Polarized Communities in Mastodon: Insights from Instance-Level Analysis

(Discussion Paper)

Lucio La Cava
DIMES Dept., University of Calabria
Rende, Italy
lucio.lacava@dimes.unical.it

Domenico Mandaglio
DIMES Dept., University of Calabria
Rende, Italy
d.mandaglio@dimes.unical.it

Andrea Tagarelli
DIMES Dept., University of Calabria
Rende, Italy
andrea.tagarelli@unical.it

***Abstract*—Centralized social media platforms are undergoing a period of change, which has fueled growing interest in alternative models such as Decentralized Online Social Networks (DOSNs). Their increasing adoption is largely driven by open-source technologies that allow users to set up independent servers, or instances, and connect them within the broader decentralized ecosystem known as the Fediverse. Although DOSNs have attracted considerable attention, limited research has examined how positive and negative ties among instances shape their structure and dynamics. In this paper, we discuss a recent work that took a first step in this direction by analyzing polarization phenomena at the instance level, with a focus on Mastodon — the most prominent decentralized platform, currently counting more than 10 million users and nearly 20K instances. Our findings indicate that polarization in the Fediverse displays distinctive characteristics, simultaneously supporting collaboration across instances while also enabling the exclusion of servers perceived as threats to the community.**

***Index Terms*—Fediverse, Polarization, Mastodon, Signed Network**

I. Introduction

Centralized social media platforms are undergoing a profound transformation, leading many users to explore alternative modes of online interaction. In this context, Decentralized Online Social Networks (DOSNs) [1] are attracting growing attention [1], [2]. The appeal of DOSNs lies in their open-source foundations, which empower individuals to set up their own *instances* (i.e., servers) and interconnect with others. This results in a decentralized ecosystem, known as the *Fediverse*, which operates in a way that resembles email networks.

Among the different decentralized platforms introduced over the past years, Mastodon has become the most widely studied case [3], [4]. With more than 10 million users and nearly 20K instances, Mastodon stands out as the reference model for investigating the dynamics of the Fediverse.

The growth of the platform has fostered interactions on two levels: user-to-user relations across instances and instance-to-instance relations emerging from them. Most studies have examined the former, showing that Mastodon's follower–followee distribution is more balanced than Twitter's [3], with fewer bots and weaker disassortative trends. Others explored how decentralization shapes user behavior, such as the emergence of strategic roles [5] or the balance between information production and consumption [6]. In contrast, research on the latter—relations among instances—remains limited, with only initial attempts to map and analyze their structure [7]. An overlooked dimension concerns the semantics of relationships. Links can be *positive* (support, agreement, affinity) or *negative* (conflict, opposition, distrust), and their coexistence relates to *polarization*, where users split into groups with opposing views on issues such as politics, religion, or sports. While polarization in social media has been widely studied [8], [9], DOSNs introduce a distinctive setting where it may occur both at the user and instance levels. Instances can form positive ties through collaboration or shared interests, and negative ones through blocking or moderation. Hence, in Mastodon, polarization extends beyond individual behavior to the ecosystem itself. Yet, polarization in DOSNs remains understudied, especially regarding the interplay between user- and instance-level dynamics. Understanding how it emerges in Mastodon is key to explaining community formation in decentralized environments and to developing strategies for healthier, less divisive online ecosystems.

Contributions. In this paper, we discuss our recent work [10] that properly addresses the aforementioned gap in the literature by conducting the first investigation of the polarization phenomena at the instance level within Mastodon. The choice to focus on instances is all but arbitrary, as gaining a macro-perspective of the polarization within Fediverse serves as a key precursor for the user level, yet still bears significant implications for the underlying user base.

To investigate polarization in DOSNs, we focus on three main questions: (1) how many polarized communities can be identified in Mastodon, (2) what the overall structure of polarization looks like—namely, how groups are organized internally and how they relate to each other, and (3) what features characterize the instances belonging to these groups.

Our strategy for addressing these questions relies on two main steps: (i) modeling the network of Mastodon instances as a *signed graph*, where links carry either a positive or a negative label to represent cooperative versus antagonistic relations, and (ii) applying established techniques for the detection of polarization in signed networks.

This approach draws on prior research that frames polariza-

tion detection as the task of partitioning the nodes of a signed graph into k disjoint sets — commonly referred to as poles or groups [11]. Importantly, in line with recent advances [8], [9], we acknowledge that the identified groups may not cover the entire set of nodes. This relaxation better reflects real-world contexts, where neutrality is a natural and often necessary state.

II. Data

We relied on the `instances.social`[1] website to collect an initial set of active instances, which we then used to define a seed dataset of approximately 270K Mastodon users.

From this seed, we carried out a *breadth-first* exploration of the network by retrieving both incoming and outgoing social links through the `/api/v1/accounts/:id/followers` and `/api/v1/accounts/:id/following` endpoints. Following the procedure proposed in [7], we progressively expanded the dataset by iteratively discovering new users connected through these links. Over the course of nine months of continuous crawling, this process yielded more than 2M users and around 116M distinct connections among them.

For the purpose of our analysis, we derived *positive links among instances* from these user-level interactions, as will be detailed in the next section.

To obtain *negative links among instances*, we used the `/api/v1/instance/domain_blocks` endpoint, which provides information on moderation decisions through *DomainBlocks* objects.[2] These JSON structures include blocked domains alongside relevant metadata such as the severity and rationale for the block. By crawling this information across our seed instances between July and November, we obtained a dataset of more than 135K raw block relations.

In line with the federated nature of DOSNs, Mastodon instances also interact with other Fediverse services supporting the ActivityPub protocol (e.g., Pleroma). Our dataset therefore includes these cross-service interactions, which—despite adding some heterogeneity—were retained to enrich this initial exploration of polarization in decentralized social networks.

III. Methodology

Network Modeling. Let $\mathcal{U}$ be the set of users and $\mathcal{I}$ the set of instances extracted from our data. We construct a directed, weighted graph of positive relations among instances, denoted as $\mathcal{G}^+ = \langle V^+, E^+, w \rangle$, where $V^+ \subseteq \mathcal{I}$ is the node set, E^+ is the set of *positive* edges, and $w : E^+ \mapsto \mathbb{R}$ assigns a weight to each edge. An edge $(i, j) \in E^+$ indicates that at least one user from instance i follows a user in instance j, with $w(i, j)$ equal to the number of such follower links. The resulting network $\mathcal{G}^+$ comprises 37,529 nodes and 1,335,490 edges.

Analogously, we define a directed graph of negative relations, $\mathcal{G}^- = \langle V^-, E^- \rangle$, where $V^- \subseteq \mathcal{I}$ and E^- contains a directed edge (i, j) whenever instance i banned instance j during our crawling period. The negative network $\mathcal{G}^-$ includes 11,401 nodes and 105,465 edges.

[1]https://instances.social/

[2]https://docs.joinmastodon.org/entities/DomainBlock/

Before combining these structures, we performed a *network simplification step* on $\mathcal{G}^+$ to remove noisy or statistically irrelevant edges. Indeed, some positive links may stem from chance interactions between individual users rather than meaningful relations between instances. To address this, we adopted the Disparity Filter method [12], which leverages a generative null model based on node strengths to assess the statistical significance of each edge. Using the Disparity Filter method, we reduced $\mathcal{G}^+$ to 117,422 significant edges. No analogous filtering was needed for $\mathcal{G}^-$, since bans are explicit decisions by administrators rather than random outcomes.

Finally, we integrated the two layers into a *signed instance-network*, $\mathcal{G} = \langle V, E, s \rangle$, with $V \subseteq \mathcal{I}$, $E = E^+ \cup E^-$, and a sign function $s : E \mapsto \{+1, -1\}$ labeling each edge as positive or negative. The resulting network contains 19,738 nodes and 222,887 signed edges. During this merging step, we also resolved ambiguities by discarding edges that appeared with both signs, i.e., cases where a positive link was later contradicted by a ban discovered in our breadth-first search.

Detection of Polarized Groups. Let $E^+(P_i, P_j)$ (resp. $E^-(P_i, P_j)$) be the set of positive (resp. negative) edges between two subsets $P_i, P_j \subseteq V$. We define $E^+(P_i)$ (resp. $E^-(P_i)$) to be $E^+(P_i, P_i)$ (resp. $E^-(P_i, P_i)$) as the set of positive (resp. negative) intra-group edges, for any $P_i \subseteq V$. Given a candidate set of groups $\mathcal{P} = \{P_1, \cdots, P_k\}$, we follow [8] and quantify their quality by a function $f(P_1, \cdots, P_k)$ of the intra/inter-group connectivity, defined as $f(P_1, \cdots, P_k) = \sum_{P_i \in \mathcal{P}} \Big(|E^+(P_i)| - |E^-(P_i)| \Big) + \frac{1}{k-1} \sum_{P_i, P_j \in \mathcal{P}} \Big(|E^-(P_i, P_j)| - |E^+(P_i, P_j)| \Big)$.

To the best of our knowledge, the method proposed in [8] is the only existing approach capable of detecting an arbitrary number k of polarized groups while also allowing for the presence of *neutral nodes*, i.e., nodes not assigned to any polarized group.

The identification of polarized groups relies on the following optimization problem, which can be seen as a special case of the correlation clustering problem [13]–[15].

Problem 1 (k-conflicting groups [8]): Given a signed graph G and an integer k, find k mutually-disjoint node sets $\mathcal{P} = \{P_1^*, \cdots, P_k^*\}$ such that:

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

Under the formulation of Problem 1, the quality of the solution depends only on the polarized groups, and not on the neutral group $P_N = V \setminus \cup_{i=1}^{k} P_i$, which fulfills the above stated requirement. The proposed approach in [8] relies on interpreting the problem objective in terms of the Laplacian of a complete graph, characterizing the spectral properties of this matrix, and identifying each conflicting group as the solution to a maximum *Discrete Rayleigh Quotient* (DRQ) problem [8]. More specifically, the objective function in Eq. 1 is found in [8] as a convex combination of $k - 1$ DRQ problems, whose solution to the i-th DRQ problem characterizes the group P_i that conflicts the most with the remaining groups P_j, for

TABLE I
MAIN CHARACTERISTICS OF THE POLARIZED GROUPS [10].

	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 ≥ 1 ban	32.0	52.4	100	28.0

$j > i$, yet to be investigated. The intensity of such conflict, characterizing group P_i, is referred to as the *DRQ value*. This value, having similar rationale to Eq. 1, should be maximized. Based on this observation, [8] proposes an iterative algorithm, dubbed SCG (Spectral Conflicting Groups), executing $k-1$ iterations that involve solving a DRQ problem.

IV. RESULTS

Determining the number of polarized groups k. To estimate k, we followed the heuristic proposed in [8], which is analogous to the "elbow" method commonly used in k-means clustering. The procedure consists of running SCG with different values of k, plotting the distribution of DRQ scores in ascending order (with the i-th largest value placed at position i), and selecting k based on visible "knees" in the resulting curve. In our experiments, k was varied from 2 to 10 in steps of 1; for each setting, SCG was executed 10 times and the outcomes were averaged (see [10] for further details).

The analysis revealed an inflection suggesting the presence of 4 conflicting groups in our signed instance-network. However, when running SCG with $k = 4$, one of the groups turned out to be empty, leaving 3 actual polarized groups. These three, together with the neutral group P_N, constitute the basis of our subsequent analysis.

Characterizing main traits of the polarized groups. We started gaining insights into the obtained polarized groups by means of the statistics shown in Table I. The distribution of instances is highly skewed: the neutral pole P_N alone accounts for more than 97% of the network, consistent with the Fediverse's identity as a federation of many small but interconnected instances. Mastodon dominates overall, yet unevenly: P_1 and P_3 are almost entirely Mastodon-based, whereas P_N and P_2 contain less than half Mastodon instances.

Moderation patterns further distinguish the groups. While P_N and P_2 attract the largest share of bans, their profiles differ markedly: P_1 and P_3 show very low averages, P_N roughly doubles those values, and P_2 stands out with strikingly high levels—about 16 times higher than P_N and up to 26 times higher than the Mastodon-pure poles. This identifies P_2 as the *ban-sink* of the Fediverse, a characterization reinforced by the fact that every instance in this group has been banned at least once, unlike the far lower proportions in the other poles.

Exploring relations between the polarized groups. After identifying the poles of the Fediverse, we examined the structure of positive flows (interactions) and negative flows (bans) among their instances. As illustrated in Table II, most interactions originate from or target the neutral pole P_N, reinforcing its role as the backbone of neutrality within the network. The remaining activity is directed toward the Mastodon-pure poles, whereas the ban-sink P_2 interacts only with itself, a sign of segregation consistent with its negative profile.

TABLE II
POSITIVE/NEGATIVE FLOWS OF INTERACTION AMONG GROUPS. NORMALIZED PERCENTAGES ARE TO BE READ ROW-WISE [10].

	P_N	P_1	P_2	P_3
Positive flows				
P_N	49.85%	40.91%	1.93%	7.30%
P_1	80.39%	14.17%	0.26%	5.17%
P_2	78.12%	5.38%	14.30%	2.20%
P_3	63.31%	24.22%	0.58%	11.89%
Negative flows				
P_N	57.67%	1.28%	40.52%	0.52%
P_1	82.27%	0.62%	16.74%	0.38%
P_2	26.51%	1.20%	22.89%	49.40%
P_3	54.65%	0.76%	44.27%	0.32%

Negative flows reveal a different picture, marked by a bipartite pattern largely split between P_N and P_2. While the high number of bans affecting P_N can be explained by its sheer size, the same argument does not apply to P_2, whose isolation is further confirmed by its disproportionate share of bans. It is also worth noting that P_1 is almost untouched by bans, while P_3 absorbs nearly half of those issued by P_2, a phenomenon that calls for further investigation.

Main representative instances in polarized groups. To better understand the division of groups described above, we analyzed each pole to identify its most representative instances. Specifically, we considered the most interacted instances, measured by the in-strength of positive edges, and the most banned ones, measured by the in-degree of negative edges (see [10] for further details).

In the neutral pole P_N, `mstdn.jp` stands out as the instance receiving the highest number of positive interactions. This is consistent with its role as one of the earliest Mastodon adopters and the second largest Japanese instance in the Fediverse. Within the Mastodon-pure group P_1, `mastodon.social` unsurprisingly emerges as the key positive hub, being the first and official Mastodon instance, while `botsin.space` is the most banned—an outcome easily explained by its focus on hosting automated bots.

Within the ban-sink pole P_2, the most interacted instance is `pawoo.net`, the second largest Mastodon server in terms of users, which has recently attracted attention for hosting controversial content. The most banned instance in the same pole is `poa.st`, a non-Mastodon server self-described as the "Fediverse for shitposters,"[3] further reinforcing the negative profile of this group. Finally, in P_3, the most notable element is the banning of `aethy.com`, a server associated with potentially NSFW content.

Activity in polarized groups. We also characterized polarized groups in terms of activity via the `/api/v1/instance/activity` endpoint of the Mastodon API, by collecting for each pole, the number

[3]https://globalextremism.org/post/poast/

of statuses created in the last 12 weeks (up to the time of writing our work [10]). The neutral group P_N produced 2.37×10^7 posts in the last 12 weeks, with an average of 3,654 per user; the most active instance is `mstdn.jp`, which accounts for 7.65% of the group's volume. Group P_1 generated 2.36×10^7 posts, with an average of 139,511 per user; the largest share comes from `mastodon.social`, contributing 32.84%. Group P_2 produced 1.74×10^6 posts, with an average of 158,065 per user, and more than half of this activity (54.55%) originates from `pawoo.net`. Group P_3 reached 2.38×10^6 posts, averaging 14,356 per user, with `mstdn.ca` as the most active instance at 15.19%.

The results highlight that, although all groups generate substantial volumes of activity, two in particular stand out: the neutral pole P_N and the Mastodon-centric P_1. In P_N, the relatively low per-instance average suggests the presence of a long tail of smaller instances, while P_1 emerges as the "beating core" of the Fediverse, a role underscored by the centrality of `mastodon.social`, which alone accounts for one-third of all posts in the group. By contrast, P_2 is marked by an exceptionally high per-user activity level and an extreme concentration of content in a single instance, `pawoo.net`, responsible for more than half of the group's output—a configuration that calls for closer scrutiny.

To further interpret these patterns, we examined the most common keywords associated with instance bans across the groups (details in [10]). No clear or recurring reasons emerged for the least banned groups, P_1 and P_3. In contrast, P_2 frequently appears on ban lists for issues related to "speech" (notably hate speech), "racism", and "harassment", providing additional support for its characterization as a negativity-driven pole. The neutral group exhibits a different trend: the most frequent ban-related terms, such as "fedi" and "federate", point to moderation practices aimed at restricting federation with undesirable instances. An illustrative case is the occurrence of "facebook/meta" among the top-5 ban keywords, reflecting concerns tied to the launch of *Threads*, Meta's new social platform, which many perceive as a potential threat to the Fediverse.

V. Conclusion

In this work, we discussed our recent work [10] that conceived a signed network model aimed at detecting polarized groups in the Fediverse through the lens of Mastodon. We carried out the first exploratory analysis of polarization among instances, showing that the Mastodon-centric network is organized into four non-overlapping poles, including a predominant neutral group, two Mastodon-pure groups, and a ban-sink group. The neutral pole gathers instances from different services, while the ban-sink consistently receives negative links as a safeguard against potential threats to the Fediverse. Activity analysis further reveals that the neutral and Mastodon-pure groups dominate in terms of content production, with `mastodon.social` emerging as the central hub, whereas the ban-sink displays anomalous posting patterns and is subject to heavy moderation due to harmful or inappropriate content.

Future work might delve into user-level polarization in DOSNs, comparing different algorithms for identifying polarized groups while also considering fairness aspects [16] and other ethical dimensions throughout the detection process of polarized groups and explore LLM-assisted detection methods [17].

Acknowledgment

This work is partly supported by the PNRR Future AI Research (FAIR) project (H23C22000860006, M4C21.3 spoke 9) and by PRIN 2022 Project "AWESOME: Analysis framework for WEb3 SOcial MEdia" (H53D23003550006). These funders had no role in data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] L. La Cava, L. M. Aiello, and A. Tagarelli, "Drivers of social influence in the twitter migration to mastodon," *Scientific Reports*, vol. 13, no. 1, p. 21626, 2023.

[2] J. He, H. B. Zia, I. Castro, A. Raman, N. Sastry, and G. Tyson, "Flocking to mastodon: Tracking the great twitter migration," in *Proc. IMC Conf.*, 2023, p. 111–123.

[3] M. Zignani, C. Quadri, S. Gaito, H. Cherifi, and G. P. Rossi, "The Footprints of a "Mastodon": How a Decentralized Architecture Influences Online Social Relationships," in *Proc. InfoCom Workshops*, 2019, pp. 472–477.

[4] C. Bono, L. La Cava, L. Luceri, and F. Pierri, "An exploration of decentralized moderation on mastodon," in *Proc. 16th ACM Conference on Web Science*, ser. WebSci '24, 2024.

[5] L. La Cava, S. Greco, and A. Tagarelli, "Information consumption and boundary spanning in decentralized online social networks: The case of mastodon users," *Online Social Networks and Media*, vol. 30, p. 100220, 2022.

[6] L. La Cava, S. Greco, and A. Tagarelli, "Network analysis of the information consumption-production dichotomy in mastodon user behaviors," in *Proc. ICWSM Conf.*, 2022, pp. 1378–1382.

[7] L. La Cava, S. Greco, and A. Tagarelli, "Understanding the growth of the Fediverse through the lens of Mastodon," *Appl. Netw. Sci.*, vol. 6, no. 1, p. 64, 2021.

[8] R.-C. Tzeng, B. Ordozgoiti, and A. Gionis, "Discovering conflicting groups in signed networks," *Proc. NIPS Conf.*, vol. 33, pp. 10 974–10 985, 2020.

[9] F. Gullo, D. Mandaglio, and A. Tagarelli, "Neural discovery of balance-aware polarized communities," *Mach. Learn.*, vol. 113, no. 9, pp. 6611–6644, 2024.

[10] L. L. Cava, D. Mandaglio, and A. Tagarelli, "Polarization in decentralized online social networks," in *Proc. 16th ACM Conference on Web Science*, 2024, pp. 48–52.

[11] M. Cucuringu, P. Davies, A. Glielmo, and H. Tyagi, "SPONGE: A generalized eigenproblem for clustering signed networks," in *Proc. AISTATS Conf.*, 2019, pp. 1088–1098.

[12] M. Ángeles Serrano, M. Boguñá, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *Proceedings of the National Academy of Sciences*, vol. 106, no. 16, pp. 6483–6488, 2009.

[13] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," *Machine learning*, vol. 56, pp. 89–113, 2004.

[14] D. Mandaglio, A. Tagarelli, and F. Gullo, "Correlation clustering with global weight bounds," in *Proc. ECML PKDD Conf.*, vol. 12976, 2021, pp. 499–515.

[15] F. Gullo, D. Mandaglio, and A. Tagarelli, "A combinatorial multi-armed bandit approach to correlation clustering," *DAMI*, vol. 37, no. 4, pp. 1630–1691, 2023.

[16] F. Gullo, L. La Cava, D. Mandaglio, and A. Tagarelli, "When correlation clustering meets fairness constraints," in *Proc. DS Conf.*, vol. 13601. Springer, 2022, pp. 302–317.

[17] C. Greco and M. Ianni, "A formal framework for llm-assisted automated generation of zeek signatures from binary artifacts," *Future Generation Computer Systems*, vol. 175, p. 108086, 2026.