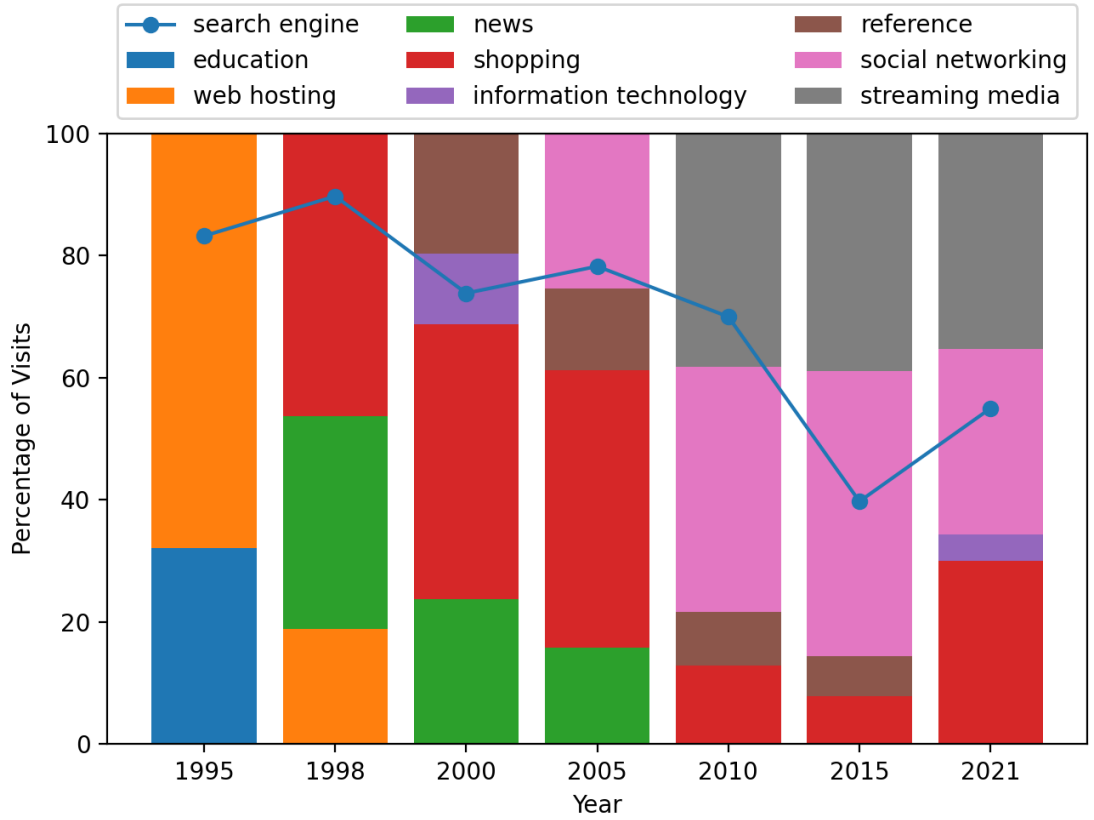
Monitoring & mitigating Online Harms

is decentralization the answer to partisanship, hate speech, tracking etc.?

NISHANTH SASTRY

UNIVERSITY OF SURREY

25 years of the Web: From News & Education → Social Media & Streaming



[V. Agarwal et al. WWW'22]

The people's revolution, 2006

"For seizing the reins of the global media, for founding and framing the new digital democracy, for working for nothing and beating the pros at their own game, Time's Person of the Year for 2006 is **you**,"

Perils of Misinformation

- Misinformation → Partisan Mistrust → Hatred of 'other'
- For the gullible: Confirmation bias enhances acceptability of misinformation
- For the sceptical: *Information* market for lemons
- For the opportunistic: Liar's dividend

End Result: A "Post Truth" society is a DIVIDED Society!

Market for Lemons: Akerlof, 1970
Liar's Dividend: Chesney and Citron, 2018
https://www.urbandictionary.com/define.php?term=Liar%E2%80%99s%20Dividend

Roadmap

1. Partisanship (and tracking)
2. Online hate and polarization
3. Is decentralization the answer?

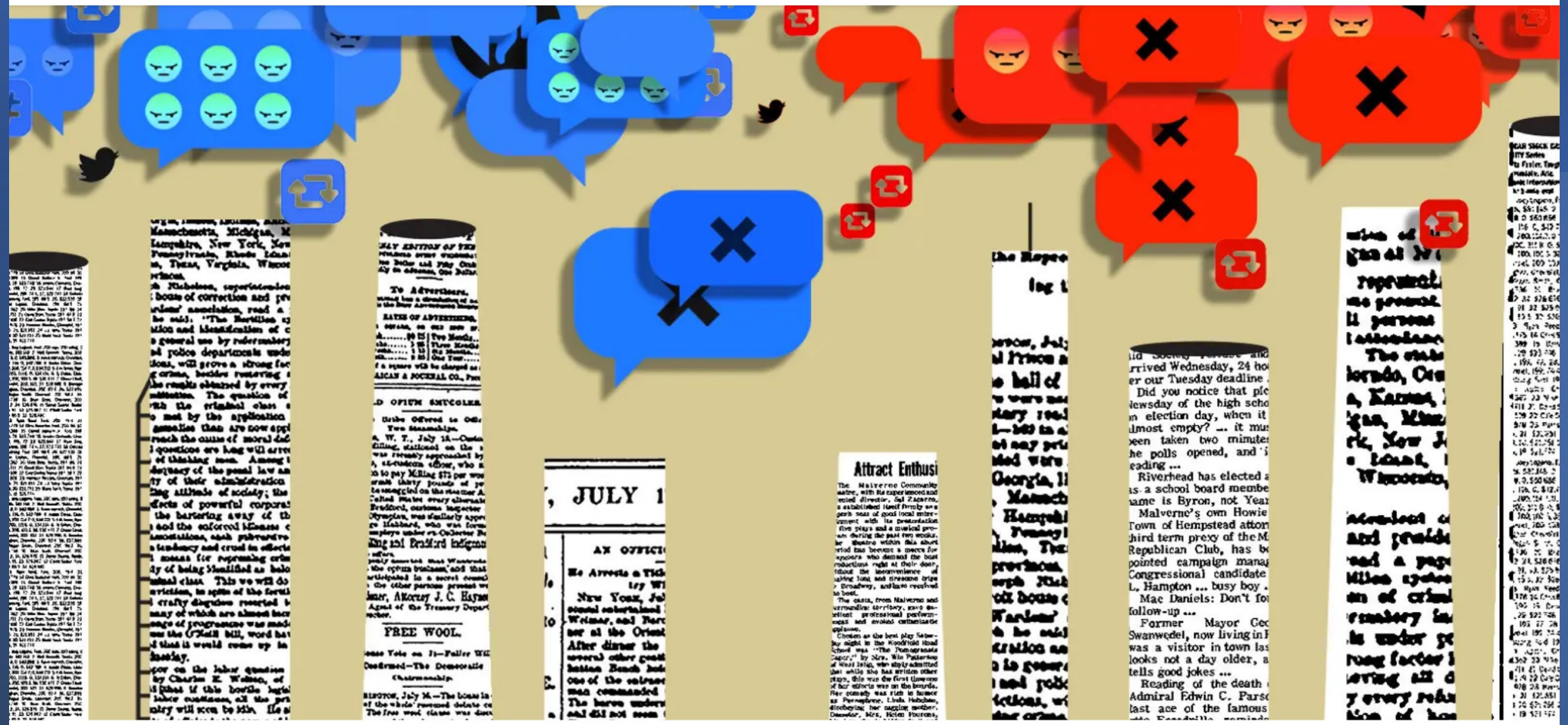
Partisan (adj): prejudiced in favour of a particular cause

1.Partisanship (and tracking)

Partisan news in the USA [*WWW'18*]

Stop Tracking me Bro: Differential Tracking on partisan sites [WWW'20]

Differential and topical tracking in Indian News Media [ICWSM'21;WebSci'21]

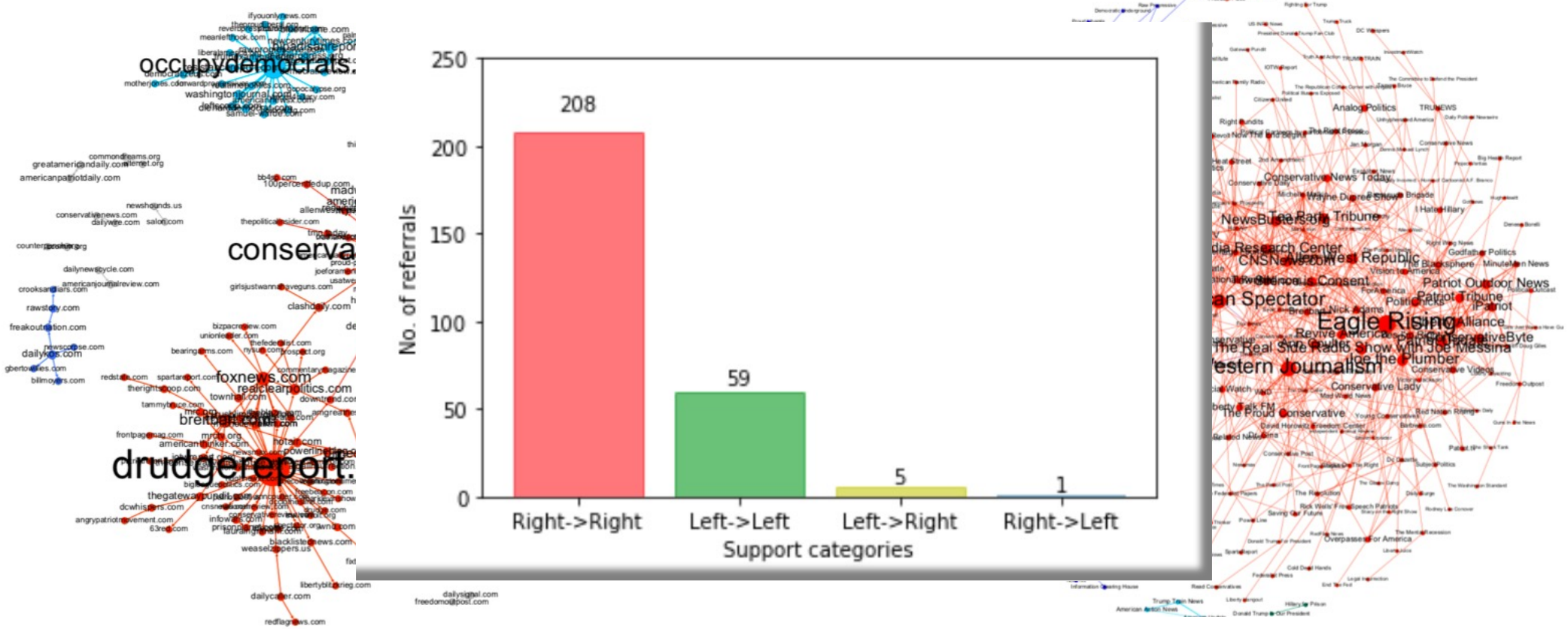


Inside The Partisan Fight For Your News Feed

How ideologues, opportunists, and internet marketers built a massive new universe of partisan news on the web and on Facebook.

https://www.buzzfeednews.com/article/craigsilverman/inside-the-partisan-fight-for-your-news-feed

Partisan referrals and filter bubbles



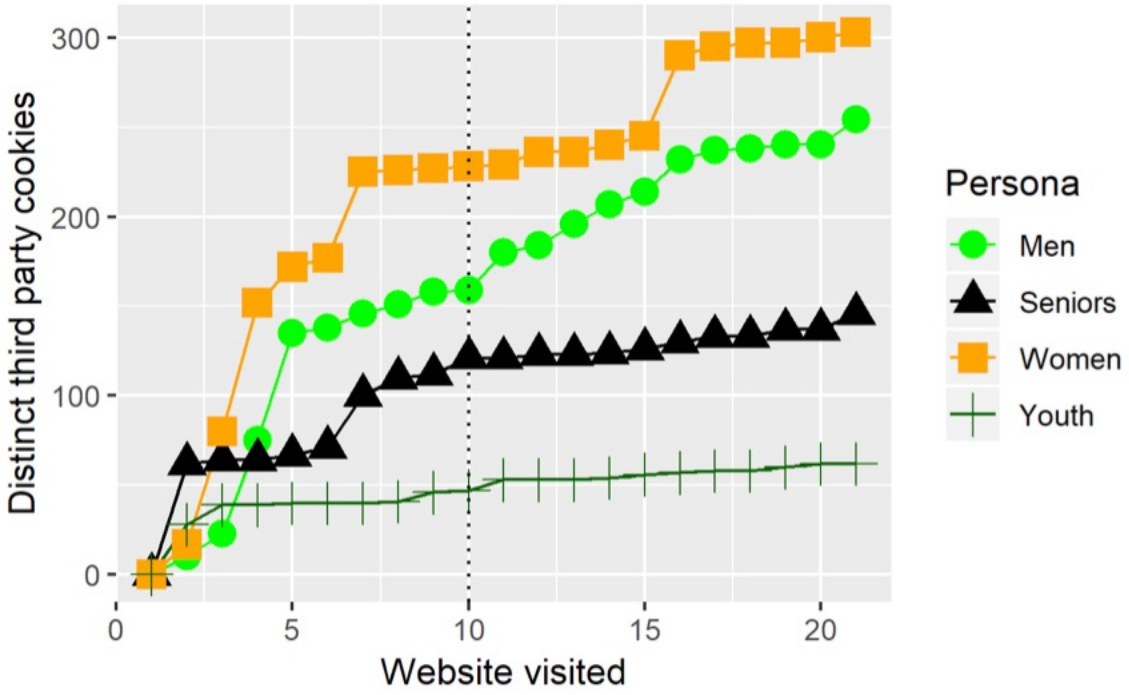
Website traffic referral

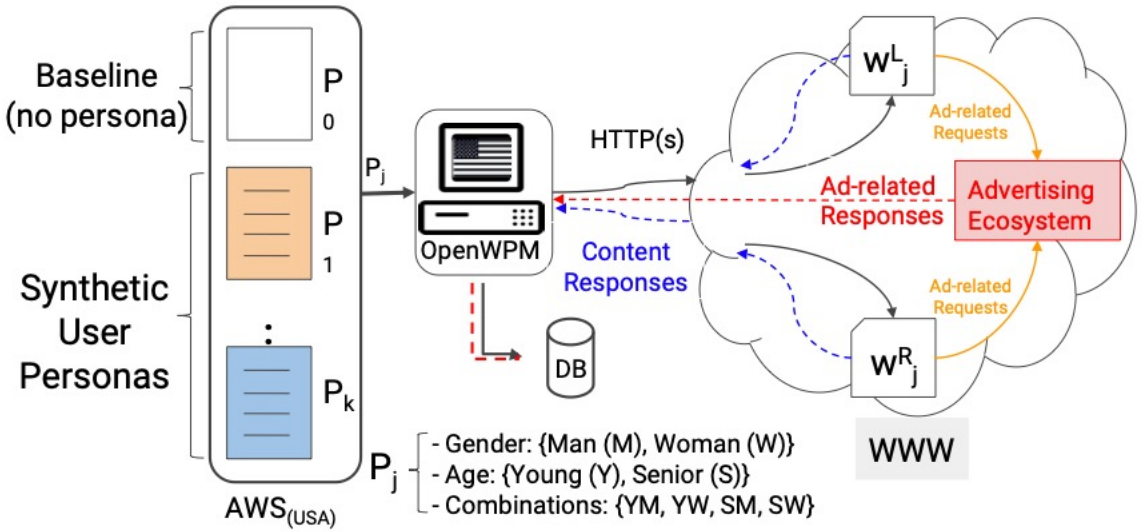
Facebook page likes

*dataset shared publicly: get in touch if interested

[Bhatt et al. WWW'18]

Stop Tracking Me Bro!: Differential tracking on partisan news sites



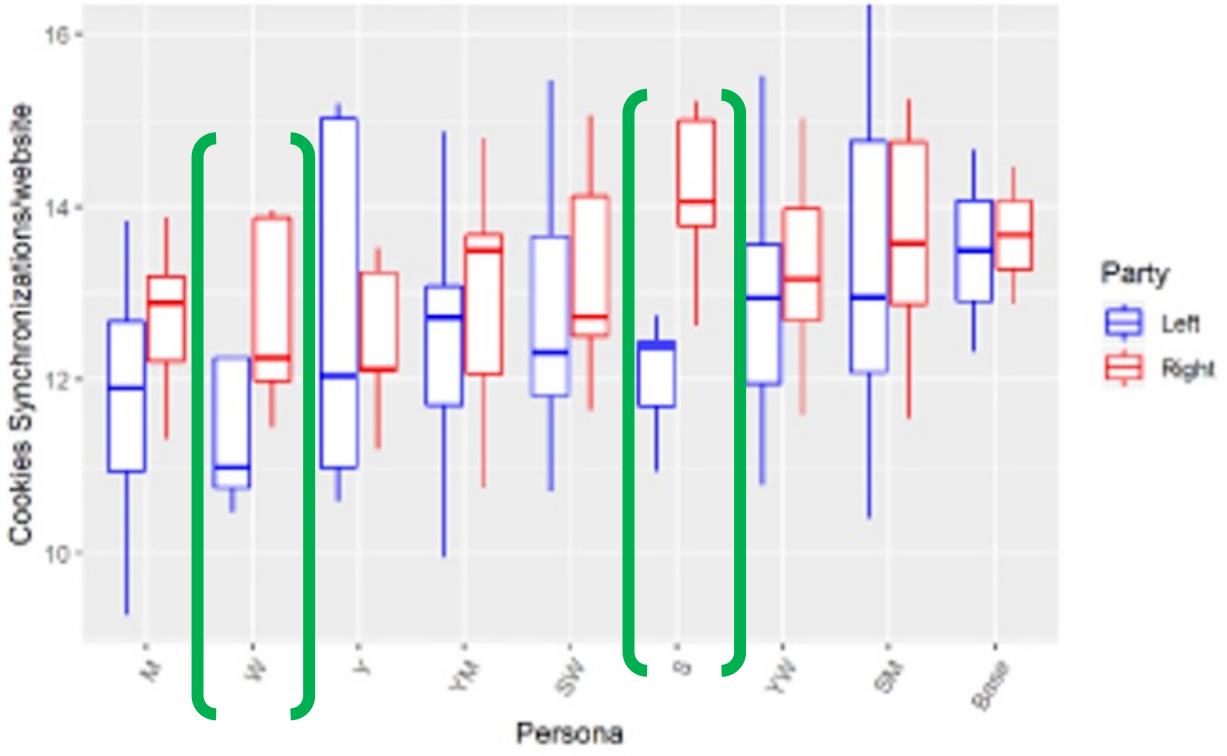


1. Incremental Persona Building

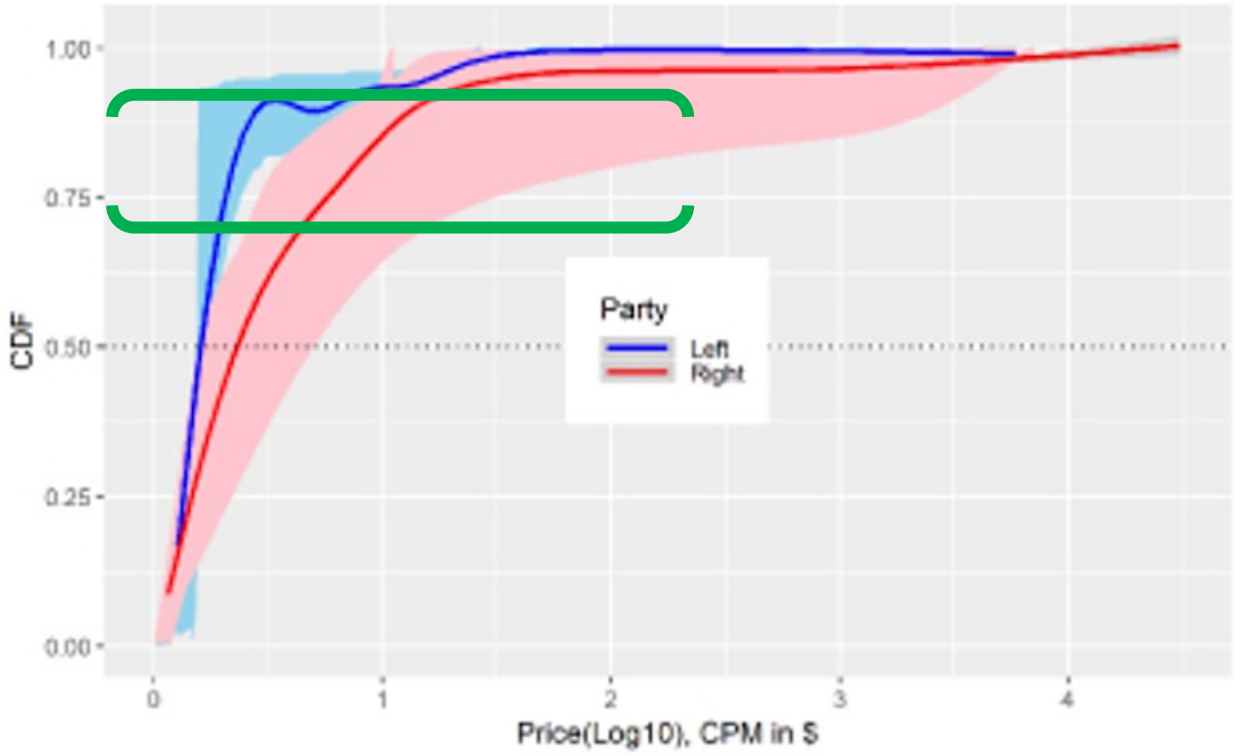
2. Tap into RTB info flowing through

[P. Agarwal et al. WWW'20]

Differential tracking on partisan news sites



(a) Cookie synchronization



(b) Distribution of Prices

[P. Agarwal et al. WWW'20]

Tangle factors of real browser histories

Tracking the ad ecosystem across 85+ countries with 10K+ installs

UNIVERSITY OF SURREY DANS

- **Loss of privacy is faster in UK than China!**
  - *China's GFW blocks Google & Facebook (and also their trackers)*

[X. Hu et al. Euro S&P'20]

- **Systematic metric for comparison of ad blockers, browsers and tracking prevention using Tangle Factor as a metric**
  - *E.g., Using Adblocker Plus and uBlock origin results in UK tracking dropping to levels experienced in China*

[X. Hu et al. WebSci'20]

- **Tracking decreased in EU after GDPR but has gone back up**
  - *Because many users simply click 'Accept all' on GDPR cookie consents*

Artefacts to share (get in touch if interested):

- Lightbeam-Chrome – Visualise and understand third-party ecosystem from your browsing history
- CookieMonster – ML to auto-classify cookies into essential/functional/ad & analytics at scale
- CookieCutter – Automatically reject third-party cookies (even outside GDPR-regulated regions)

2. Polarisation and Hate Speech

Tweeting MPs in the UK [ICWSM'19]

Hate speech in political discourse [HT'21]

An Expert Annotated Dataset for the Detection of Online Misogyny [EACL'21]

GraphNLI: Graph neural network to understand polarity and hate speech [WWW'22,TWEB'23]

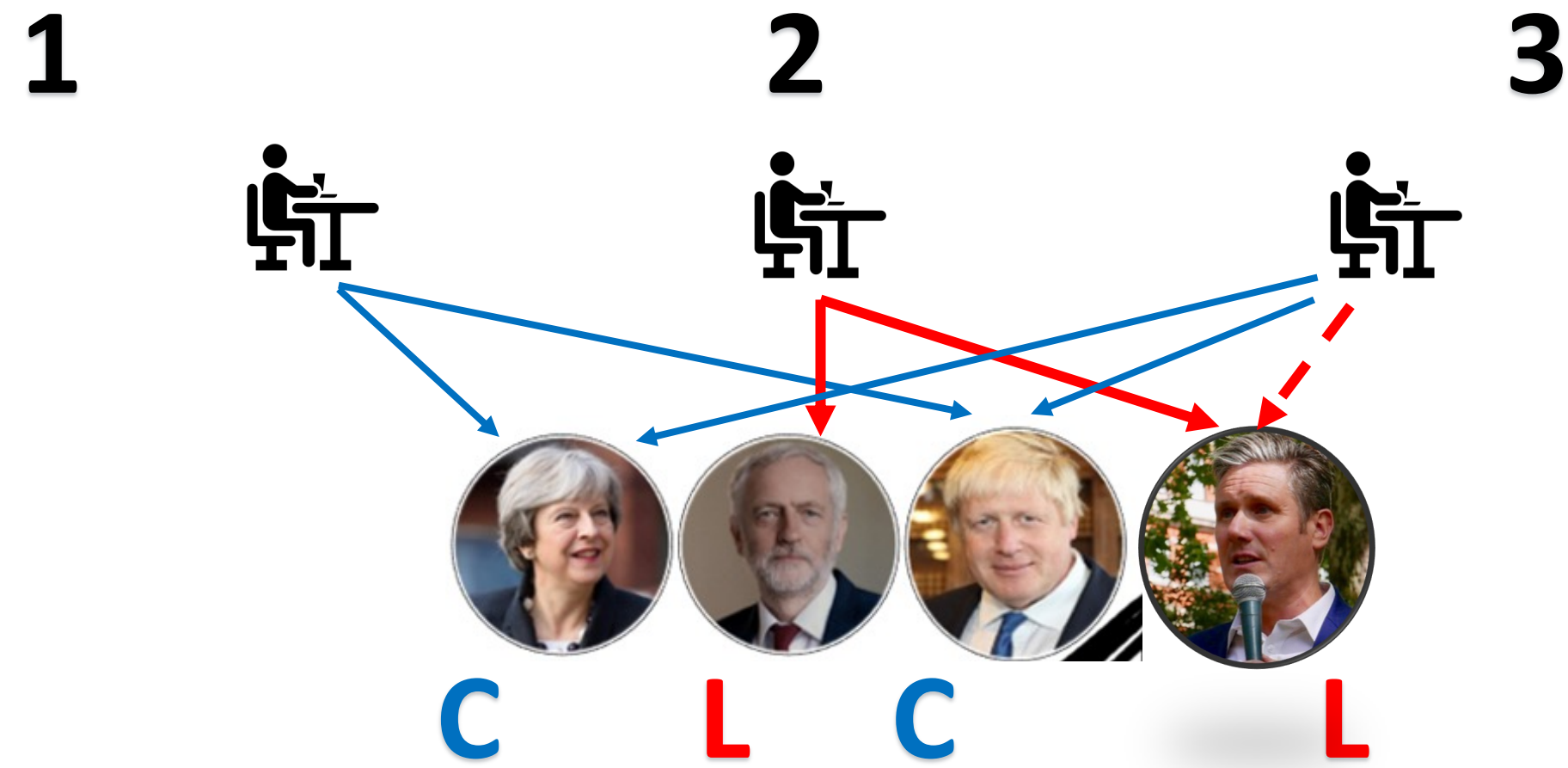
Digital Citizen Engagement

Theresa May

Jeremy Corbyn

- 553 of 650 MPs on Twitter
- 13 M Followers (4.3M unique)

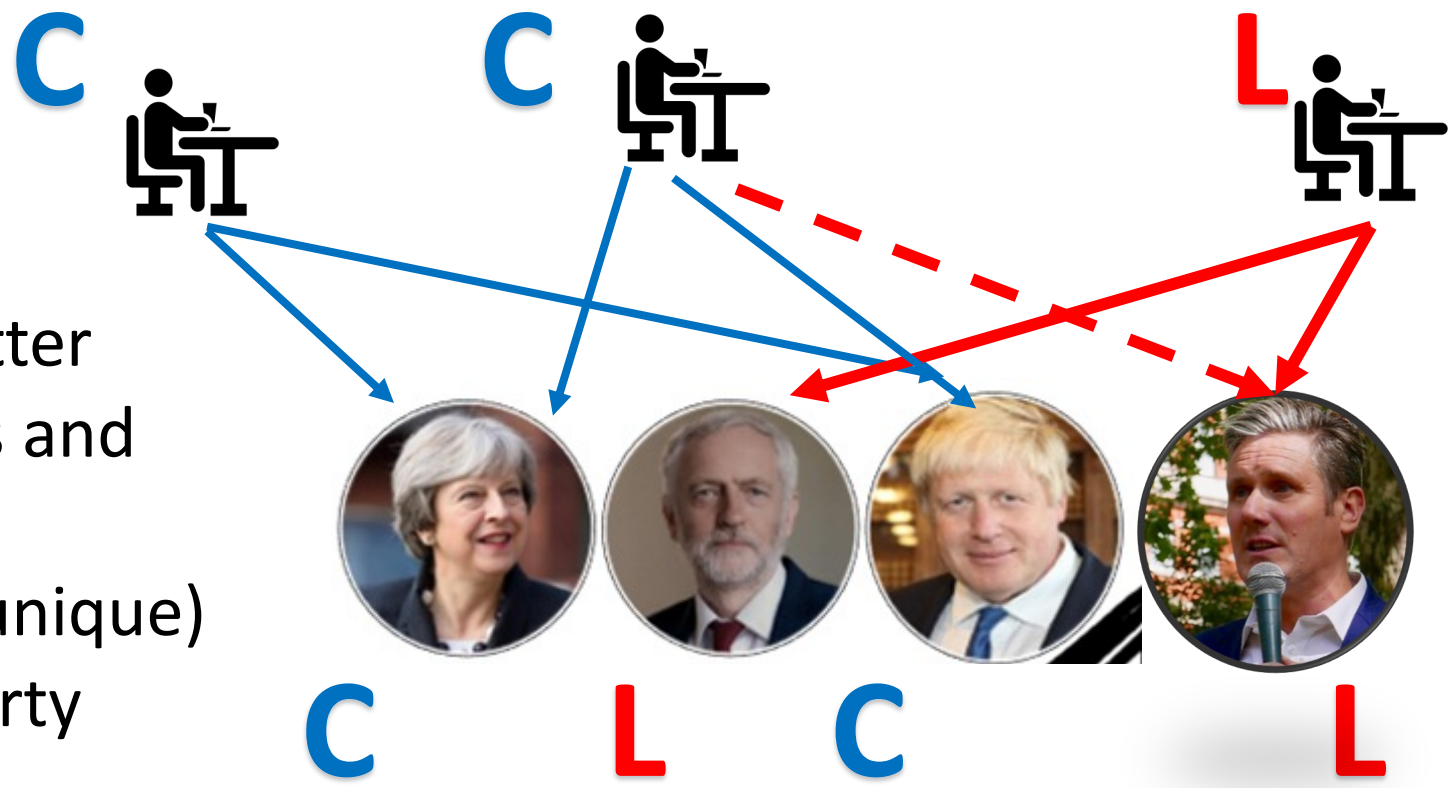
Assigning party affiliation to users



Dataset* details

Shared publicly: get in touch if interested

- ❑553 of 650 MPs on Twitter
- ❑Fetch all tweets by MPs and tweets towards MPs
- ❑13 M Followers (4.3M unique)
- ❑78% follow only one party



Significant Cross-Party Talk in UK!

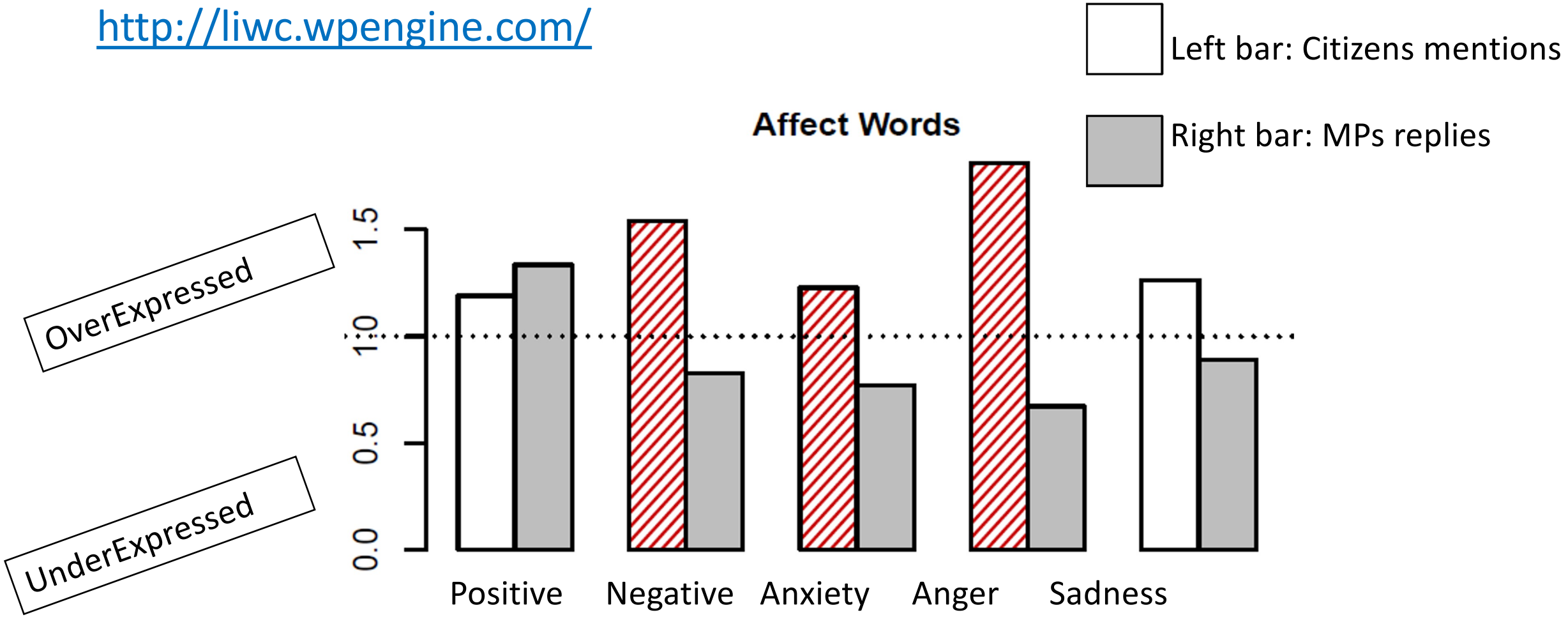
UNIVERSITY OF SURREY DANS

MP / User	Con.	Lab.	SNP	Lib. Dem.	DUP
Con.	70	26	1	2	0
Lab.	42	54	2	2	0
SNP	35	20	43	1	0
Lib. Dem.	45	25	1	28	0
DUP	26	15	2	1	56

[P. Agarwal et al. ICWSM'19]

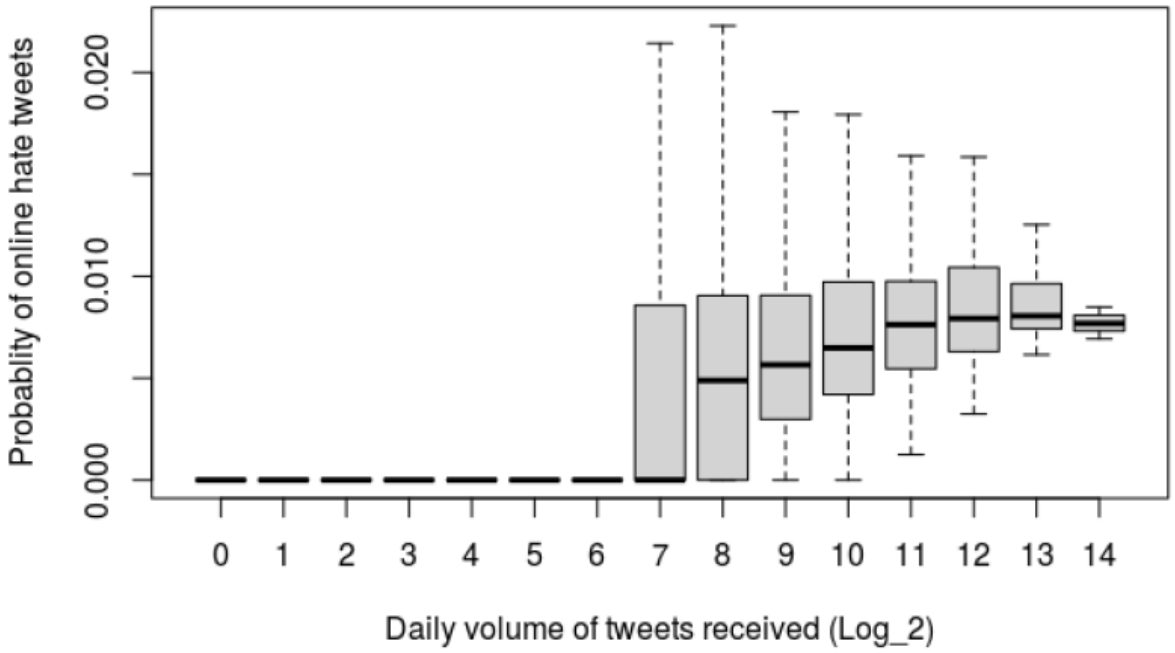
Is the tone civil?

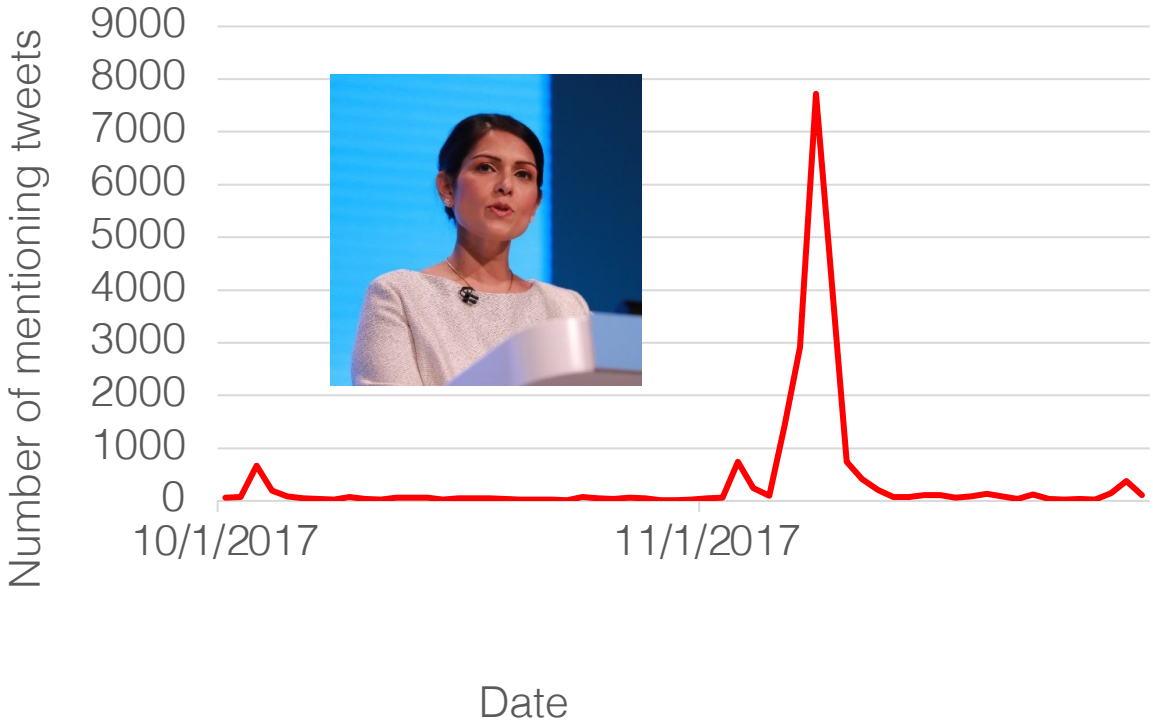
http://liwc.wpengine.com/



[P. Agarwal et al. ICWSM'19]

Hate Speech in Political Discourse





‘Pile-on’ hate

[P. Agarwal et al. ACM HT’21]

Is there hate here?

Got a few girls sweet talking since the past few days hoping to get me to take them out for valentine's day. Lmfao. A day designed to cuck more simps.

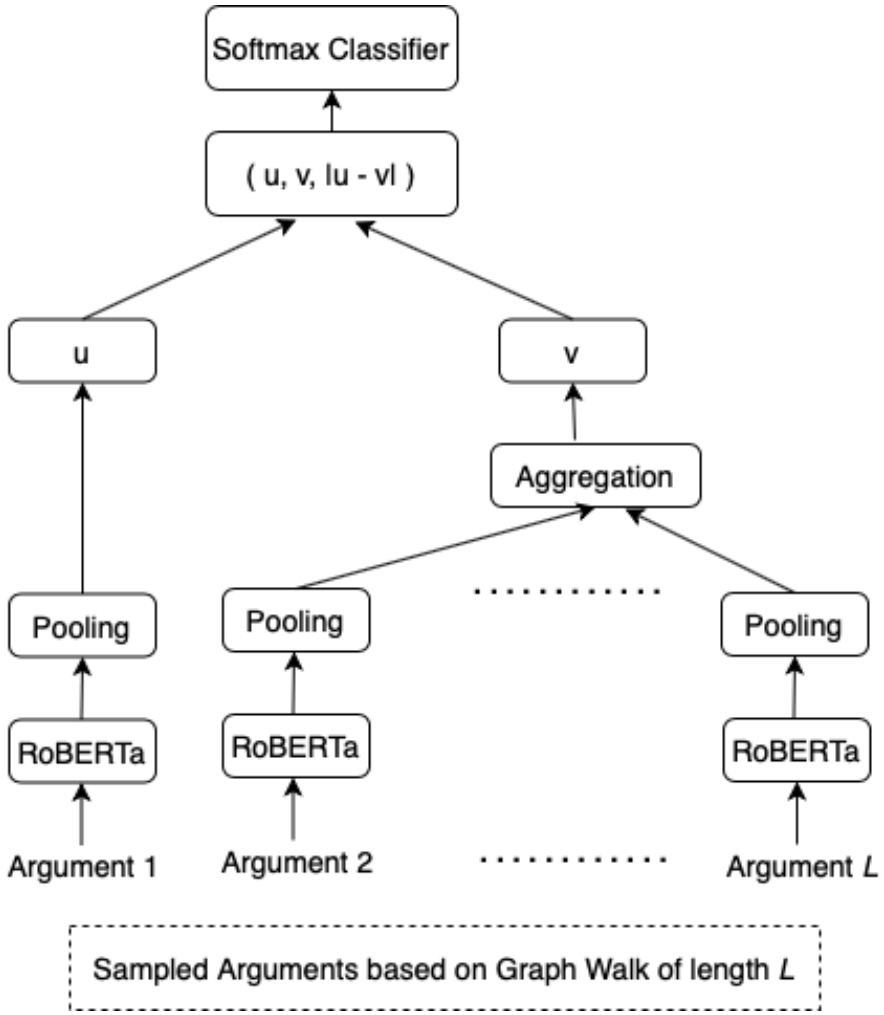
There aren't many things that are more satisfying than telling a girl, "No".

A lot of women do this to their simp boyfriends and its sad, hope when I'm older (as in 14-16) I don't fall into this trap.

It's funny to see the hamster that starts to act up in their little widdle tiny brains after saying that too.

[E. Guest et al. EACL'21]

GraphNLI: Graph Neural Network Architecture for understanding online conversations



Problem: Online conversations unfold over multiple posts and threads. Hate or polarity can depend on far away context.

Idea: Use random walks to capture surrounding context, not just post + reply

Works better than SoT baselines (for polarity – WWW'22, and for hate speech – TWEB'23)

*dataset and model shared publicly: get in touch if interested

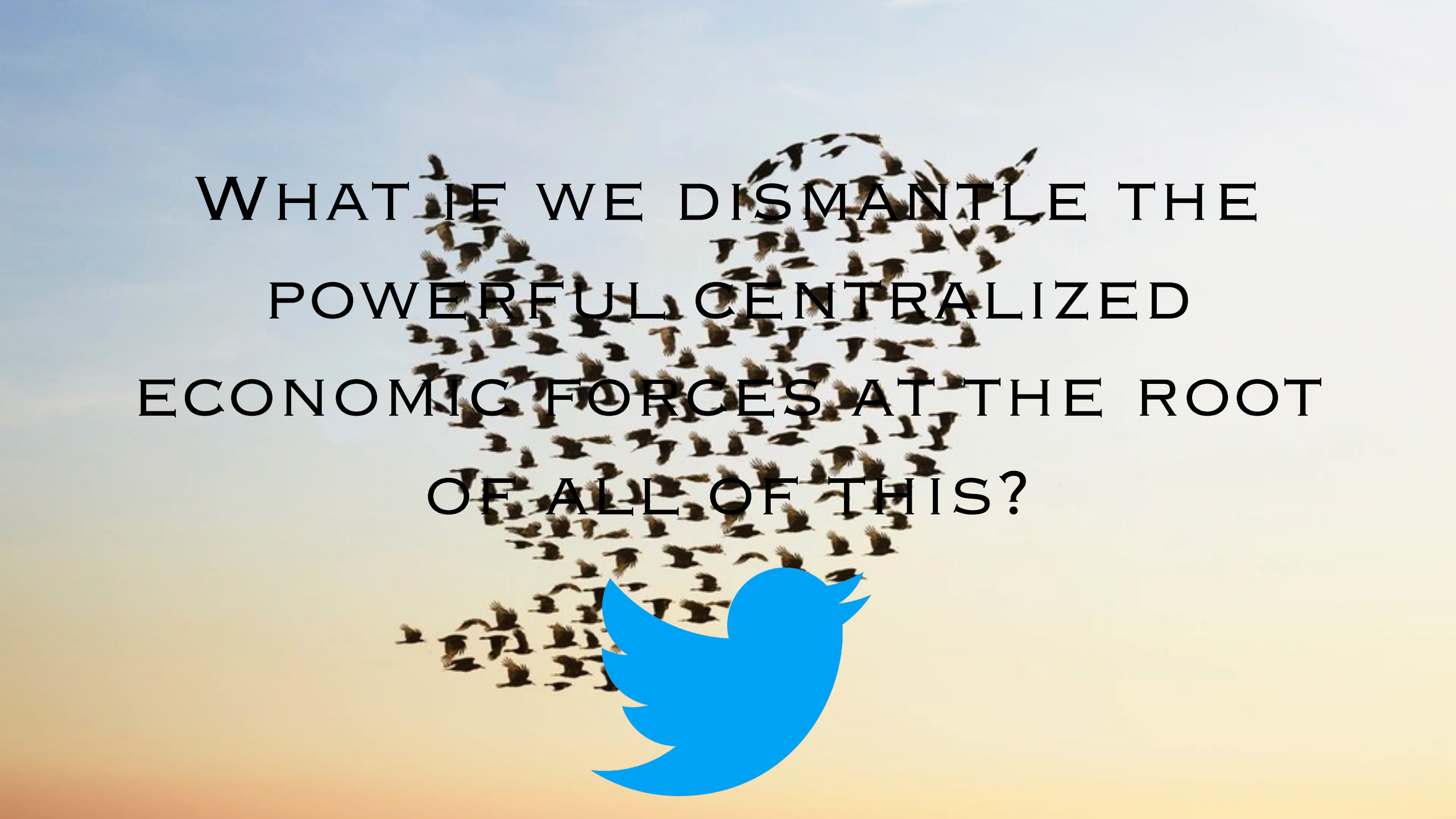
[V. Agarwal et al. WWW'23; TWEB'23]

3. Can we cure the web by Redecentralizing it?

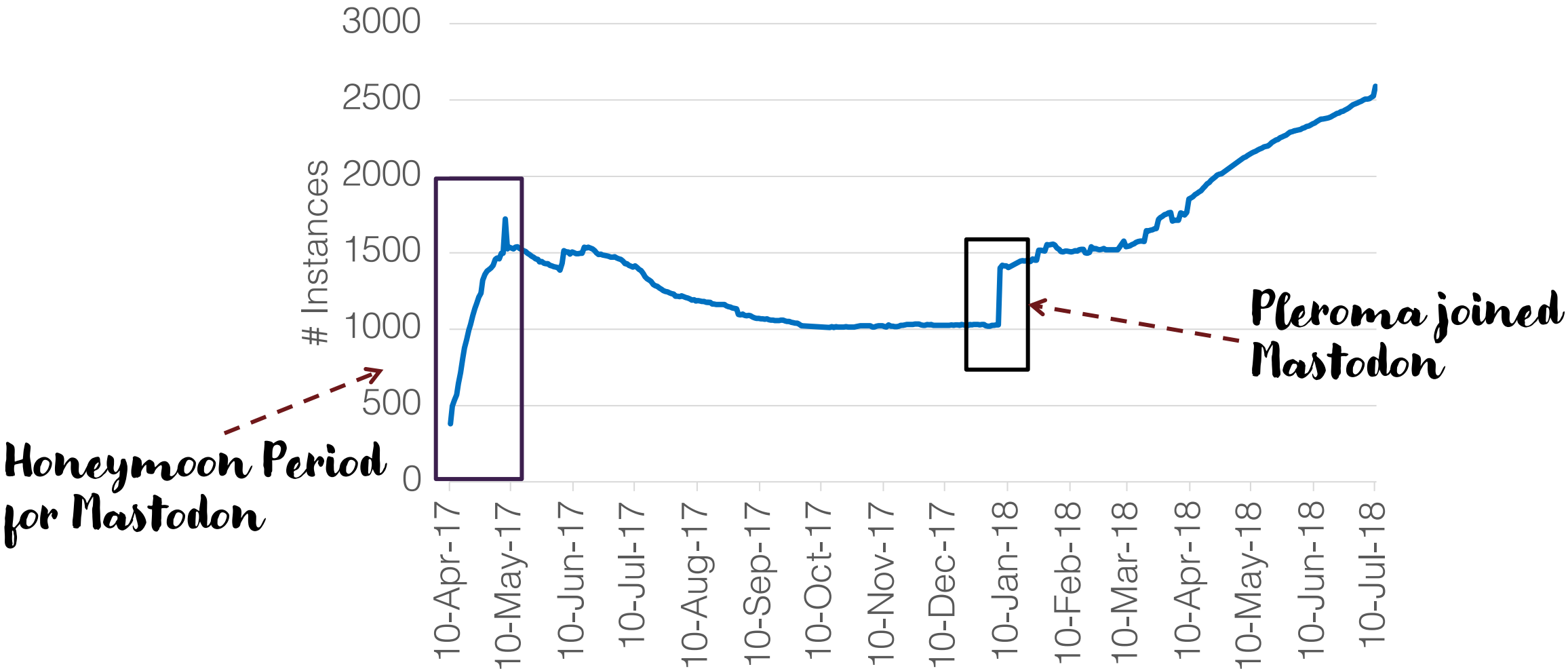
Challenges in the Decentralised Web [IMC'19]

Exploring Content Moderation in the Decentralised Web [Conext'21]

Toxicity in the Decentralized Web and the Potential for Model Sharing [Sigmetrics'22]

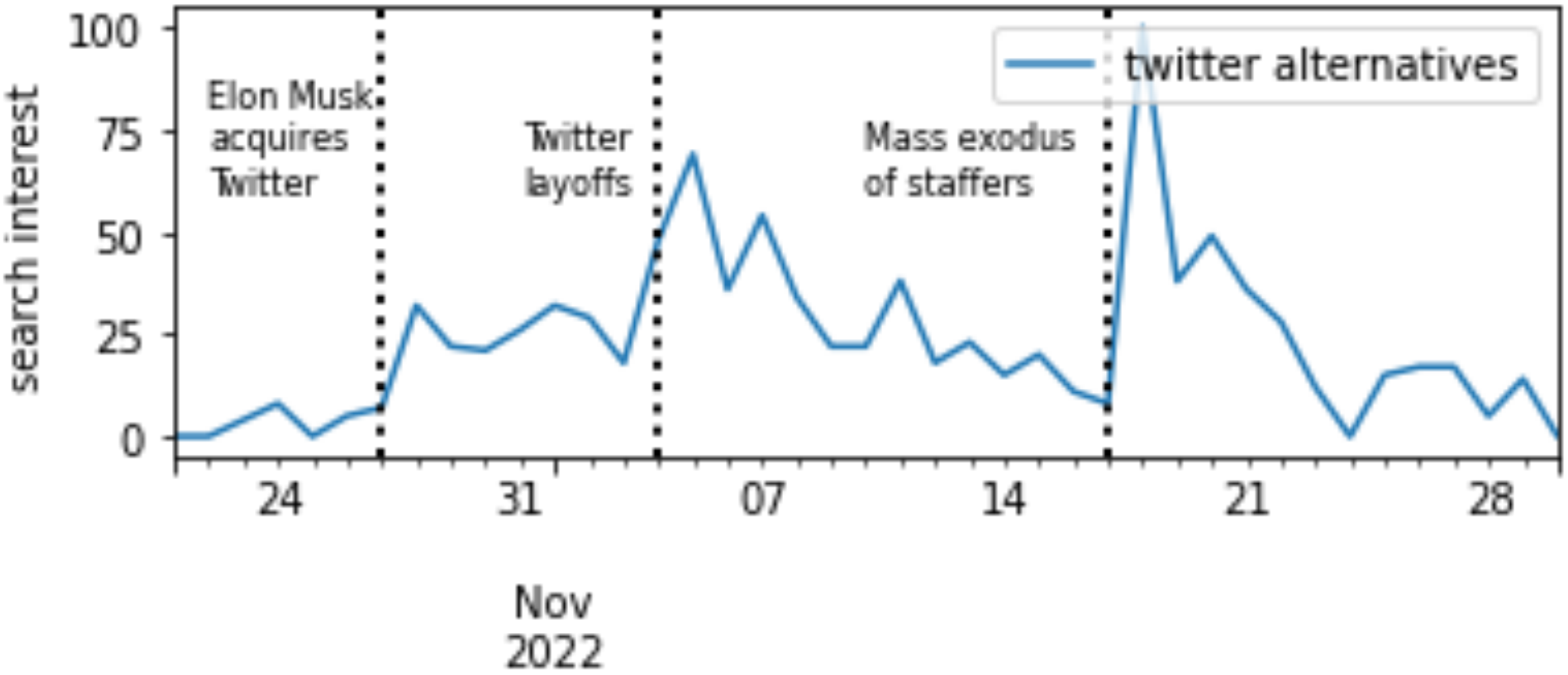

What if we dismantle the powerful centralized economic forces at the root of all of this?

Mastodon has become popular!

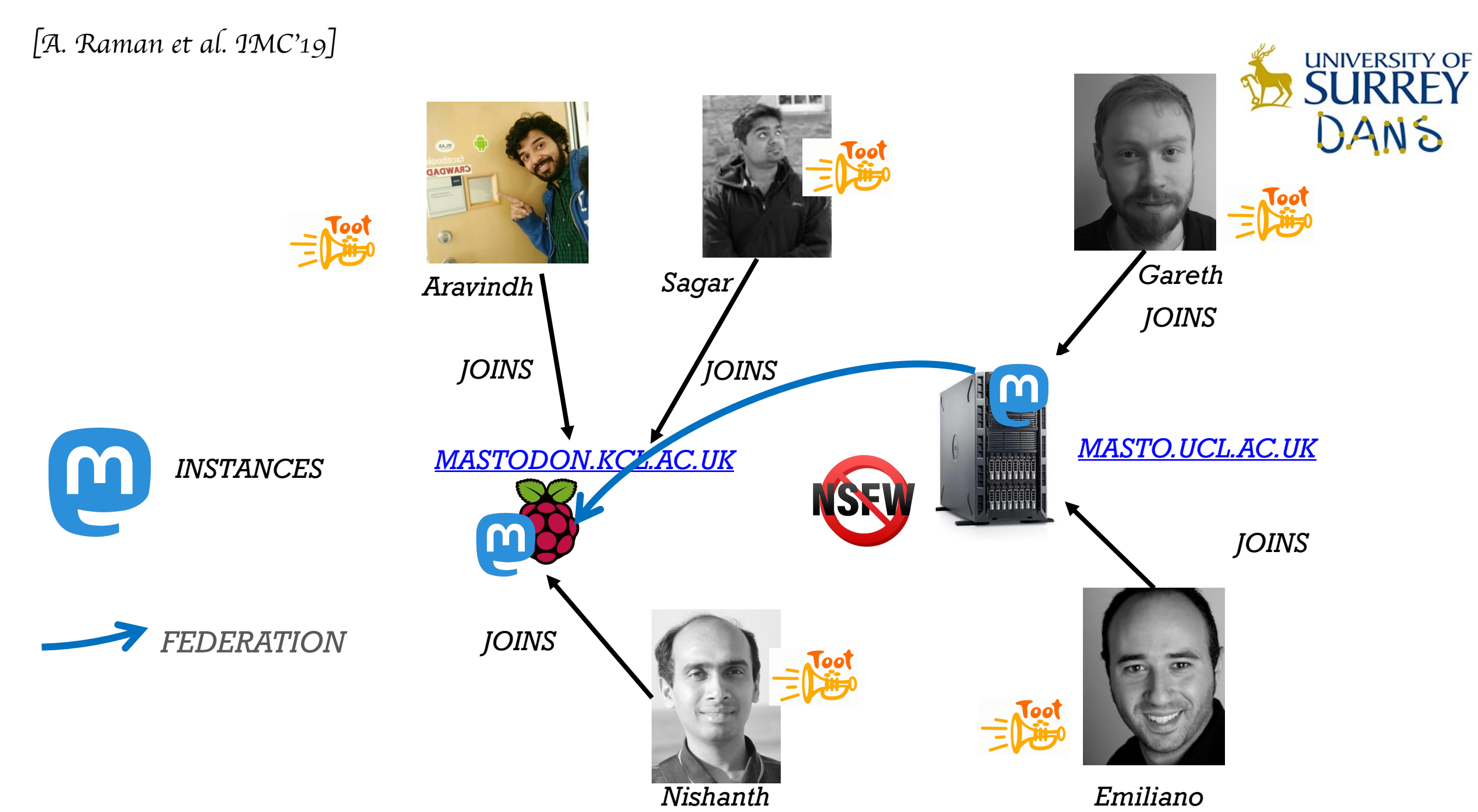


[A. Raman et al. IMC'19]

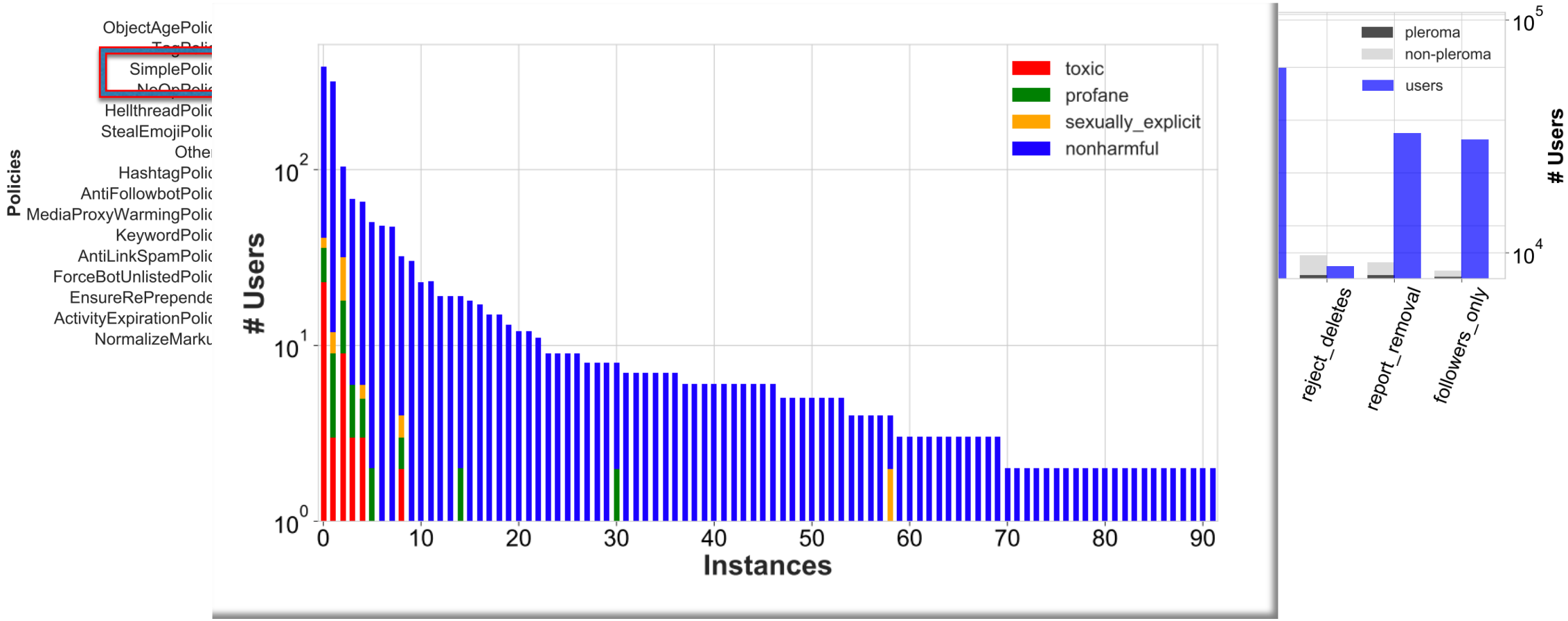
Leaving Twitter is now cool!



[A. Raman et al. IMC'19]



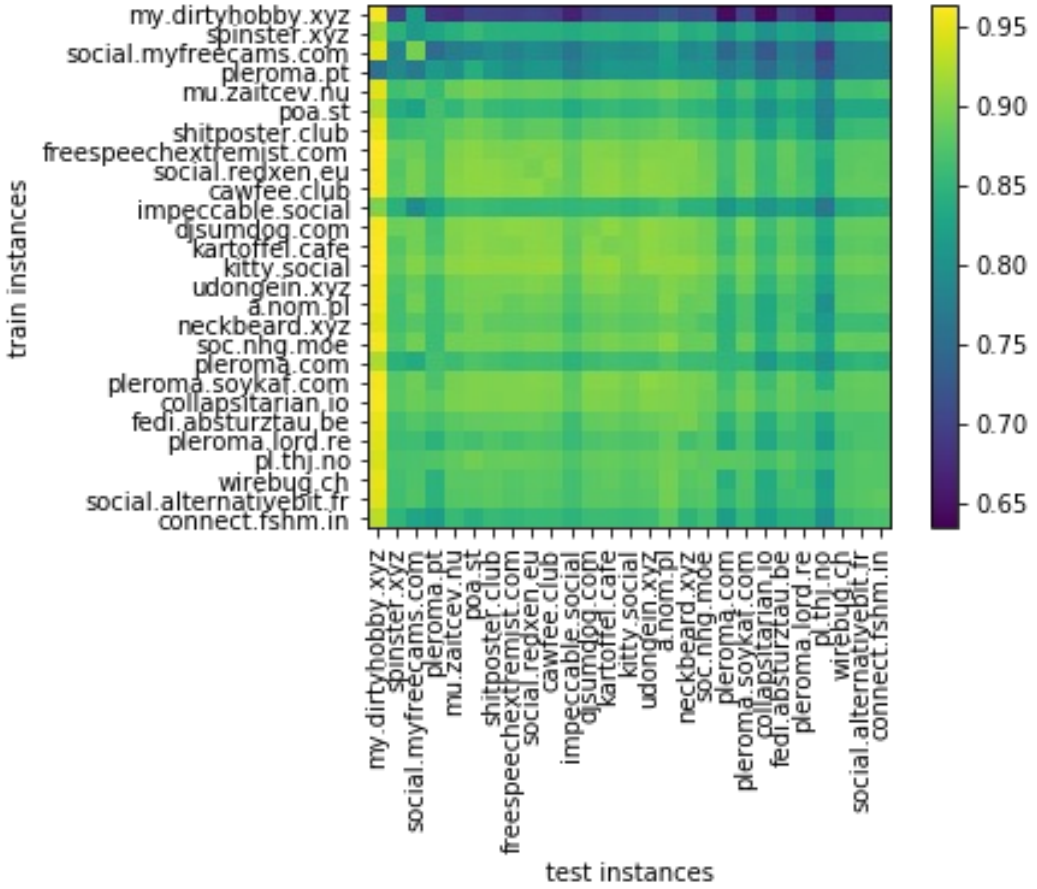
Moderation is difficult



[A. Hassan et al. CoNext'21]

MobPair: Model Sharing

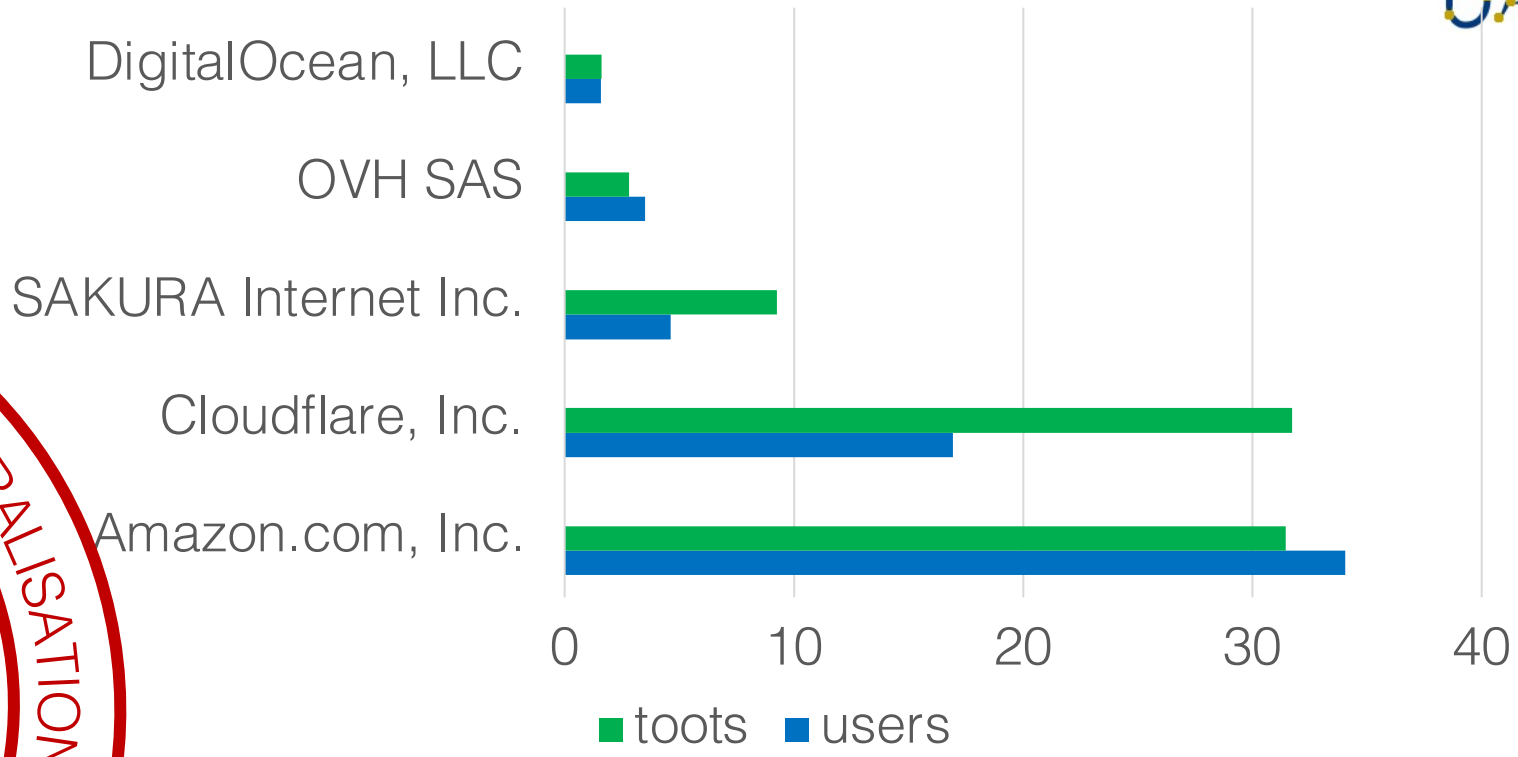
Allow better-resourced instances (in terms of annotations) to share their models with other instances



Model sharing can work (if instances are matched well)

[H. Bin Zia et al. Sigmetrics'22]

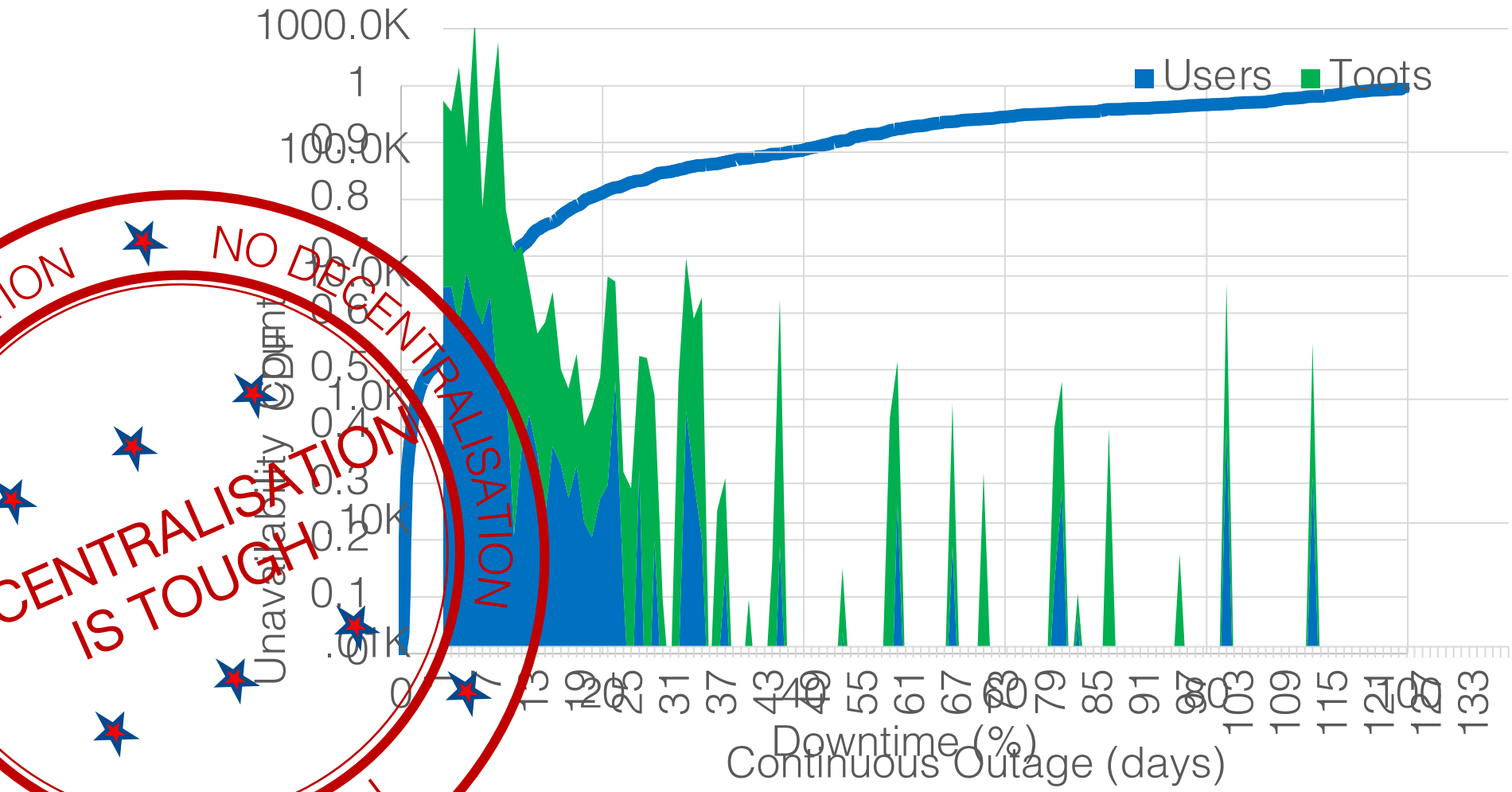
Where are instances hosted?



- ½ the users reliant on just 3 Autonomous Systems
- 85% of toots come from two countries (JP and US)

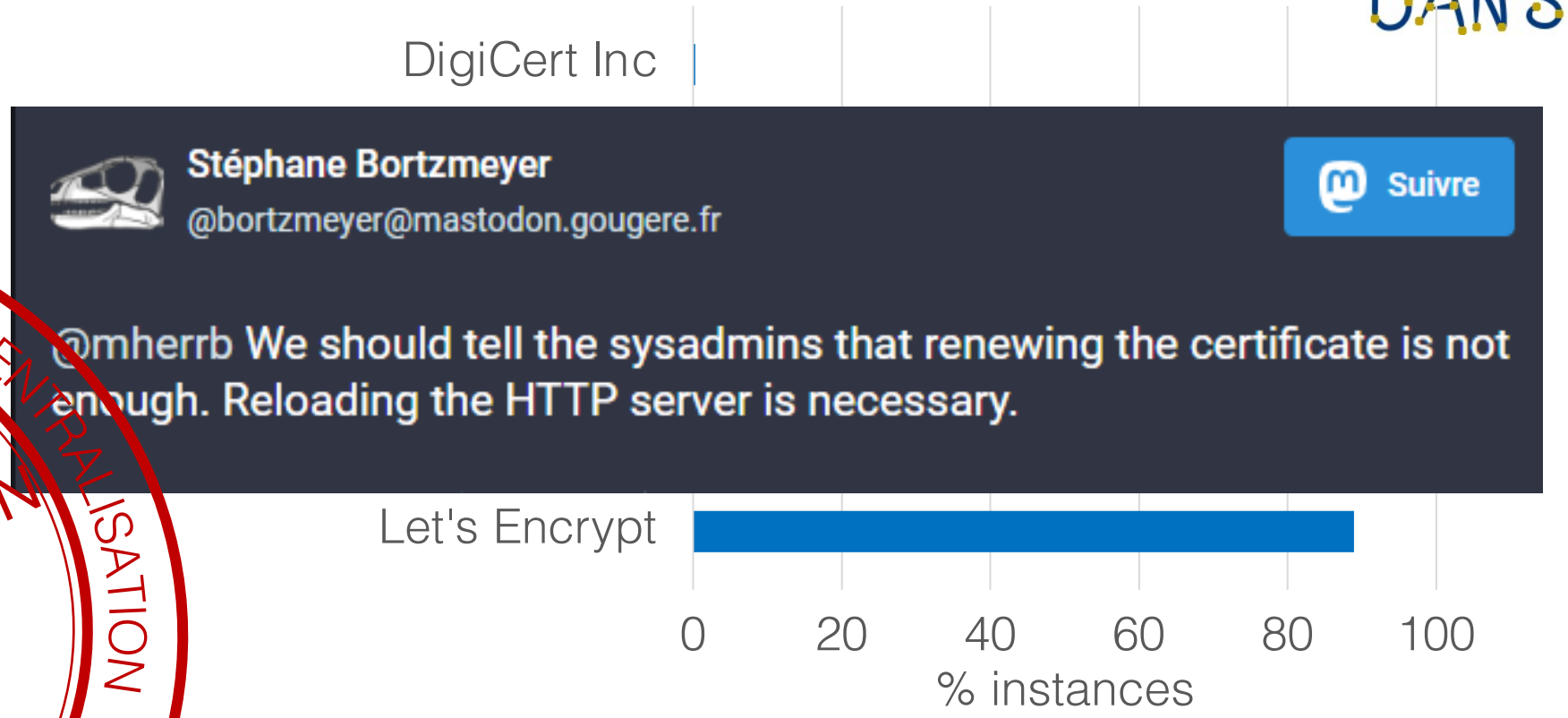
[A. Raman et al. IMC'19]

Reliability and availability



- 11% instances inaccessible for half of the time

[A. Raman et al. IMC'19]



- 85% of instances use **one** Certificate Authority

[A. Raman et al. IMC'19]

Summary

1. Partisanship (and tracking) across the world
2. Online hate and polarization (UK Politics test case)
3. Decentralisation is promising, but beware pitfalls!

What Next?

- The AP4L Project: Embedding privacy & online safety *into* people's lives during life transitions

House of Commons

The Alan Turing Institute

UKRI Engineering and Physical Sciences Research Council