Discovering the Landscape of Decentralized Online Social Networks through Mastodon

(Discussion Paper)

Lucio La Cava[1], Sergio Greco[1] and Andrea Tagarelli[1]

[1]*Dept. Computer Engineering, Modeling, Electronics, and Systems Engineering (DIMES), University of Calabria, 87036 Rende (CS), Italy*

Abstract

Decentralized Online Social Networks (DOSNs) are gaining popularity in the social media landscape as a concrete alternative to the centralized and profit-driven counterparts, such as Facebook or Twitter. By leveraging open-source software and specific protocols, DOSNs allow users to create their own instance (i.e., server) and federate to an extensive interconnected social network called Fediverse, where users can transparently communicate with each other, even if registered to different instances. Mastodon represents the most successful service in the Fediverse to date, and in recent years, it has drawn great attention from the research community. In this paper, we discuss our recent study [1], which contributed to advance research on Mastodon and the Fediverse. First, we built the most up-to-date and representative dataset of Mastodon. Upon this dataset, we defined the network of Mastodon instances and exploited it to shed light on the key macroscopic and mesoscopic structural features of Mastodon to unveil the fundamental pillars of the underlying federative mechanism; the backbone of the network, to unveil the essential interrelations between the instances; and the growth of Mastodon, also accounting for instances belonging to other services.

Keywords

decentralized online social networks, Mastodon instances, structural network analysis, community detection, core decomposition, graph pruning

1. Introduction

Nowadays, social networks represent the primary infrastructure to keep inter-personal relationships alive through the Internet, allowing us to cross the world in the blink of an eye. However, the considerable popularity earned by centralized platforms — i.e., owned by a single company, as in the case of Facebook or Twitter — has eventually determined a rapid transition from a user-centric approach to a profit-driven vision concerning user relationships based on social-marketing goals or advertisement mechanisms. As a result, we witnessed the emergence of information bubbles and echo chambers, along with privacy concerns, as users increasingly share their lives within these platforms. The need for bringing the user back to the center of the stage has led to the rise of a new paradigm known as Decentralized Online Social Networks (DOSNs) [2, 3], where privacy control and spontaneous interactions among users are favored

SEBD 2022: The 30th Italian Symposium on Advanced Database Systems, June 19-22, 2022, Tirrenia (PI), Italy

lucio.lacava@dimes.unical.it (L. La Cava); greco@dimes.unical.it (S. Greco); tagarelli@dimes.unical.it (A. Tagarelli)

and unbiased from external factors. The main components of DOSNs are the availability of open-source software, which allows anyone to set up their own *instance* (i.e., server), and the adoption of specific communication protocols that ensures seamless communication between (users registered on) different servers. The result is a *federated* and extensive social network, known as the *Fediverse*, where users can use their accounts to interact with peers on other instances, even belonging to different services. There is a large variety of services composing the Fediverse, which include *Mastodon* and *Friendica* for microblogging, *PeerTube*, *Funkwhale* and *PixelFed* for multimedia hosting.

To date, Mastodon is the most well-established service in the Fediverse, and the platform that has drawn most attention from the research community [4, 5, 6, 7, 8, 9]. Mastodon borrows some interesting concepts from Twitter, recognizing itself as the decentralized alternative for publishing short texts dubbed *toots*. Moreover, in analogy with Reddit, Mastodon emphasizes niche — independent yet linked — communities and content moderation, e.g., instances' administrators can declare the main topics and prohibited contents of their instances, or close registrations to ensure effective content moderation. Besides, users can exploit *content warnings* mechanisms, to temporarily hide potentially sensitive content behind a textual summary, dubbed *spoiler*. Furthermore, leveraging the *ActivityPub* protocol, coupled with a subscription-based mechanism, Mastodon allows the aforementioned cross-instance (and service) "extended" followship, leading to a peculiar manifold timeline structure: *home-timeline*, containing toots generated by the followed users, *local-timeline*, containing toots created within the membership instance, and *federated-timeline*, which gathers all public toots from the users known to the membership instance of a user, being them from the same instance or not.

Zignani et al. [6, 7] are among the first to study Mastodon from a network-science perspective, focusing on topological aspects, including degree distribution, triadic closure, and assortativity, also evaluating similarities and differences w.r.t. Twitter. In this regard, they spotted a more balanced behavior between followers and followees in Mastodon, as well as a limited fraction of social bots, which is lower than the one observed in Twitter [10]. Besides, they unveiled that the clustering coefficient of Mastodon lies between those of Facebook and Twitter. Zignani et al. also found differences in the degree assortativity in Mastodon compared to the ones shown by well-known social networks, and shed light on the influence that home-instances have on users' hubiness. Finally, in [7] the authors discuss about the development of individual instances' footprint and how it impacts on the underlying relationships between users.

Despite the recent proliferation of studies on Mastodon, however, the main contributions only concern user-level interactions, leaving the linkage mechanism between instances unexplored. Recently, in [1], we contributed to the understanding of the Mastodon instances network from unprecedented perspectives, including various macroscopic to mesoscopic aspects, the network backbone and temporal evolution, through the building and exploitation of the most up-to-date and representative dataset concerning Mastodon relationships. More specifically, in [1] we addressed the following research questions:

RQ1 – *Network data and models*: How are the Mastodon instances detected and modeled as a network?

RQ2 – *Structural features*: What are the salient structural features of the network of Mastodon instances, at *macroscopic* as well as *mesoscopic* level?

RQ3 – *Fingerprint*: Are there any clues to the presence of notable phenomena that distinguishes Mastodon from centralized OSNs? How does a federative mechanism arise from the Mastodon instances?

RQ4 – *Network backbone*: What is the backbone of the network of Mastodon instances, and does it preserve the structural features of the whole network?

RQ5 – *Growth*: How has the shape of the network of Mastodon instances evolved during the last few years?

In the remainder of this paper, we summarize and discuss the main findings drawn from the investigation of each of the above outlined research questions. For detailed analysis and results, the interested reader is referred to [1].

2. Data extraction and modeling

Mastodon crawling. Social networks evolve continuously and the validity of the observed phenomena can be ephemeral. Within this view, at the time of writing of [1], the only publicly available dataset [6] concerning the relationships between Mastodon users observed between 2017 and 2018 could have been potentially obsolete. Therefore, to answer our first research question (**RQ1**), we developed a privacy-friendly crawler upon the official Mastodon APIs to create a fresher and more representative dataset, while avoiding any scraping techniques. The decentralized nature of Mastodon posed some obstacles, as the proliferation of distributed instances is not easily traceable. Nonetheless, we leveraged the de-facto Mastodon instances tracker *instances.social* [1] to retrieve the online instances at the time of the crawling, obtaining authentication from about 900 of them. We hence traversed such instances, building a seed-set of approximately 81,000 users, upon which we performed an incremental breadth-first-search crawling task, i.e., we progressively extracted new users to explore during the outgoing/incoming links collection. From a technical perspective, we ensured efficient crawling by adopting a caching mechanism (i.e., *Redis*) equipped with a MongoDB database. Moreover, to prevent computational bottlenecks, we demanded the check for duplicate edges to an offline data refinement step at the end of our crawling phase, using efficient data and network manipulation software libraries. Also, to meet strict privacy principles, we avoided crawling information from instances which did not provide us with an authentication token, and suddenly anonymized the collected links. In this regard, we also resorted to minimal descriptive textual data fetching, to generate the seed-user set by discovering them through toots we read (but not stored) from the timelines of the seed instances, to solve the "cold-start" issue of our crawling process.

Overall, we detected about 28M raw links which, after the removal of duplicate links, led to the discovery of 1.4M and 18M unique users and links, respectively, with an overall coverage of more than 16k instances. Nonetheless, since as previously stated, Mastodon allows users to interact also with services external to the platform, we discerned the Mastodon instances using a combination of information available on the *instances.social* and *fediverse.party* [2] platforms, resulting in 9,433 known Mastodon instances. We used such information to validate our dataset,

[1] https://instances.social/
[2] https://fediverse.party/

eventually obtaining 6,960 out of 9,433 Mastodon instances (both online and offline), and 1,116 out of 1,193 online instances. These values testify our coverage of most of the online Mastodon instances to date, up to doubling the earlier state-of-the-art in terms of currently online instances. As a side yet relevant remark, we emphasize that our dataset includes an exceptional amount (9,322) of non-Mastodon instances, i.e., belonging to other Fediverse platforms, paving the way to an in-depth study of the role of Mastodon in the Fediverse.

Network modeling. Given the set $\mathcal{U}$ of users and the set $\mathcal{I}$ of instances available in the extracted Mastodon data, we denote with $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ a directed network modeling the Mastodon data, where the node set $\mathcal{V}$ contains pairs (u, i), with $u \in \mathcal{U}$ and $i \in \mathcal{I}$, and the edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ corresponds to the set of following relations, such that any $(x, y) \in \mathcal{E}$ with $x = (u, i)$ and $y = (v, j)$ means that user u in instance i follows user v in instance j. Note that u may coincide with v provided that $i \neq j$. To answer the second point of **RQ1**, we derived from $\mathcal{G}$ three Mastodon networks at instance level, which are formally defined as follows.

To model the relations between all the instances in $\mathcal{I}$, we defined the INSTANCES network as the directed weighted graph $G_{\mathcal{I}} = \langle V, E, w \rangle$, where $V = \mathcal{I}$ is the set of nodes, E is the set of edges such that $(i, j) \in E$ if there exists at least one user in instance i that follows another user in instance j, and $w : E \mapsto \mathcal{R}$ is an edge weighting function such that, for any $(i, j) \in E$, $w(i, j)$ stores the multiplicity of the following relation from i to j (i.e., number of users in i following users in j).

Our second network is induced from the set of instances that were detected as online at the time of the crawling process. By denoting with $V^o \subseteq \mathcal{I}$ the set of online instances, the ONLINE-INSTANCES network $G^o_{\mathcal{I}} = \langle V^o, E^o, w^o \rangle$, with edge-set $E^o = E \cap (V^o \times V^o)$ and edge weighting function $w^o : E^o \mapsto \mathcal{R}$, is defined to model the connections between the online instances only. Our third network generalizes the first one by accounting for instances that have been recognized outside Mastodon. Actually, every link extracted during our crawling process is by definition incident with at least one instance that belongs to Mastodon. Therefore, we also defined an expanded network to explore the boundary of the Mastodon network to the rest of the Fediverse. By denoting with $V^* \supset \mathcal{I}$ such expanded set of instances, i.e., the whole set of crawled instances, the EXPANDED-INSTANCES network is defined as $\mathcal{G}^*_{\mathcal{I}} = \langle V^*, E^*, w^* \rangle$, where $E^* = E \cup \{(i, j) \mid (i \in V \wedge j \in V^* \setminus V) \vee (i \in V^* \setminus V \wedge j \in V)\}$, and the weighting function $w^* : E^* \mapsto \mathcal{R}$ follows analogous definition as for the INSTANCES network.

Considering the size of the three networks, EXPANDED-INSTANCES contains 16,282 nodes and 318,218 edges, INSTANCES contains 6,960 nodes and 216,504 edges, and ONLINE-INSTANCES contains 1,115 nodes and 75,046 edges. Furthermore, we emphasize that about 80% of the instances in the network modeled by Zignani et al. [6], hereinafter referred to as *Earlier*, are also contained in our INSTANCES.

3. Structural analysis of the Mastodon instances network

We answered our second research question (**RQ2**) by performing an extensive analysis of macroscopic and mesoscopic properties of the INSTANCES network (Table 1).

From a macroscopic perspective, we spotted that, in contrast to what is commonly observed

Table 1
Summary of structural characteristics of the Instances network, including details on community structure and core decomposition. (* Edge orientation discarded, ** Weighted edges) [1]

	Instances	Instances inner-most core		
		deg	*in-deg*	*out-deg*
#nodes	6 960	189	208	196
#edges	216 504	25 790	28 690	26 463
reciprocity	65.1%	88.4%	85.7%	88.2%
density	0.004	0.726	0.666	0.692
avg. deg*	41.966	152.328	157.702	150.98
avg. in-deg	31.107	136.455	137.933	135.015
% sources	12%	0%	0%	0%
% sinks	6.6%	0%	0.005%	0%
deg. assort.*	-0.274	-0.117	-0.158	-0.135
deg. assort.	-0.253	-0.14	-0.171	-0.151
avg. path len.	2.330	1.270	1.330	1.310
diameter	5	2	2	2
#SCC	1 305	1	2	1

	Instances	Instances inner-most core		
		deg	*in-deg*	*out-deg*
transitivity*	0.128	0.832	0.798	0.807
clust. coeff.*	0.836	0.837	0.810	0.816
clust. coeff. *(full avg)**	0.687	0.837	0.810	0.816
modularity	0.289	0.032	0.039	0.037
#comm. by *Louvain**	5 (5)	3 (3)	3 (3)	3 (3)
modularity	0.353	0.242	0.246	0.246
#comm. by *Louvain***	6 (8)	4 (5)	3 (4)	4 (6)
#comm. by *Infomap***	6 (54)	1 (3)	1 (4)	1 (3)

in centralized OSNs which tend to exhibit a power-law fitting of their degree distributions, Mastodon better fits a lognormal degree distribution. Also, we detected few instances that have higher average degree than the rest of the network.

Mastodon instances can inherently be bounded to specific topics, as suggested by their decentralized nature. However, users tend to look for a broader and complementary range of topics, thus we observed interactions spanning across multiple instances. We ascribe this trait to a mutual reinforcement mechanism, aimed at reducing the *sectorization bias*, reasonably associated with individual instances. This trait certainly matches the capabilities provided by the underlying shared protocol (i.e., ActivityPub) between instances, and leads to a concept of *federation* — manifested through a set of independent yet cooperating instances. We found structural evidences of such phenomena in the Instances network, as suggested by the high clustering coefficient values and percentage of reciprocal edges (cf. Table 1). Concerning such a mutual reinforcement, we also observed degree *disassortativity* (i.e., negative degree correlation) [11], which means that users pertaining to different instances with heterogeneous degrees tend to interact with each other, thus aiming at a better user experience and increasing the speed of information transfer. This represents another distinctive trait of Mastodon compared to what is commonly observed in centralized OSNs, even when they are built upon shared memberships of group [12].

Distinctive traits of Mastodon have also emerged from our mesoscopic-level analysis. By resorting to two well-known community detection methods, namely Louvain [13] (both in its original, undirected implementation and directed implementation) and (directed weighted) Infomap [14], we unveiled the modular structure within instances, which strengthens the previous observations concerning the existence of a federated and close-knit framework among instances. By delving into the nature of such modules, we gained additional insights into their patterns. Particularly, we report topics, languages and temporal processes (e.g., instances' creation time) as the primary factors for the community formation. Moreover, we leveraged core decomposition [15] to shed light on further mesoscopic structural features of our Instances network [16, 17], spotting a surprising and remarkable number of connections from the inner

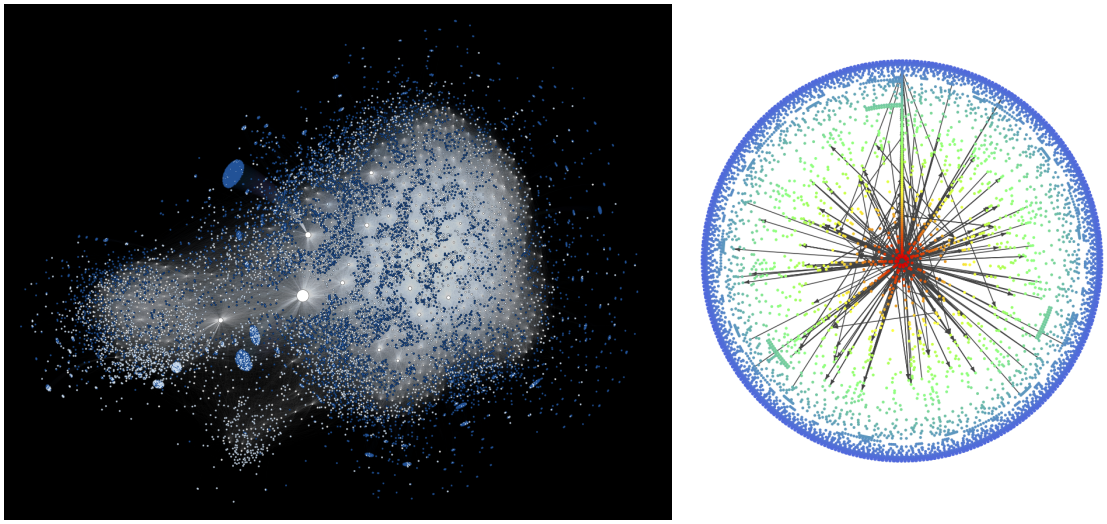

Figure 1: (*Left*) Illustration of the EXPANDED-INSTANCES network, with layout based on the force-directed drawing ForceAtlas2 model. White and light blue indicate online and offline Mastodon instances, respectively, whereas dark blue corresponds to non-Mastodon instances. (*Right*) Core decomposition of the INSTANCES network, based on node in-degrees. Nodes having the same core-index are assigned the same color (inner-most, resp. outer-most core correspond to red, resp. blue). To avoid cluttering, only edges having a weight greater than the first quartile of (unique) edge weights are displayed. [1]

cores to the peripheral ones (Fig. 1); further investigation (results not shown) revealed a balance between links outgoing, resp. incoming, from instances with intermediate core-index values. Also, we noticed that the majority of links between instances involve the inner-most core.

As a result, we can state that Mastodon has several unique traits, which constitute its "fingerprint" — as in the case of the federative mechanism — and make it clearly distinguishable from well-known centralized OSNs. This also answered our third research question (**RQ3**).

In addition, we leveraged two theoretically well-principled *graph pruning* approaches based on probabilistic generative null models, namely *Disparity Filter* [18] and *Marginal Likelihood Filter* [19], to detect and remove noisy edges, with the ultimate objective of unveiling the "backbone" on the INSTANCES network (**RQ4**). Regardless of the significance thresholds we considered, and leaving out some minor fluctuations in the measures due to pruning, we surprisingly observed that the main structural characteristics in the pruned networks remain comparable to those of the original network — or even more emphasized — further strengthening our previously identified characteristic features of Mastodon.

4. Following the temporal evolution of Mastodon

To answer our fifth research question (**RQ5**), we investigated the temporal evolution of Mastodon through the instance perspective. In particular, we compared the main structural traits of our INSTANCES network with the *Earlier* one [6], referring to more than three years ago. Surprisingly, although the two networks differ in size and in crawling time, they were found to be consistent

according to several properties: both networks share the same average path length and diameter, which shows that Mastodon instances act in a *small-world* fashion; the *Earlier* network shows the same disassortative trait as the Instances network, which can be regarded as another distinctive trait of Mastodon w.r.t. centralized OSNs; and even the mesoscopic point of view yields some confirmed traits: when accounting for different sizes, in both the *Earlier* and the Instances networks, Mastodon instances exhibit high degeneracy values.

We also analyzed the current status of the Mastodon landscape, narrowing the focus on our defined Online-Instances network, which exhibits a remarkable reduction in size w.r.t. the Instances one (-84% instances and -65% links among them). We ascribed this shrinking to the achievement of a balance in the number of online Mastodon instances, due to the overcoming of a first initial phase in which the proliferation of new instances is driven by a feeling of novelty. Interestingly, even the Online-Instances network shows the previously spotted disassortative trait as well as the high degeneracy in its core decomposition, and the noticeable concentration of instances in the inner-most core, further validating our findings.

Finally, to investigate the role of the instances through the temporal growth of Mastodon, we assessed the strength of relatedness between the PageRank solutions obtained on the above networks according to *Kendall correlation coefficient* [20] and *Fagin's intersection metric* [21]. As a result, the instance-rankings computed over the different pairs of networks show good or very high correlation, indicating that the most prestigious instances firmly settle in their roles, and they do consistently over time. Overall, our investigation on the evolution of Mastodon allowed us to state that it appears to have reached its structural stability.

5. Conclusions and ongoing work

The user-centric vision and the numerous novelties (e.g., self-hosting and content management) introduced by the DOSN paradigm has led to the development of many services in the Fediverse. Among these, Mastodon stands out as the most adopted decentralized social platform and the most studied by the research community to date. In this paper, we discussed the main findings of our recent research work [1], where we built the largest and most up-to-date Mastodon dataset, and we analyzed macroscopic and mesoscopic structural aspects and the growth of the network of Mastodon instances.

To complement our findings with user-level insights, we are currently studying the underlying network of Mastodon users [22, 23]. Our goal here is to assess the impact of decentralization on user behaviors and information flow, and to provide the first in-depth analysis of how users shape their roles in a decentralized context, by exploiting the dualism between information consumption and boundary spanning. To foster an ever greater understanding of the decentralized landscape through Mastodon, our Mastodon data can be made available upon request.

References

[1] L. La Cava, S. Greco, A. Tagarelli, Understanding the growth of the Fediverse through the lens of Mastodon, Appl. Netw. Sci. 6 (2021) 64. doi:10.1007/s41109-021-00392-5.

[2] B. Guidi, M. Conti, A. Passarella, L. Ricci, Managing social contents in decentralized online social networks: A survey, Online Soc. Networks Media 7 (2018) 12–29.

[3] A. Datta, S. Buchegger, L.-H. Vu, T. Strufe, K. Rzadca, Handbook of social network technologies and applications, Springer, 2010, pp. 349–378.
[4] C. Cerisara, S. Jafaritazehjani, A. Oluokun, H. T. Le, Multi-task dialog act and sentiment recognition on Mastodon, in: Proc. COLING, 2018, pp. 745–754.
[5] J. Trienes, A. T. Cano, D. Hiemstra, Recommending users: Whom to follow on federated social networks, CoRR abs/1811.09292 (2018).
[6] M. Zignani, S. Gaito, G. P. Rossi, Follow the "Mastodon": Structure and Evolution of a Decentralized Online Social Network, in: Proc. ICWSM, 2018, pp. 541–551.
[7] M. Zignani, C. Quadri, S. Gaito, H. Cherifi, G. P. Rossi, The Footprints of a "Mastodon": How a Decentralized Architecture Influences Online Social Relationships, in: Proc. IEEE INFOCOM Workshops, 2019, pp. 472–477.
[8] A. Raman, S. Joglekar, E. D. Cristofaro, N. Sastry, G. Tyson, Challenges in the Decentralised Web: The Mastodon Case, in: Proc. ACM IMC, 2019, pp. 217–229.
[9] D. Zulli, M. Liu, R. Gehl, Rethinking the 'Social' in 'Social Media': Insights into Topology, Abstraction, and Scale on the Mastodon Social Network, New Media & Society 22 (2020) 1188–1205.
[10] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, A. Flammini, Online human-bot interactions: Detection, estimation, and characterization, in: Proc. ICWSM, 2017, pp. 280–289.
[11] M. E. J. Newman, Assortative mixing in networks, Physical Review Letters 89 (2002).
[12] D. N. Fisher, M. J. Silk, D. W. Franks, The perceived assortativity of social networks: Methodological problems and solutions, CoRR abs/1701.08671 (2017).
[13] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment 10 (2008) P10008.
[14] M. Rosvall, C. T. Bergstrom, Maps of information flow reveal community structure in complex networks, Proc. Natl. Acad. Sci. (PNAS) 105 (2008).
[15] S. Seidman, Network structure and minimum degree, Social Networks 5 (1983) 269–287.
[16] F. D. Malliaros, C. Giatsidis, A. N. Papadopoulos, M. Vazirgiannis, The core decomposition of networks: theory, algorithms and applications, VLDB J. 29 (2020) 61–92.
[17] A. Caliò, A. Tagarelli, F. Bonchi, Cores matter? an analysis of graph decomposition effects on influence maximization problems, in: Proc. ACM Web Science, 2020, p. 184–193.
[18] M. Á. Serrano, M. Boguñá, A. Vespignani, Extracting the multiscale backbone of complex weighted networks, Proceedings of the National Academy of Sciences 106 (2009) 6483–6488.
[19] N. Dianati, Unwinding the hairball graph: Pruning algorithms for weighted complex networks, Physical Review E 93 (2016) 012304.
[20] H. Abdi, The Kendall Rank Correlation Coefficient, in: Encyclopedia of Measurement and Statistics, 2007.
[21] R. Fagin, R. Kumar, D. Sivakumar, Comparing Top k Lists, SIAM Journal on Discrete Mathematics 17 (2003) 134–160.
[22] L. La Cava, S. Greco, A. Tagarelli, Information Consumption and Boundary Spanning in Decentralized Online Social Networks: the case of Mastodon Users, Online Social Networks and Media (2022).
[23] L. La Cava, S. Greco, A. Tagarelli, Network Analysis of the Information Consumption-Production Dichotomy in Mastodon User Behaviors, in: Proc. AAAI Conference on Web and Social Media (ICWSM), 2022.