UNIVERSITÀ DELLA CALABRIA

AI & Data Science Lab
@DIMES Dept.
MACHINES, LANGUAGES & NETWORKS team

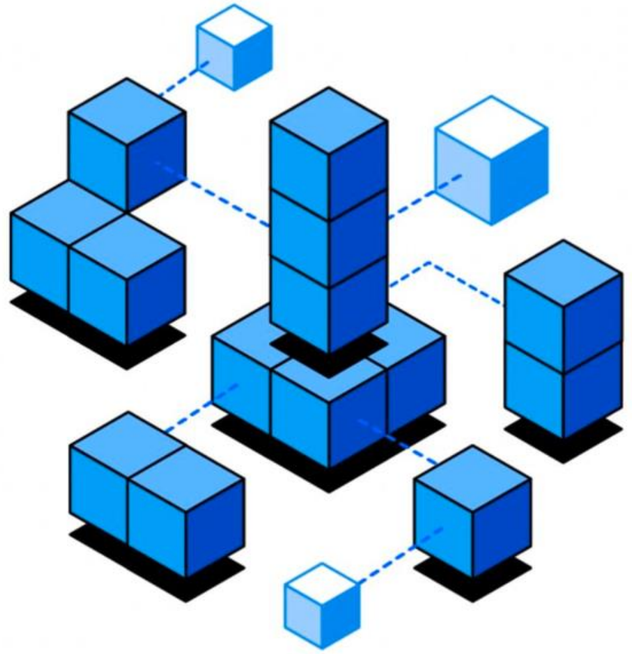
*SAI4OID workshop@WI-IAT 2025,*
*17 November 2025, London (UK)*

# Polarized Communities in Mastodon: Insights from Instance-Level Analysis

**Lucio La Cava, Domenico Mandaglio, Andrea Tagarelli**

DIMES Dept., University of Calabria, Italy

# Decentralized Online Social Networks



*Source: blueskyweb.xyz*

## User-Centric

Foster spontaneous and unbiased interactions, advertisement free

## Decentralized Growth

Independent yet cooperating servers to escape from individual owners

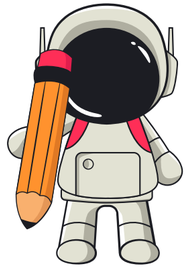
# How is decentralization achieved?

## Open Source Software

Allows anyone to create a new server, or instance, thus favoring the emergence of communities guided by spontaneous interest towards certain topics

## Communication Protocols

Enable seamless communication between (users registered on) different instances, even if pertaining to heterogeneous services



Development of the **Fediverse**
the federated universe of decentralized instances

# Here comes Mastodon

- Decentralized alternative to Twitter
- Niche communities and content moderation (cf. Reddit)
- Content policies/rules
- Fine-grained instance controls

# Interactions within DOSNs instances

## Positive Interactions

**Followship** relations among people across different servers reflecting on interactions between servers

## Negative Interactions

Enforcement of **moderation** policies toward servers (e.g., bans, suspensions, etc)

## What are the effects of such interactions?

- Emergence of **polarized** and **conflicting groups**
  - **Intra-group** prevalence of **positive** interactions
  - **Inter-group** prevalence of **negative** interactions
  - High **intra-group density**

# Unveiling Polarization in DOSNs instances

**(RQ1)** *How many polarized groups can be found in Mastodon?*

**(RQ2)** *What is the polarization structure in Mastodon, that is, how are polarized groups linked internally and to each other?*

**(RQ3)** *What are the main characteristics of the instances within the detected polarized groups?*

# Data Crawling

## Detecting positive links among instances

- Seed set of instances from instances.social
- Seed set of **270K Mastodon users**
- *Breadth-first search* to incrementally expand known users
- Identification of **incoming** and **outgoing links**
  - `/api/v1/accounts/:id/followers`
  - `/api/v1/accounts/:id /following`
- **9+ months** of crawling
- **2M users** and **116M unique links** among them

# Data Crawling

## Detecting negative links among instances

- List of all tracked instances from instances.social
- Crawling of **moderation rules** established from each instance
  - `/api/v1/instance/domain_blocks`
  - **DomainBlocks** JSON objects containing **blocked instances** and **associated metadata** (e.g., the severity and motivation of the block).
- Crawling between July and November 2023
- More than **135K raw enforced blocks** among instances

# Network Modeling

## Our directed positive instances network

- Nodes represent instances
- Edges represent links between instances deriving from those among users
- Edge weights code the multiplicity of interactions between instances

## Our directed negative instances network

- Nodes represent instances
- Edges represent moderation enforced from the source instances to the target ones

$\mathcal{G}^+ = \langle V^+, E^+, w \rangle$ **contains 37,529 nodes and 1,335,490 edges**

$\mathcal{G}^- = \langle V^-, E^- \rangle$ **contains 11,401 nodes and 105,465 edges**

# Network Modeling

## Simplifying our positive network

- Prune **noisy** or **statistically irrelevant edges** due to spurious interactions
- ***Disparity Filter*** method* to prune edges w.r.t. significance thresholds
  - **117,422** remaining **edges** with $\alpha = 0.05$
- Not needed for the negative network, as blocks among instances are **explicitly declared** by instances' administrators and not due to randomness

## Creating our signed instance-network

$$\mathcal{G} = \langle V, E, s \rangle \qquad V \subseteq \mathcal{I},\ E = E^+ \cup E^- \qquad s : E \mapsto \{+1, -1\}$$

$$s(i,j) = +1 \text{ if } (i,j) \in E^+,\ -1 \text{ otherwise}$$

**Our resulting network contains 19,738 nodes and 222,887 pos/neg edges**

*M. Ángeles Serrano, Marián Boguñá, and Alessandro Vespignani. 2009. Extracting the multiscale backbone of complex weighted networks. Proceedings of the National Academy of Sciences 106, 16 (2009), 6483–6488.

# Detecting Polarized Groups

- **Problem:** Given a signed graph G and an integer k, find k mutually-disjoint node sets $\mathcal{P} = \{P_1^*, \cdots, P_k^*\}$ such that:

$$P_1^*, \cdots, P_k^* = \underset{P_1, \cdots, P_k \subseteq V}{\arg\max} \frac{f(P_1, \cdots, P_k)}{\left| \cup_{i=1}^{k} P_i \right|}$$

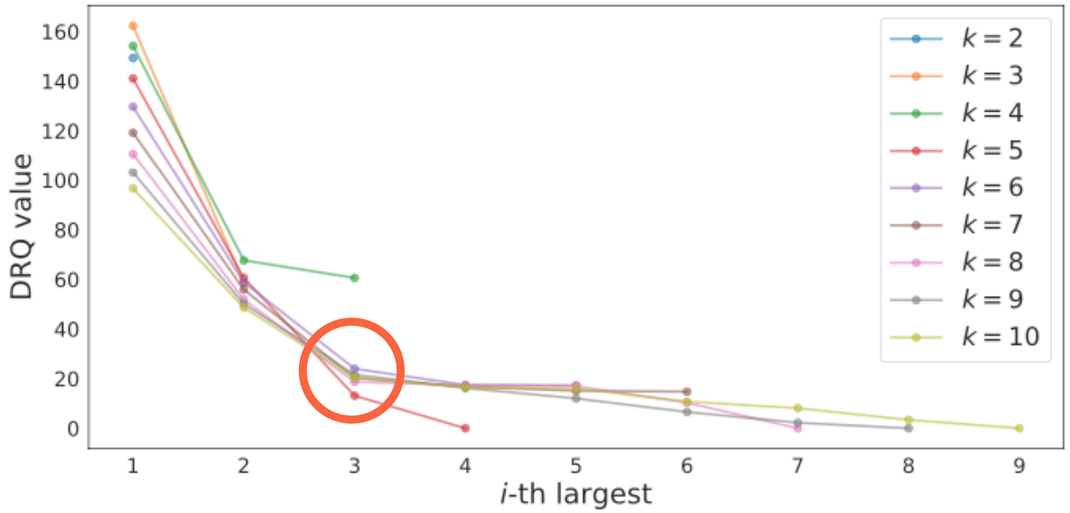
$$f(P_1, \cdots, P_k) = \sum_{P_i \in \mathcal{P}} \left( |E^+(P_i)| - |E^-(P_i)| \right) + \frac{1}{k-1} \sum_{P_i, P_j \in \mathcal{P}} \left( |E^-(P_i, P_j)| - |E^+(P_i, P_j)| \right).$$

- **Spectral Conflicting Groups (SCG)** algorithm
  - Only method admitting *neutral nodes*
  - For each group, it solves **Discrete Rayleigh Quotient (DRQ) problem**
  - The solution to the $i$-th DRQ problem characterizes the group $Pi$ that conflicts the most with the remaining groups $Pj$ , for $j > i$
  - **DRQ value** representing the **intensity** of such a conflict

*Ruo-Chun Tzeng, Bruno Ordozgoiti, and Aristides Gionis. 2020. Discovering conflicting groups in signed networks. Proc. NIPS Conf. 33 (2020), 10974–10985.

# Determining the number of polarized groups

- **Elbow**-like approach
  - Run of SCG with different k values
  - Plotting DRQ values in ascending order, i.e., the $i$-th largest DRQ value is at the $i$-th position
  - Determining $k$ to be one of the discernible “knees” on the resulting curve
- $k$ = 4 **conflicting groups** with an empty group




**3 polarized** *groups* + **1 neutral** *group*

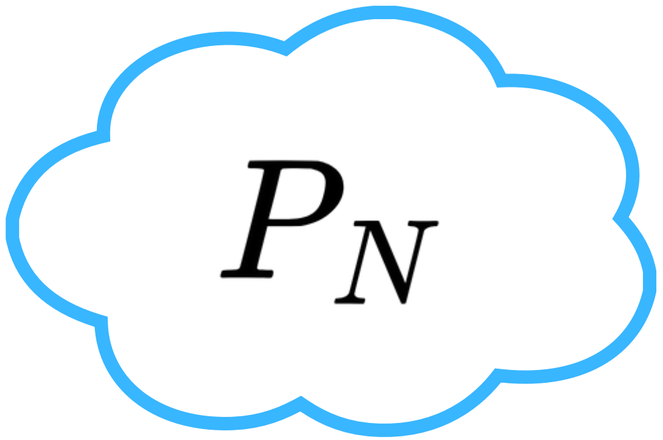
# Characterizing polarized groups

- 97% instances in the **neutral group PN** matching the idea of Fediverse
- P1 and P3 are **Mastodon-pure** polarized groups
- **Non-Mastodon** instances dominating the neutral group and P2

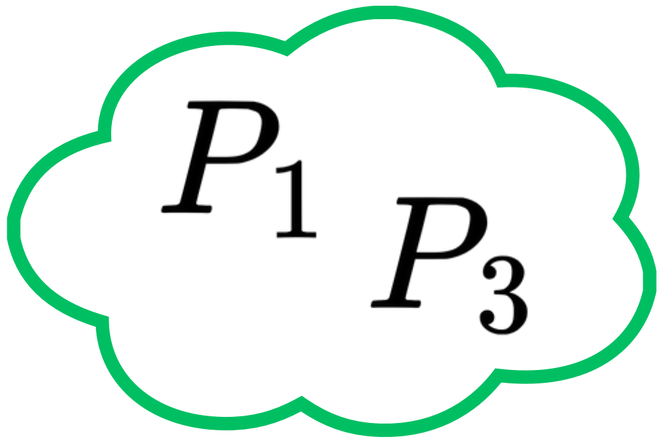
| | $P_N$ | $P_1$ | $P_2$ | $P_3$ |
|---|---|---|---|---|
| # Instances | 19,241 | 189 | 122 | 186 |
| % Mastodon | 43.6 | 92.6 | 36.1 | 91.4 |
| # Incoming bans | 79,690 | 728 | 24,651 | 396 |
| Avg. # bans | 12.94 | 7.35 | 202.06 | 7.62 |
| % Instances $\geq$ 1 ban | 32.0 | 52.4 | 100 | 28.0 |

- **PN** and **P2** are the **most banned** within the Fediverse
- **P2** exhibits **higher** (more than 16x) **avg #bans** with **100% instances** receiving at least **one ban**, thus becoming the **“ban-sink”** pole of the Fediverse

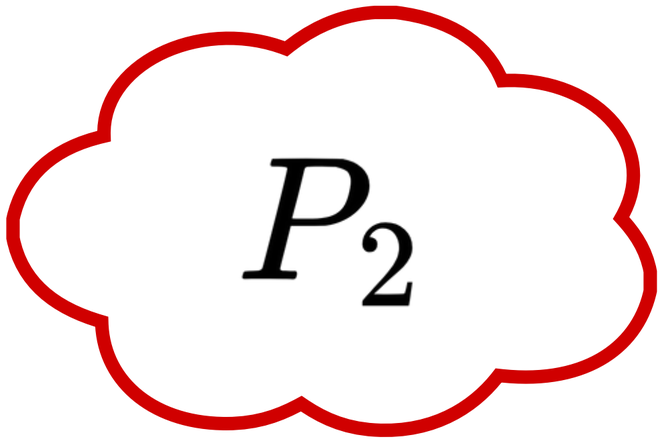
# Characterizing polarized groups




Neutral Group

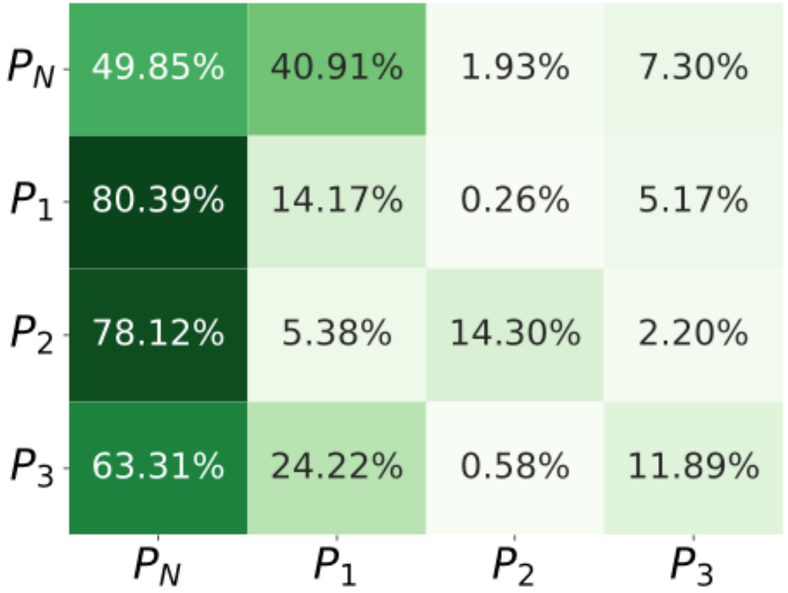



Mastodon-pure Group

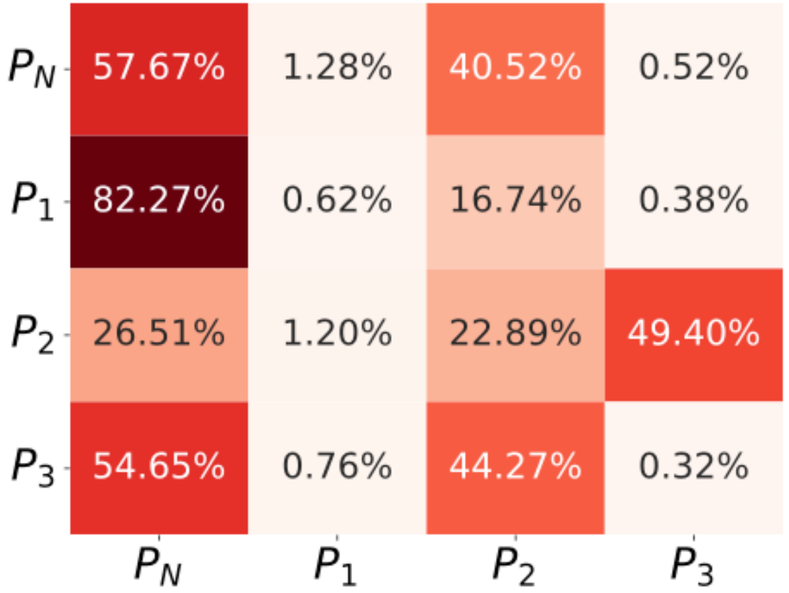



Ban-sink Group

# Relations between polarized groups




- Most interactions involve **PN**
- Ban-sink **P2** receiving only interactions from itself, **is it segregation?**




- **Bipartite** banning involving **PN** and **P2**
- Further hints at a **P2 segregation**
- Anomalous bannings from P2 to P3 deserving more attention

# Main instances in polarized groups

- ***mstdn.jp*** is among the oldest Mastodon instances and the second-largest Japanese one
- ***mastodon.social*** is the official instance of the Mastodon project
- ***botsin.space*** is the reference instance for running bots on Mastodon

| | Most interacted | Most banned |
|---|---|---|
| $P_N$ | mstdn.jp | geofront.rocks |
| $P_1$ | mastodon.social | botsin.space |
| $P_2$ | pawoo.net | poa.st |
| $P_3$ | det.social | aethy.com |

- ***pawoo.net*** is the second-largest Mastodon instance in terms of users, recently under the spotlight due to the hosting of controversial content
- ***poa.st*** is a *non-Mastodon* instance advertising itself as the "*Fediverse for shitposters*" *

*Source: https://globalextremism.org/post/poast/

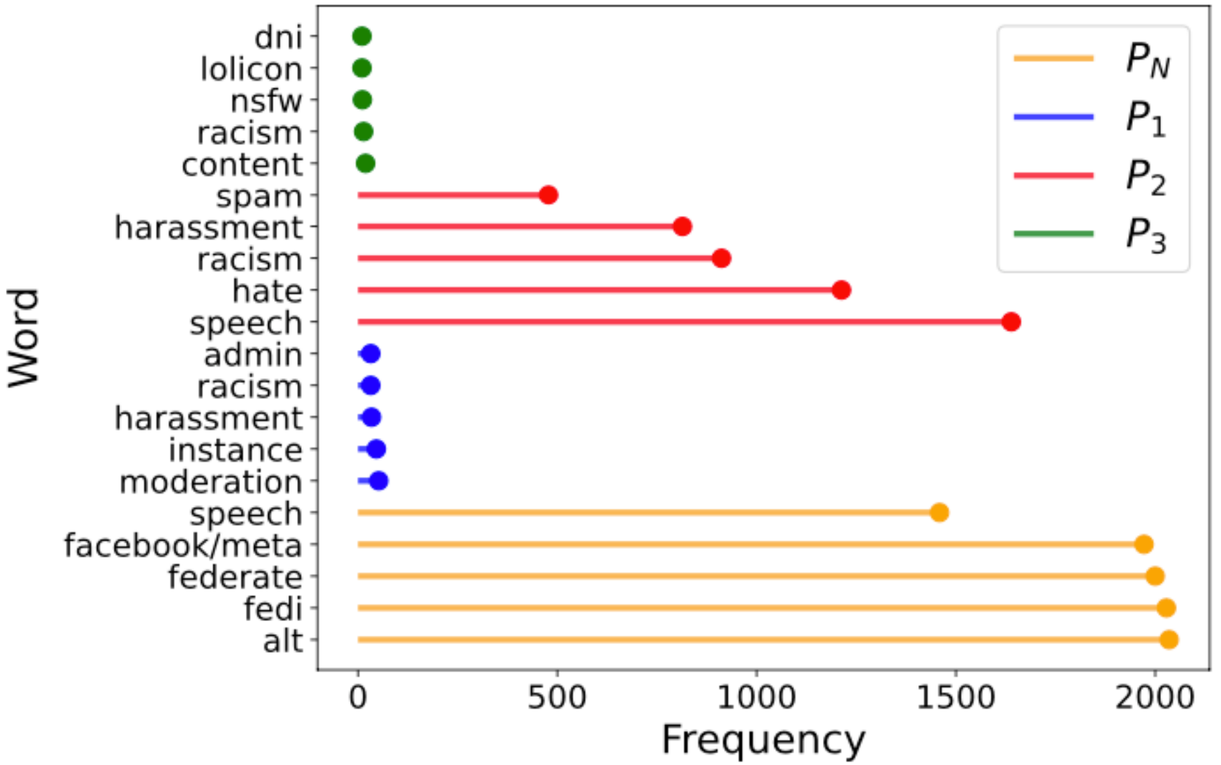
# Activity in polarized groups

| | Volume | Avg | Top-active Instance | % Volume |
|---|---|---|---|---|
| $P_N$ | $2.37 \times 10^7$ | 3,654 | `mstdn.jp` | 7.65% |
| $P_1$ | $2.36 \times 10^7$ | 139,511 | `mastodon.social` | 32.84% |
| $P_2$ | $1.74 \times 10^6$ | 158,065 | `pawoo.net` | 54.55% |
| $P_3$ | $2.38 \times 10^6$ | 14,356 | `mstdn.ca` | 15.19% |

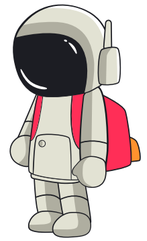
- Collecting, for each pole, the **number of statuses** created in the last **12 weeks**
  - `/api/v1/instance/activity`
- All groups producing **large** volumes of data, especially **PN** and **P1**
- Tail of **small** instances in PN (small avg number of posts)
- **P1** emerging as the **“beating core”** of the Fediverse
  - High avg number of posts and % volume
- **Anomalous volume** in **P2**, considering the concerns about *pawoo.net*

# Banning reasons in polarized groups

- No particularly evident motivations in P1 and P3 bannings
- Banning motivations for P2 hint at the ***negativity*** of the group
- PN witnesses bannings due to the moderation of instances that **federate** with **unwelcome** ones (e.g., **Threads**)



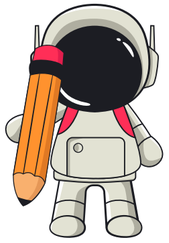
# Take Home Messages



**(RQ1)** *The Mastodon-centric Fediverse instance network encompasses four non-overlapping groups of instances identified as poles*

**(RQ2)** *There is a unique polarization structure with a predominant neutral group, the remainder includes Mastodon-pure groups and a “ban-sink” one, which receives negative links as a protective measure*

**(RQ3)** *The ban-sink group exhibits anomalous trends in content production, receiving strong moderation due to harmful and inappropriate content*



**Future Work:** *Characterizing user-level polarization in DOSNs and exploring alternative strategies for building the signed network.*

# Thanks for your attention!

Questions?