iDRAMA Lab

Established 2016

What Social Media Means for Future History

Jeremy Blackburn

blackburn@cs.binghamton.edu / @jhblackb

https://mrjimmyblack.com / https://idrama.science

BINGHAMTON UNIVERSITY

STATE UNIVERSITY OF NEW YORK


WARNING
The contents of this talk covers some potentially disturbing subject matter

How a racist, s
mob forced Les
off Twitter

By Kristen V. Brown

SHARE

Over the past 24 hours
hateful vitriol on Twit

INTERNET 8 DECEMBER 2016

Pizzagate: How a 4Chan conspiracy went mainstream

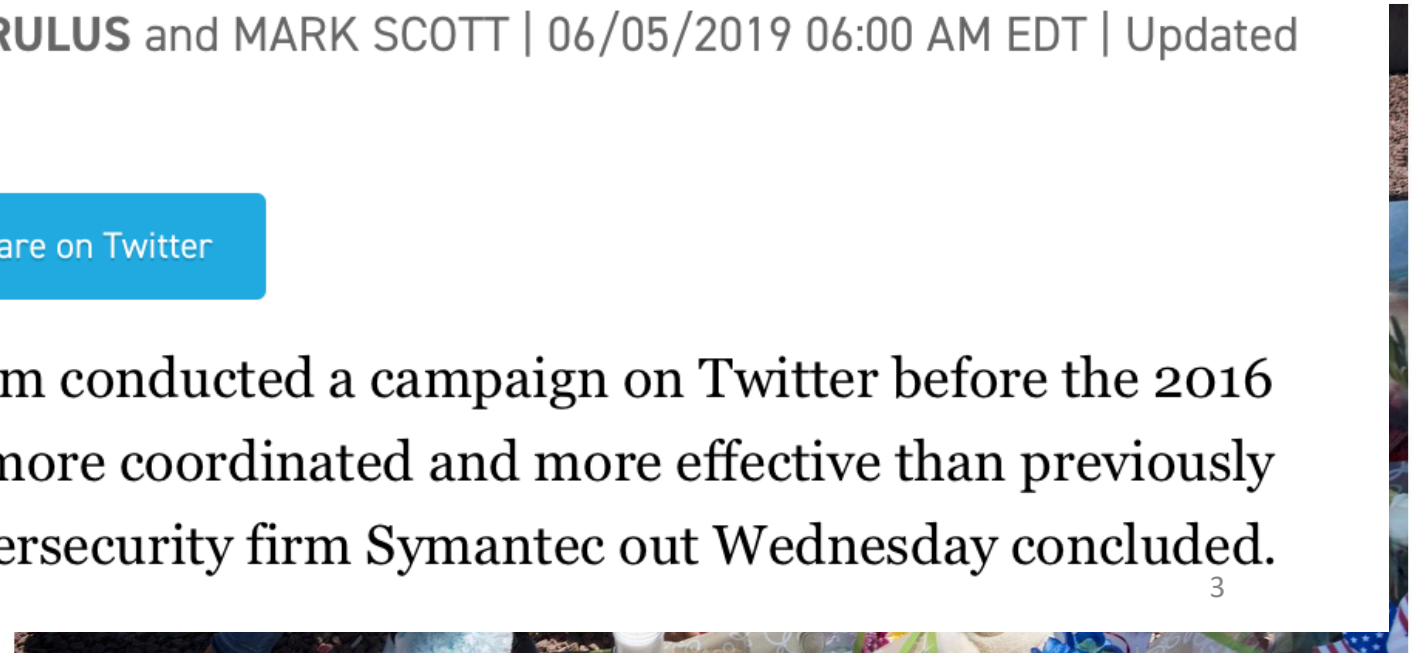
CYBERSECURITY

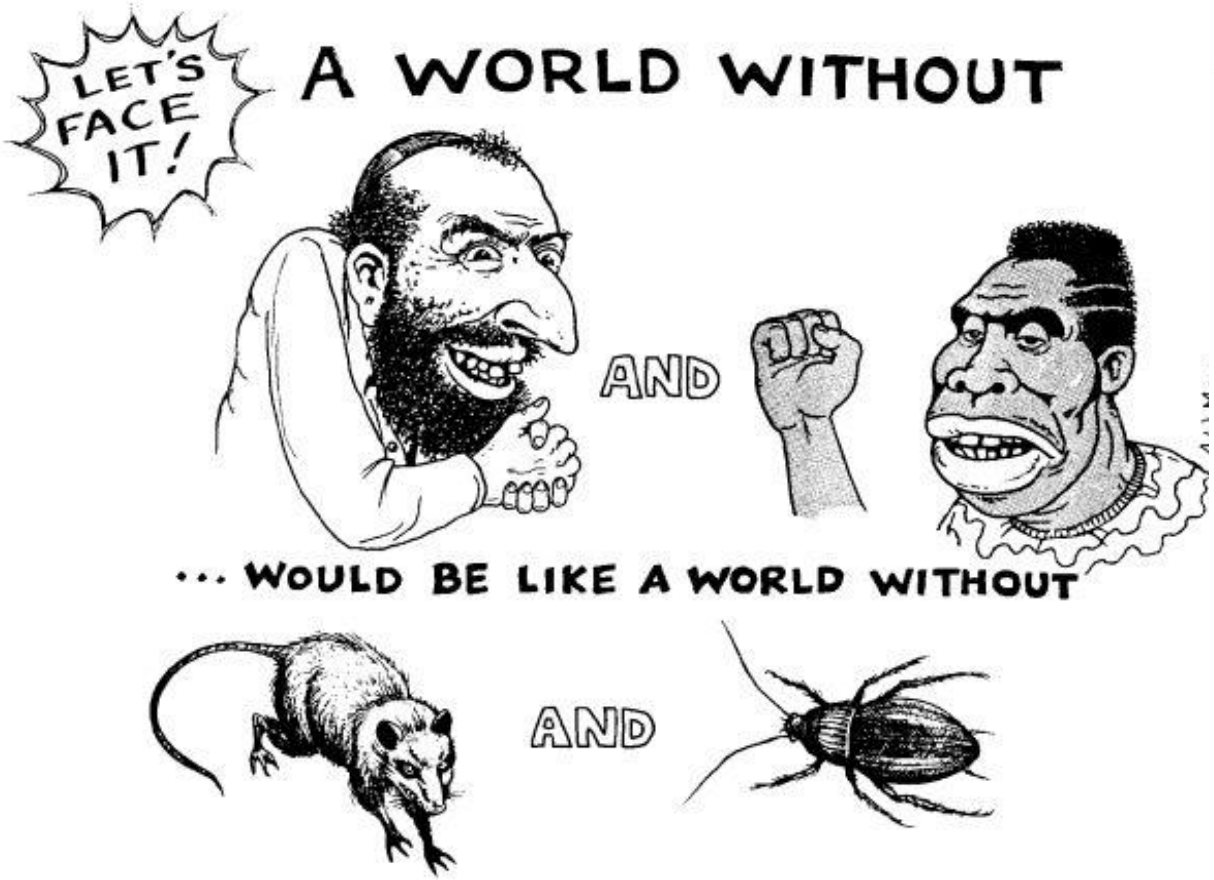
Russia's manipulation of Twitter was far vaster than believed

By **TIM STARKS**, **LAURENS CERULUS** and MARK SCOTT | 06/05/2019 06:00 AM EDT | Updated 06/05/2019 08:13 PM EDT

Share on Facebook Share on Twitter

Russia's infamous troll farm conducted a campaign on Twitter before the 2016 elections that was larger, more coordinated and more effective than previously known, research from cybersecurity firm Symantec out Wednesday concluded.




LET'S FACE IT!
A WORLD WITHOUT
AND
... WOULD BE LIKE A WORLD WITHOUT
AND

How the Christchurch shooter used memes to spread hate

Learning from years of online right-wing extremism, the shooter made his manifesto a weaponized shitpost.

By Aja Romano | @ajaromano | Mar 16, 2019, 4:30pm EDT

TECHNOLOGY

The Shooter's Manifesto Was Designed to Troll

The violent rhetoric was written for an audience.

TAYLOR LORENZ MAR 15, 2019

Shitposting is a slang term used to describe the act of posting trollish and usually ironic content designed to derail a conversation or elicit a strong reaction from people who aren't in on the joke. Certain aspects of the shooter's manifesto fall into this category. He includes Navy Seal Copypasta, a meme that originated on 4chan. He claims that Spyro: Year of the Dragon, a video game, taught him ethno-nationalism and that Fortnite taught him to "floss on the corpses," referring to a viral dance move from the game. These absurd references are meant to troll readers.

TECHNOLOGY

The Shooter's Manifesto Was Designed to Troll

The violent rhetoric was written for an audience.

TAYLOR LORENZ MAR 15, 2019

Before people can even begin to grasp the nuances of today's internet, they can be radicalized by it. Platforms such as YouTube and Facebook can send users barreling into fringe communities where extremist views are normalized and advanced. Because these communities have so successfully adopted irony as a cloaking device for promoting extremism, outsiders are left confused as to what is a real threat and what's just trolling. The darker corners of the internet are so fragmented that even when they spawn a mass shooting, as in New Zealand, the shooter's words can be nearly impossible to parse, even for those who are Extremely Online.

China / Diplomacy

Ukraine: photo of Chinese army heading to Russia is fake news, Beijing says

- Photo circulating on Twitter is a cropped version of a picture first published in 2021, China's internet watchdog says
- There had been reports in the US that Russia had asked China for military support in its Ukraine invasion – which both governments deny

South China Morning Post

SUBSCRIBE

China / Diplomacy

Ukraine: photo of Chinese army heading to Russia is fake news, Beijing says

- Photo circulating on Twitter is a cropped version of a picture first published in 2021, China's internet watchdog says
- There had been reports in the US that Russia had asked China for military support in its Ukraine invasion – which both governments deny

NEWS RUSSIA-UKRAINE FULL COVERAGE LIVE UPDATES PLAN YOUR VOTE POLITICS COVID U.S. NEWS OPINION WATCH NOW

RUSSIA-UKRAINE CONFLICT

U.S. soldiers alive, despite Russian 'fake news' report, U.S. military says

"They are accounted for, safe and not, as the article headline erroneously states, U.S. mercenaries killed in Donetsk People's Republic," the Tennessee National Guard said.

South China Morning Post

SUBSCRIBE

China / Diplomacy

Ukraine: photo of Chinese army heading to Russia is fake news, Beijing says

- Photo circulating on Twitter is a cropped version of a picture first published in 2021, China's internet watchdog says
- There had been reports in the US that Russia had asked China for military support in its Ukraine invasion – which both governments deny

Fact check: Queen Elizabeth death hoax stemmed from tabloid site

Bayliss Wagner
USA TODAY

Published 6:20 p.m. ET Feb. 23, 2022

NEWS RUSSIA-UKRAINE FULL COVERAGE LIVE UPDATES PLAN YOUR VOTE POLITICS COVID U.S. NEWS OPINION WATCH NOW

U.S. soldiers alive, despite Russian 'fake news' report, U.S. military says

They are accounted for, safe and not, as the article headline erroneously states, U.S. mercenaries killed in Donetsk People's Republic," the Tennessee National Guard said.

South China Morning Post

SUBSCRIBE

China / Diplomacy

Ukraine: photo of Chinese army heading to Russia is fake news, Beijing says

- Photo circulating on Twitter is a cropped version of a picture first published in 2021, China's internet watchdog says
- There had been reports in the US that Russia had asked China for military support in its Ukraine invasion – which both governments deny

Fact check: Queen Elizabeth death hoax stemmed from tabloid site

Bayliss Wagner
USA TODAY

Published 6:20 p.m. ET Feb. 23, 2022

NEWS RUSSIA-UKRAINE FULL COVERAGE LIVE UPDATES PLAN YOUR VOTE POLITICS COVID U.S. NEWS OPINION WATCH NOW

U.S. soldiers alive, despite Russian 'fake news' report, U.S. military says

They are accounted for, safe and not, as the article headline erroneously states, U.S. mercenaries killed in Donetsk People's Republic," the Tennessee National Guard said.

Lots of Challenges for Future Historians to Understand!

- Cyberbullying
- Cyberaggression
- Coordinated Hate Campaigns
- Radicalization
- Misogyny

Mis/Disinformation

Propaganda

Intentionally Sowing Discord

Conspiracy Theories

Election Interference

Computational methods
to understand and mitigate
information weaponization
at large scale

Data-driven analysis of complex and intertwined communities

Building proactive and preventive systems to counter it

The rest of this talk

- Part 1: Jerks online
- Part 2: State-sponsored actors
- Part 3: Mitigation techniques

Online Jerks

4chan


I CAN HAS
CHEEZBURGER?


I CAN HAS
CHEEZBURGER?


ANONYMOUS
Because none of us are as cruel as all of us.


I CAN HAS
CHEEZBURGER?


ANONYMOUS
Because none of us are as cruel as all of us.


I CAN HAS
CHEEZBURGER?


ANONYMOUS
Because none of us are as cruel as all of us.


I CAN HAS
CHEEZBURGER?


CLAIMS TO BE
BETTER THAN 4CHAN
USES 4CHAN MEMES
CONSTANTLY

ANONYMOUS
Because none of us are as cruel as all of us.


I CAN HAS
CHEEZBURGER?


CLAIMS TO BE
BETTER THAN 4CHAN


Donald J. Trump
@realDonaldTrump
Follow
"@codyave: @drudgereport @BreitbartNews
@Writeintrump "You Can't Stump the Trump"
youtube.com/watch?v=MKH6PA… "

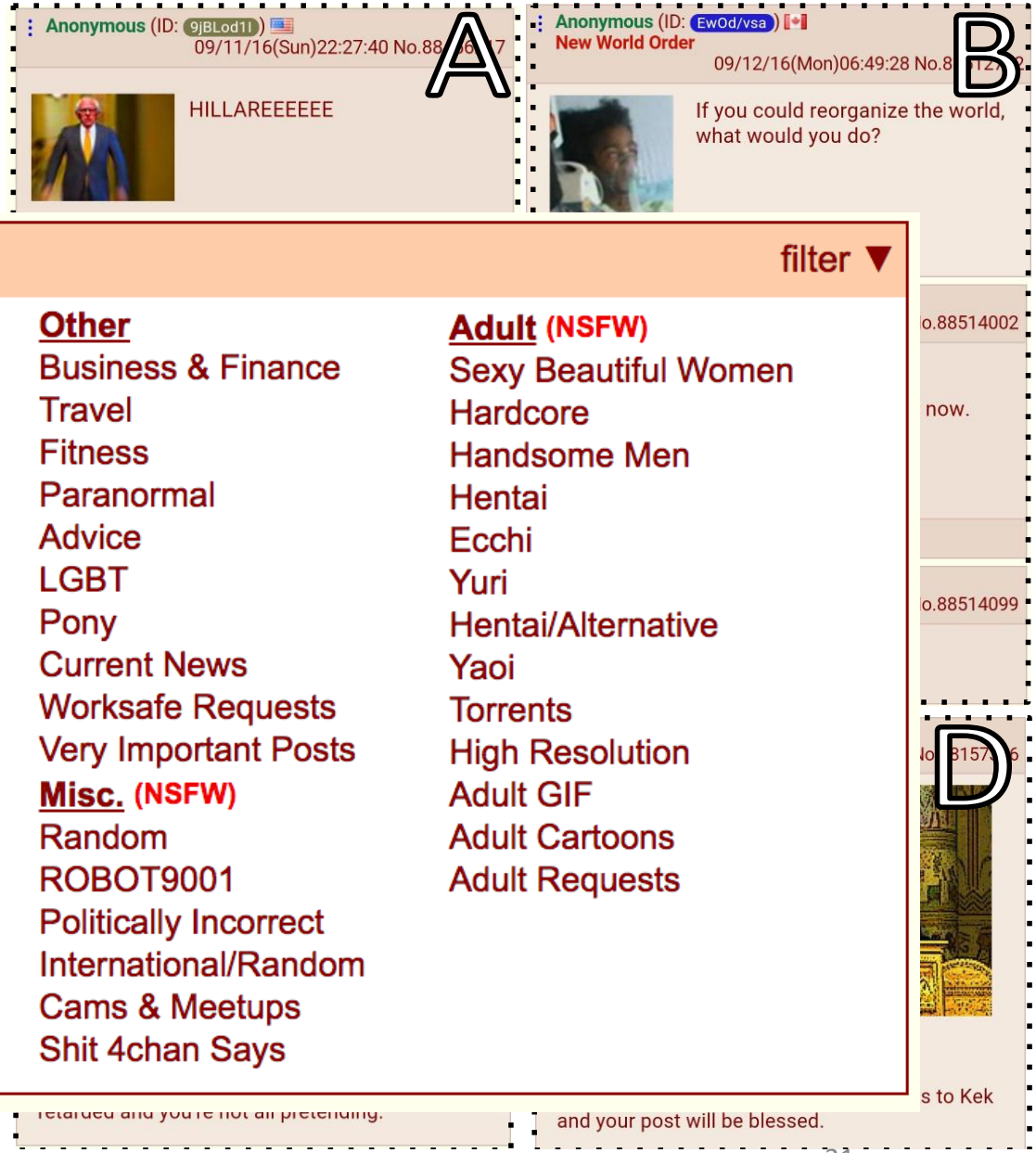
N MEMES
ANTLY
ANONYMOUS
Because none of us are as cruel as all of us.


I CAN HAS
2004
2012
2014
2016
WHERE DID EVERYTHING
GO SO REICH

What Exactly *Is* 4chan

- An image board
- Conversations grouped into *threads*
- An “original poster” (OP) creates a new thread by making a post
  - Single image attached
- Other users can reply
  - With or without images

What Exactly *Is* 4chan



/pol/ - Politically Incorrect

File: sticky.jpg (571 KB, 1600x1131)

Welcome to /pol/ - Politically Incorrect

Anonymous (ID: Iv9eTFJC) 01/08/15(Thu)23:40:44 No.40489590

[Reply]

This board is for the discussion of news, world events, political issues, and other related topics.

Off-topic and /b/-tier threads will be deleted (and possibly earn you a ban, if you persist). Unless they are quality, well thought out, well written posts, the following are examples of off-topic and/or /b/-tier threads:

>Red pill me on X. (with no extra content or input of your own)
>Are X white?
>Is X degeneracy?
>How come X girls love Y guys so much?
>If X is true, then how come Y? Checkmate Z.

The variety of threads allowed here are very flexible and we believe in freedom of speech, but we expect a high level of discourse befitting of the board. Attempts to disrupt the board will not be tolerated, nor will calls to disrupt other boards and sites.

/pol/ - Politically Incorrect

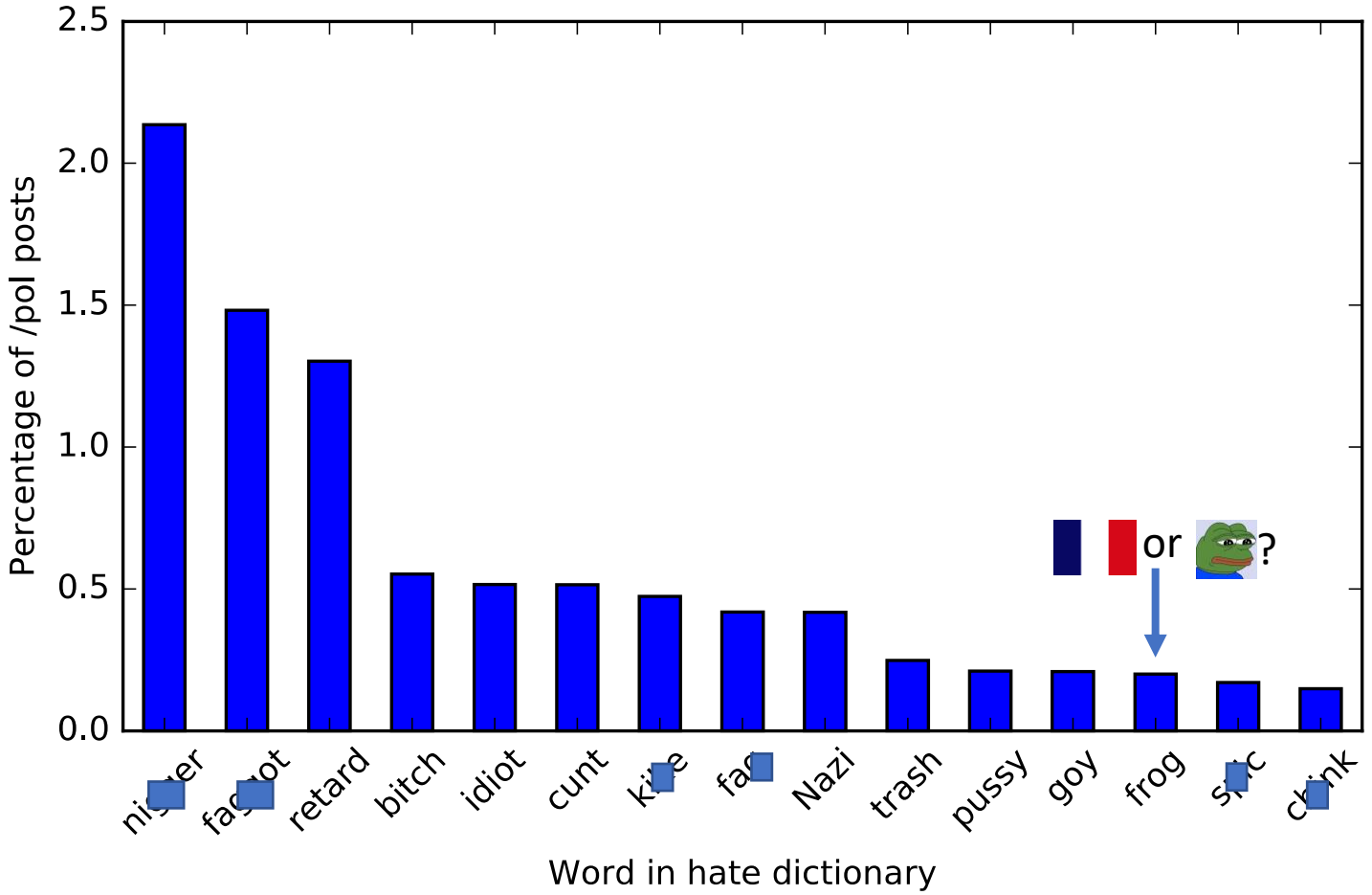
Anonymity and Ephemerality

- Users don’t need to register
  - Anonymity is default (and preferred) behavior
- There is a degree of permanence and identifiability
  - Can enter a user name along with a post (still no accounts)
- Threads are deleted after a while
  - All posts deleted after about a week

Are These Guys Racists?

- Crowdsourced dictionary
  - Manually filtered a bit

- /pol/ by far most hate speech use
  - /pol/ 12%
  - /sp/ 7.3%
  - /int/ 6.3%
  - Twitter 2.2%



Raids

- Attempts to disrupt another site
- *Not* a DDoS
- Disrupts *community* that calls service home, not the service itself
- Raids are a favorite past time on 4chan
  - “Pool’s closed!”
- Have become less “funny” and more “scary” lately
- It’s a *socio-technical* problem

Raids

- Attempts to disrupt another site
- *Not* a DDoS
- Disrupts *community* that calls service home, not the service itself
- Raids are a favorite past time on 4chan
  - “Pool’s closed!”
- Have become less “funny” and m e “s
- It’s a *socio-technical* problem

This is something new…

Using technology to exploit fundamental aspects of human nature

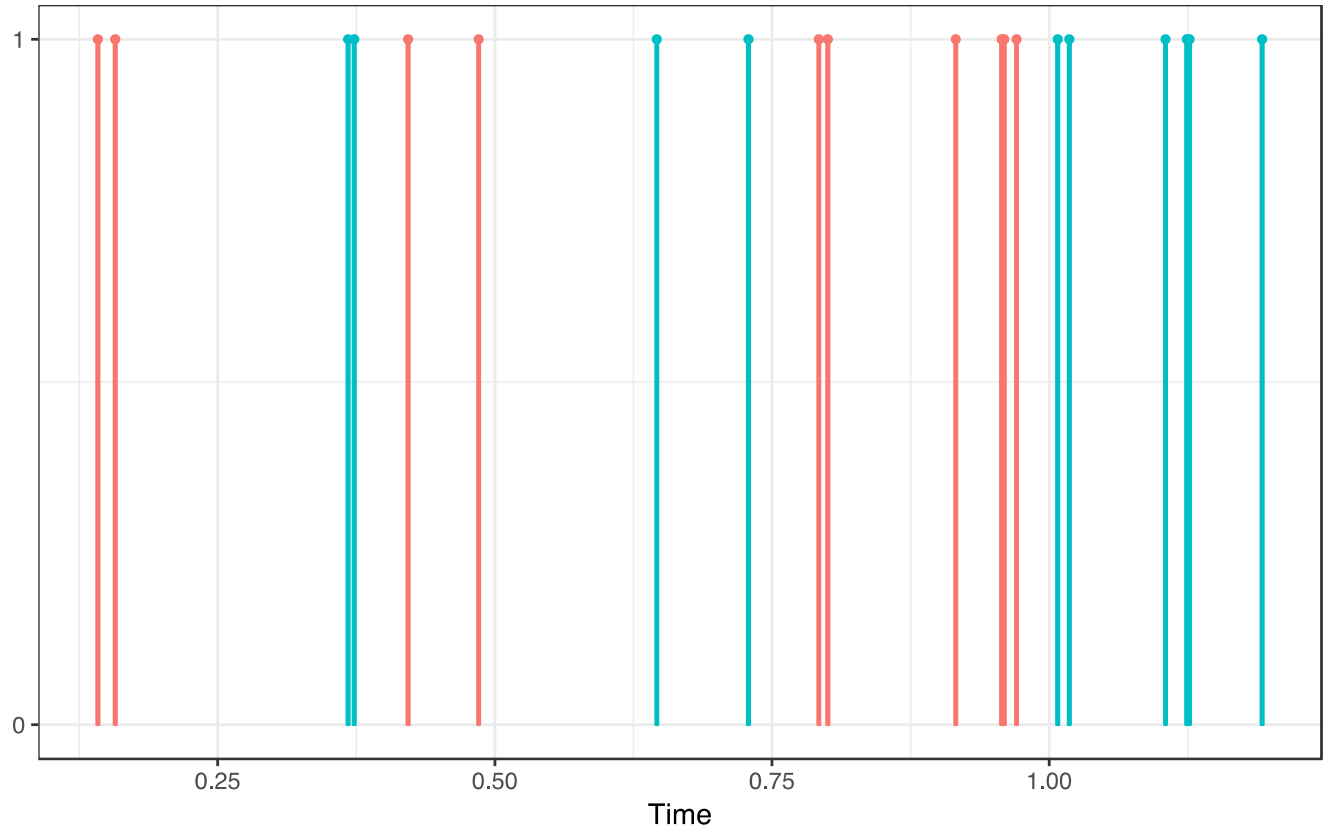
Raids on YouTube Videos

- YouTube is most commonly linked domain on /pol/
- We anecdotally observed them being raided
  - Plural of anecdote is not data…
- Can we find *quantitative* evidence of raids?!

How Might A Raid Happen?

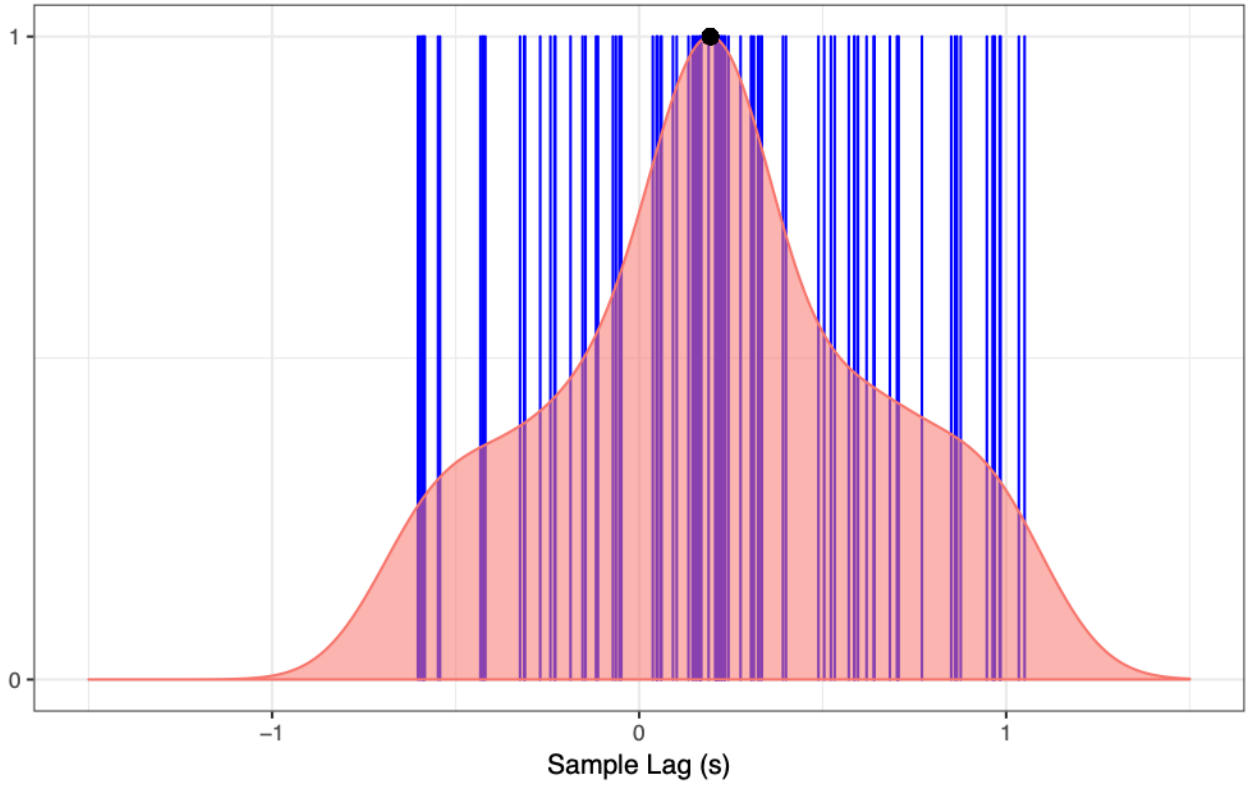
- Someone posts a YouTube link
  - Maybe with a prompt like “you know what to do”
- Thread is an aggregation point for raiders
  - E.g., “Hah! I called that person a n**ger!”
- If raid is taking place:
  - /pol/ thread and YT comments *synchronized?*

What Do We Mean By “Synchronization?”



- Two time series
- Second randomly shifted from first by 0.2 on avg

What Do We Mean By “Synchronization?”



Blue lines →

per-sample lag

Red area →

density of the lags

Peak of density curve = 0.2

Anonymous (ID: DaWZlbeh) 10/13/16(Thu)20:08:55 No.92772435

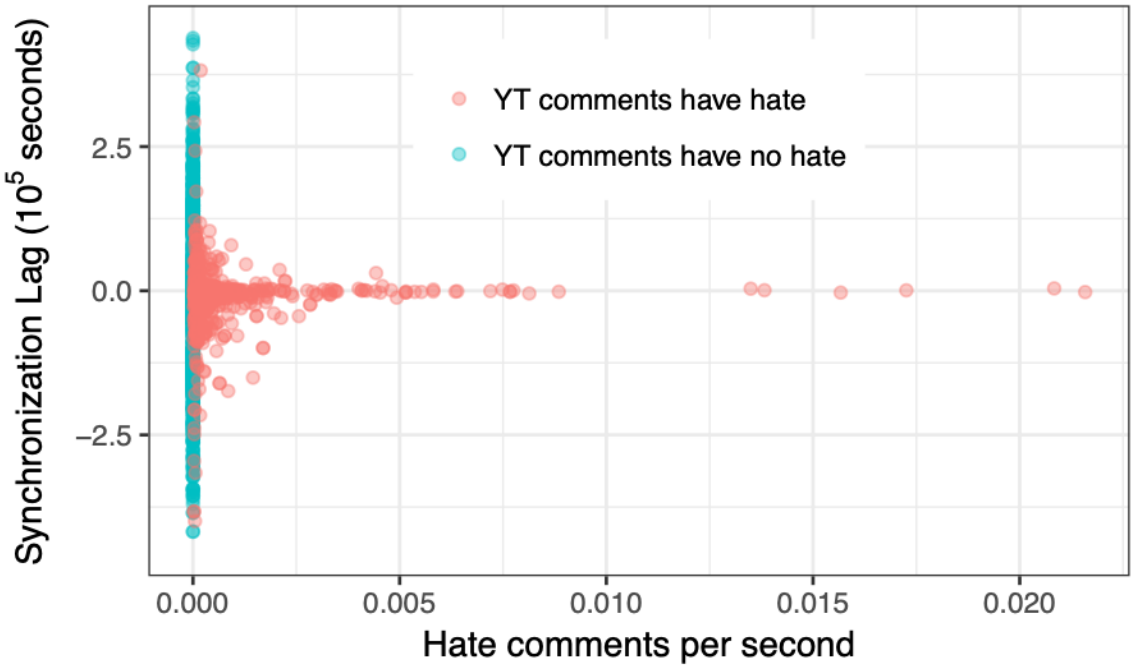
>>92765569

Physicist finalist here, can confirm the maths is legitimate. Holy fuck

How Can We *Validate?*

- Just because there is synchronization, doesn't mean a raid is going on
  - In fact, we expect *some* background noise
- We need to come up with metric to validate
- If a raid is happening, we would expect to see elevated levels of hate speech
  - → more hate comments per second

Evidence For Raids



Anonymous (ID: dvirZvG7) 10/15/16(Sat)02:00:33 No.92962384
File: tmp_24950-147365898299716(...).gif (389 KB, 320x240)

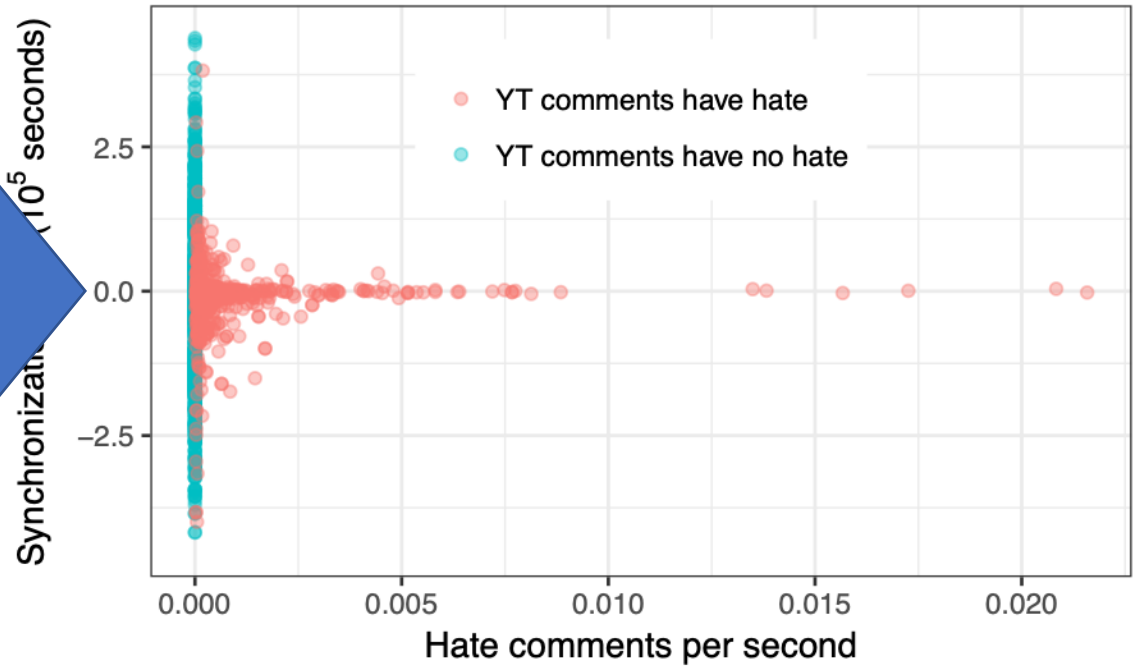
>>92957853 (OP)

The SI unit describing one hate comment per second (1hc/s) is a Kek. Using complex mathematics, we have discovered that the internet in general has a baseline of about 6 Kk, but introduce /pol/, and that number jumps up over a thousand fold, resting around 8.814 kKk. This number increases still further when you factor in the Australian Coefficient.

As a scientist, I find this shocking and disgusting. They simply cannot be allowed to keep getting away with this.

Evidence For Raids

As lag approaches zero (i.e., more synchronization) we see more hate comments per second on YT



Anonymous (ID: dvirZvG7) 10/15/16(Sat)02:00:33 No.92962384
File: tmp_24950-147365898299716(...).gif (389 KB, 320x240)

>>92957853 (OP)

The SI unit describing one hate comment per second (1hc/s) is a Kek. Using complex mathematics, we have discovered that the internet in general has a baseline of about 6 Kk, but introduce /pol/, and that number jumps up over a thousand fold, resting around 8.814 kKk. This number increases still further when you factor in the Australian Coefficient.

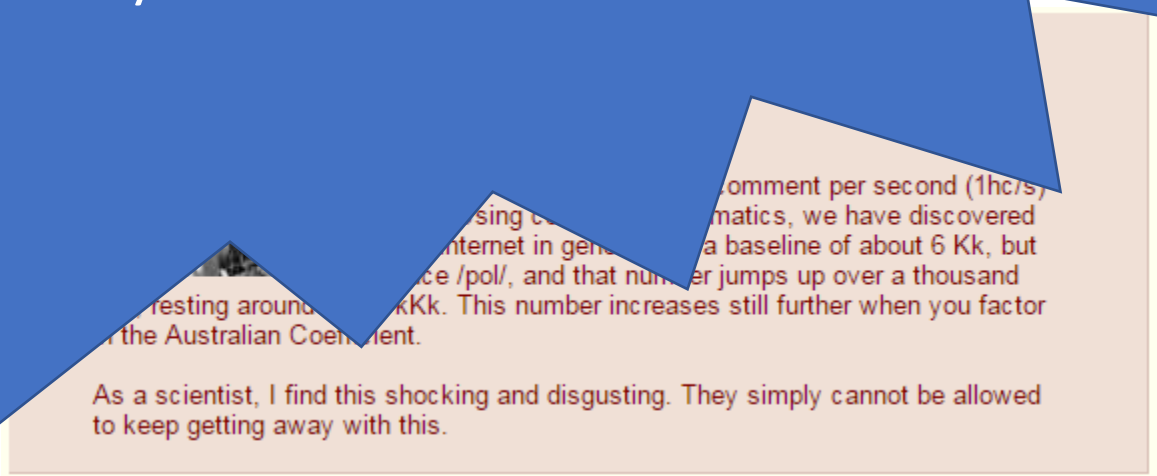
As a scientist, I find this shocking and disgusting. They simply cannot be allowed to keep getting away with this.

Evidence For Raids

Bottom line:

Fringe communities on the Web are not self contained

They effect other Web communities!



Summary

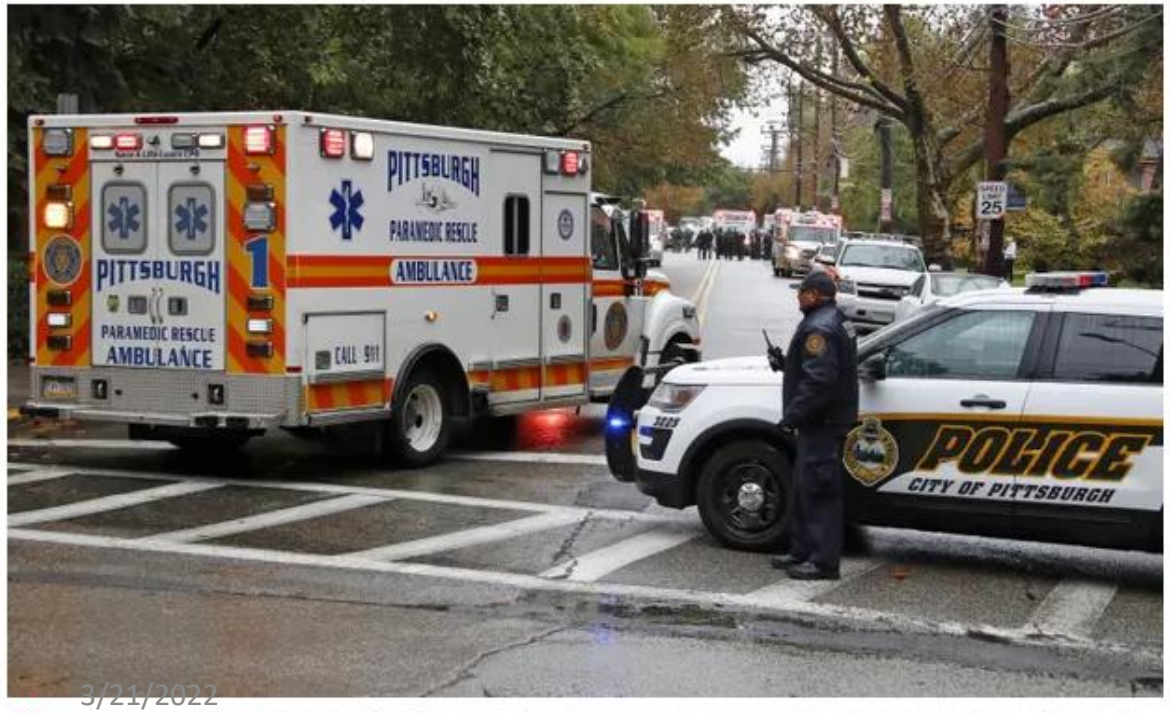
- 4chan is a notorious fringe Web community
- While it has always been a not very nice place, it has gotten worse
  - Hate speech, disinformation, conspiracy theories, violence…
- We have quantitative evidence they are affecting the rest of the Web
- They are not the *only* extremist community on the Web

Antisemitism is on the rise

Antisemitism is incubated on social media platforms

Pittsburgh shooting: suspect railed against Jews and Muslims on site used by 'alt-right'

Robert Bowers appears to have used the platform Gab to accuse Jews of bringing 'evil Muslims' into US



▲ An ambulance arrives at the Tree of Life synagogue where a shooter opened fire. Photograph: Gene J Puskar/AP

Robert Bowers @onedingo
2 hours ago

HIAS likes to bring invaders in that kill our people.

I can't sit by and watch my people get slaughtered.

Screw your optics, I'm going in.

Comments Repost Quote

Current efforts to understand and combat Antisemitism

- There are numerous organizations who deal with antisemitism

- They have a lot of “historical” knowledge and expertise
- But, their methodology faces problems in the online world
  - It does not scale
  - It is slow
  - It does not deal well with emerging trends

What can we **really** do about this?

- Use scalable data-driven methods to detect, quantify, and understand
  - Antisemitic imagery (i.e., memes)

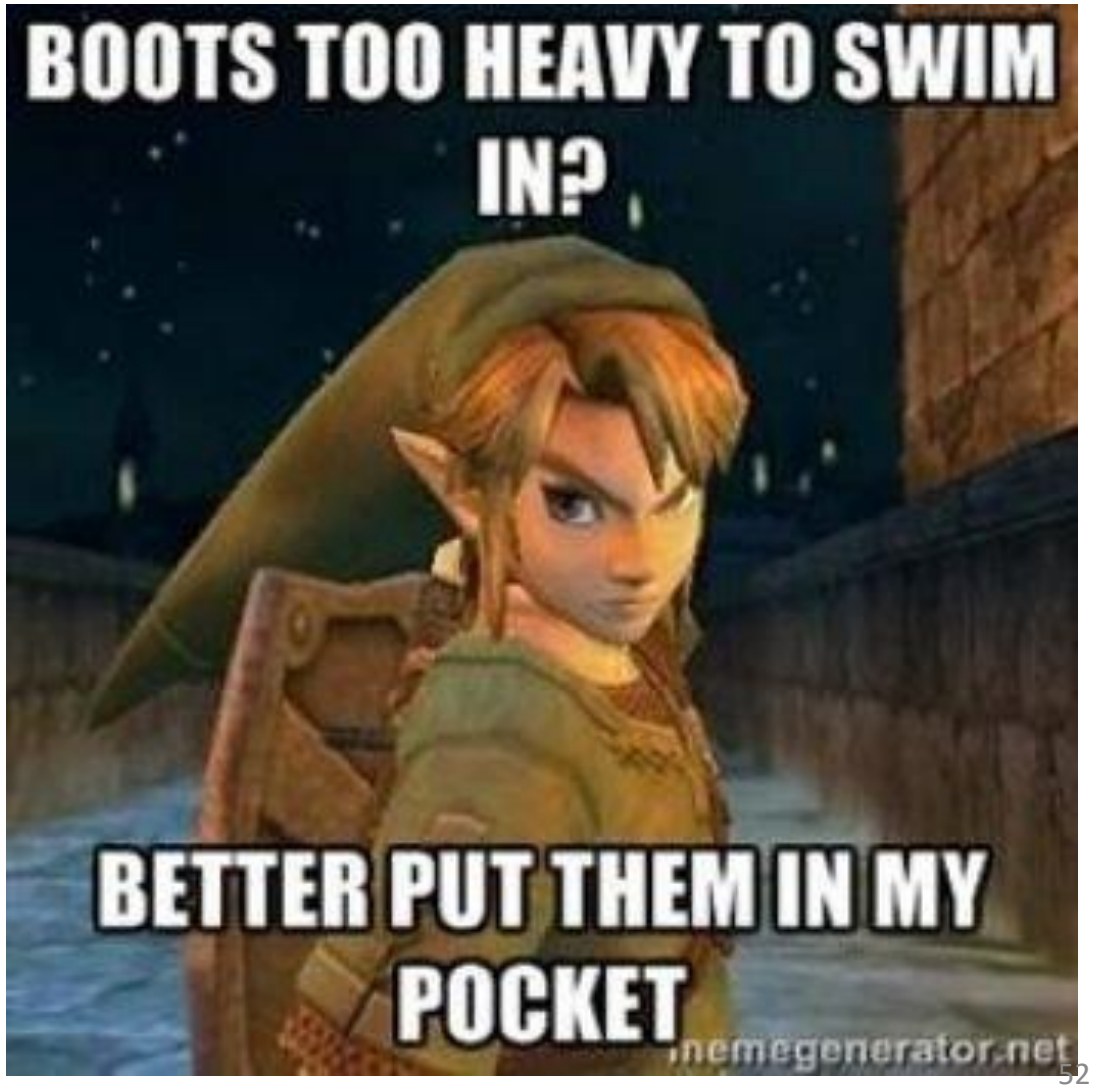
- Collected data, between July 2016 and January 2018, from two fringe social media platforms

67M posts and 5.8M images

35M posts and 1.1M images

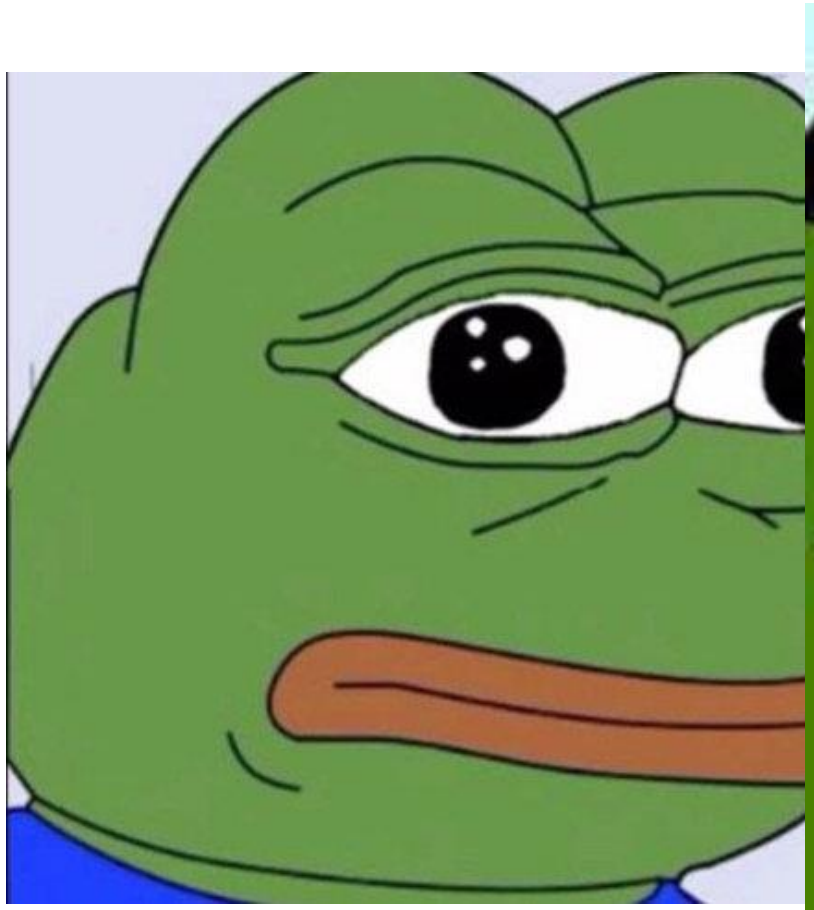

MEMES...
MEMES EVERYWHERE
makeameme.org

Memes are fun…

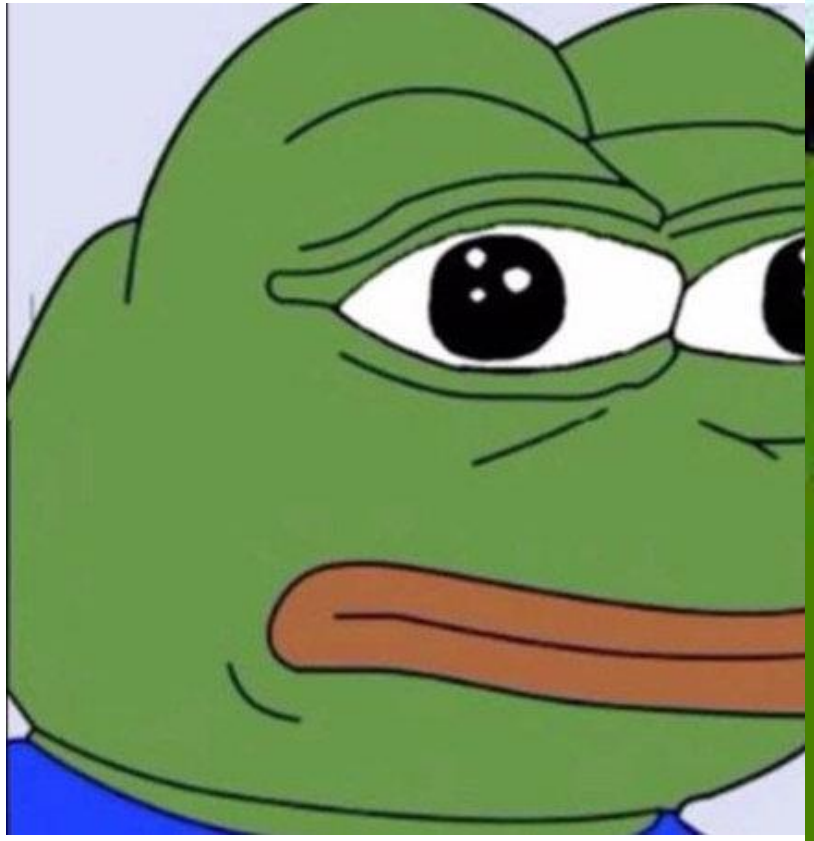


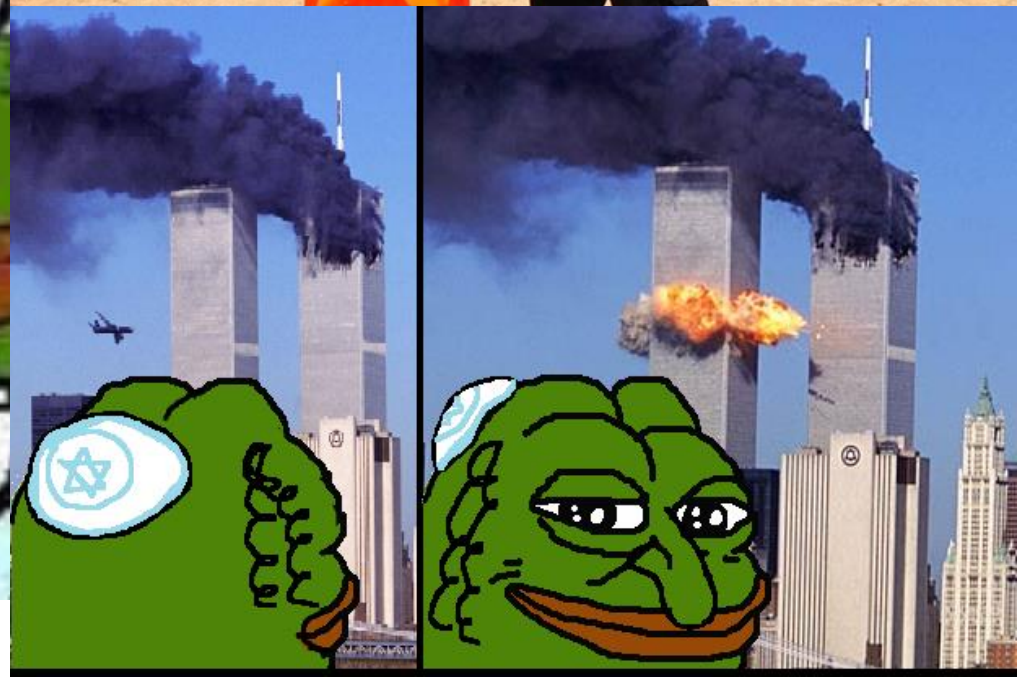
Not always though…

Not always though…

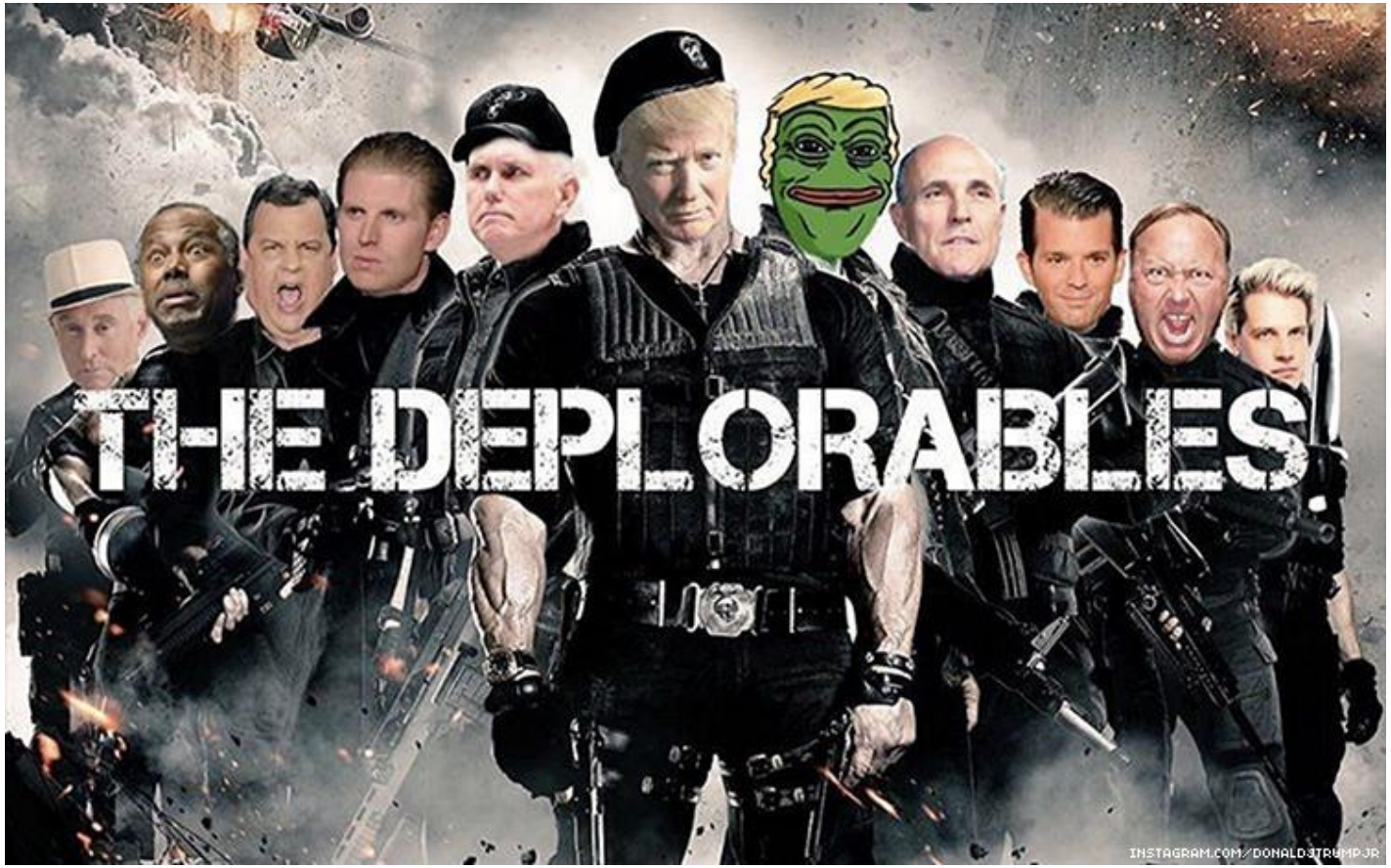


Not always though…

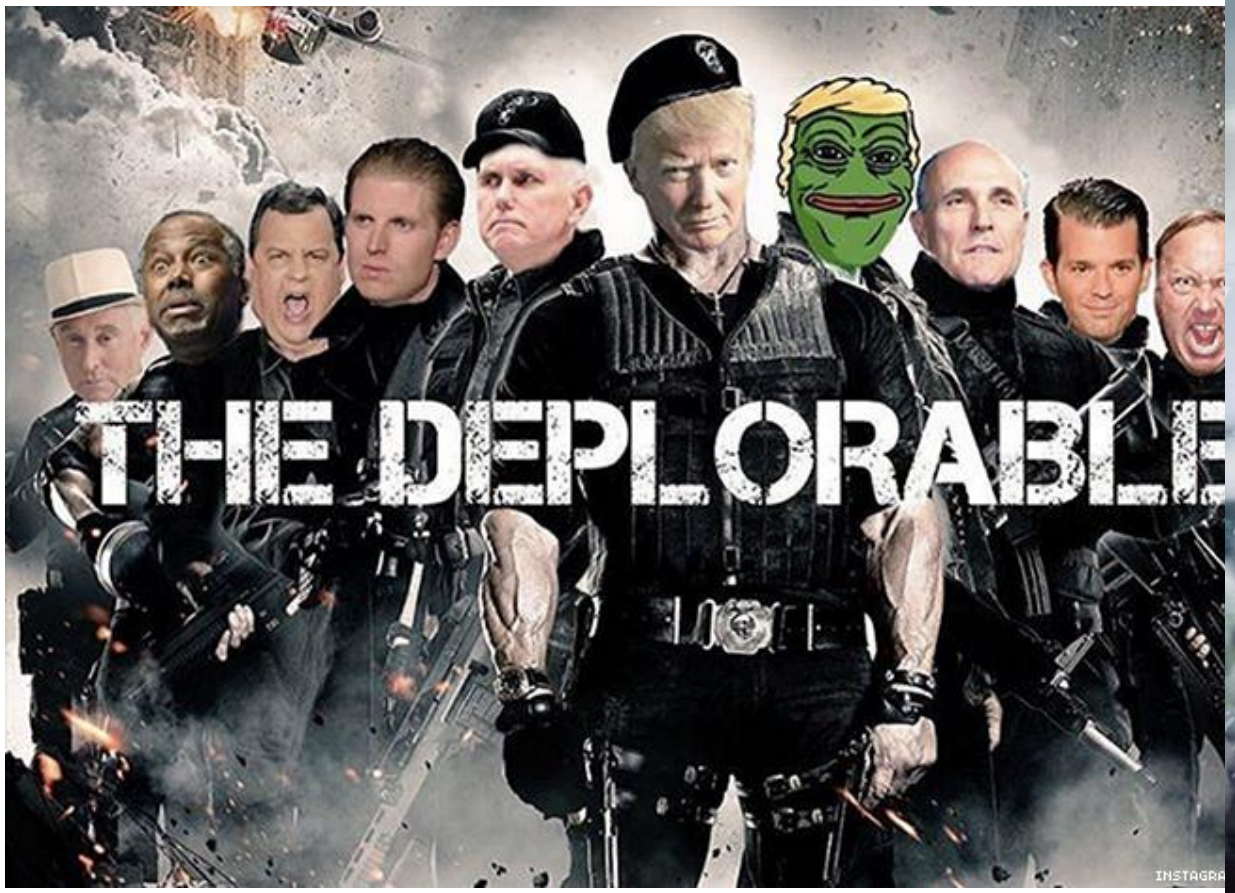




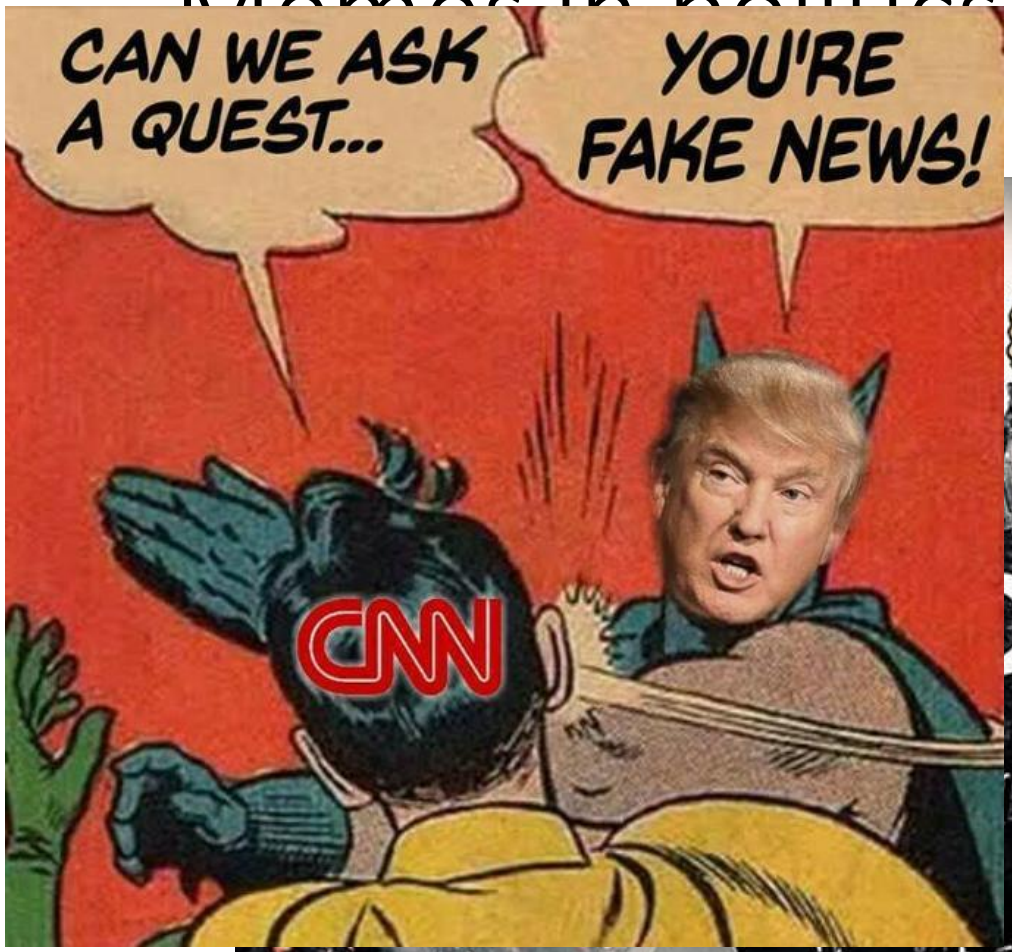
Memes in politics



Memes in politics



Memes in politics



jobs Sign in Search US edition

Sport Culture Lifestyle More

The Guardian

US midterms 2018 Business Tech Science

Annotated: the Trump memes stuck to Cesar Sayoc's van

The vehicle seized by law enforcement in Florida was covered in strange propaganda

Cesar Sayoc's van is seen in Boca Raton, Florida, on 18 October 2018 in this picture obtained from social media. Photograph: Social Media/Ed Kennedy via Reuters

most viewed in US


Pittsburgh synagogue shooting: suspect held after at least four killed


'I bet $500 they are lying': Trump fans sceptical about pipe bomb arrest

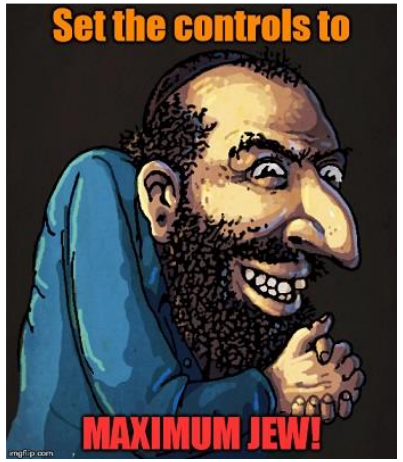

'Why is that racist?' Trump greets young black leaders with freewheeling rant


Barack Obama takes aim at Donald Trump for 'making stuff up'

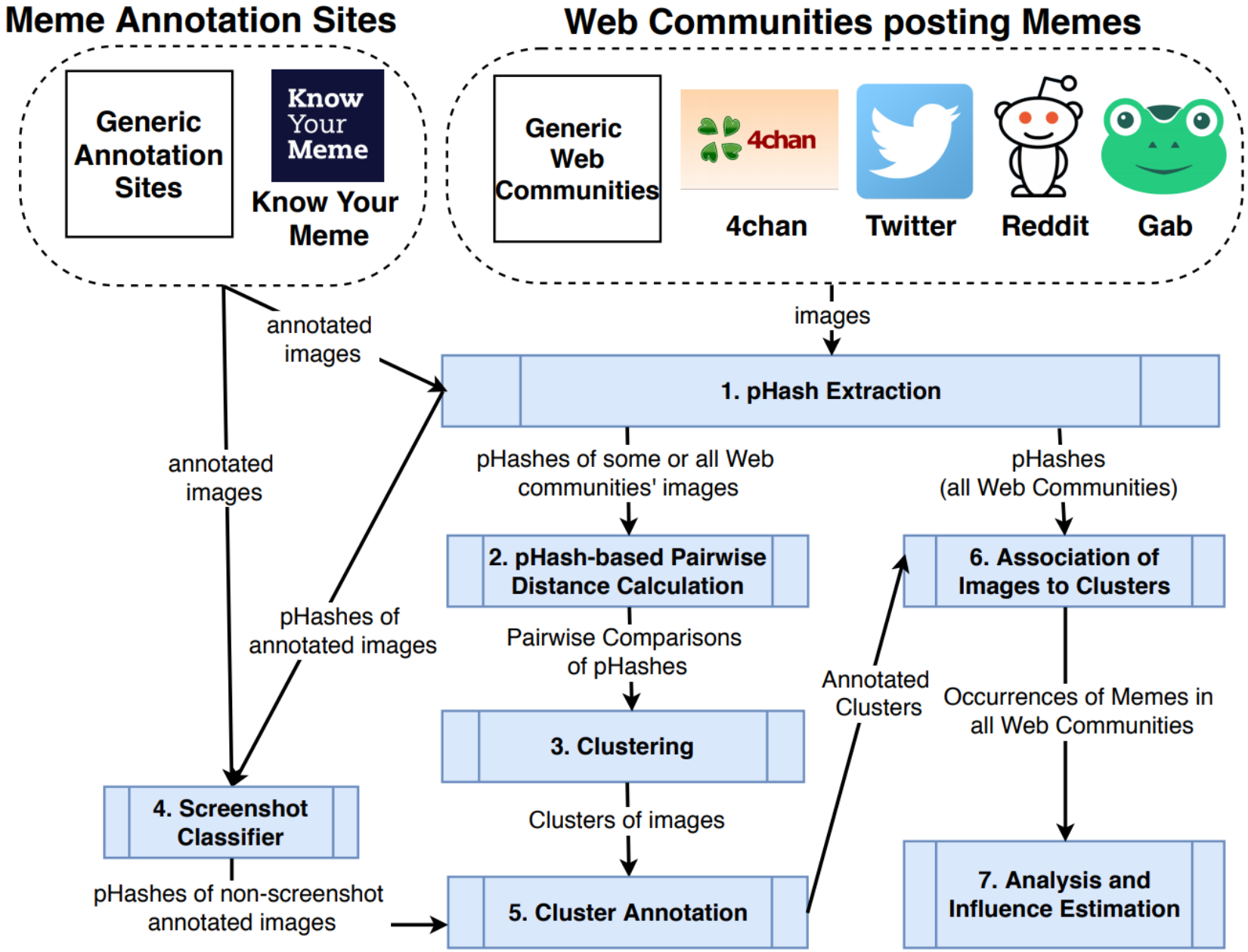

Muncy hits walk-off homer as Dodgers win longest ever World Series game

Understanding Antisemitic imagery

- Memes processing pipeline
  - Perceptual hashing
  - Clustering techniques
  - Ground truth data from Know Your Meme for annotation
- Detect instances of the Happy Merchant meme



Memes processing pipeline



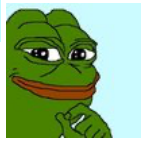
Absolutely Disgusting

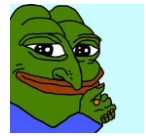
DemotivationalPosters

Apu Apustaja

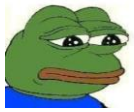
Make America Great Again

Smug Frog



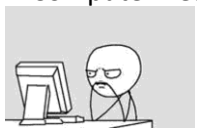


Sad Frog

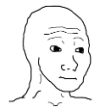


Dubs/Check'em

Computer Reaction Faces



Wojak/Feels Guy



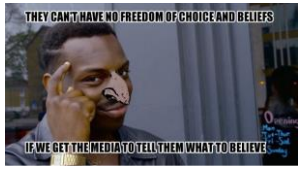
Feels Good

Bait this is Bait

I Know that Feel Bro

Happy Merchant

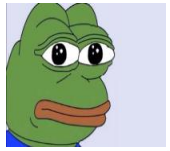
Roll Safe



Sweating Towel Guy

Spurdo Sparde

Pepe the Frog



Autistic Screeching

Summary

- Data-driven approaches to quantify and better understand
  - Antisemitic Imagery (i.e., memes)

- These approaches are important because:
  - They can scale
  - They are fast
  - They can capture evolving and emerging trends

What is the Manosphere?

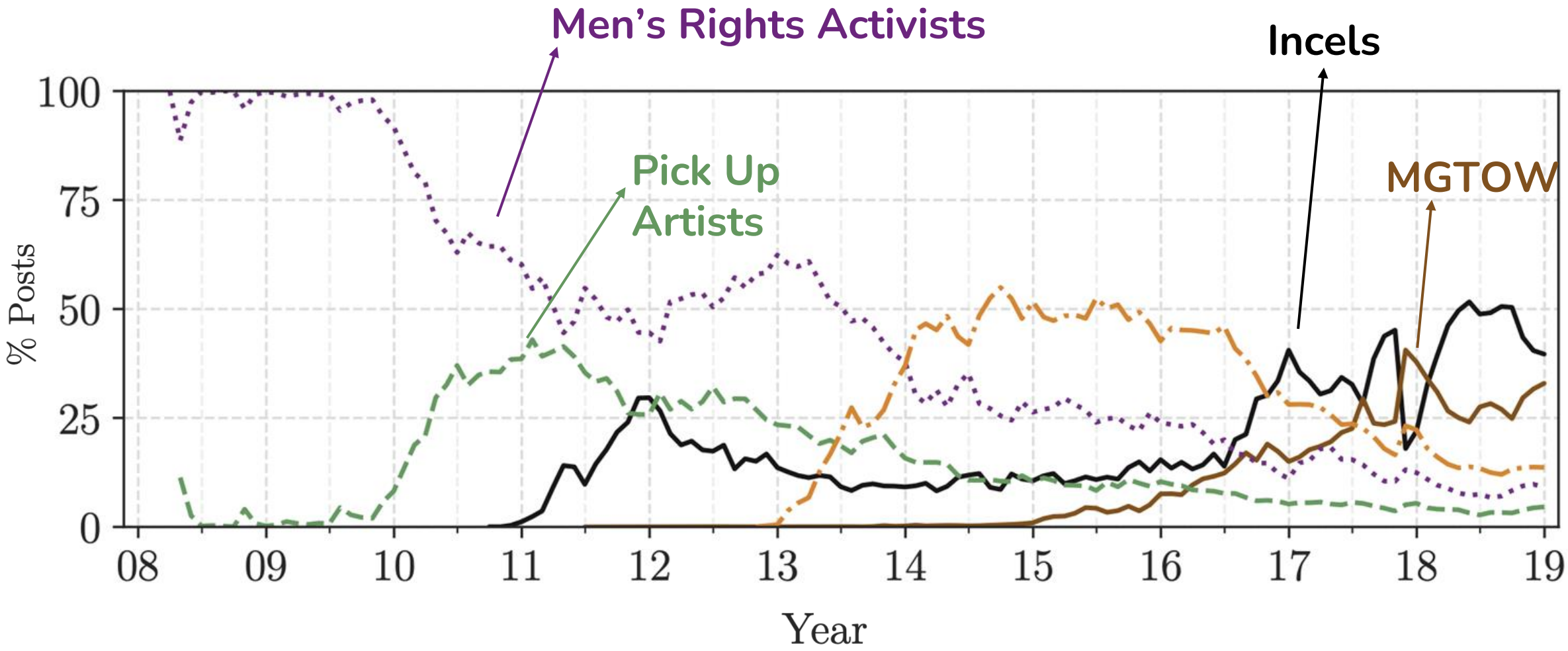
Why is it important to study it?

Data Used

52 subreddits
6 standalone forums
835k users
22M posts

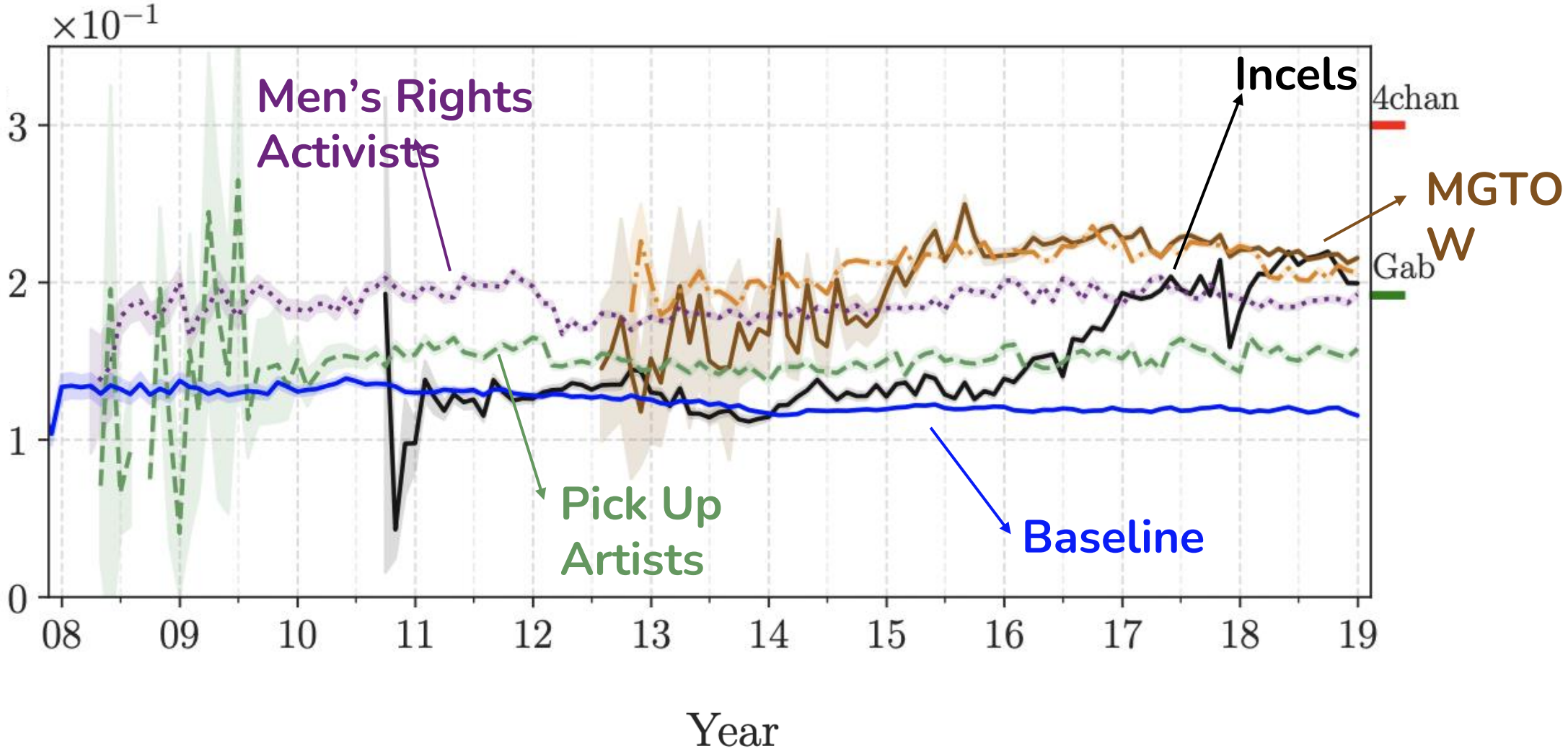
RQ1

How has the popularity/levels of activity of the different Manosphere communities evolved over time?



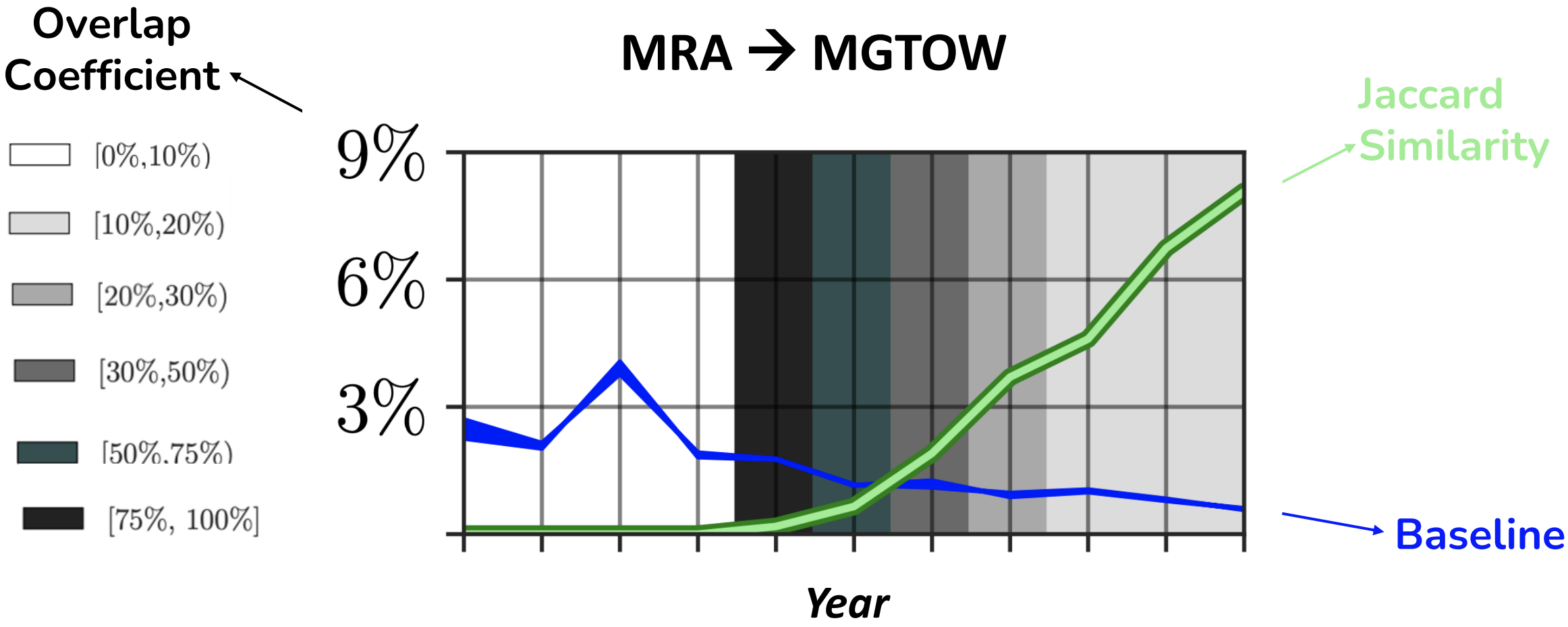
RQ2

Has speech become more toxic and/or misogynistic over time?



RQ3

Has there been substantial migration of, or intersection in, users across communities?



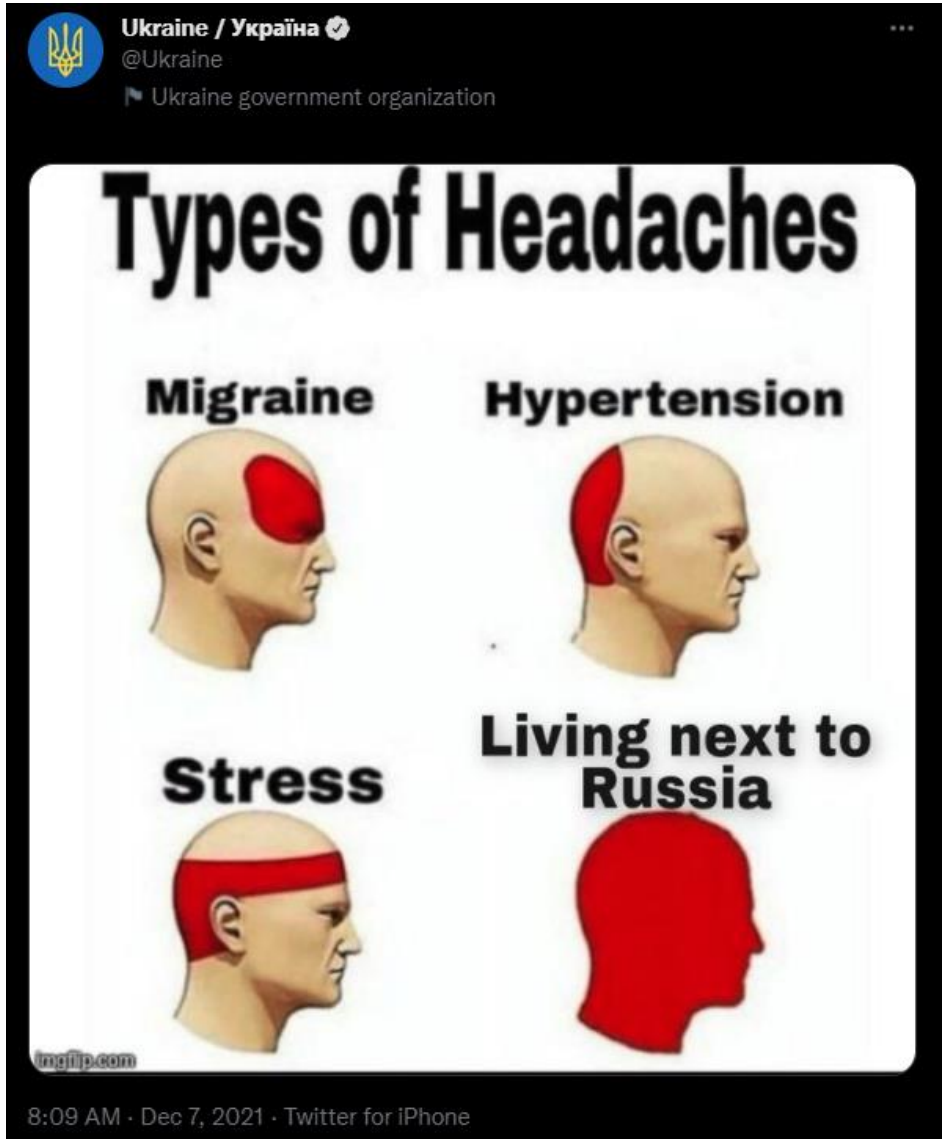
Summary

- Older communities, such as MRAs and PUAs are becoming less popular and active, while newer communities, like Incels and MGTOW, are thriving.

- These communities do not exist in a vacuum! They are connected by members who migrate between them and who participate in multiple communities!

- These communities are becoming increasingly more toxic!

State-sponsored actors



THE**RUSSIA**INVESTIGATION

How Russian trolls manipulated American politics

By Marshall Cohen, CNN

Updated 0220 GMT (1020 HKT) October 20, 2018

ARGUMENT

How Russia Sows Confusion in the U.S. Vaccine Debate

Not content to cause political problems, Moscow's trolls are also undermining public health.

BY KATHERINE KIRK | APRIL 9, 2019, 2:48 PM

Twitter publishes data on Iranian and Russian troll farms

18 OCT 2018 5

Social networks, Twitter

TC

Login

Reddit has banned 944 accounts linked to the IRA Russian troll farm

Lucas Matney @lucasmtny / 1 year ago

Comment

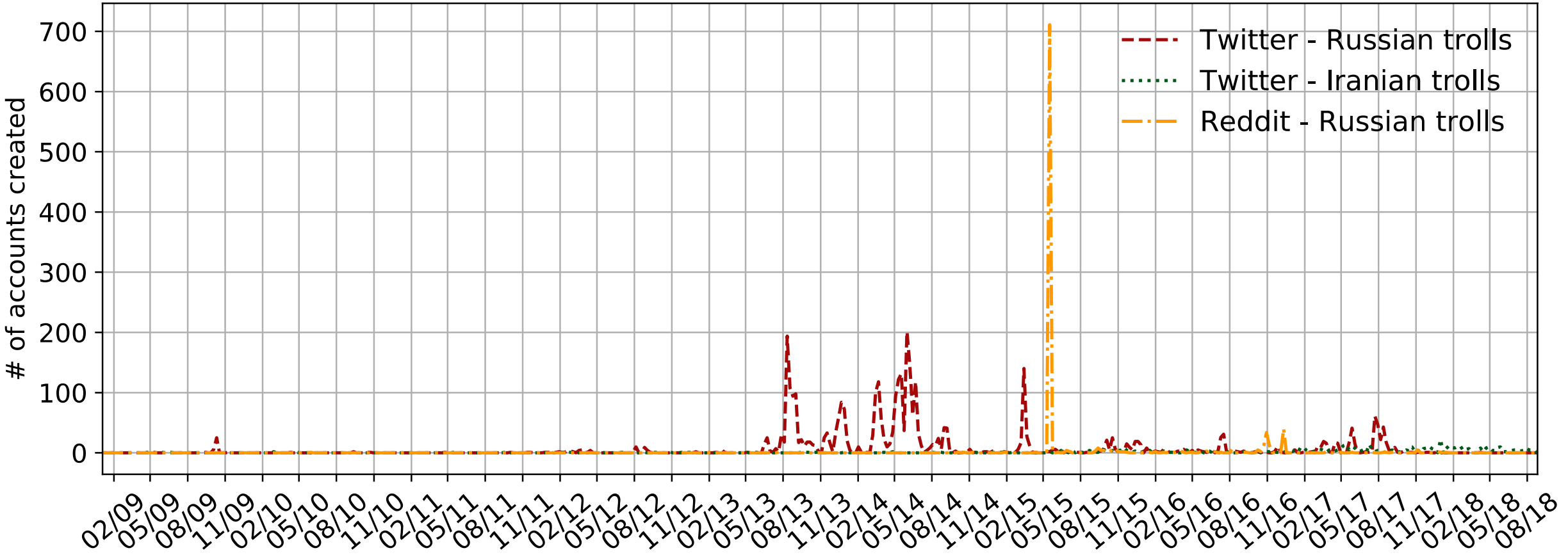
Datasets

Platform	Origin	# trolls	# trolls posts	# posts
Twitter	Russia	3.83K	3.66K	9.04M
Twitter	Iran	0.77K	0.66K	1.12M
Reddit	Russia	0.94K	0.33K	21.32K

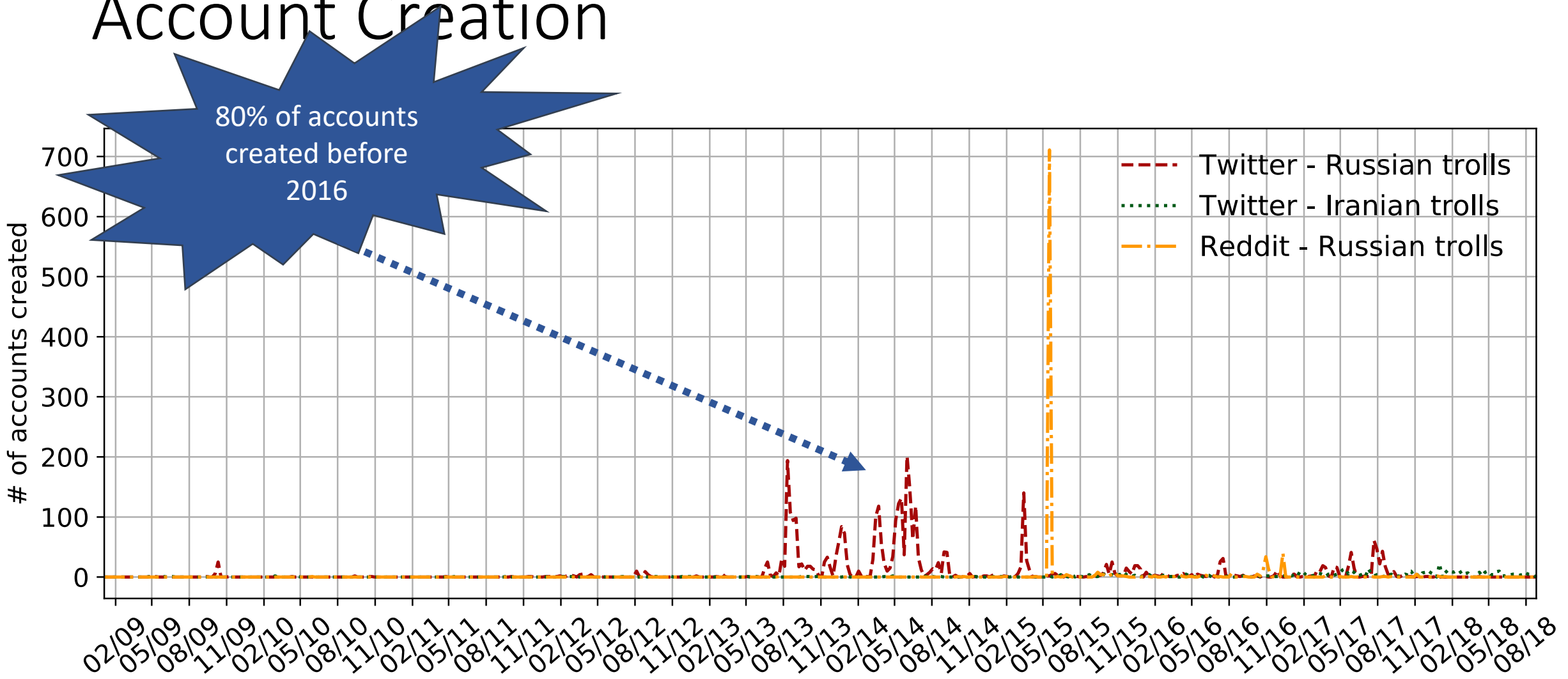
Questions

- How does their behavior change over time?
- What do they talk about?

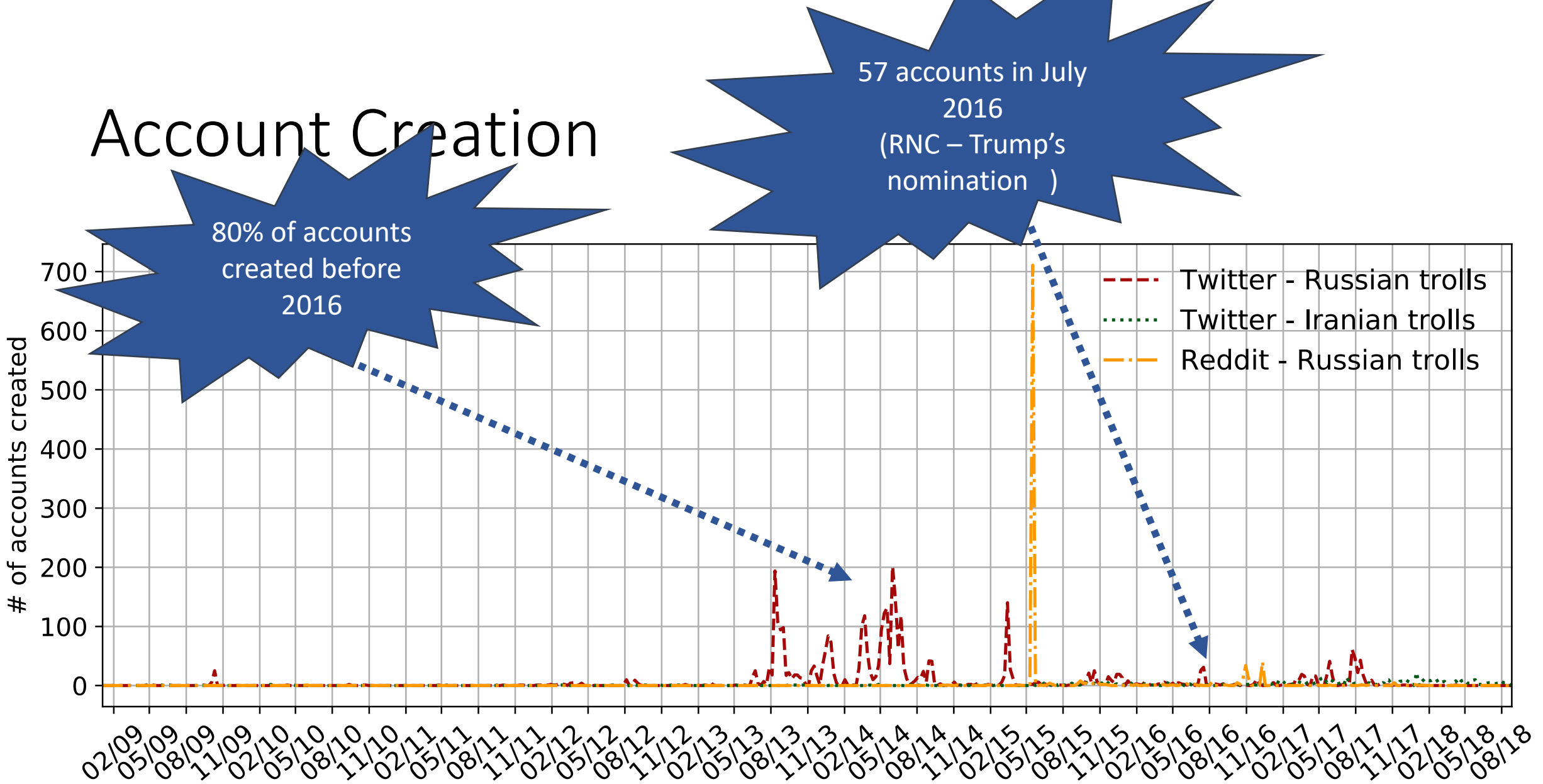
Account Creation



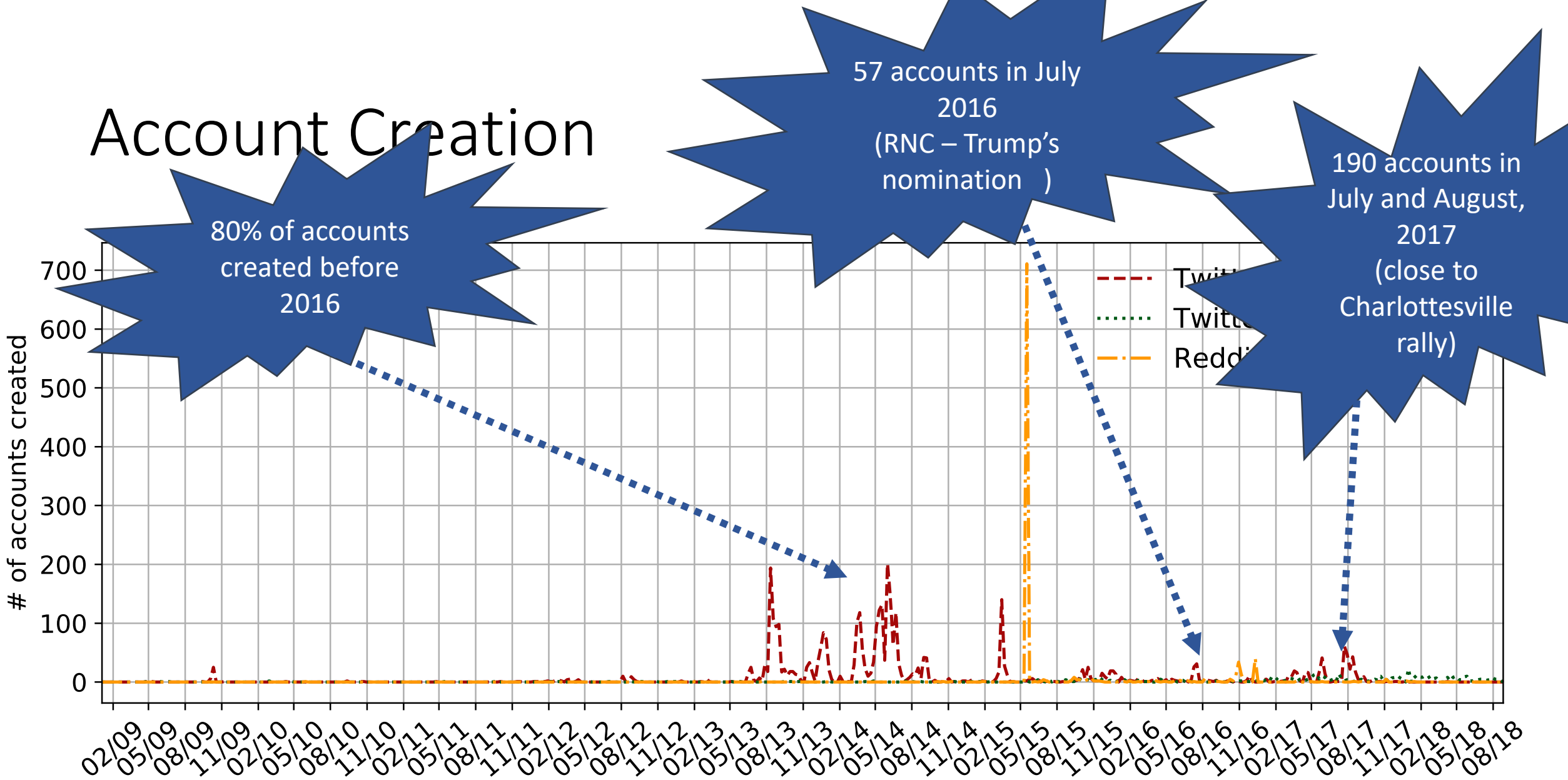
Account Creation



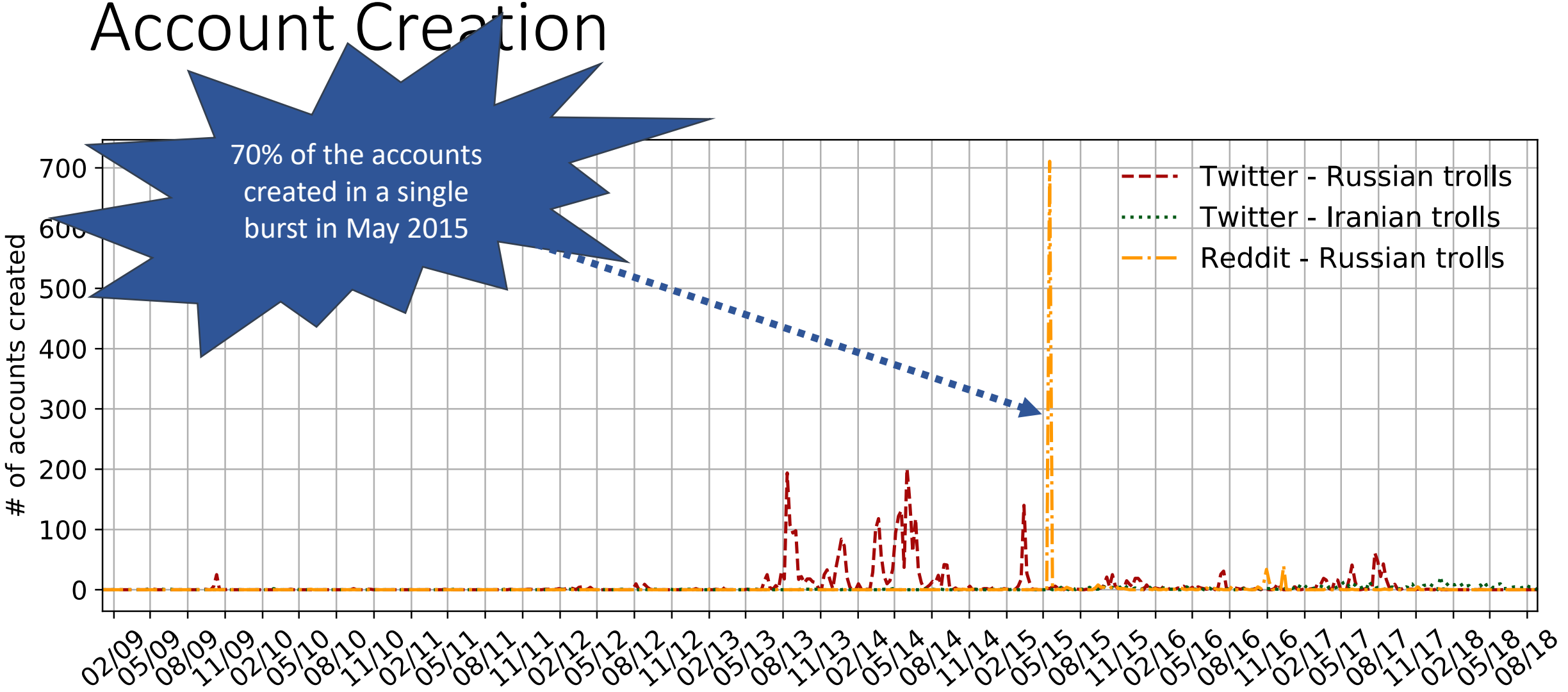
Account Creation



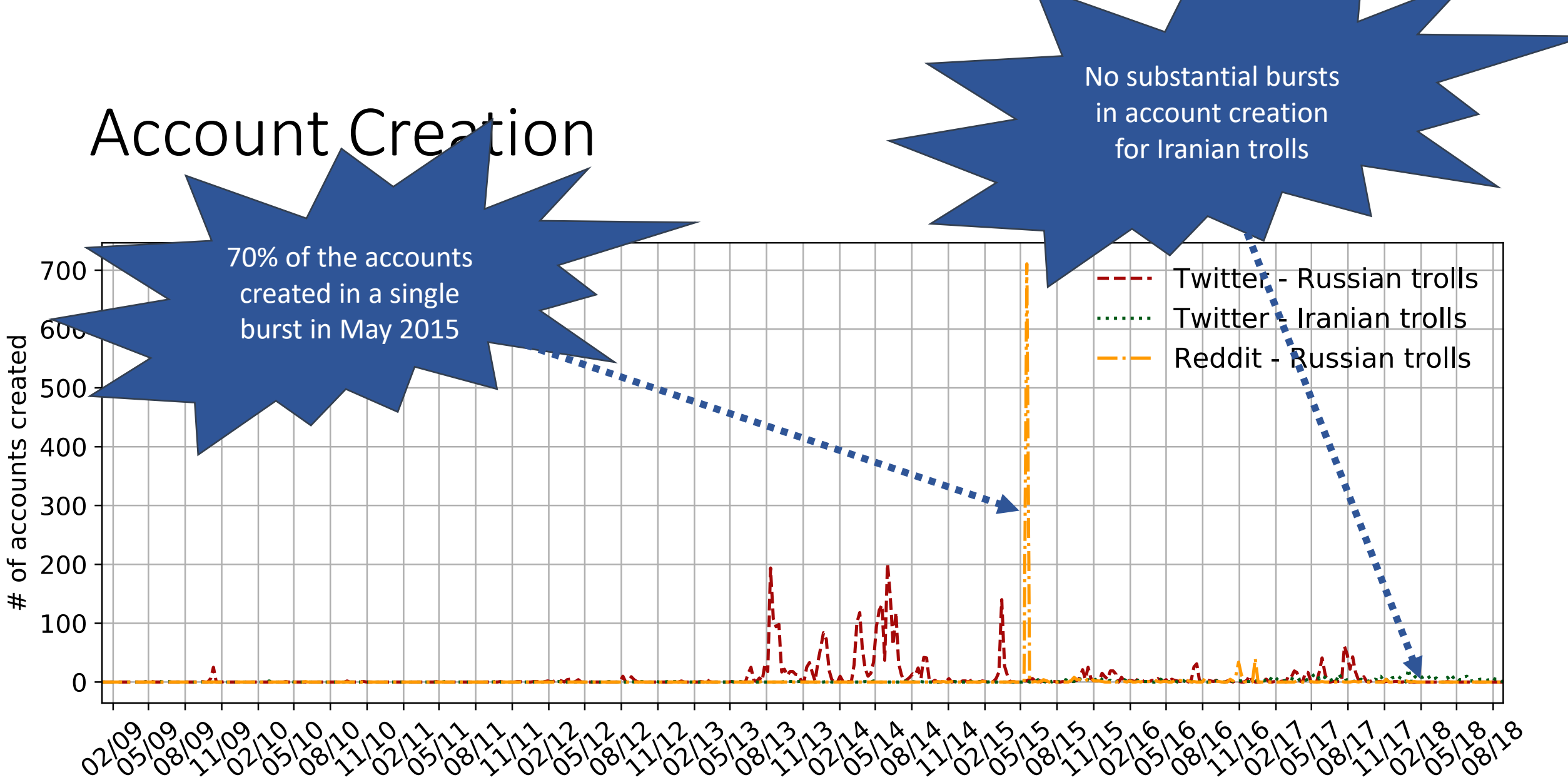
Account Creation



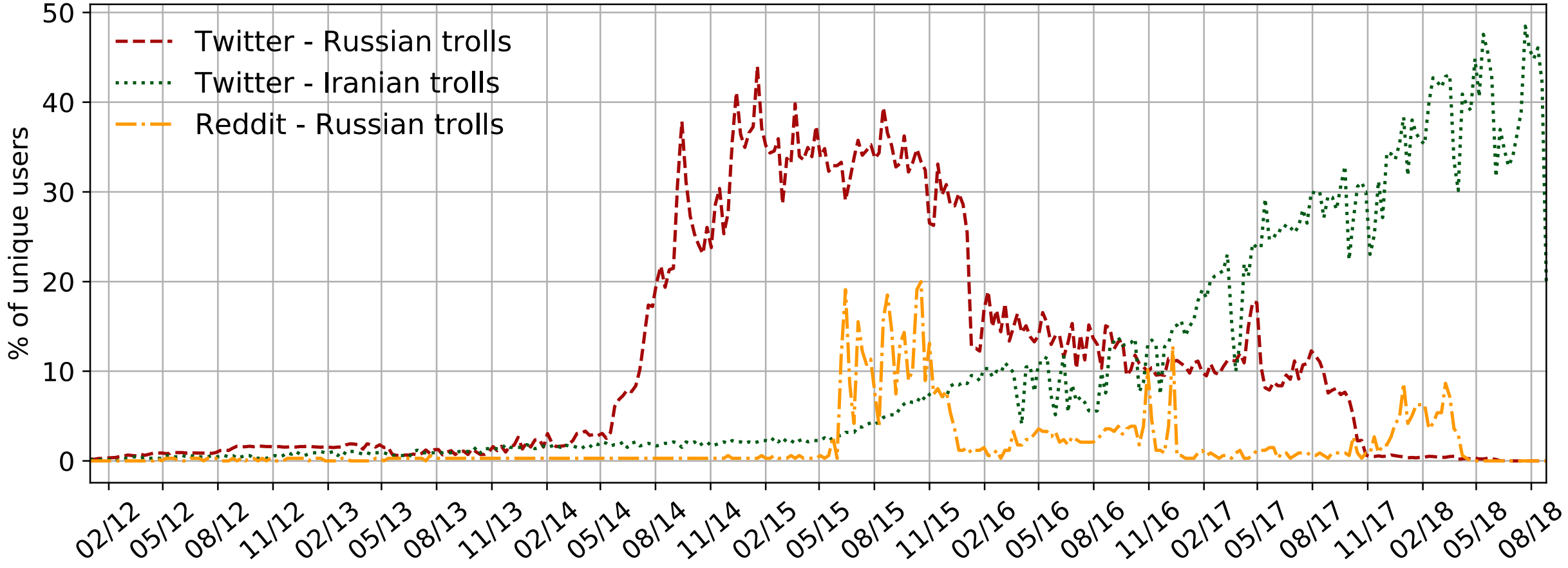
Account Creation



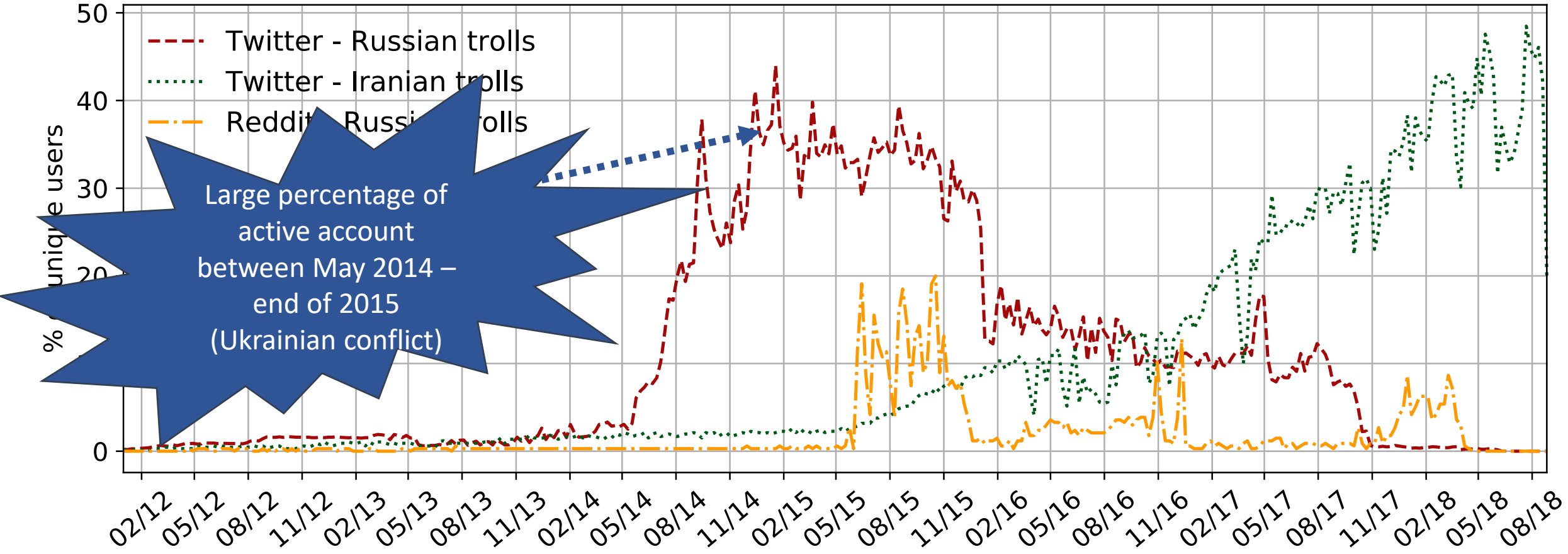
Account Creation



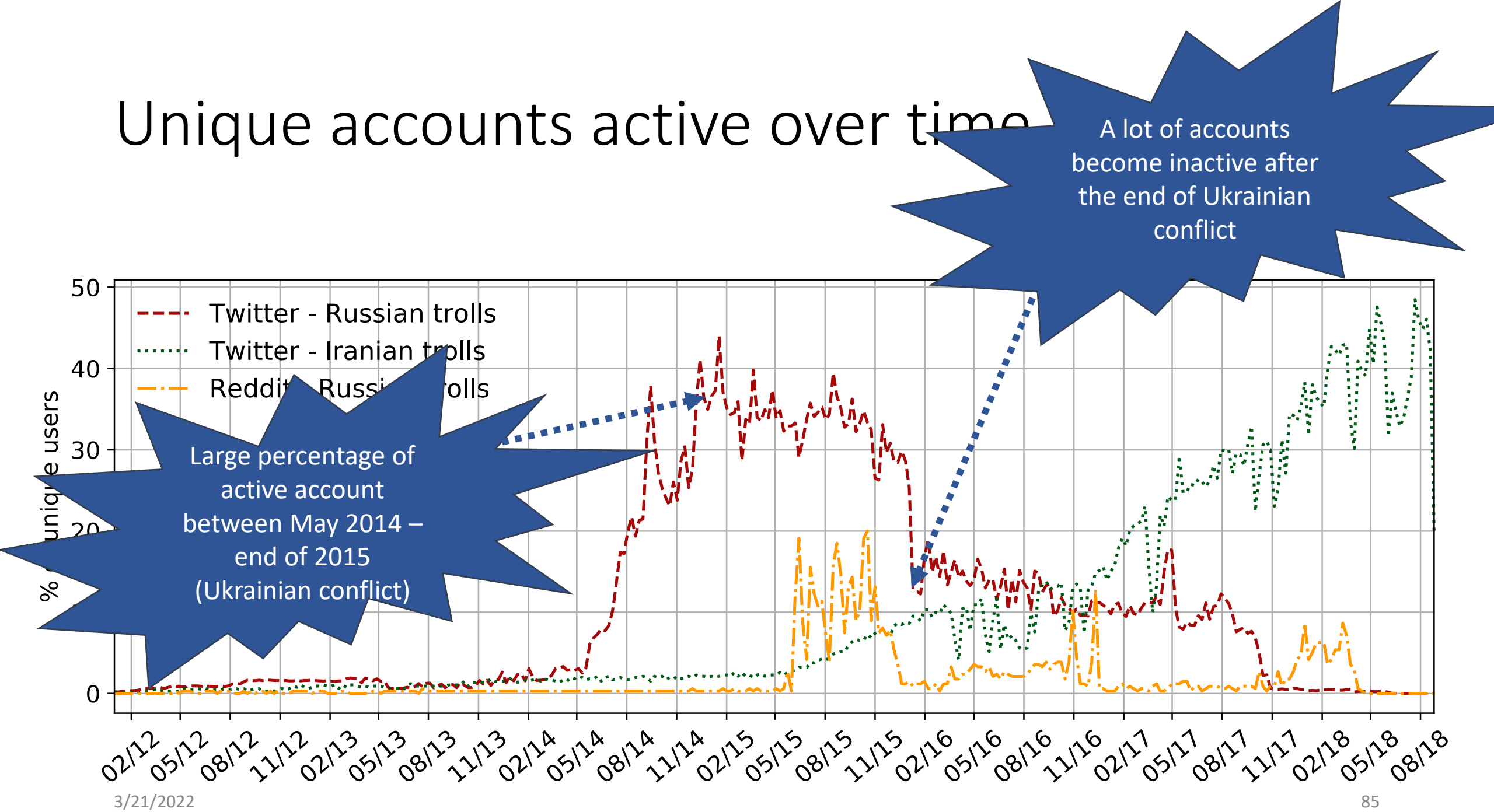
Unique accounts active over time



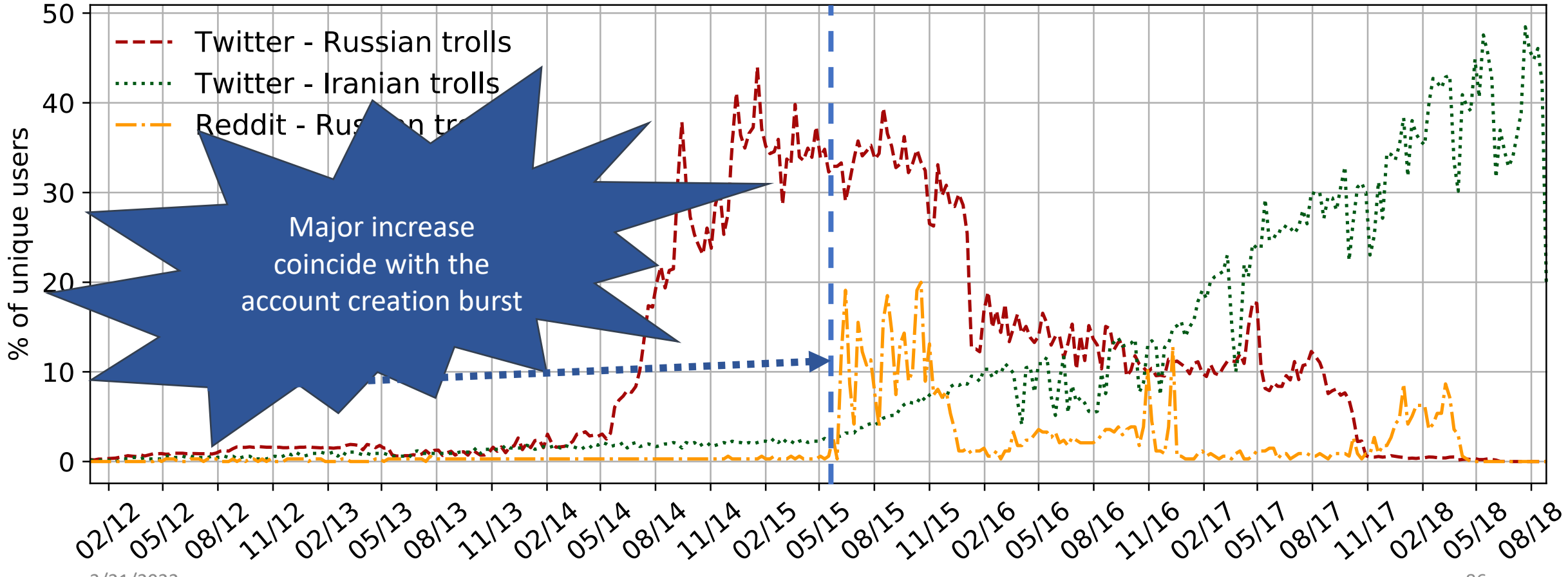
Unique accounts active over time



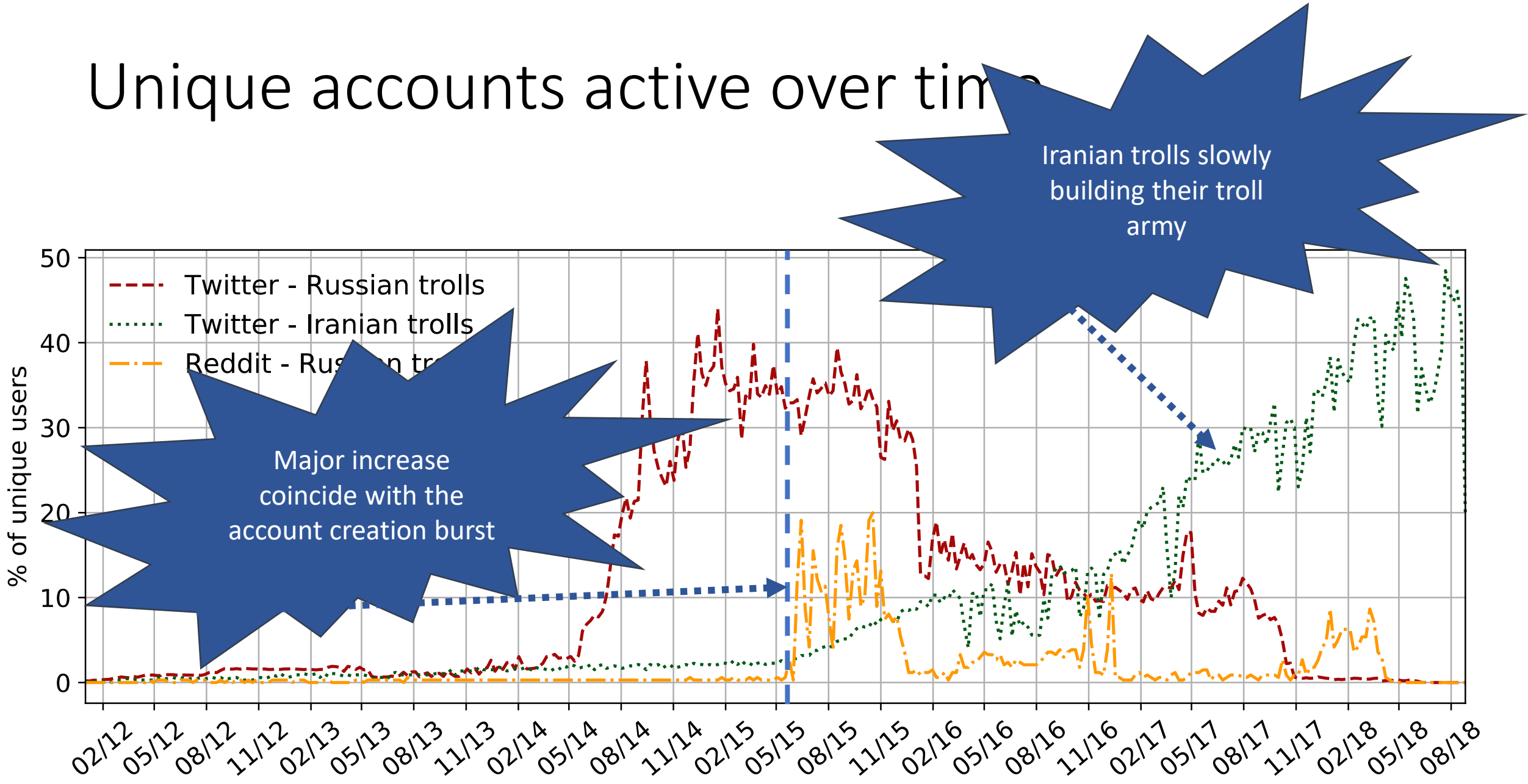
Unique accounts active over time



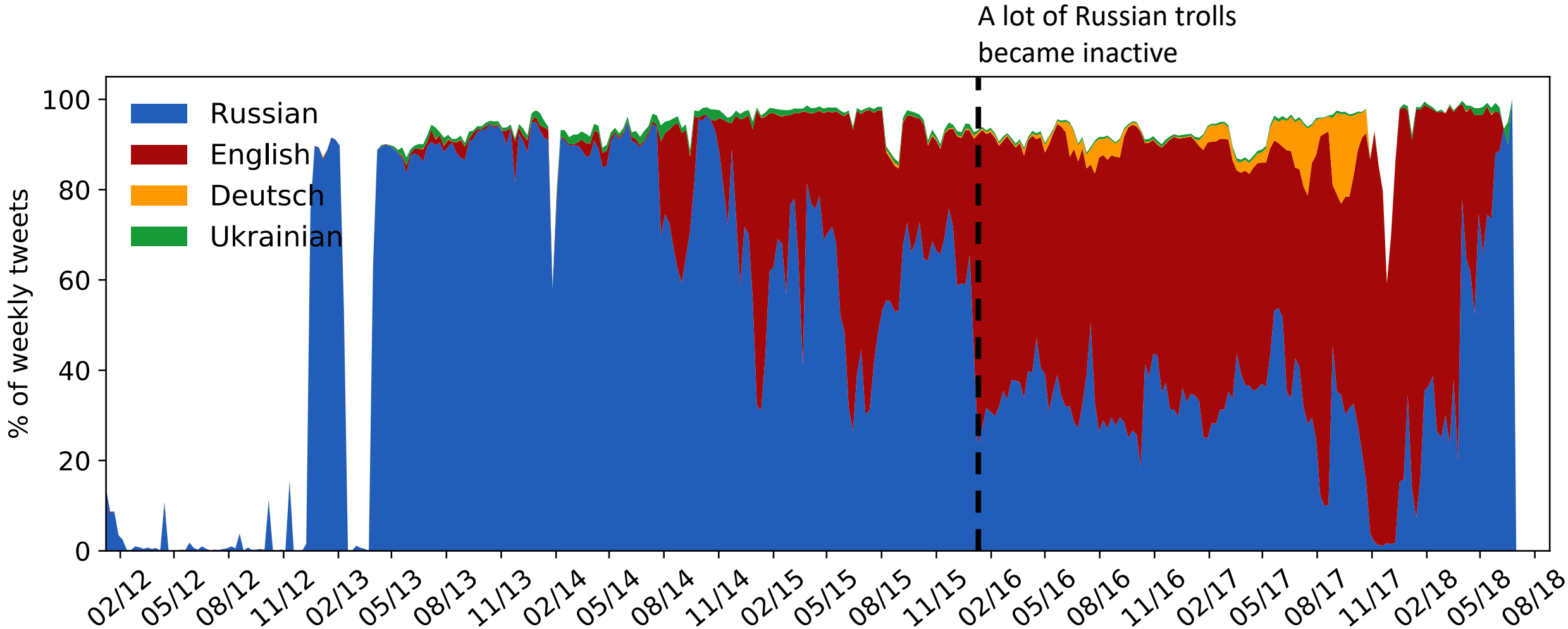
Unique accounts active over time



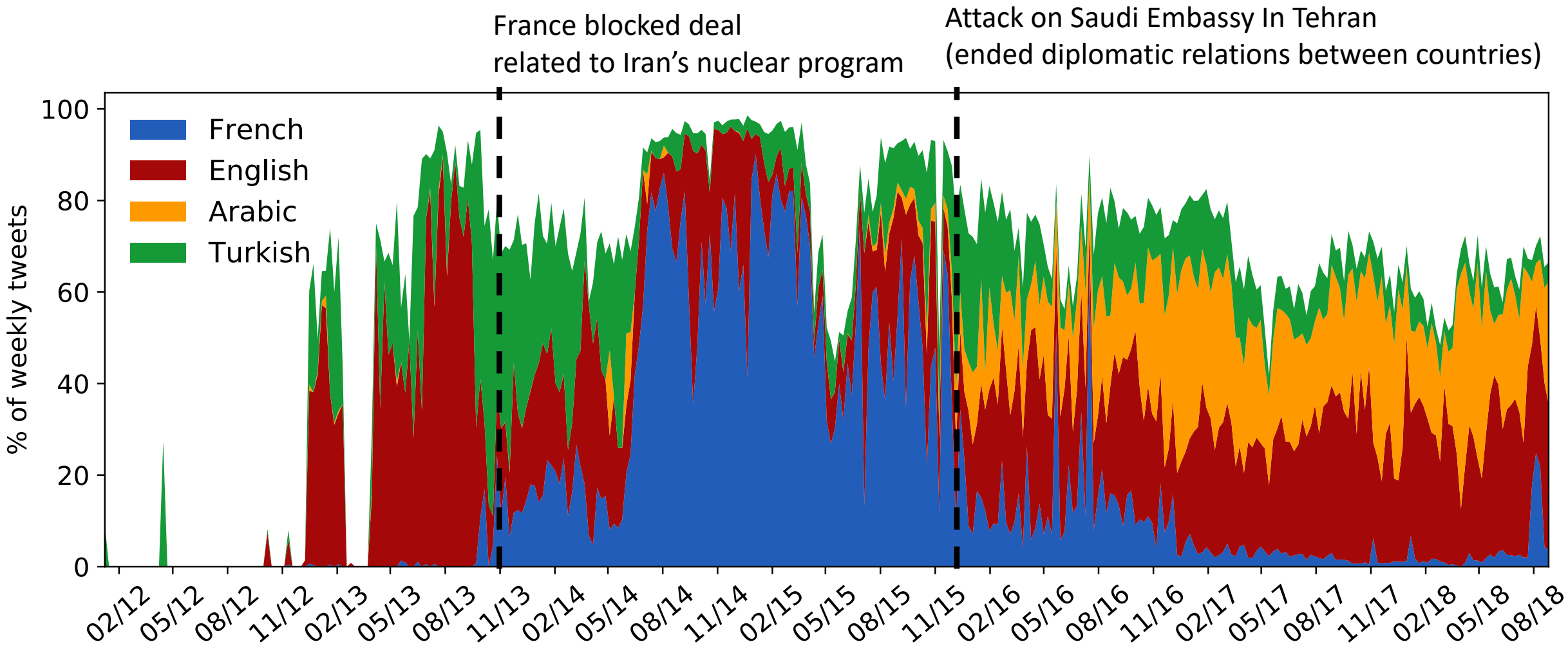
Unique accounts active over tim



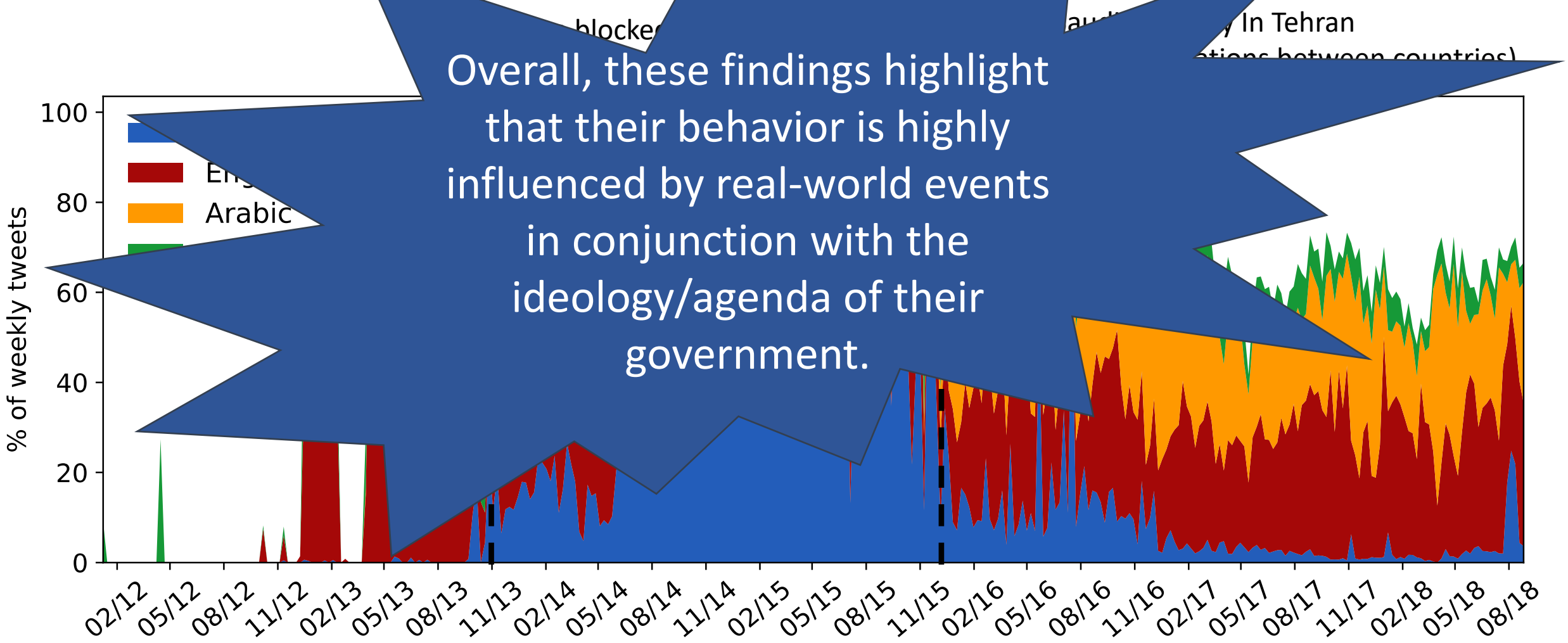
Use of language – Russian trolls (Twitter)



Use of language – Iranian trolls (Twitter)



Use of language – Ira[illegible] tr[illegible] (Twitter)



Overall, these findings highlight that their behavior is highly influenced by real-world events in conjunction with the ideology/agenda of their government.

Content Analysis

Build two word2vec models on English tweets posted by Russian and Iranian trolls

Each model generates a multi-dimensional vector for each word

- Allow us to assess whether words are used in similar contexts by providing cosine similarity between words

An example with the word “maga”

Russian trolls on Twitter		Iranian trolls on Twitter	
Word	Cosine Similarity	Word	Cosine Similarity
trumparmi	0.68	impeachtrump	0.81
trumptrain	0.67	stoptrump	0.80
votetrump	0.65	fucktrump	0.79
makeamericagreatagain	0.65	trumpisamoron	0.79
draintheswamp	0.62	dumptrump	0.79
trumppenc	0.61	ivankatrump	0.77
wakeupamerica	0.59	theresist	0.76
@realdonaldtrump	0.58	trumpresign	0.76
thursdaythought	0.57	notmypresid	0.76
realdonaldtrump	0.57	worstpresidentev	0.75

Next, we look into hashtags…

n trolls on Twitter				Iranian trolls on Twitter			
Hashtag	**%**	**Hashtag**	**%**	**Hashtag**	**%**	**Hashtag**	**%**
news	9.5%	USA	[illegible]%	Iran	1.8%	Palestine	0.6%
sports	3.8%	breaking	[illegible]%	Trump	1.4%	Syria	0.5%
politics	3.0%	TopNews	0.6%	Israel	1.1%	Saudi	0.5%
local	[illegible].1%	BlackLivesMatter	0.6%	Yemen	0.9%	EEUU	0.5%
world	1.1%	true	0.5%	FreePalestine	0.8%	Gaza	0.5%
MAGA	1.1%	Texas	0.5%	QudsDay4Return	0.8%	SaudiArabia	0.4%
business	1.0%	NewYork	0.4%	US	0.7%	luvm	0.4%
Chicago	0.9%	Fukushima2015	0.4%	realiran	[illegible].6%	InternationalQudsDay2018	0.4%
health	0.8%	quote	0.4%	ISIS	0.6%	Realiran	0.4%
love	0.7%	Foke	0.4%	DeleteIsrael	0.6%	News	0.4%

News

MAGA

Social movements

US politics

Anti-Israel

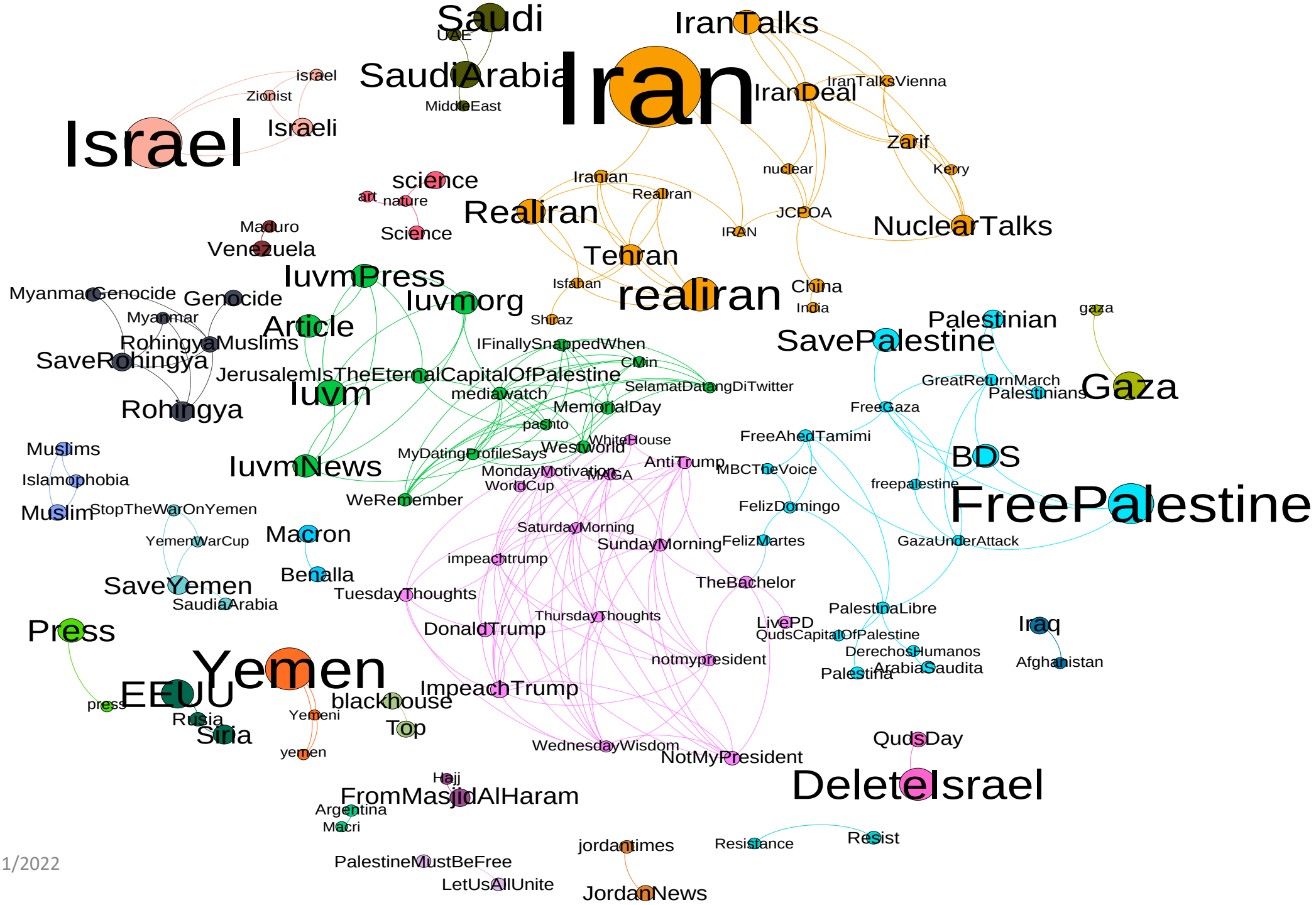
Middle East -related

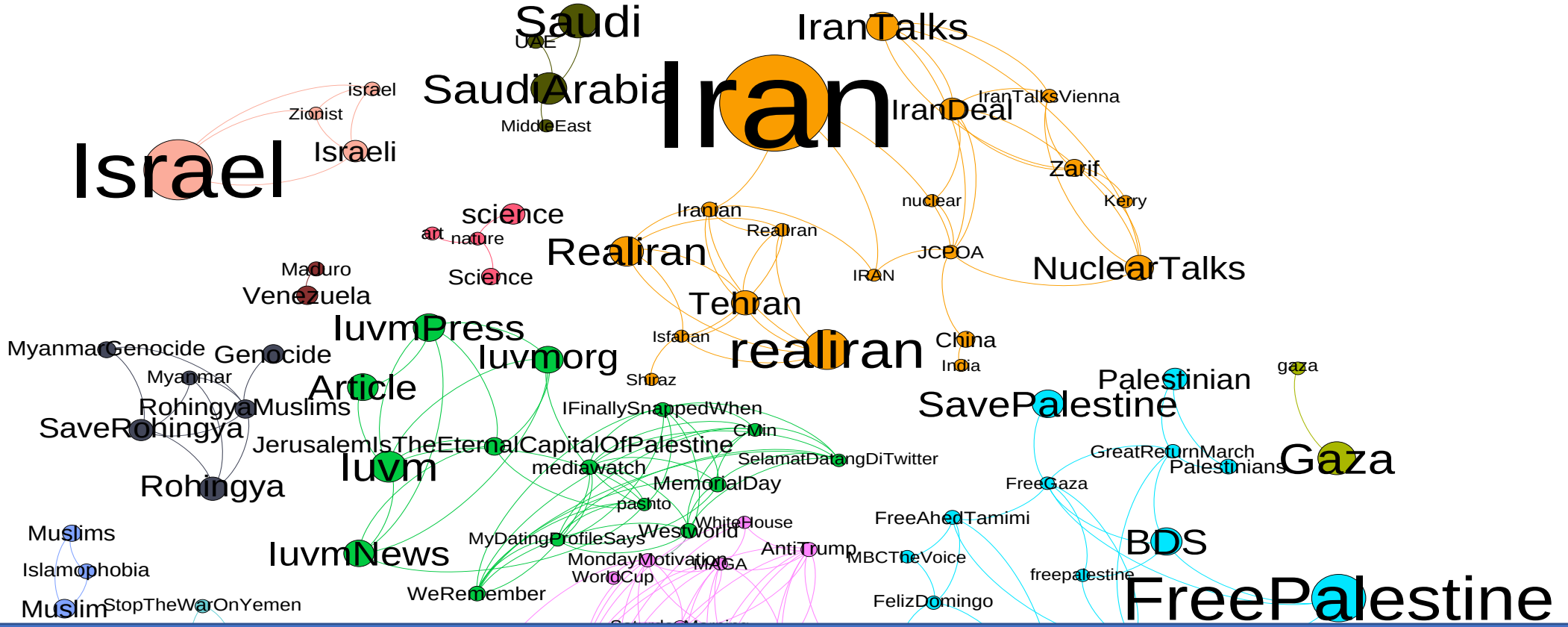
Fake News Network

How can we better visualize the usage of hashtags while taking into account the context?

Build a graph from the word embeddings:

- Each hashtag is a node
- An edge exists between two nodes in the graph if their distance (as determined by their vector) is less than some pre-defined threshold

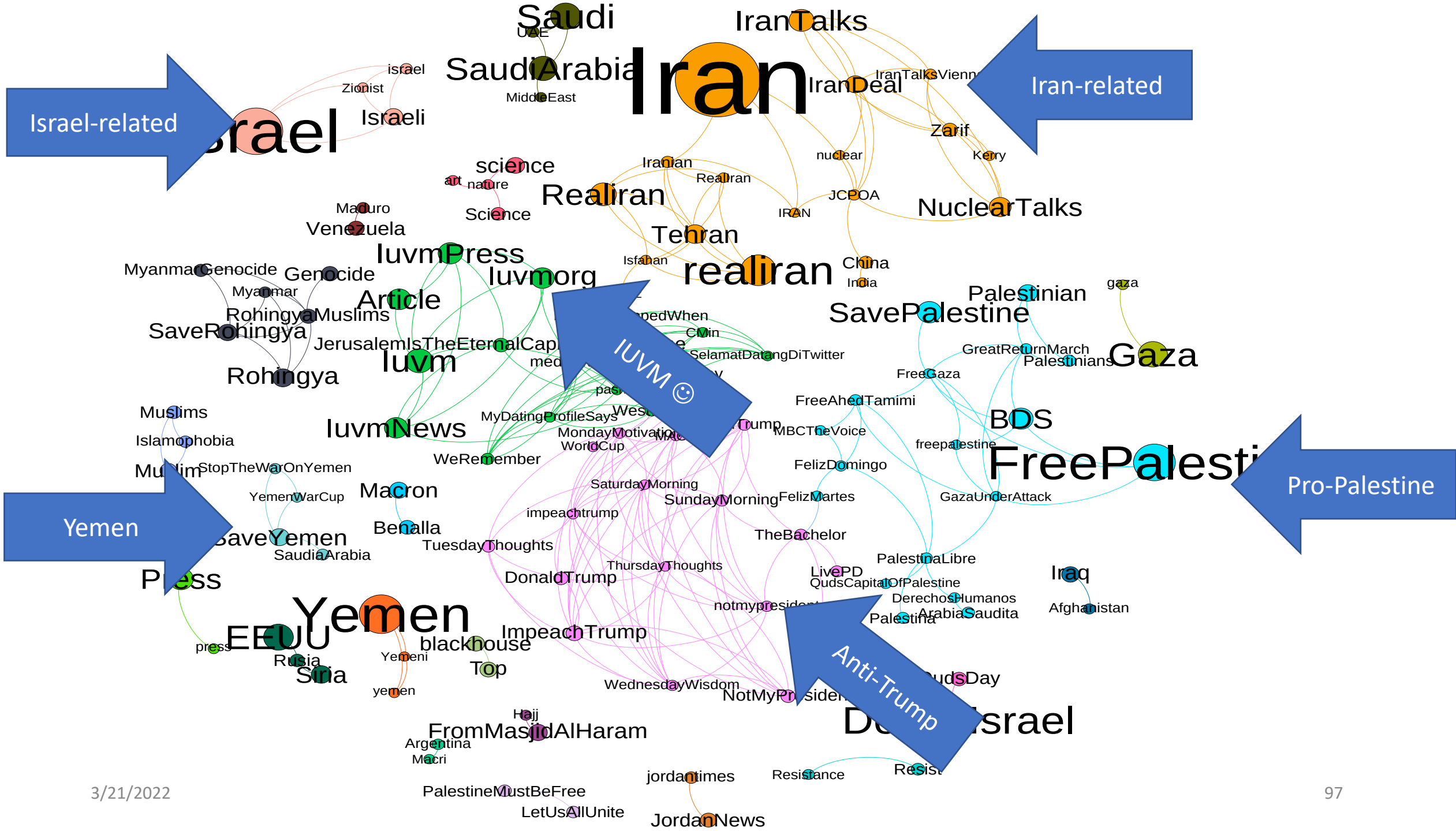

Saudi
UAE
IranTalks
Iran
SaudiArabia
IranDeal
IranTalksVienna
israel
Zionist
MiddleEast
Israeli
Israel
Zarif
nuclear
Kerry
science
art
nature
Iranian
RealIran
Realiran
JCPOA
IRAN
NuclearTalks
Maduro
Venezuela
Science
Tehran
luvmPress
Isfahan
China
India
realiran
luvmorg
MyanmarGenocide
Genocide
Myanmar
Article
Shiraz
Palestinian
gaza
SavePalestine
RohingyaMuslims
IFinallySnappedWhen
SaveRohingya
CMin
JerusalemIsTheEternalCapitalOfPalestine
SelamatDatangDiTwitter
GreatReturnMarch
Palestinians
Gaza
luvm
mediawatch
MemorialDay
FreeGaza
Rohingya
pashto
WhiteHouse
FreeAhedTamimi
Muslims
MyDatingProfileSays
Westworld
BDS
luvmNews
AntiTrump
MBCTheVoice
Islamophobia
MondayMotivation
MAGA
freepalestine
WorldCup
FreePalestine
Muslim
StopTheWarOnYemen
WeRemember
FelizDomingo
SaturdayMorning
Macron
SundayMorning
FelizMartes
GazaUnderAttack
YemenWarCup
impeachtrump
Benalla
TheBachelor
SaveYemen
TuesdayThoughts
SaudiArabia
PalestinaLibre
Press
ThursdayThoughts
LivePD
Iraq
DonaldTrump
QudsCapitalOfPalestine
DerechosHumanos
notmypresident
Yemen
Afghanistan
ArabiaSaudita
Palestina
EEUU
ImpeachTrump
blackhouse
press
Yemeni
Rusia
Top
Siria
yemen
QudsDay
WednesdayWisdom
NotMyPresident
DeleteIsrael
Hajj
FromMasjidAlHaram
Argentina
Macri
jordantimes
Resistance
Resist
PalestineMustBeFree
LetUsAllUnite
JordanNews



Size of node relative to the number of occurrences in dataset

Graph laid out in space according to distances (more similar words are closer together)

Run community detection → Color nodes according to their community

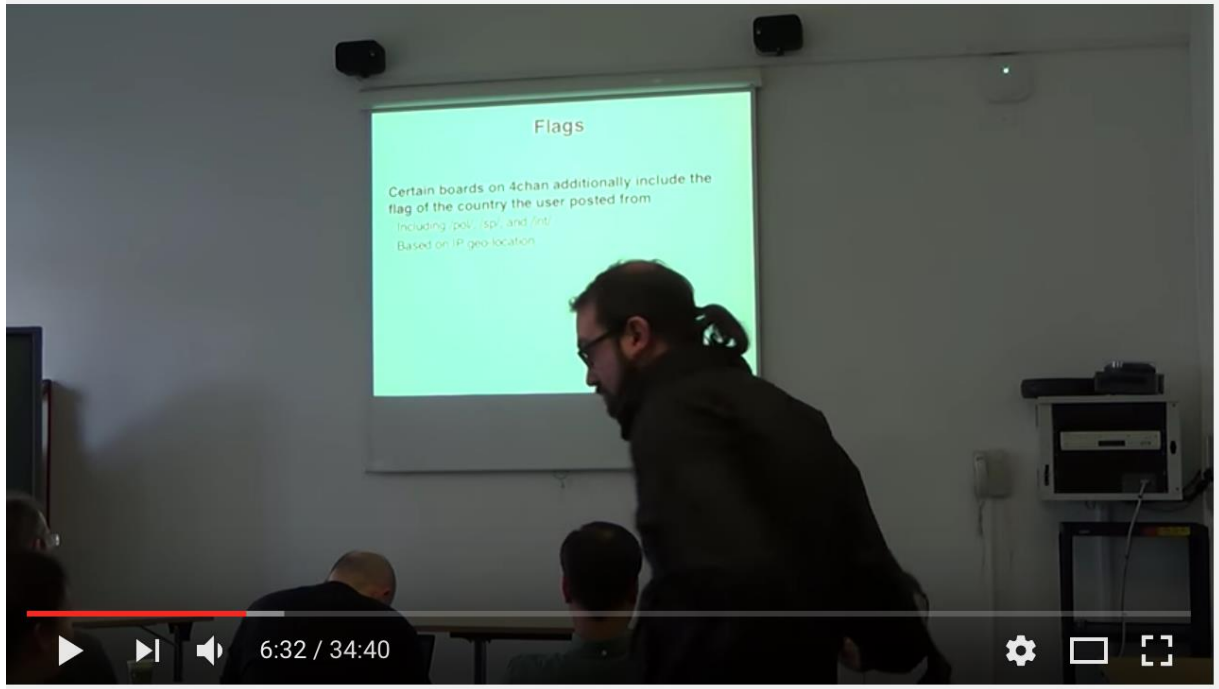

Israel-related
Iran-related
IUVM ☺
Pro-Palestine
Yemen
Anti-Trump
Israel
Iran
Saudi
UAE
SaudiArabia
IranTalks
IranDeal
IranTalksVienn
Zarif
nuclear
Kerry
JCPOA
IRAN
NuclearTalks
israel
Zionist
Israeli
MiddleEast
science
art
nature
Science
Iranian
RealIran
Realiran
Tehran
Isfahan
realiran
China
India
Maduro
Venezuela
luvmPress
luvmorg
Article
MyanmarGenocide
Genocide
Myanmar
RohingyaMuslims
SaveRohingya
Rohingya
JerusalemIsTheEternalCap
luvm
luvmNews
Muslims
Islamophobia
StopTheWarOnYemen
YemenWarCup
Macron
Benalla
SaveYemen
SaudiArabia
Press
Yemen
EEUU
Rusia
Siria
press
Yemeni
yemen
blackhouse
Top
ImpeachTrump
FromMasjidAlHaram
Hajj
Argentina
Macri
PalestineMustBeFree
LetUsAllUnite
jordantimes
JordanNews
Resistance
Resist
WednesdayWisdom
NotMyPresident
DonaldTrump
impeachtrump
TuesdayThoughts
ThursdayThoughts
SaturdayMorning
SundayMorning
FelizMartes
FelizDomingo
TheBachelor
MyDatingProfileSays
MondayMotivation
WorldCup
WeRemember
MBCTheVoice
FreeAhedTamimi
SelamatDatangDiTwitter
FreeGaza
GreatReturnMarch
Palestinians
Palestinian
SavePalestine
gaza
Gaza
BDS
freepalestine
FreePalesti
GazaUnderAttack
PalestinaLibre
LivePD
QudsCapitalOfPalestine
DerechosHumanos
ArabiaSaudita
Palestina
Iraq
Afghanistan

Mitigation

Ok, lots of new understanding of what's up…

- Measurements and modeling of what is going on is important
  - *But it is only the first step!*
- At some point, we need to make use of this new understanding
- How can we *mitigate* these problems?
- How do we assess the success of deployed mitigation strategies?

What Does A Raid Look Like?



A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web

E

Emiliano De Cristofaro

Subscribe 2

10,594 views

Add to Share More

78 512

Hans2k4ever 1 month ago
When will the normies learn?
Reply • 181
View all 3 replies

Jack Harper 1 month ago
mfw he accidentally redpills his whole class, they all found the memes funny even the so called "hate speech" he talks about
Reply • 13

Jack Harper 1 month ago
I guess I haven't tightened my fedora enough to be as enlightened by my own enlightenment as you sir. Carry on brave knight!
Reply •

Adem Adiyaman 1 month ago
23:21 "the victory of donald trump and the rise of ultra nazi people"
you dense fuck, nice brain washing
Reply • 227
View all 16 replies

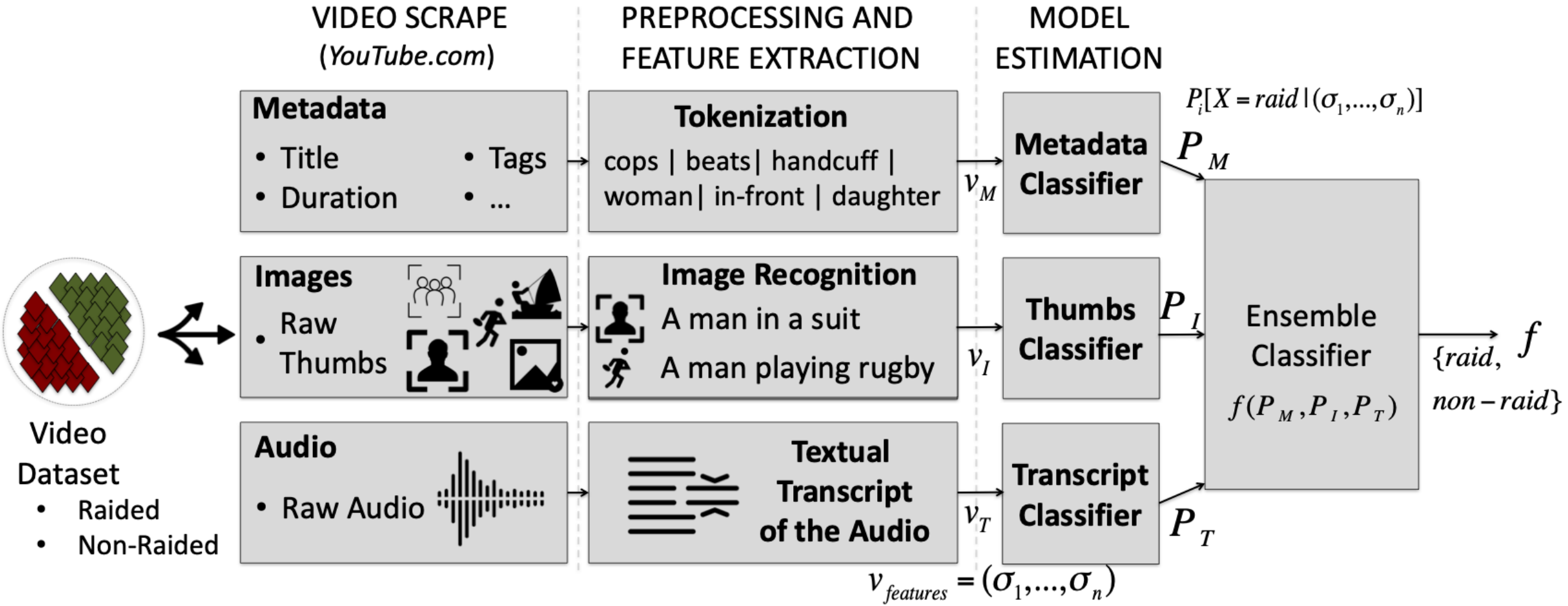
Szgerle 1 month ago
Not even just your regular nazi but ULTRA NAZI!
Reply • 1

Drakewood 1 month ago
+Potato Jenova A FUCKING LEAF!
Reply •

Thought Criminal 1 month ago
Holy fuck, ((((they)))) are actually intimidated by the powers of Kek.
Reply • 169

What if we could predict a raid will happen?

- We can definitely detect if a raid *has* happening
  - But the damage is already done at this point…
- Can we maybe predict that a raid *will* happen?
  - There is some intuition that certain “types” of videos are likely to be raided
- What if, at upload time, we could determine if a video will be raided?
  - It would be great if we could do this by only looking at the video itself!

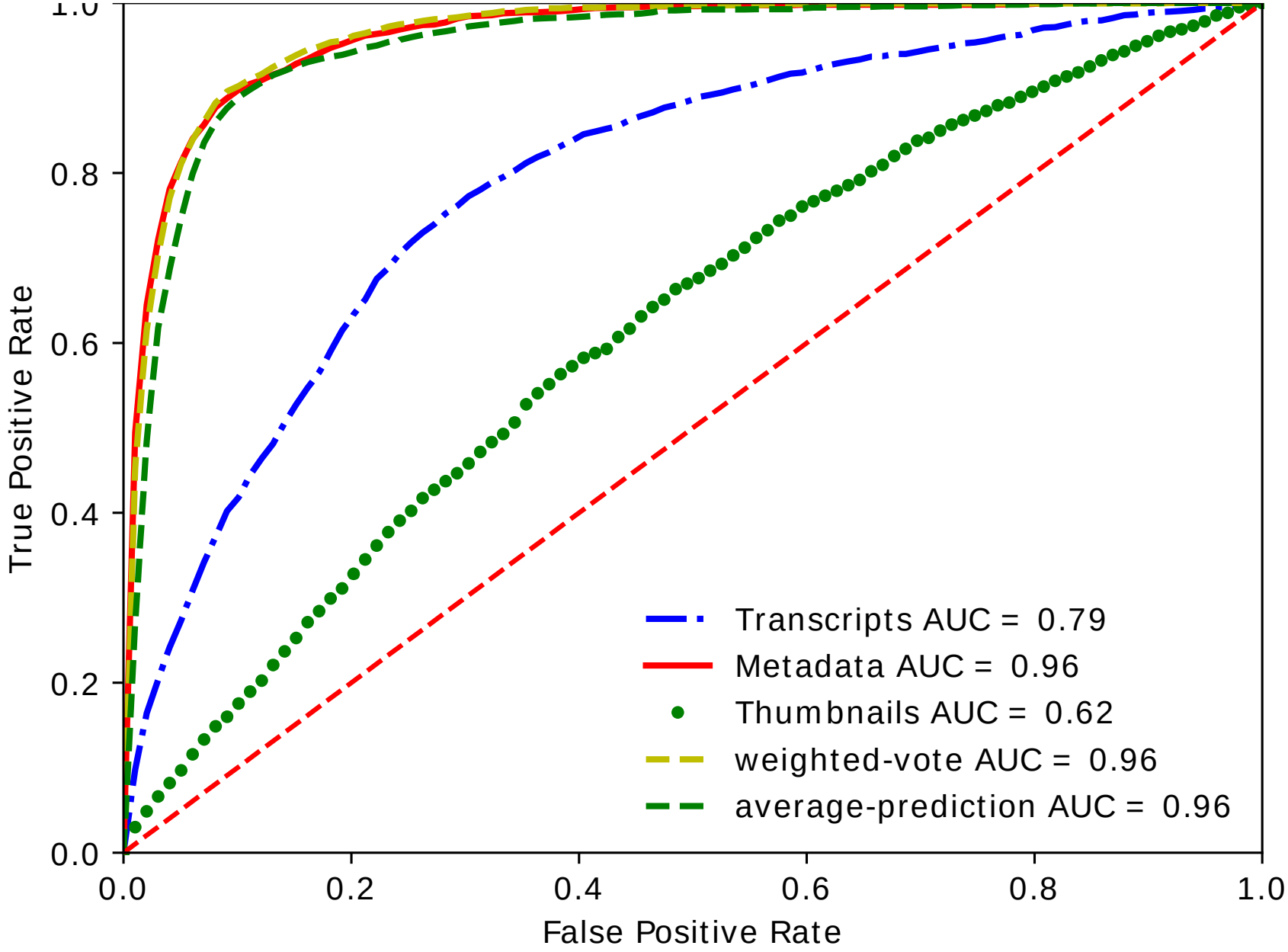

VIDEO SCRAPE
(YouTube.com)
PREPROCESSING AND
FEATURE EXTRACTION
MODEL
ESTIMATION
Metadata
• Title
• Duration
• Tags
• …
Tokenization
cops | beats| handcuff |
woman| in-front | daughter
v_M
Metadata
Classifier
$P_i[X = raid \mid (\sigma_1,\ldots,\sigma_n)]$
P_M
Images
• Raw
Thumbs
Image Recognition
A man in a suit
A man playing rugby
v_I
Thumbs
Classifier
P_I
Ensemble
Classifier
$f(P_M, P_I, P_T)$
$\{raid,$
$non-raid\}$
f
Video
Dataset
• Raided
• Non-Raided
Audio
• Raw Audio
Textual
Transcript
of the Audio
v_T
Transcript
Classifier
P_T
$v_{features} = (\sigma_1,\ldots,\sigma_n)$

Three Experiments

- Experiment 1
  - Can we predict that a video has been linked to on 4chan in the first place?
- Experiment 2
  - Can we discriminate between raided and non-raided videos?
  - Regardless if the video was linked to on 4chan or just randomly sampled
- Experiment 3
  - Can we predict if a video linked to on 4chan will be raided?

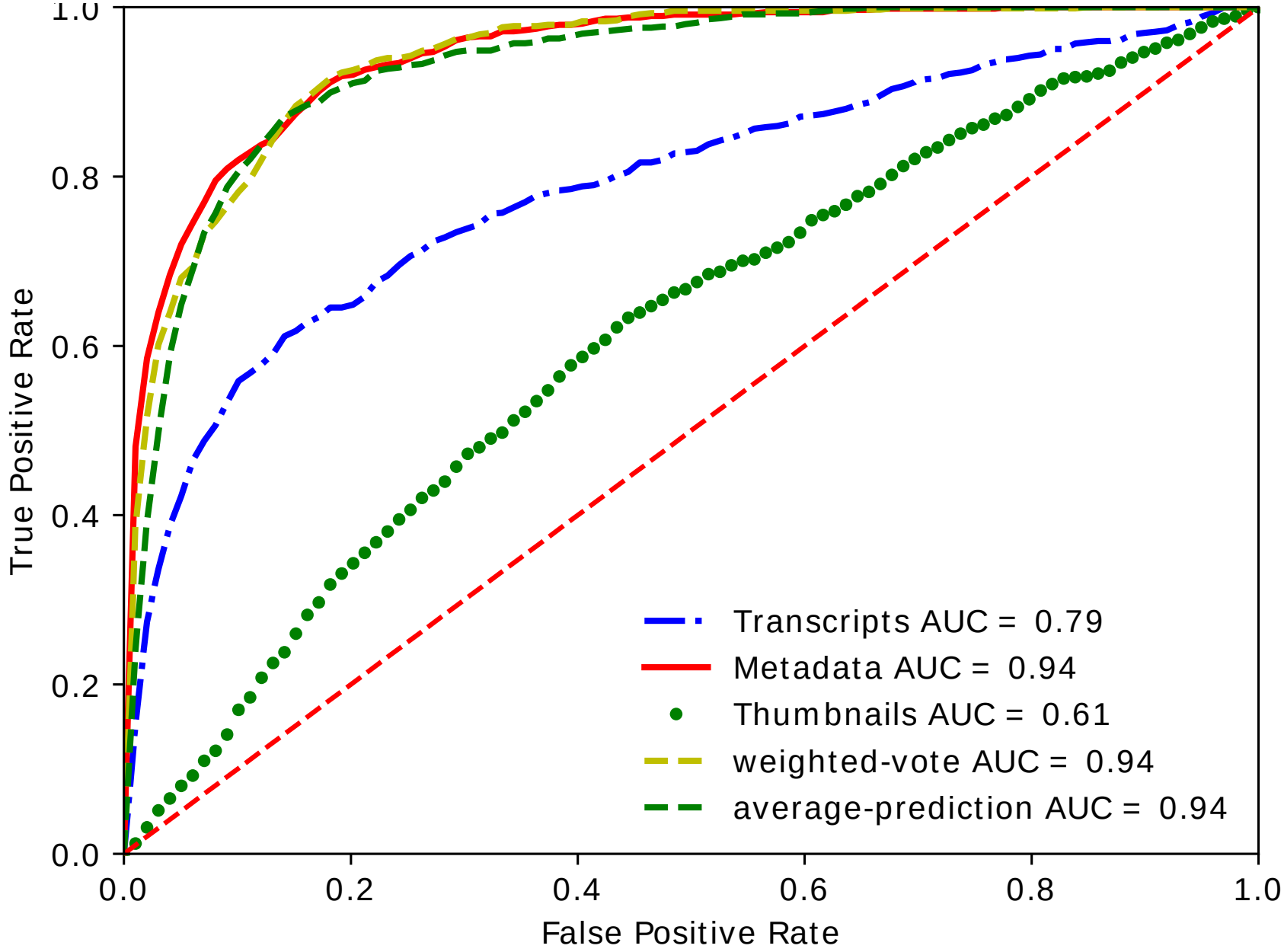
Experiment 1

Can we predict that a video has been linked to on 4chan in the first place?



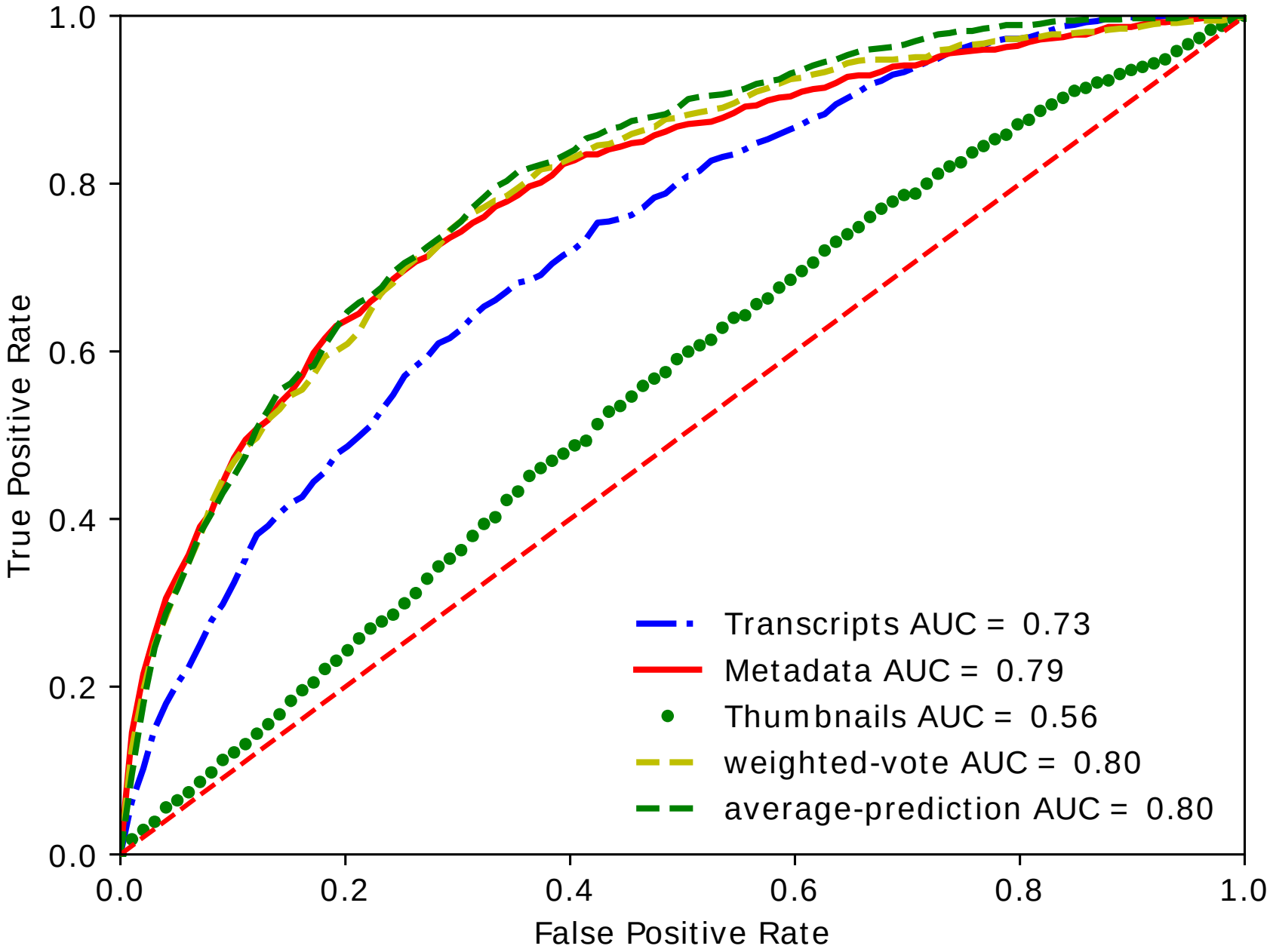
Experiment 2

Can we discriminate between raided and non-raided videos?



Experiment 3

Can we predict if a video linked to on 4chan will be raided?



Detecting State Sponsored Troll accounts on social media

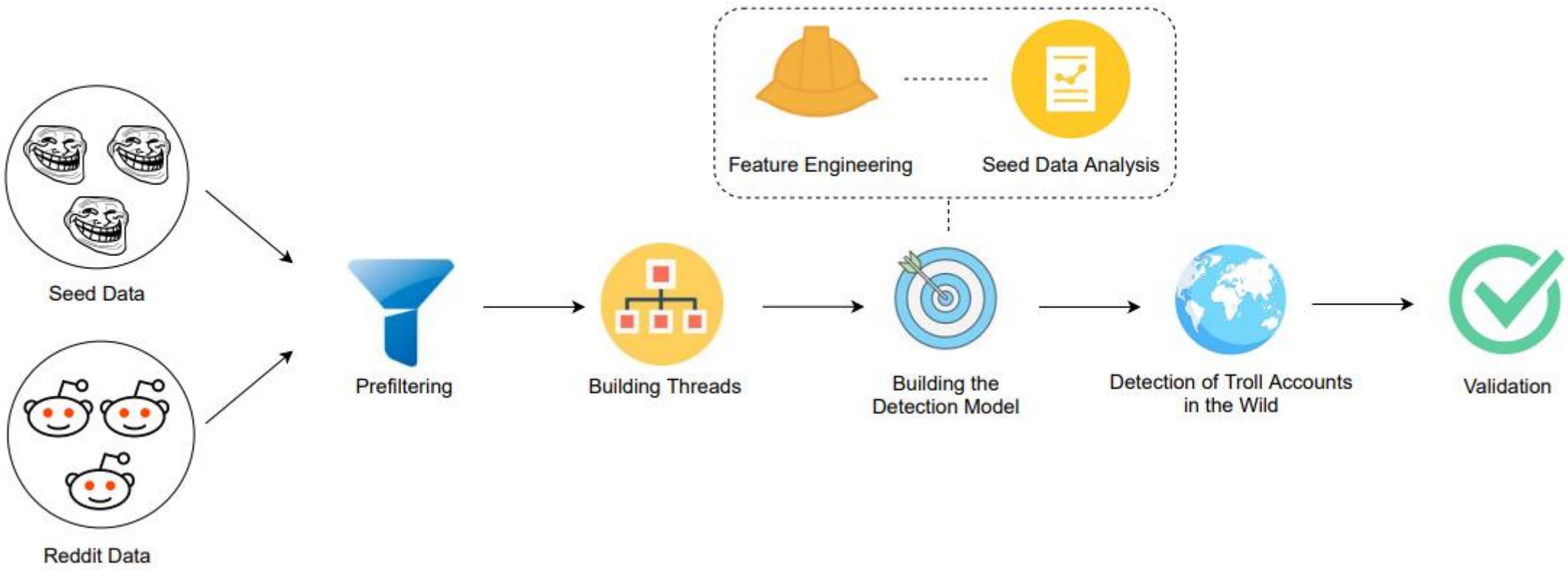
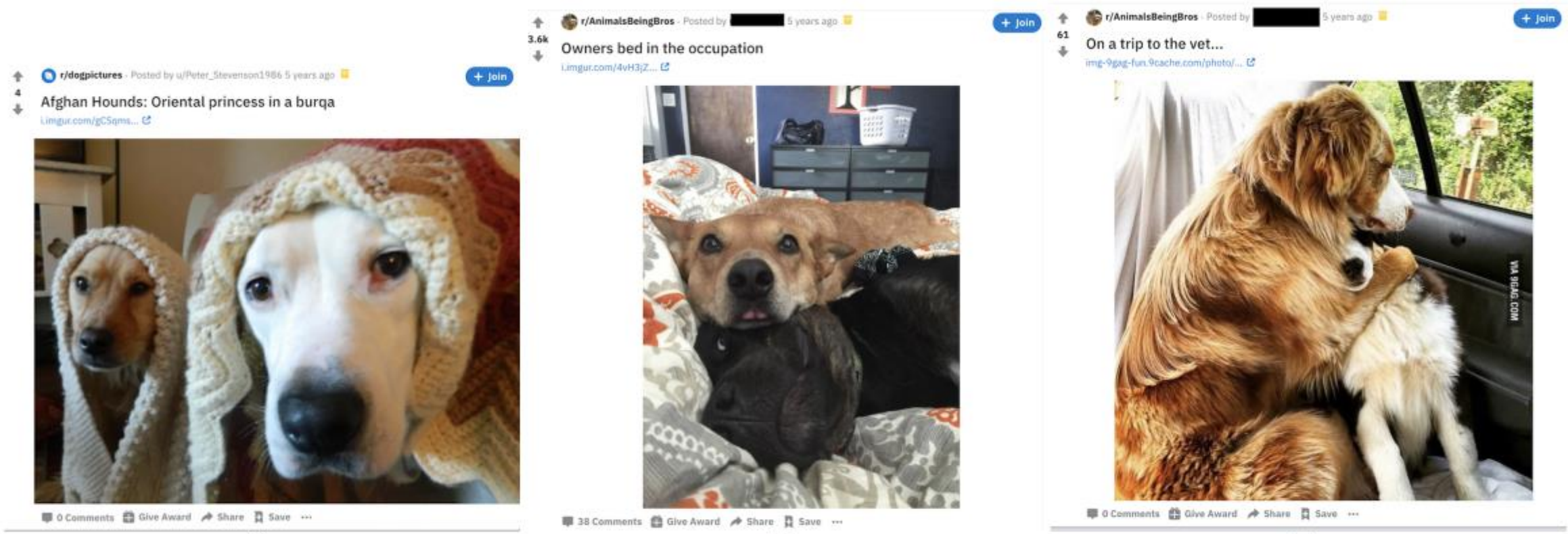
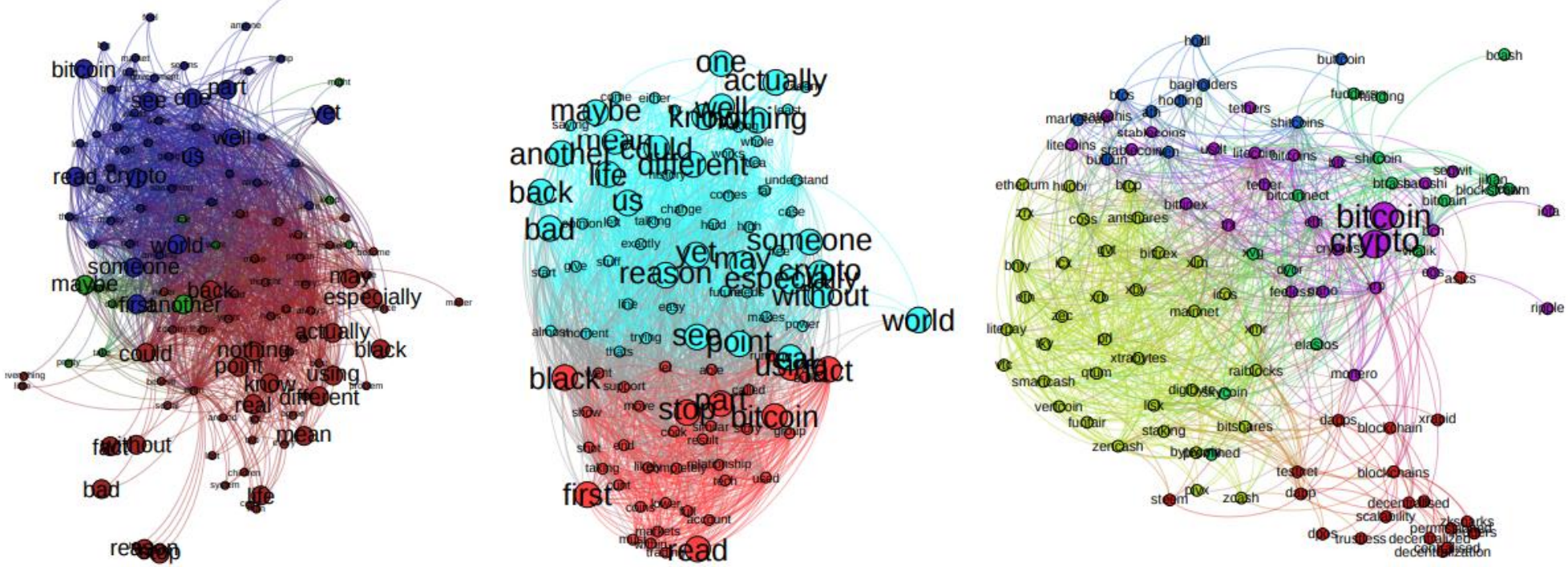


Figure 5: An example of manufactured conflict between known trolls and accounts detected by TROLLMAGNIFIER.



(a) (b) (c)

Figure 6: The figure shows the similarity of posts made by known trolls and accounts detected by TROLLMAGNIFIER. The left-most post is made by a known-troll and the other two are from accounts detected by TROLLMAGNIFIER.

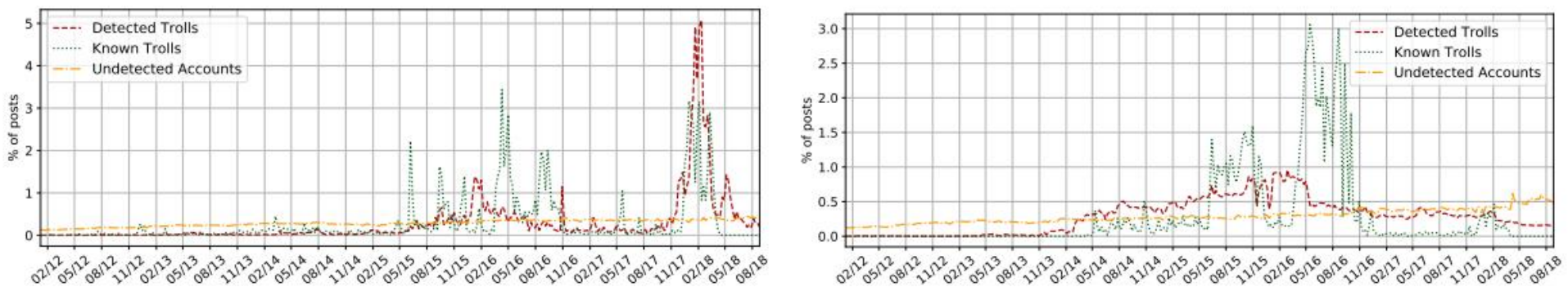


(a) *Known trolls*. Larger nodes represent words common with (b).

(b) *Detected trolls*. Larger nodes represent words common with (a).

(c) *Non-trolls*. Larger nodes ("crypto" and "bitcoin") are common with (a).

Figure 3: A visualization of language usage in relation to the keyword "crypto" where nodes from the same community (detected using the Louvain community detection method [5]) are depicted with the same color. It is evident that known trolls and trolls detected by TROLLMAGNIFIER have more words in common than known trolls and non-trolls.



(a) Comments

(b) Submissions

Figure 4: Time Series of Comments and Submissions.

Great! We can actually predict something!

- We can probably do something *before* a raid happens!

- But what do we do?!
  - We could maybe temporarily disable comments
  - We could maybe put extra human moderators on it
  - We could …

- How would we even measure the effectiveness of any of these?
  - We are not YouTube, we have limited access to data…

There *is* a deployed mitigation strategy

Facebook imposes major new restrictions on QAnon, stepping up enforcement against the conspiracy theory

A person wears a QAnon sweatshirt during a pro-Trump rally Saturday in New York City. (Stephanie Keith/Getty Images)

By **Craig Timberg** and **Isaac Stanley-Becker**

Oct. 6, 2020 at 11:15 p.m. EDT

ation strategy

Facebook imposes major new restrictions on QAnon, stepping up enforcement against the conspiracy theory

A person wears a QAnon sweatshirt during a pro-Trump rally Saturday in New York City. (Stephanie Keith/Getty Images)

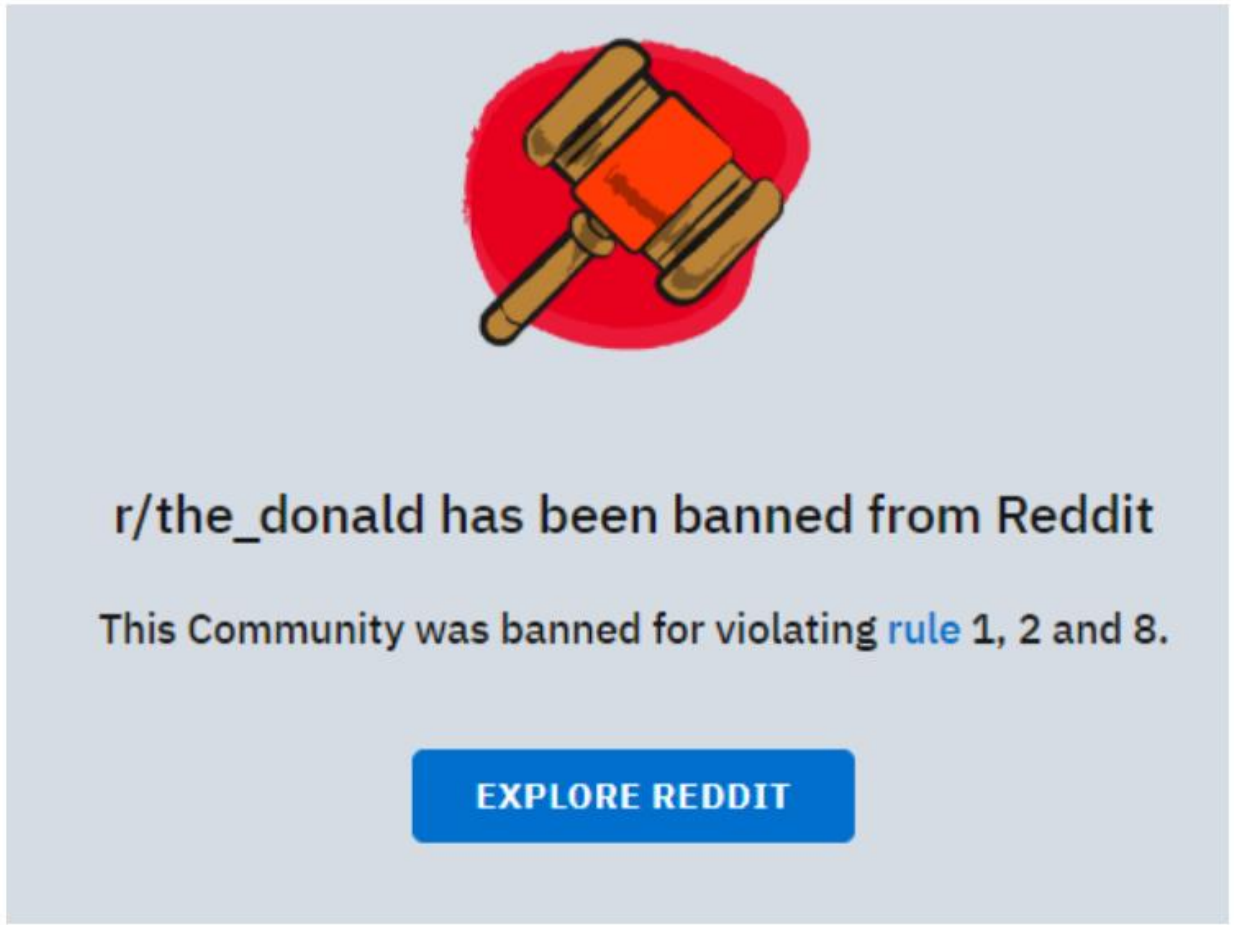
By **Craig Timberg** and **Isaac Stanley-Becker**

Oct. 6, 2020 at 11:15 p.m. EDT

Reddit Bans The_Donald, Forum Of Nearly 800,000 Trump Fans, Over Abusive Posts

June 29, 2020 · 5:10 PM ET

BOBBY ALLYN



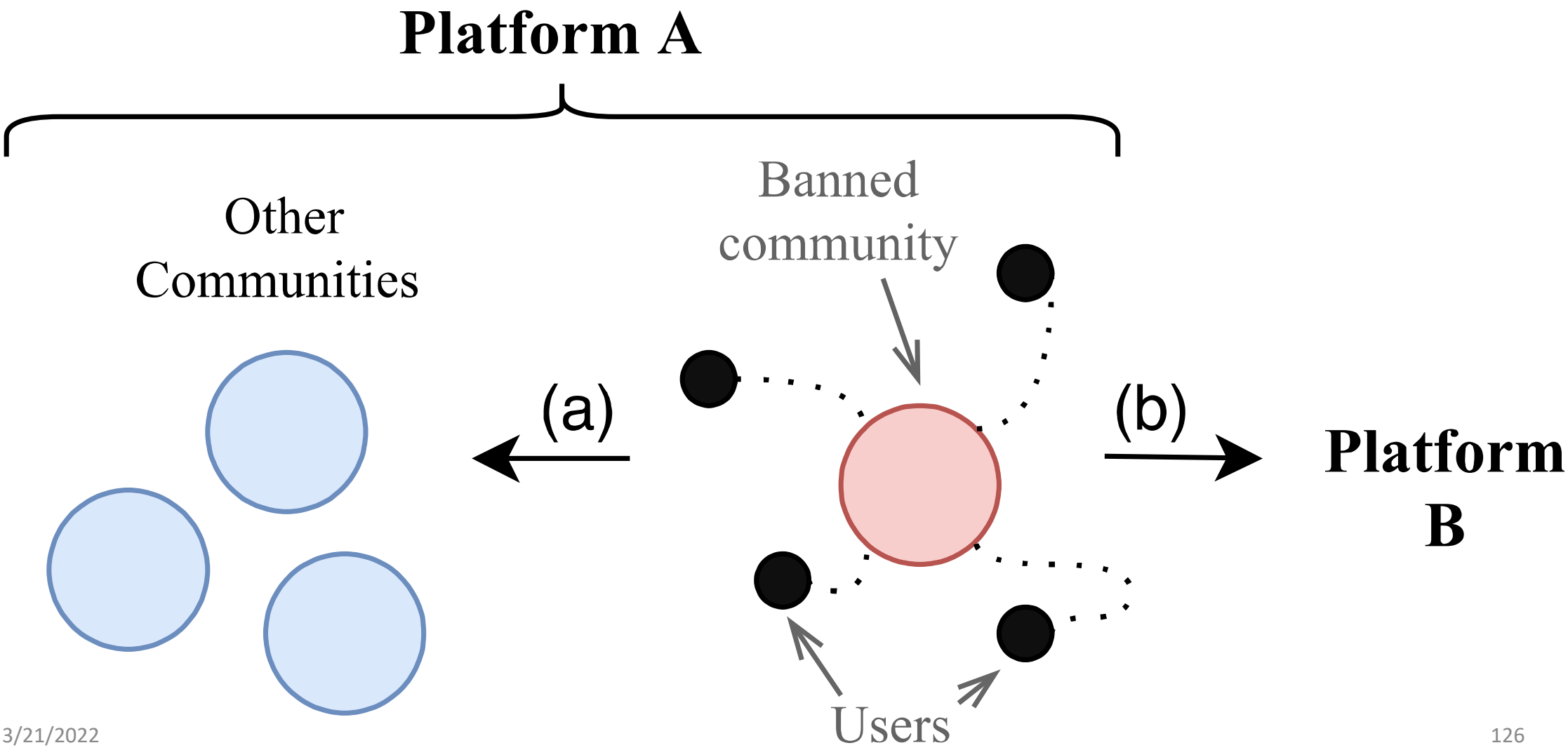
Reddit announced on Monday that it has banned the popular subreddit for Trump fans called The_Donald. Reddit previously had taken action against the forum over posting content that violated its rules.

Reddit.com

Deplatforming/Banning Communities

- Social media platforms have made serious efforts here
- The idea is that if some community is being toxic, excise it
- Seems pretty straight forward
  - Some research shows it does actually clean up the platform a bit
- But, turns out nothing is that simple…

Yup, they just make their own sites!



Far-right finds new online home in TheDonald.win

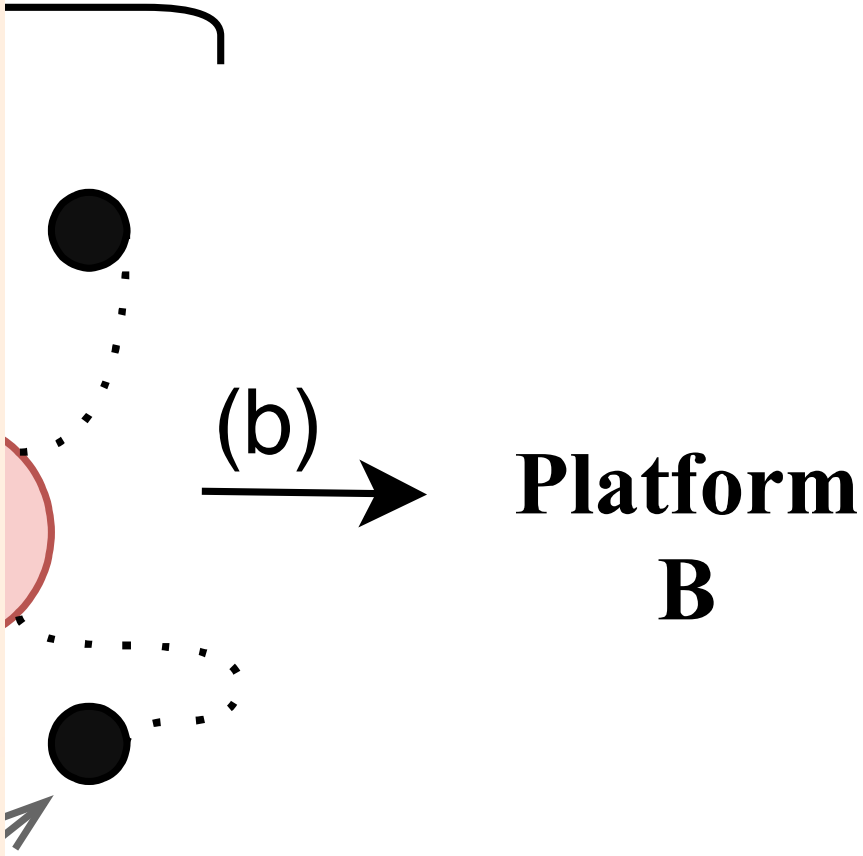
Pro-Trump meme website gains popularity as critics express concern over its role in the spread of hateful content

The far-right has its own social media spaces, including the forum TheDonald.win, which originated in Reddit © FT montage

Siddharth Venkatramakrishnan in London AUGUST 10 2020

344

sites!



Far-right finds new online home in TheDonald.win

Pro-Trump meme website gains popularity as critics express con… role in… spread of hateful content

The far-right has its own social media spaces, in… forum TheDonald.win, which originated in Reddit © FT montage

Siddharth Venkatramakrishnan in London AUGUST 10 2020

344

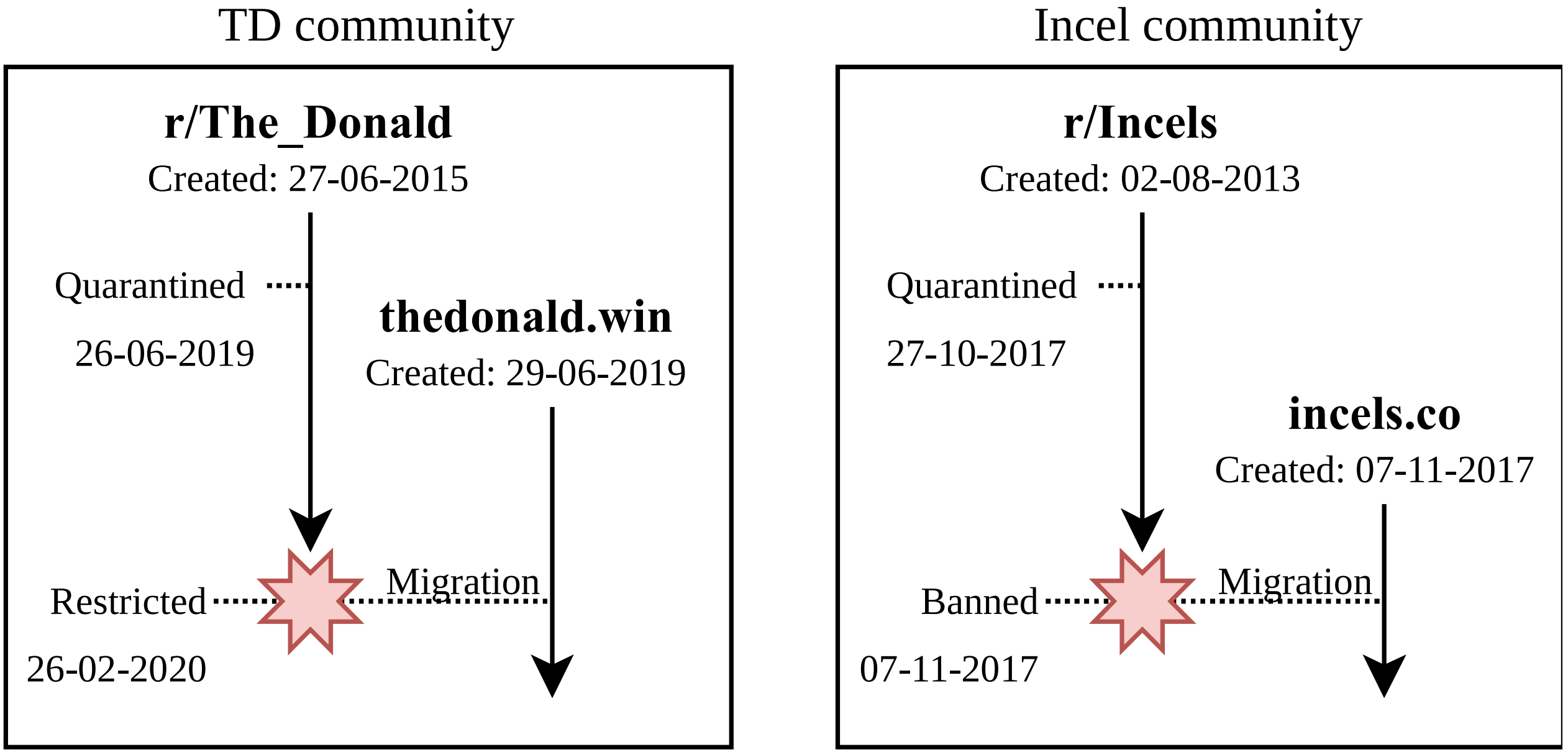
We know that fringe-communities like 4chan affect the rest of Web…

Just because Reddit is a bit cleaner, what happens when these communities move offsite?!?

…es!

…latform B

We Look at Two Such Communities


TD community
r/The_Donald
Created: 27-06-2015
Quarantined
26-06-2019
thedonald.win
Created: 29-06-2019
Restricted
26-02-2020
Migration
Incel community
r/Incels
Created: 02-08-2013
Quarantined
27-10-2017
incels.co
Created: 07-11-2017
Banned
07-11-2017
Migration

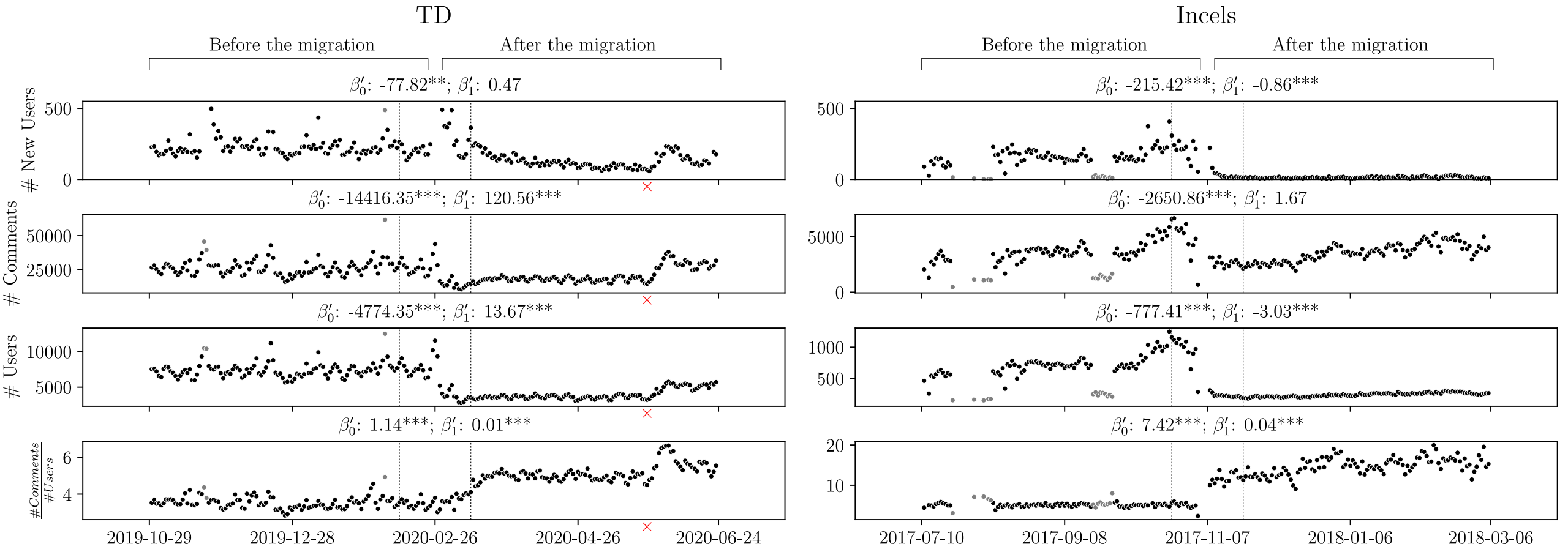
We Look at Two Such Communities

Platform	Community	Submissions	Comments	Users
Reddit	/r/Incels	17,403	340,650	18,088
	/r/The_Donald	251,090	2,703,615	80,002
Forums	Incels.co	25,138	385,765	2,270
	thedonald.win	280,156	2,390,641	38,767

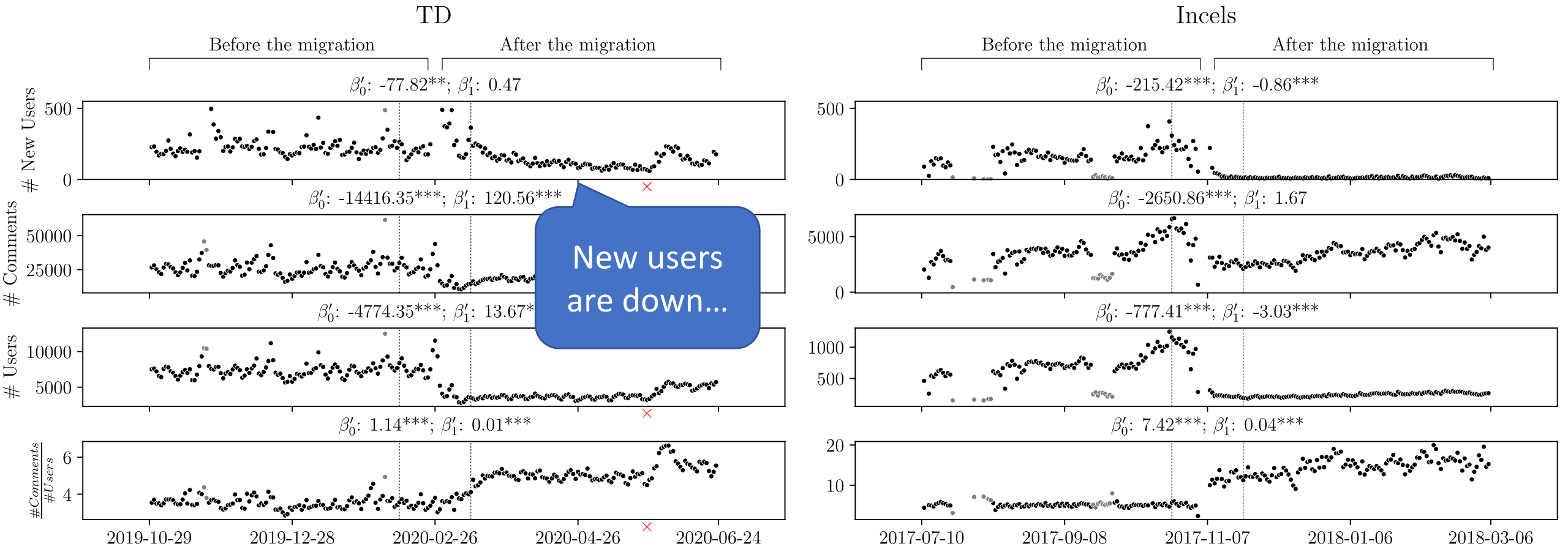
Measure Across Several Axes

- Activity levels
  - Were the communities dealt a serious blow?
- Fixation dictionary
  - Previous work shows that fixation is a warning sign of radicalization
- Group identification
  - Intensification of in-/out-group identification is also a warning sign
- Toxicity level
  - A major justifications for banning is that these communities are *toxic*
  - I.e., they cause harm long lasting harm to the entire platform

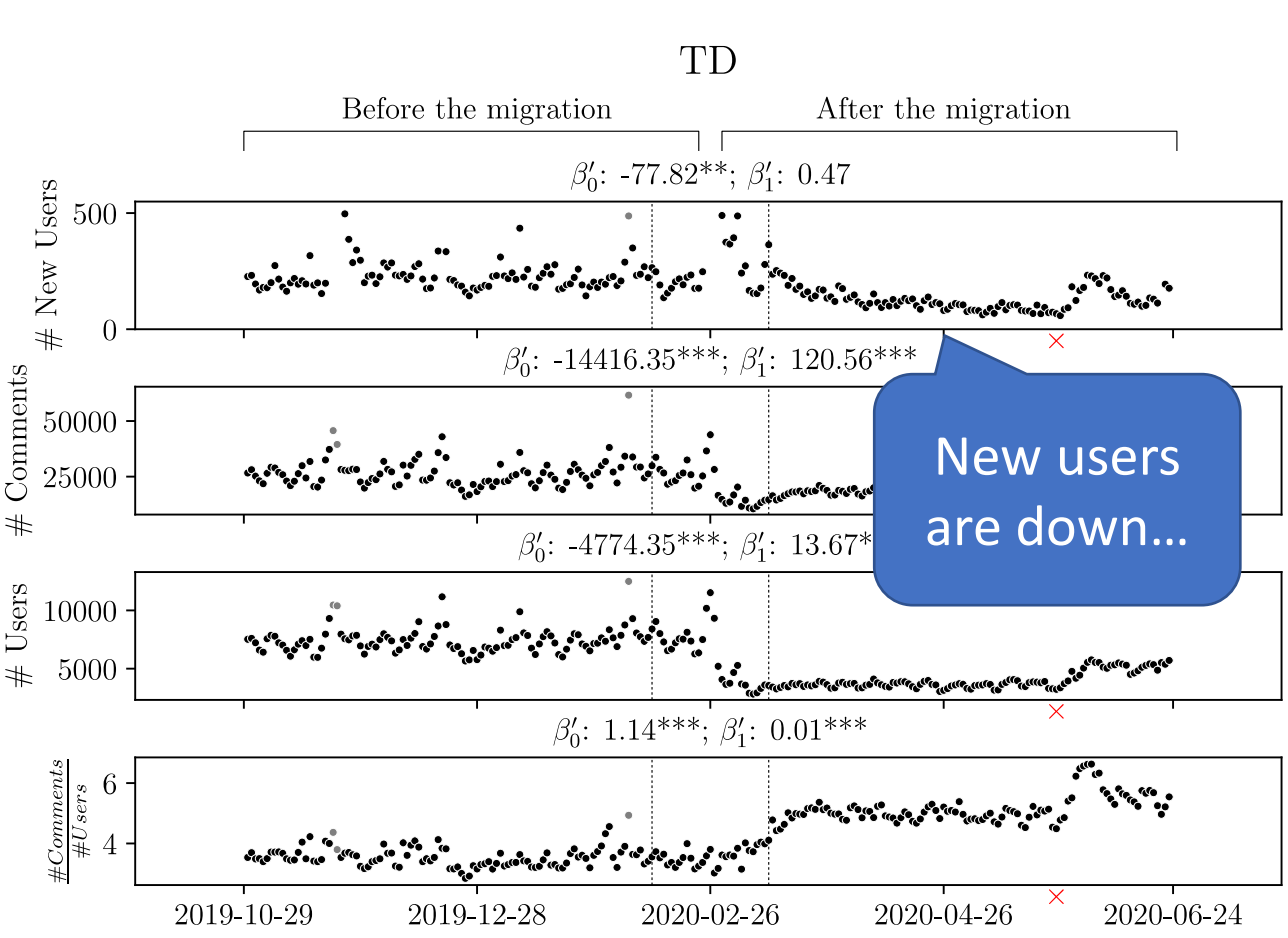
Have the communities become less active?

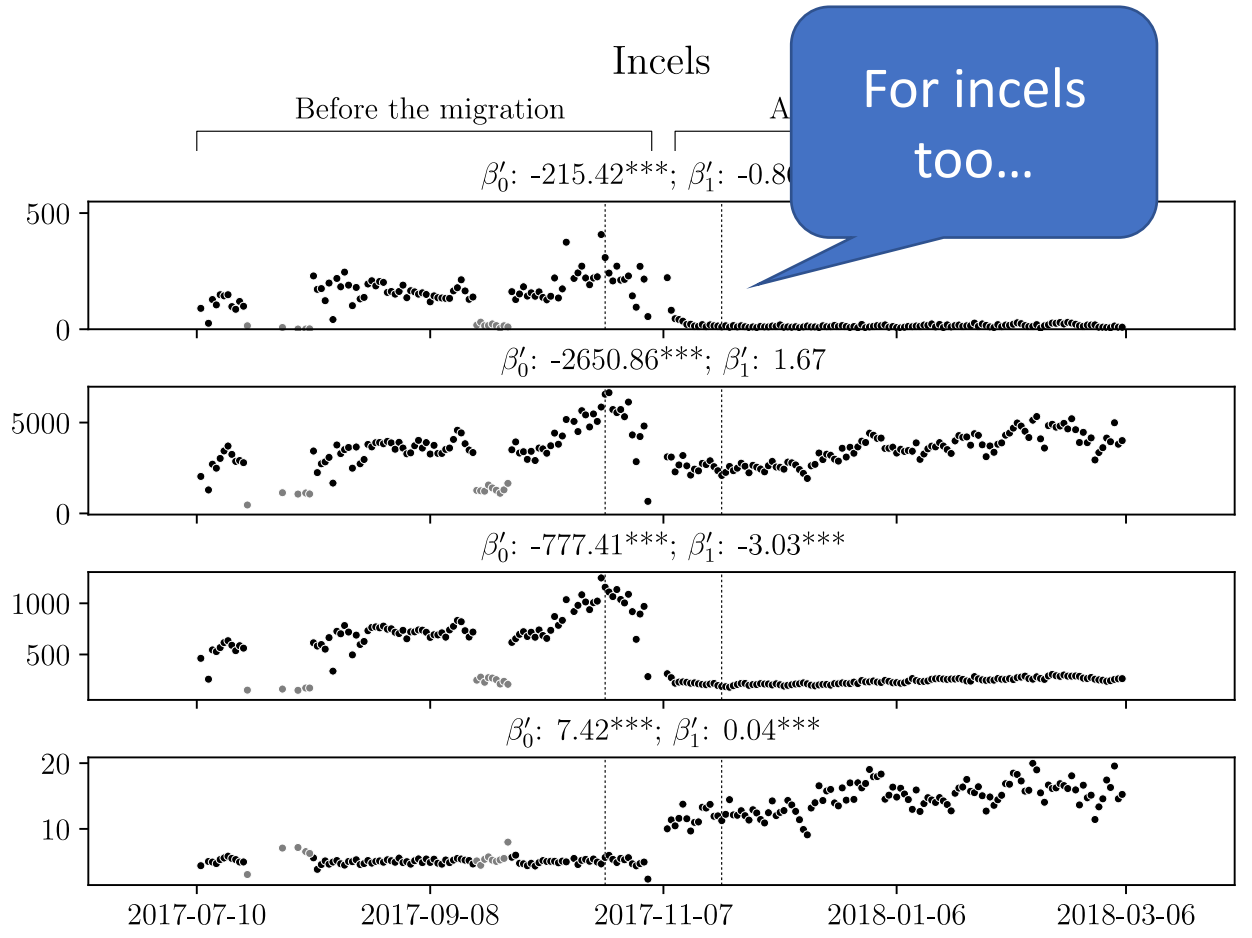


Have the communities become less active?

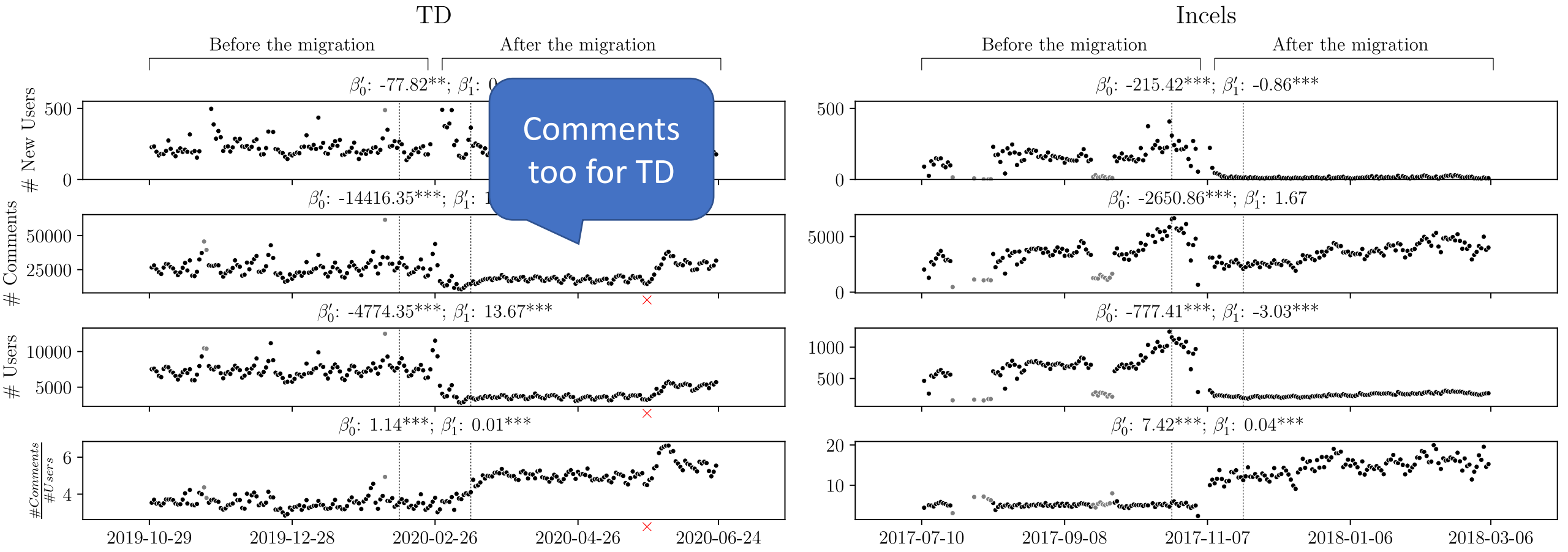


Have the communities become less active?

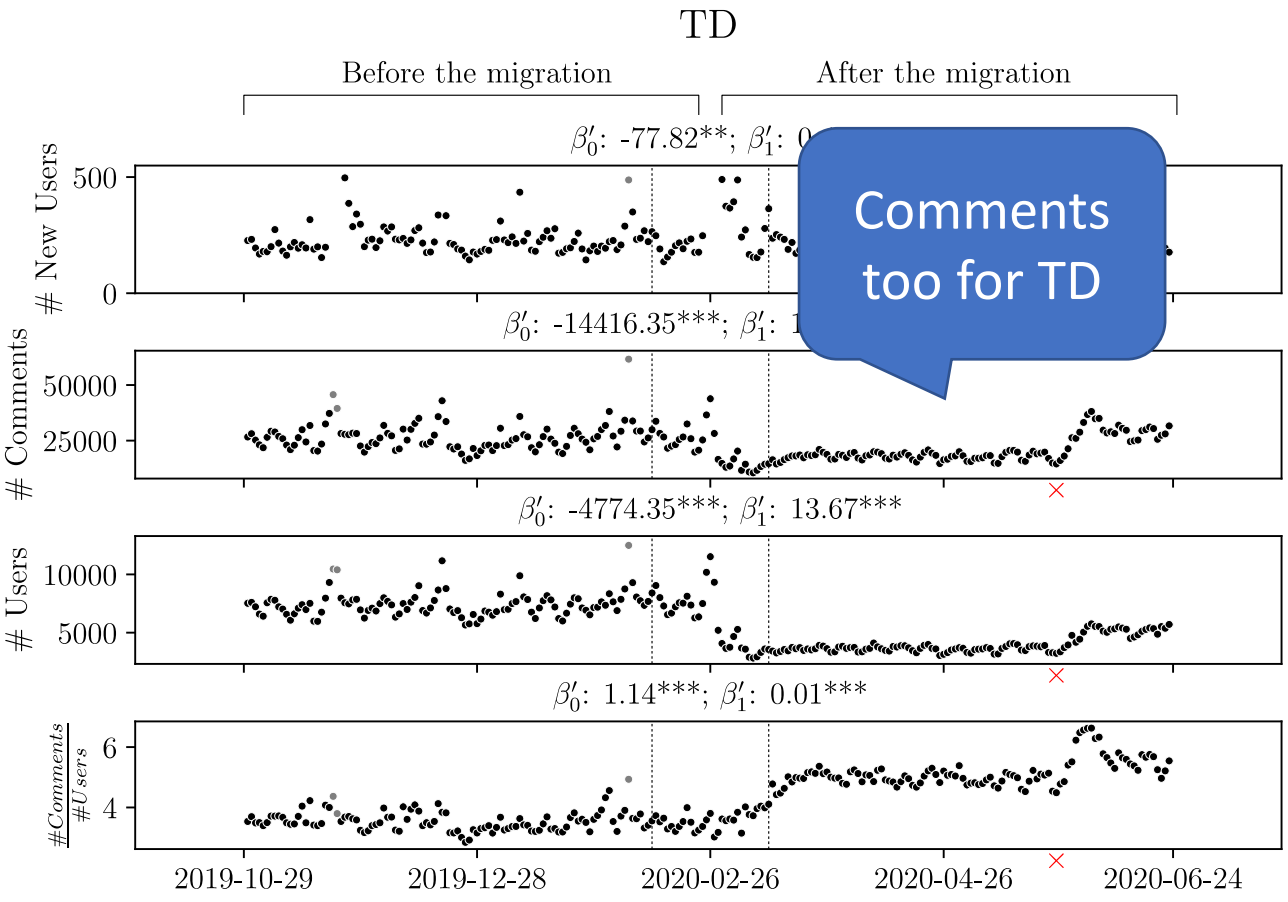


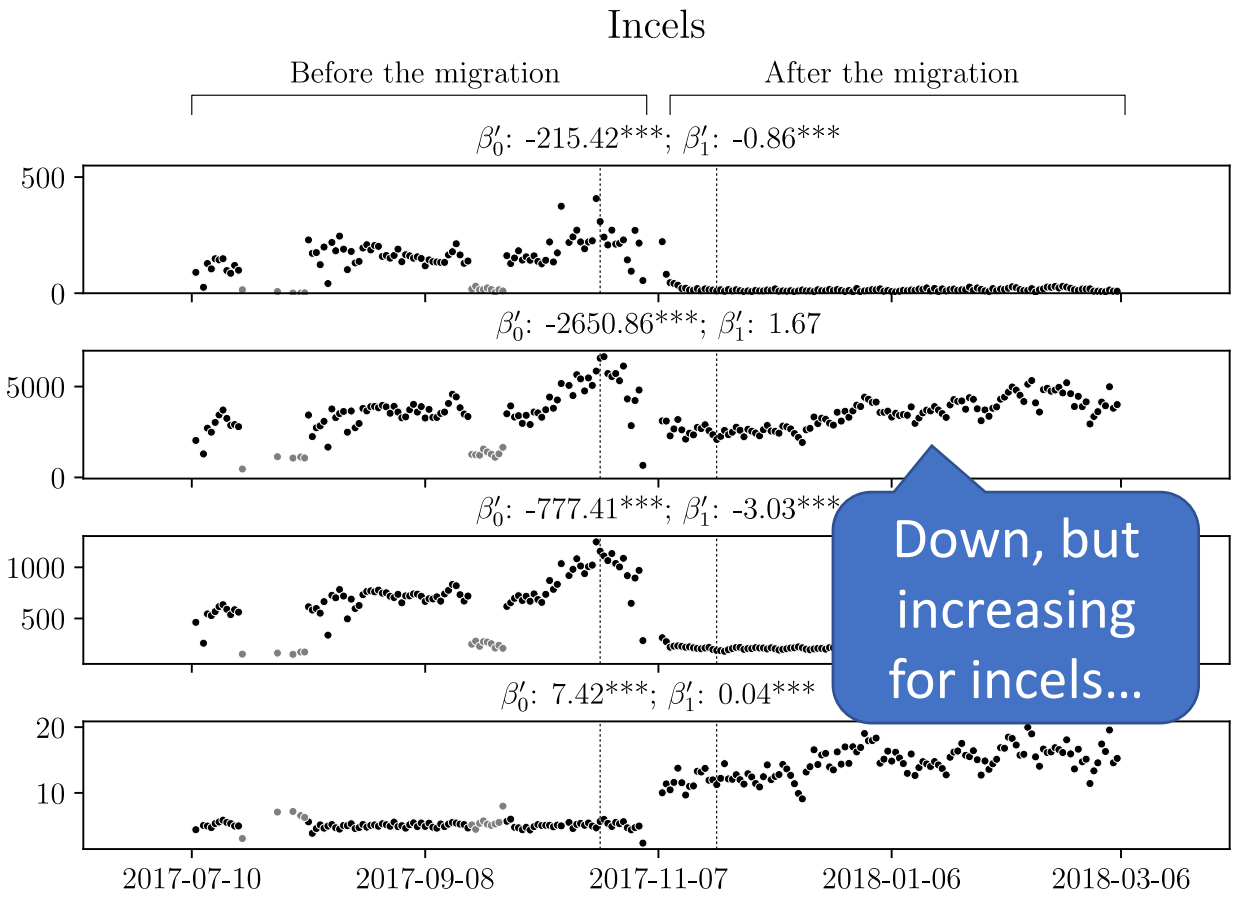


Have the communities become less active?

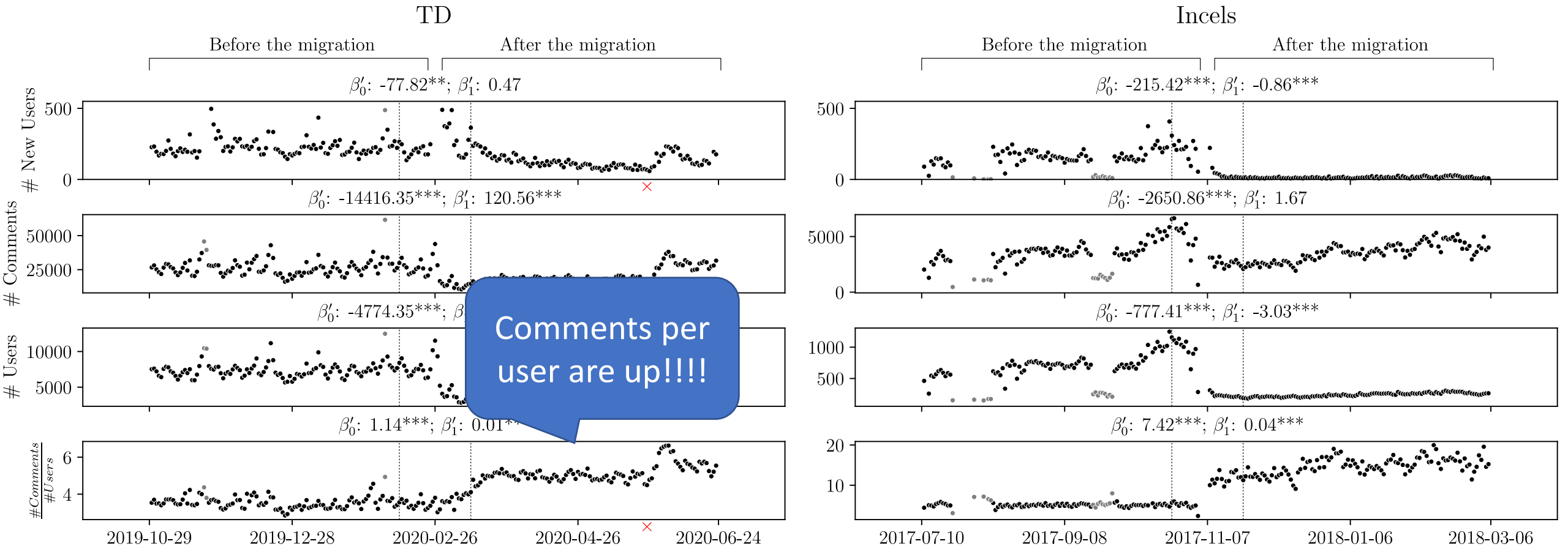


Have the communities become less active?

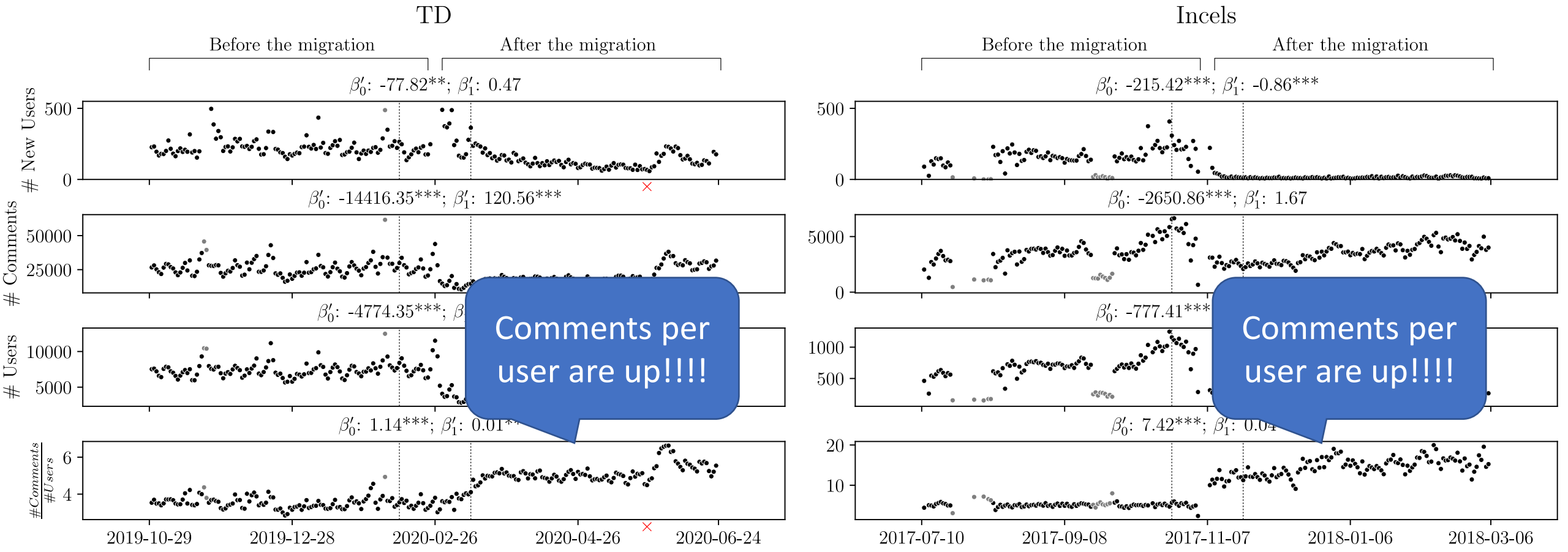




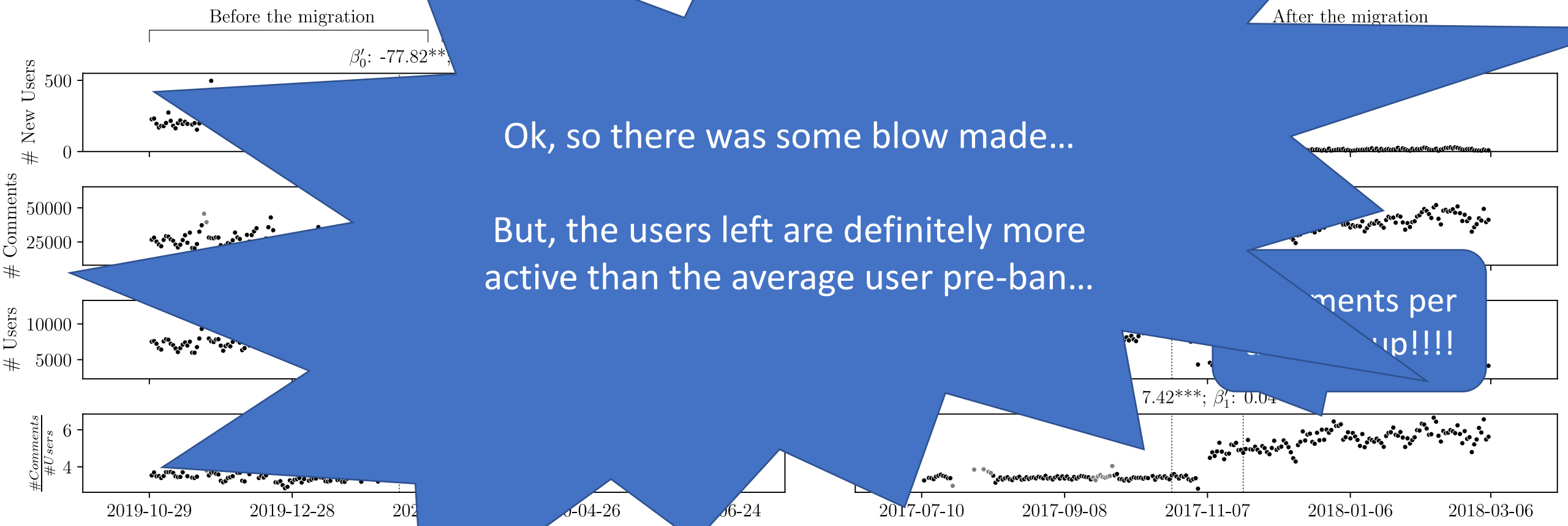
Have the communities become less active?



Have the communities become less active?



Have the communities become less active?



Ok, so there was some blow made…

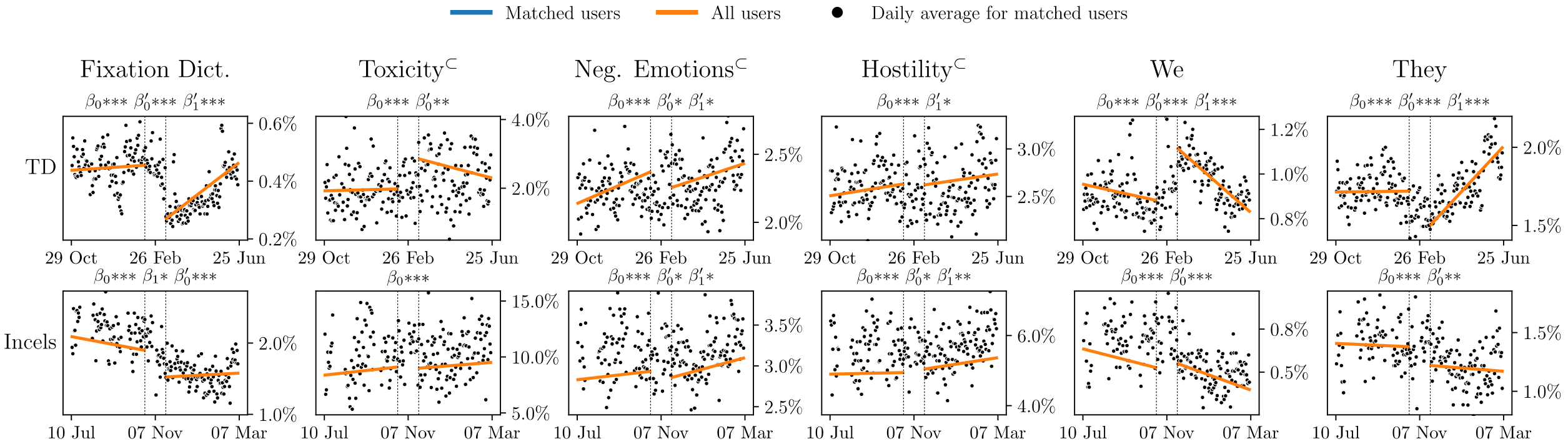
But, the users left are definitely more active than the average user pre-ban…

ments per
up!!!!

Ok, about these users that stayed…

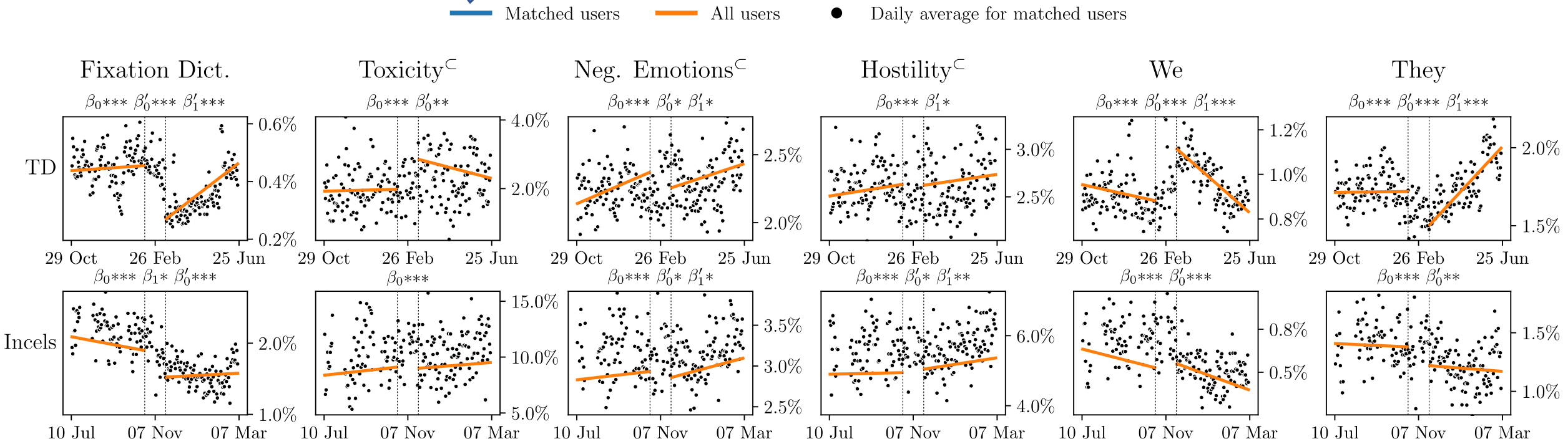
- It’s great that Reddit did disrupt the community
- There is definitely reduced activity
- Fewer users overall
- Less new users
- But, have the users that migrated changed?

Changes in individual users

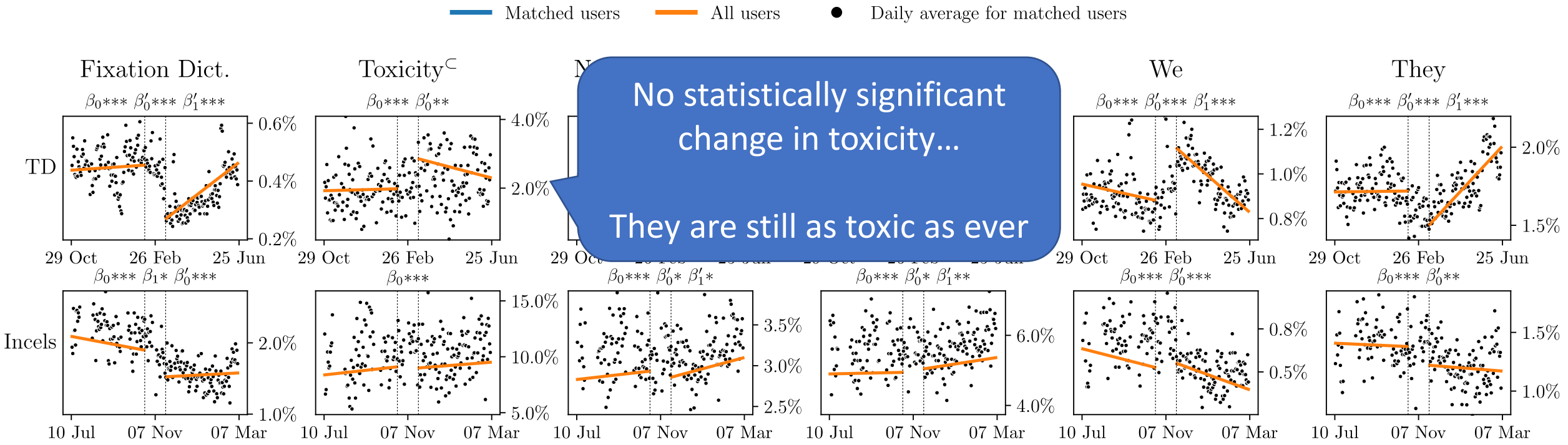


Cha vidual users

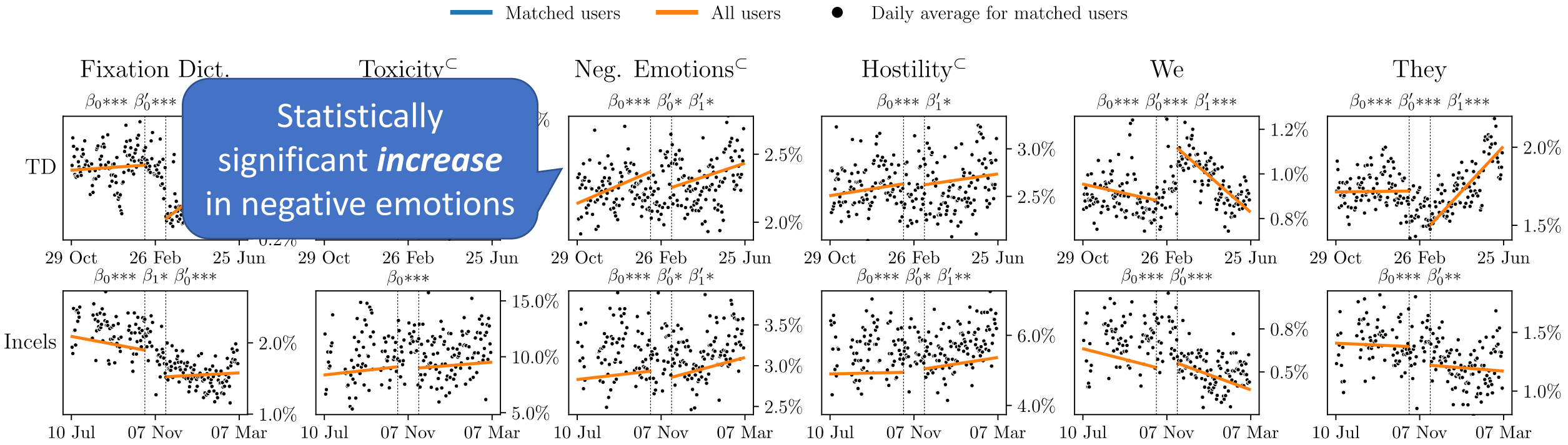
We matched reddit user names to new platform user names



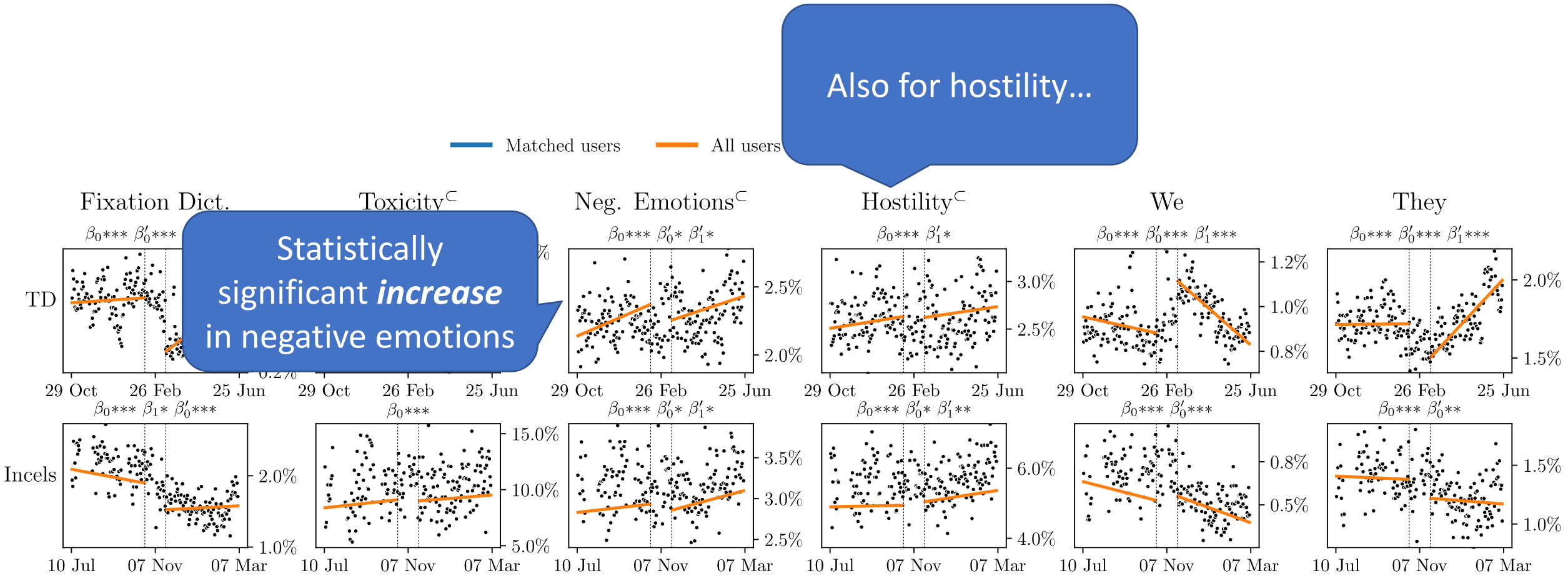
Changes in individual users



Changes in individual users



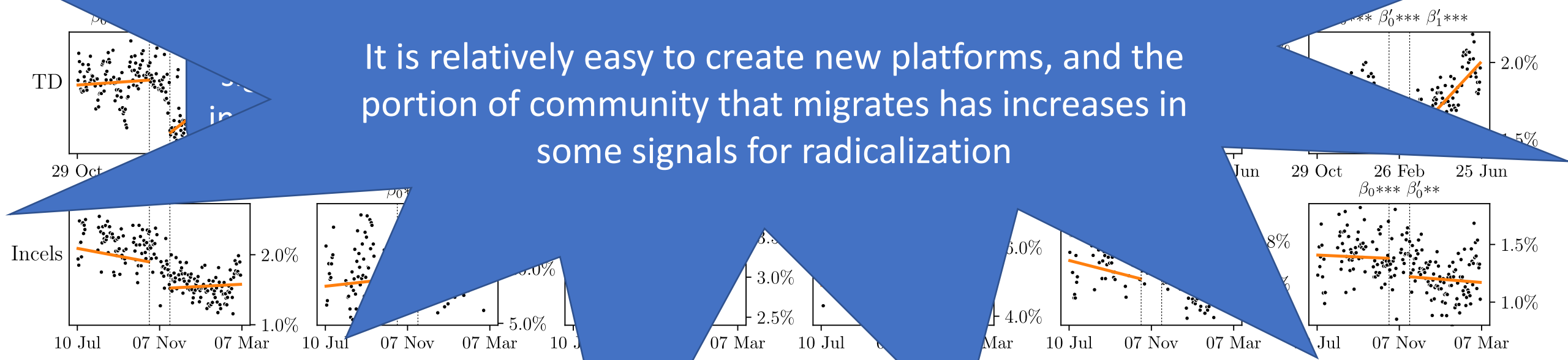
Changes in individual users



Changes in individual users

Large-scale banning/deplatforming of communities might backfire to some extent.

It is relatively easy to create new platforms, and the portion of community that migrates has increases in some signals for radicalization



Summary

- We have entered a new historical age
- There are a host of emerging problems
  - We've only just begun to identify and understand these
- State-sponsored actors are in the game
- There are tons of challenges to understanding things
  - There's only so much content any person can look at
- Mitigation techniques are in their infancy
  - We are just learning how to automatically detect this behavior
  - It's a *socio*-technical problem; solutions can have unforeseen consequences