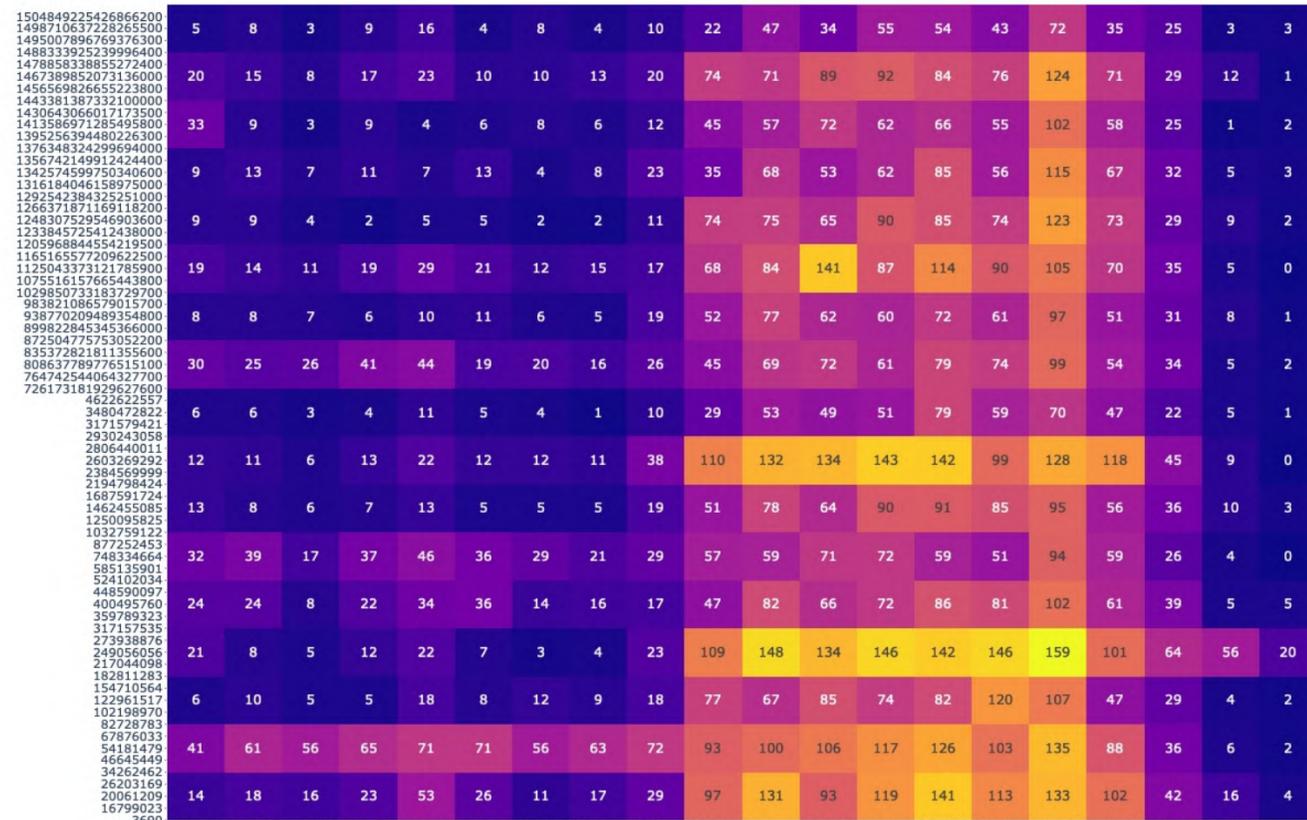


How often do users share an article
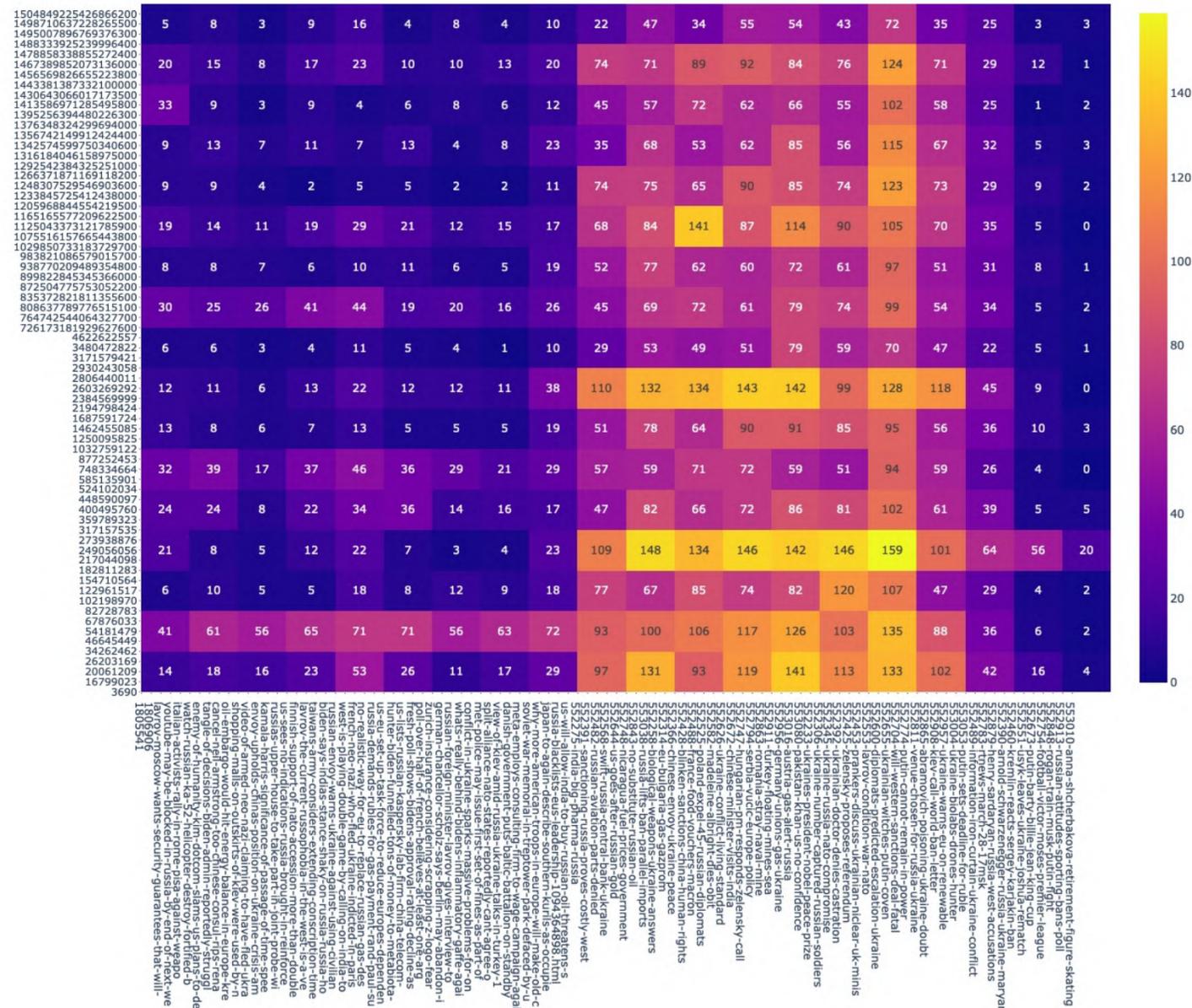
Heatmap showing user groups (x-axis) and chronologically ordered articles (y-axis) indicating a marked uptick in sharing of news around the time of the Russian invasion of Ukraine (late 2021–2022)

# Analysis Reveals Coordination Patterns



How often do users share an article

How often do users share an article

23
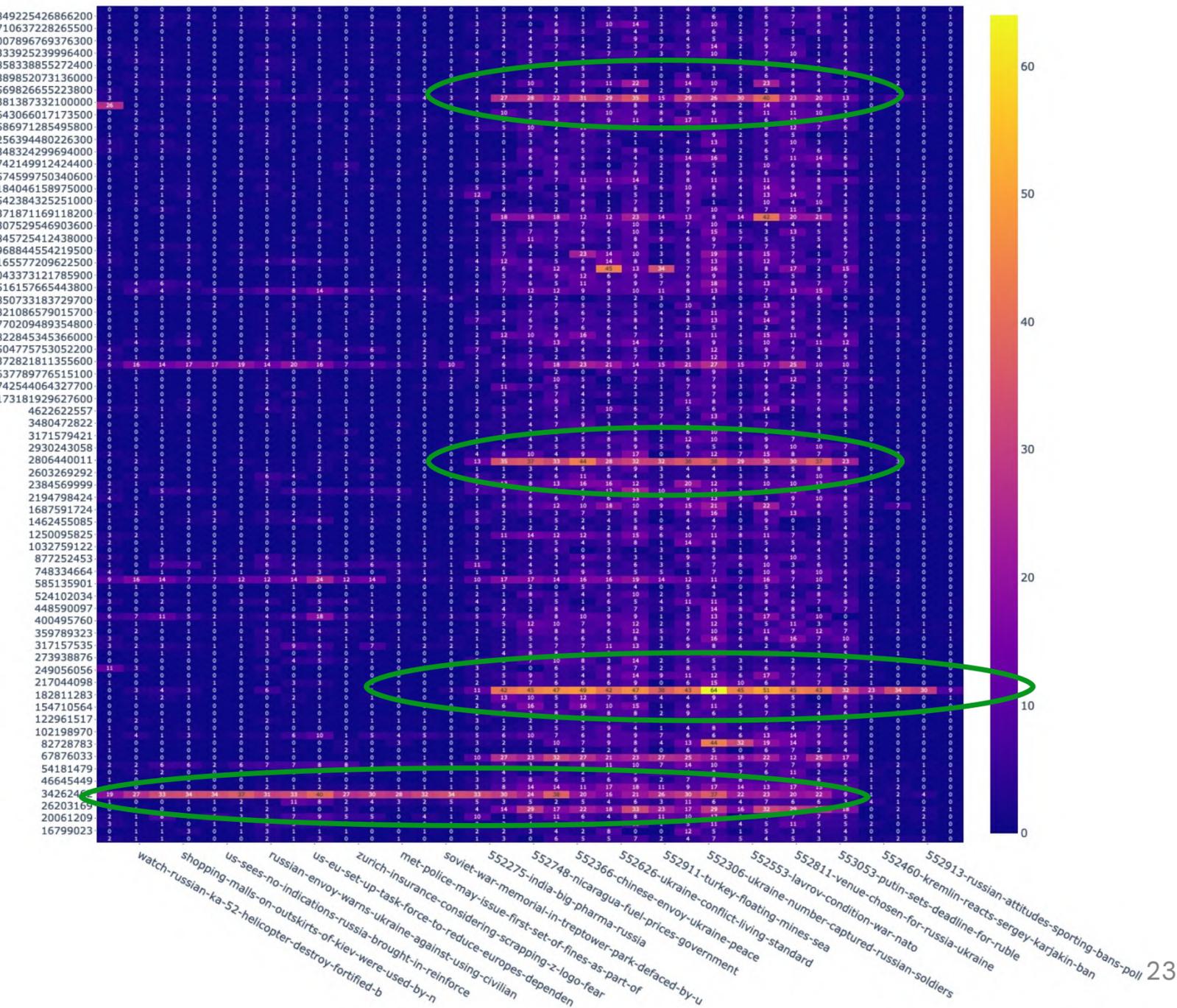
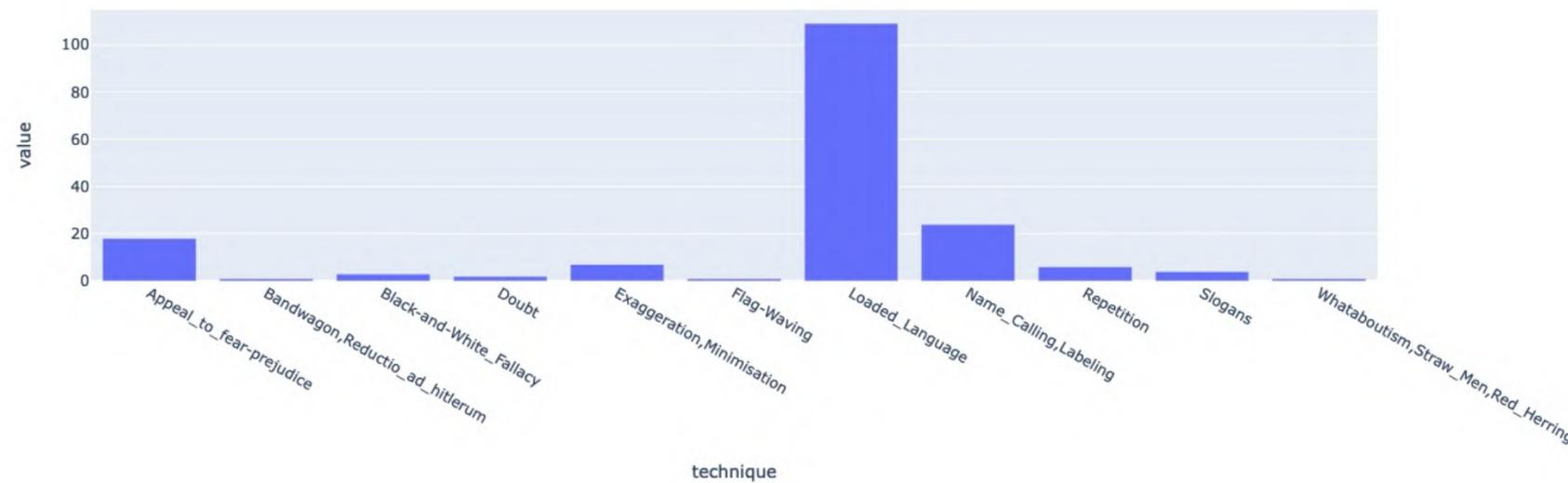# Identify Candidate Narratives with Unusual Trends

toxicity/tweet trend with time (size represents no. of tweets)



**Ukraine Aid Bill is a License to Kill**

**Russian Officials Speculating on FIFA Decision**

# Detecting Propaganda in the Content

# Making Identification and Verification Easier

**Governance**

How to go about all 3 steps in a principled manner?



**Identification**

What information is being shared online?



**Verification**

Is this misleading?
Is this coordinated?



**Actioning**

What can we do to limit its harms?

# Discover Optimal Policies with Simulations

**Problem**

## It's an Adversarial Space

Platforms slowly make policies that bad actors quickly adapt to bypass, undermining public trust and safety.

**Solution**

## Adaptive Governance

**Anticipate** adversarial behavior from coordinated networks.
Use simulation intelligence to test policy effectiveness and plug loopholes.

# Better Governance against Online Harms

01 **Track Superspreaders**

Identify inorganic content–sharing behavior to track superspreaders of misinformation.

02 **Model Information Flows**

Probabilistic AI recreates algorithm-driven information propagation across online communities.

03 **Predict Future Risks**

Detect emergent narratives in fringe and alt communities going beyond mainstream media.

04 **Deploy Interventions**

Determine when and where to intervene in order to optimally counter influence operations.

# Help Newsrooms and Policymakers Understand Audiences



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



| 14 | 23 | 86 ↑ |
| --- | --- | --- |
| Politics | Elections | Bernie Sanders |

@nytimes

## Community Membership

Estimate what communities are at risk for polarized debate and disinformation.
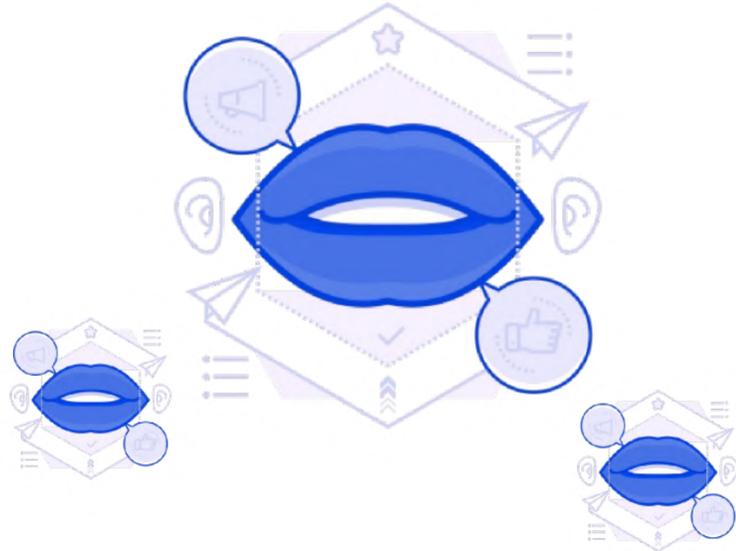
## Demographic Variance

Model trends, preferences, and topics of conversation for each demographic via cohort analysis.
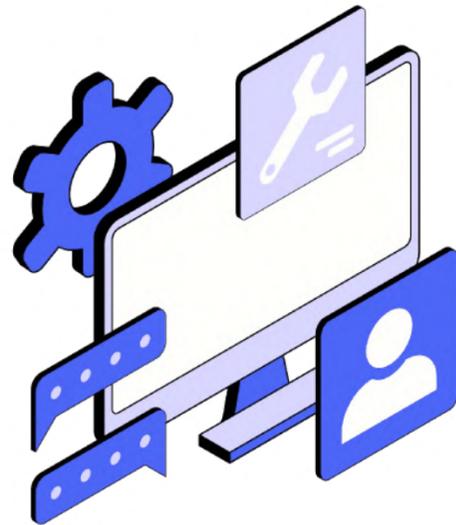
## Social Media Analysis

Track what other content audiences like to engage with to personalize content.

# Help Defend the Public against Information Operations

**Track Misleading Cliques**

Identify actors that frequently participate in the promotion of policy–violating content.
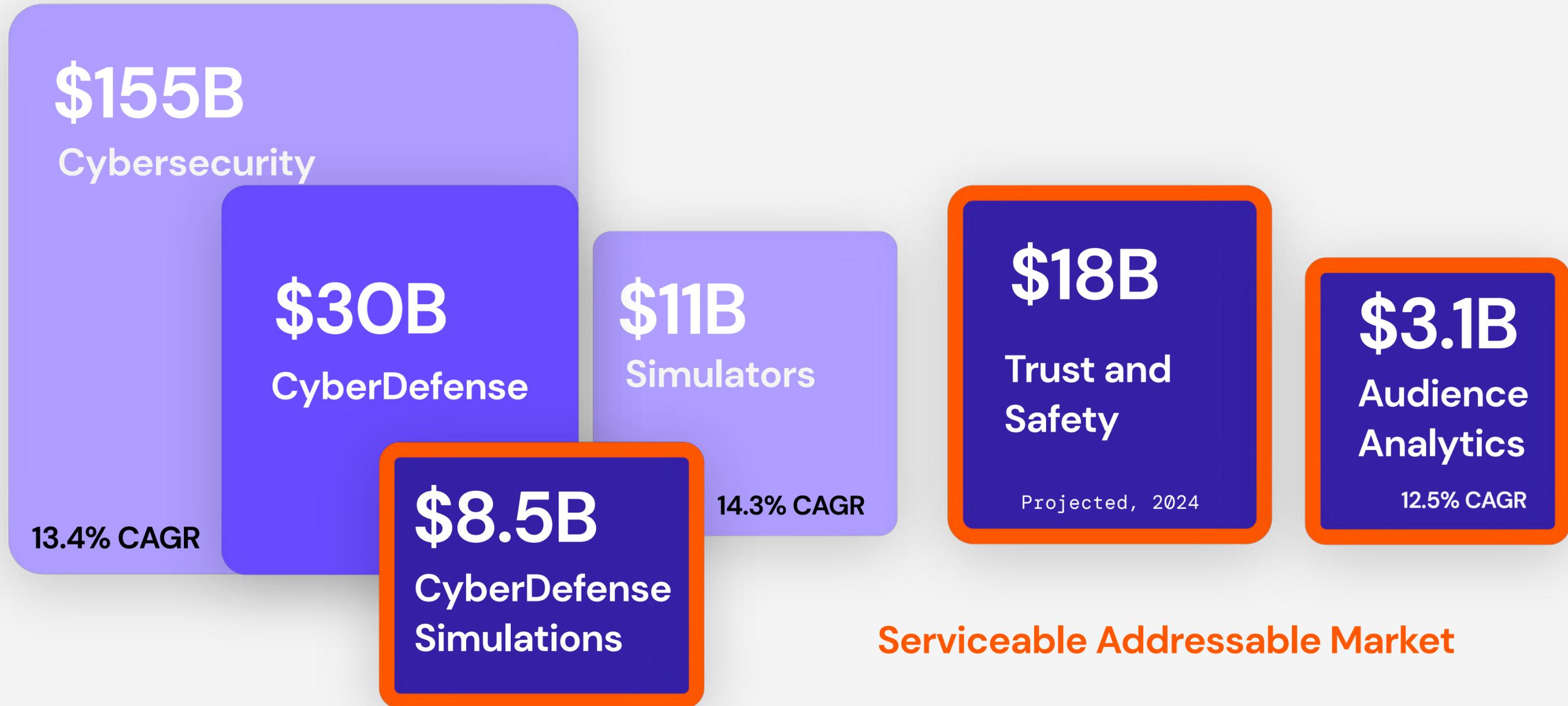
**Analyse Bad Actor Behavior**

Find patterns within strategic behavior exhibited by coordinated networks of actors.

**Deploy Countermeasures**

Provide safeguards and improve defense against information operations.

# Market Opportunity

**$155B**
Cybersecurity

13.4% CAGR

**$3OB**
CyberDefense

**$11B**
Simulators

14.3% CAGR

**$8.5B**
CyberDefense Simulations

**$18B**
Trust and Safety

Projected, 2024

**$3.1B**
Audience Analytics

12.5% CAGR

**Serviceable Addressable Market**

# Next Steps

### 1 Newsrooms

Working with journalists to make it easy to track the impact of digital publications.

### 2 Disinformation Tracker

Help organizations like the Wikimedia Foundation track coordinated behavior propagating misleading information online.

### 3 Cyberdefense Operations

Calibrate and deploy simulation models to track influence ops in the wild, online abuse working with defense agencies.

# Team

## Swapneel Mehta
Ph.D. @ NYU Data Science

**Founder**

### Founding Team

**Shwe Han**
DS @ B of A

**Jay Gala**
Resident @ IIT Madras

**Deep Gandhi**
MSc. @ UAlberta

**Dhara Mungra**
DS II @ Bombora

**Raghav Jain**
NLP @ IIT Patna

**Jhagrut Lalwani**
Ex – Goldman Sachs

### Shwetanshu Singh
DS @ Leaf Logistics

### Harshit Agarwal
Health ML @ Twitter

**Early Team / Tech. Advisors**

# What's Next

**Team Parrot**

Release public tool to track and analyse online propaganda.

**Govt. Funding**

Non-dilutive funding for cutting-edge technology ventures.

**Research**

Deploy and publish our simulator to study online disinformation.

**Cyber Defense**

Scale our product to automate detection of information operations.