Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)

Follow the "Mastodon": Structure and Evolution of a Decentralized Online Social Network

Matteo Zignani, Sabrina Gaito, Gian Paolo Rossi
Computer Science Department, University of Milan, Milan, Italy

Abstract

In this paper we present a dataset containing both the network of the "follow" relationships and its growth in terms of new connections and users, all which we obtained by mining the decentralized online social network named Mastodon. The dataset is combined with usage statistics and meta-data (geographical location and allowed topics) about the servers comprising the platform's architecture. These server are called instances. The paper also analyzes the overall structure of the Mastodon social network, focusing on its diversity w.r.t. other commercial microblogging platforms such as Twitter. Finally, we investigate how the instance-like paradigm influences the connections among the users.

The newest and fastest-growing microblogging platform, Mastodon is set to become a valid alternative to established platforms like Twitter. The interest in Mastodon is mainly motivated as follows: *a*) the platform adopts an advertisement and recommendation-free business model; *b*) the decentralized architecture makes it possible to shift the control over user contents and data from the platform to the users; *c*) it adopts a community-like paradigm from both user and architecture viewpoints. In fact, Mastodon is composed of interconnected communities, placed on different servers; in addition, each single instance, with specific topics and languages, is independently owned and moderated.

The released dataset paves the way to a number of research activities, which range from classic social network analysis to the modeling of social network dynamics and platform adoption in the early stage of the service. This data would also enable community detection validation since each instance hinges on specific topics and, lastly, the study of the interplay between the physical architecture of the platform and the social network it supports.

Introduction

In the last 15 years we have witnessed the stunning growth of a plethora of web platforms based on the social relationships among their users. Most of them, such as Orkut or MySpace, have disappeared despite periods of success; some are still in limbo, widespread in some countries yet without reaching the expected success, e.g. Google+; meanwhile a few have been adopted worldwide and become the reference social network of millions/billions of people. The latter include Facebook, Twitter, Instagram or LinkedIn. In becoming the most popular social networks, most have drifted away from their original goals and changed their business model as they faced problems like data monetization and data privacy. In fact, the utilization of user data for advertisement purposes is at the core of the public debate and impacts consumer trust in these online services.

As a reaction to these concerns many tecno-activists and software developers have created various forms of online social platforms that put social communication and user content at the heart of their actions. In doing so, they provide no advertisement and recommendation algorithms and leave to users the ownership of their data. Among these alternative social networks we focus on Mastodon since it is the newest and fastest-growing microblogging platform. Mastodon offers some interesting characteristics that make it eligible as a data source in different research fields: *a*) unlike Twitter, Mastodon is not centralized but is made up of interconnected communities located in different servers, called *instances*; *b*) independently owned, operated and moderated, each instance supports specific interests, languages, and needs; *c*) Mastodon provides anti-abuse tools to protect the members of an instance against unwanted contents; and *d*) users have a more detailed control of the visibility of their posts; in fact, each post can alternatively be public and visible on local (instance) and global (federated) timelines, public but not visible on the timelines, or totally private. Thus, we provide a new dataset capturing this new kind of online social network (OSN) far from the usual centralized platform.

A data collection campaign on Mastodon may overcome the following issues researches are facing when dealing with well established and company-based online social networks. First, data confidentiality policies of the major social network platforms severely limited the access to information exploitable for the purpose of reconstructing the structure and evolution of the social network. To protect this kind of data the major platforms continuously act on the API [1] and on the spider policies (*robots.txt*) removing any entry-point. One example we can mention is that, in moving from version 1 to version 2 of its API, Facebook has dismissed the resource which returned the list of friendships of a user. On the contrary, Mastodon facilitates the access to information re-

[1] Up to now Twitter provides the less restrictive API.

lated to the instances and the social communications among its members: *a*) the information about the status of the instances is available at registration time and is easily accessible on the home page [2]; *b*) most of the Mastodon servers do not ban requests to the "follower" and "following" pages of the instance members; and *c*) the decentralized architecture of Mastodon allows us to reach a higher page rate, while also respecting the politeness factor, since the request load is distributed among the instances.

Second, in order to continuously engage their customers and increase relative social communications, the major online social networks come with recommendation algorithms which suggest new contents or users, more or less sponsored. However, these algorithms have significant effects on the structure of the underlying social graph. For instance, the "Who To Follow" algorithm abruptly changed the following graph of Twitter (Su, Sharma, and Goel 2016), as well as what happened after the introduction of the "People You May Know" functionality in Facebook (Zignani et al. 2014). Both studies have shown that the recommender boosted the growth of already popular users and speeded up the triadic closure process, increasing the number of triangles in the social network. On the contrary, the bias given by the recommendation systems is missing in Mastodon. In fact, the only way to establish a connection is by searching an already known account through the search functions or by exploring the feeds of the instances in search of users with similar interests. So being a recommendation-free social network makes Mastodon a data source for understanding the growth mechanisms that drive online social networks, without the influence of external factors. Moreover, the decentralized structure based on defined communities (instances) well suits with the tendency of people to gather and form groups. The understanding of the formation and evolution dynamics of the communities is a central topic in the network analysis as they represent the building blocks of the entire network. Information tracking the temporal dynamics of these groups is essential in understanding how communities grow and overlap; and Mastodon provides this kind of data: we can track how connections evolve and we can gather the communities, eventually along with the topics they are focused on or the behaviors they prohibit.

After summarizing the services offered by the platform, this paper presents the methodology adopted for collecting the meta-data and the network structure of Mastodon social network; it also discusses the main properties of the dataset, making a comparison with the most similar centralized counterpart, Twitter. Our contributions are as follows:

- We release a dataset describing the structure and the growth of a decentralized online social network, Mastodon. The initial snapshot of the network contains more than 400K users and 5.5 millions links among the members of the platforms located in about 1,700 instance around the world. We also collect the meta-data associated to each instance and enrich them by providing the geographical position (country) and topics, the instances focus on.
- We analyze the network structure of Mastodon and compare its main properties against Twitter, i.e. its most widespread centralized counterpart. We show that in Mastodon bidirectional relationships are more likely, i.e. the links are less "weak" than in Twitter, and that, as in Twitter, there are hub-users who attract many other users. However the presence of spambots is marginal, unlike Twitter, where social bots have been an issue. Finally, we find that an instance-based organization highly impacts on the tightly clustered structure of the social network.
- We investigate how the decentralized architecture of Mastodon would impact the relationships among users sited in different instances. In fact, instances built around interests may result in well-separated and scarcely interconnected groups. By analyzing the connections across the instances, we observe that the three major instances are well interconnected, but weakly overlapped. So, despite the decentralized and fragmented architecture, Mastodon users keep connected to the core of the network and are able to search for friendships in other instances, even if a friendship suggestion function is still lacking.

[2] https://joinmastodon.org/

Related Work

The characterization and the analysis of the structure of the online social networks has attracted the attention of research work in different disciplines, from social sciences to computer science and physics. Thus, in the following overview we only include works which focused mainly on data gathering and which eventually released the network they analyze. Specifically, we limit our attention to studies on network structure, network evolution and groups in online social networks.

OSN Structure Due to its huge popularity from the start, Facebook has been the subject of data collection campaigns. However major large-scale studies have been conducted on proprietary data (Traud, Mucha, and Porter 2012), capturing the entire Facebook network(Ugander et al. 2011). The same has happened with Twitter, where the first massive studies (Cha et al. 2010; Kwak et al. 2010) have shown its nature not as a social network but as an information one. This is a characteristic which has been confirmed recently by the Twitter data science team (Myers et al. 2014); at the moment the reference point w.r.t. the structural properties of the famous microblogging platform. Further studies have focused on other popular OSNs, such as Google+ (Magno et al. 2012), Instagram (Manikonda, Hu, and Kambhampati 2014), YouTube (Cheng, Dale, and Liu 2008), MySpace(Ahn et al. 2007), Renren (Jiang et al. 2013) and Weibo (Zhang et al. 2014); or messaging services, such as Microsoft Messenger (Leskovec and Horvitz 2008). Finally, (Mislove et al. 2007) have illustrated most of the issues researchers deal with when they gather data from online social networks.

OSN Evolution and Longitudinal Data In all the above studies researchers have conducted their analyses on a single snapshot of the network. However, the evolution of these online social networks has also been the focus of different works. (Mislove et al. 2008) have presented one of the first

massive studies on the evolution of specific network elements in Flickr. This was followed by other measurements on the network evolution of Yahoo! 360 (Kumar, Novak, and Tomkins 2010), Facebook (Viswanath et al. 2009) and Google+ (Gonzalez et al. 2013): the latter two represent the major studies on the evolution of network connectivity and user activity on a gathered dataset. Finally, (Zhao et al. 2012; Gaito et al. 2012) have investigated the network dynamics in the early stage of the Renren social network by exploiting proprietary data. With respect to the above data sources, the Mastodon dataset captures the evolution of a different social platform paradigm, the decentralized one; it has a day-granularity; it is enriched by different kinds of meta data which allow a more detailed analysis of the mechanisms driving it microscopical evolution; and its evolution is not biased by any friend recommendation algorithm. Finally, the dataset is publicly available [3].

Groups in OSN Since groups represent a fundamental element in the formation of any social networks, from the very beginning researchers on OSN have collected social, collaboration, and information networks where membership to a particular group is explicit. (Backstrom et al. 2006) published one of the seminal works in this area, dealing with the formation and evolution of user-defined groups in LiveJournal and DBLP. Then, the collected datasets have been included in the comprehensive collection of community-annotated networks released in (Kairam, Wang, and Leskovec 2012; Yang and Leskovec 2015). Recently other studies have introduced community-annotated datasets based on deviant communities in Tumblr (Coletto et al. 2016) and categories gathered from Reddit (Zhang et al. 2017). With respect to the previous datasets, Mastodon data provide the explicit membership to an instance along with the instance meta-data, such as the geographical position and the topics, and the evolution of the connections within and among the groups.

All these studies have focused on centralized online social platforms; to the best of our knowledge this is the first dataset on a decentralized social network which includes the network structure and its evolution, and the meta-data about its elements.

Mastodon Overview

Mastodon is a federated social network with microblogging features. It is organized as a decentralized federation of independently operated servers running open source software. The goal of the project, which dates to 2016, is to offer a decentralized alternative to commercial social media. The basic aim is to return control of the content distribution channels to the people by avoiding the insertion of sponsored users or posts in the feeds.

From an architectural viewpoint, the platform is organized into two layers implementing the ActivityPub protocol [4], as shown in Figure 1. The ActivityPub protocol allows it

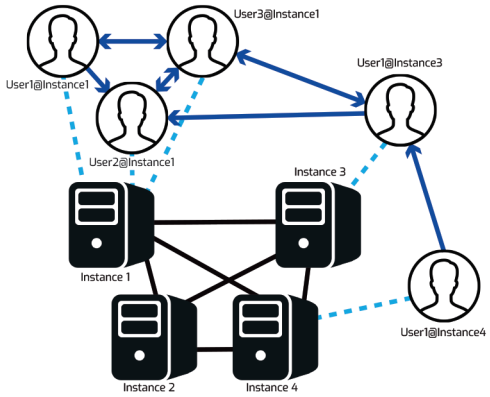

Figure 1: The architecture of Mastodon as a decentralized online social network. The architecture distinguishes between two layers: the server-to-server layer, made up of all interconnected (black links) instances (*InstanceN*); and the social network layer, formed by the 'follow' relationships (blue links) between users (*UserK@InstanceN*) hosted (dotted cyan links) by the different instances. The figure also shows how usernames are composed.

both to manage the communications (black links) among the servers - *instances* - comprising the federation and to offer a client-to-server interface which enables interactions (blue links) among the users having their accounts on the instances. In the server-to-server layer the instances are connected as nodes in a network, and each of them administrates its own rules, account privileges, and whether or not to share messages coming to and from other instances. Unlike a centralized social media, anyone can run a server of Mastodon. Each server hosts individual user accounts, the content they produce, and the content they subscribe to. Consequently, a user joins a specific Mastodon instance and is univocally identified by a unique name (e.g User1@Instance1), consisting of the local username (User1) and the domain of the instance it is on (Instance1).

From a user experience viewpoint, Mastodon releases the major features of a microblogging platform and uses an interface similar to TweetDeck, a professional Twitter application. Like in other microblogging services, users can follow one another, whether or not they are hosted on the same instance - when a local user follows a user on a different server, the server subscribes to that user's updates. Users belong to a single Mastodon instance but can communicate with users on other instances as well. Users post short messages consisting of up to 500 text characters, called 'toots', 'noots' or 'awoos' for others to read. The toots are aggregated in local and federated timelines. The local timeline shows messages from users hosted on a singular instance, while the federated timeline aggregates the messages across all participating Mastodon instances. The non-commercial purposes of the social network only allow a chronological ordering in both timelines, thus avoiding any ranking mechanism based on advertisement or other recommendation algorithms. The service also includes a number of privacy features which im-

[3]Data hosted at https://dataverse.mpi-sws.org/dataverse/icwsm18

[4]https://www.w3.org/TR/activitypub/

pact the contents shown in the timelines. In fact, each message has a privacy option, and users can choose whether the post is public or private. Public messages can be displayed on the federated timeline, and private messages are only shared on the timelines of the user's followers. Messages can also be marked as unlisted from timelines or direct between users. Finally, users can also mark their accounts as completely private, so their posts never appear on any timeline.

Another key point which distinguishes Mastodon from other commercial microblogging platforms is its orientation towards small communities and community-based services, a consequence of its decentralized nature. To this aim, each instance declares in its description the topics their users should be interested in and the maximum number of users the instance can handle. So prior to registration, a user is encouraged to choose the instance more suited to her/is own tastes. If the instance has no room for her/him, the user may choose another similar instance or run a new server supporting specific contents or applying different moderation policies. In fact, the community-orientation strongly impacts the moderation procedures since each instance can limit specific contents. For instance, the flagship instance Mastodon.social bans contents that are illegal in Germany or France, including Nazi symbolism and Holocaust denial. In the mind of Mastodon's founder, small and close communities would defeat unwanted behavior more effectively than a centralized solution based on an operation team screening harming contents. Finally, the organization in decentralized communities is harder for government to censor, since the migration of the community on a server placed in a safer country is an easy task for the instance administrator.

Other than making Mastodon the newest and fastest-growing microblogging platform to emerge in the last year, we can exploit the above characteristics to build a valid data source usable in different research contexts:

- The information about the status of the instances is already available when a user has to register; also, it is easily accessible through a specific web resource [5]. This way, we are able to collect a set of meta-data related to the setting of the instances. The meta-data are useful in studies on the formation of groups and communities (Backstrom et al. 2006; Kairam, Wang, and Leskovec 2012) and in the validation of online social network models where the structure of the social network is strictly related to the interests of its members (Coletto et al. 2016; Zhang et al. 2017).
- The spider policies of most of the Mastodon instances are soft and allow requests towards the resources which return the 'follower' and 'following' relationships for the instance members. This way we can build the Mastodon social network. The network structure, combined with the instances' meta-data, paves the way to a number of research activities, ranging from classical social network analysis to the overlapping community formation and detection (Xie, Kelley, and Szymanski 2013) since people have connections to several instances expressing diverse interests and tastes. Moreover, the fact that each community corresponds to a physical server raises new questions with regards to the understanding of socio-technological systems since there is an explicit interplay between the physical network architecture and the overlaid social network (Schneider et al. 2009).
- The easy access to the information made available by the open platform allows us to track the evolution of the large social network, whereas in most established social media data confidentiality policies impede the gathering of this information. Through a continuous monitoring of the users, in Mastodon we are able to obtain data about the new links during a period of fast growth and adoption of the media. This kind of data is usable in various research fields. Temporal data are fundamental in the mining and modeling of the massive network dynamics (Zhao et al. 2012; Gaito et al. 2012) or in studies about the principles governing link formation and recommendation. In the latter field, Mastodon temporal data offer an undeniable advantage w.r.t. current datasets on link creation: the link creation process is not biased by any friend recommendation algorithm, as occurs in other modern online social networks such as Twitter (Su, Sharma, and Goel 2016) or Facebook (Zignani et al. 2014). Thus, the influence on the network's evolution on the part of the company which owns and handles the social platform is minimal. Finally, temporal data might be combined with instance meta-data supporting research project that range from the formation and evolution of social groups to the dynamics of the adoption and the spreading of emerging platforms.

[5] https://instances.social/

Data Collection

One choice we must address before starting data collection from an online social network concerns how to access enough data to get a representative snapshot of the system. In literature three different approaches are usually adopted to cope with this problem: *a*) a direct access to the data from the system (Jiang et al. 2013; Zhao et al. 2012); *b*) a passive measurement, i.e. a tracking of the communication between users and the platform through click-stream or monitoring applications installed on the users' devices (Schneider et al. 2009; Benevenuto et al. 2009); and *c*) an active measurement by actively querying the OSN platform through the API provided by the system or by parsing web pages. In this study we adopt the third methodology since the first and second options impose strong limitations in terms of number of users willing to install an application and release their private data, in terms of data representativeness, and in terms of number of instance administrators who should be contacted and willing to share the logs of their servers and users' data. Thus, the dataset has been completely gathered by combing queries to the Mastodon API and a custom web crawler. In this section we describe the three main elements of the dataset: the instance meta-data, the static Mastodon graph (the initial snapshot) and the dynamic Mastodon graph, i.e. its growth. We also discuss the choices we made and the tools we developed to gather the data.

Instance meta-data

One of the signature elements in the Mastodon platform is the idea of instance, the cornerstone of the server-to-server layer in Figure 1. The collection of information about elements within a decentralized system may raise some issues in the search for the servers, since each server is independent and the ActivityPub protocol does not provide specifications for polling the other servers. To overcome the latter limitation and to facilitate access to the Mastodon instances, the home page - `joinmastodon.com` - shows a list of all the available instances, along with their status. Moreover, during the collection campaign, Mastodon has introduced an API to query different kinds of information about the instances. Thus, we start from scraping the web resource used to fill the instance list of the home page and then exploit the API. In fact, APIs provide more information about the instances, like a full description and a list of topics, and are based on a registration procedure (namely, instance administrators have to subscribe to API for being inserted into the query results).

We gather the list of all the instances by querying the resource *instances.social/api/1.0/instances/list*, after receiving an access token which prevent an improper usage of the API. For each instance, we obtained the following information:

- Name: domain name of the instance.
- Users: the number of users registered on the instance.
- Connections: the number of connections between the instance and the other ones. A connection corresponds to the black links in Figure 1.
- Statuses: the number of posts published by users hosted by the instance.
- Full description: a description of the instance. Usually, it introduces the Mastodon platform and illustrates what are the topics and the prohibited contents and behaviors.
- Topics: a short list of topics the instance focuses on.

In addition to the above information, API provide data about the version of the software or the uptime of the server; not released in the dataset.

Figure 2 shows the trend of the number of instances, users, connections and posts (statuses) along a six-month period, namely from July 19, 2017 to January 23, 2018 and highlights the actual status of the platform. As reported in Figure 2a the number of instances on January 23 is 1733, an increase of about 450 servers w.r.t. the beginning of the monitoring. By observing the trend, we note a slight decrease during the summer and a stable tendency for the period from the beginning of September to 21 December 2017, date on which we moved from the scraping of the home page to the API. After this date, we observe a remarkable increase in the instances. Finally, we note in both Figure 2a) and b), an artefact on 5 October whose reasons are unknown, yet may be ascribable to failure in the procedure used by the "instance-list" resource for collecting the instance information. As shown in Figure 2b, an increasing trend characterizes the number of Mastodon users and connections among the instances, too. In particular, at the end of December the platform achieved an important milestone: it hit the one million user mark. Lastly, we also monitored the activity of Mastodon instances by the number of post-per-day. In Figure 2c we focus on the activity of the most used instances. In general we observe that the activity level depends on the instance. In fact, pawoo.net and mstdn.jp users are very active (25K and 50K post-per-day on average), while the third most used instance (mastodon.social) is much less active (1K post-per-day).

We also enriched the instance meta-data by adding the geographical location of the instances at a country-granularity. To this aim, we exploited the geo-lookup service provided by *freegeoip.net* for assigning to each server the country it is in. The service relies on a database of IP addresses associated to cities and countries along with other information we overlooked. Thus, we introduce in the released meta-data the field *Country* which contains the ISO 3166-1 alpha-3 [6]code of the country hosting the server. By the geo-lookup service we found the geographical position of 93% (1588) of the instances.

Having the geographical position of the instances allows us to understand where the instance and the users are distributed all over the world; and to quantify the strength of the interplay between the geographical position of the servers and the overlaid social network. The measurements of instances and users positioning are shown on the maps displayed in Figures 2d and 2e. One third of the instances is located in Japan, while the remaining are distributed in North America, in Europe, most notably the France, and in other countries as China, Australia, Brazil and India. We also note that in many countries there are no instances, e.g. there are none in any of the African countries and throughout most of Central and South America and Russia. Thus, so far the Mastodon platform has not reached a worldwide diffusion, despite its decentralized nature and the ease of setting up new instances. Similarly, distribution of the users (Figure 2e) follows instance one; indeed we do not find countries with few instances and many users.

As a further manipulation of the instance meta-data we acted on the field "Topic", which contains the topics the users on an instance should be interested in. In fact, a first check of the "Full description" and "Topic" fields has shown that most instance administrators do not apply the good practice of describing their instance and indicating the topics. Practically, we found that only 18% of instances have a description and 8% also have a not-empty topic list. To sanitize the topic field we manually insert the topics by examining the full description. Usually the instance description contains a paragraph describing the allowed topics. We extract the paragraph and summarize it, keeping the same words authors used to indicate the topics. For instance, from the sentences "Octodon.social is a general purpose instance, where topics are mostly technical and scientific, but there's a bit of everything"; we extracted the topics: technology, science and general. In the case where the instance description does not specify the topic, we assign the word 'general'; this is a common practice in many labelled instances. To summarize the topics, in Figure 2f we visualize them by a word cloud. By visually inspecting the figure, we observe that: *a*) most

[6]https://www.iso.org/iso-3166-country-codes.html

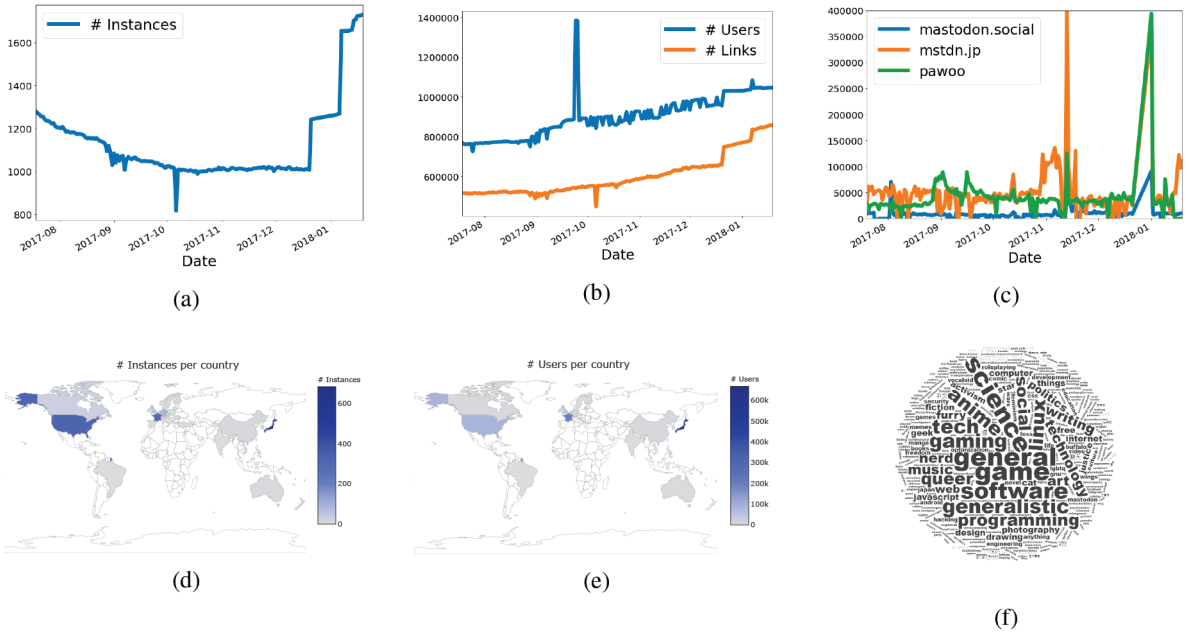

Figure 2: Trends, localization of the instances and topics. In (a) the number of instances day-by-day, in (b) the number of users registered in Mastodon and the number of connections among Mastodon instances during the six-month period, and in (c) the number of post-per-day for the three main instances. In (d) and (e) we report the number of instances and the number of users grouped by the countries where servers are sited, respectively. In (f) the word cloud of the topics assigned to Mastodon instances.

instances are not focused on something particular, they are generalist; *b*) programming and technology, and in general science, are the most common among the specific topics; and *c*) there are also instances dedicated to the arts, creativity and gaming. Despite the above results, the information on topics should be combined with a text analysis on the published statuses in order to fill the missing data. This aspect will be dealt with in further work of ours.

Mastodon Network: Structure and Evolution

The second element of our released dataset is the structure of the social network resulting from the "follow" relationships among the Mastodon users. The asymmetry of the relationship turns the gathering of the network structure into the visit of a directed graph. The visit of a graph, directed or not, is a well-studied problem and there are many off-the-shelf tools to retrieve information from networked data on the Web, such spiders or crawler. Nevertheless, the development of the tools to gather this kind of data requires some choices which depend on the features of the platform:

1. how to access the information on the connections among the users;
2. which users the graph visit should start from;
3. which connections to follow and which policy to implement during the graph visit.

As for the first point, Mastodon offers a rich API to create third-party application, meanwhile providing an access point to user data. As in the case of most of the current API implementation in modern social network, a user must be logged into the system before accessing her/his data and the information returned by the API concerns the logged user only. That is, we are able to collect the in/out connections only if we log into Mastodon and limit ourselves to our own user profile. To overcome these limitations, we developed a web spider targeted to the web pages of the platform. From each profile page we extract the URLs which return both the followers and the followees. Then, by scraping the web pages linked to the above URLs we gather the in-going and out-going relationships of a user. This is an advantage in building the network, since the crawl of a directed network using out-going links only, as Flickr does (Mislove et al. 2008), may not result in the entire weakly connected component. We also highlight that the information in following/follower web pages are also available to visitors who are not logged on.

Once we have identified how to access the data, we have to define the seed set, i.e. the set of users the crawler starts to visit. To build a seed set as large as possible we exploit both the global and the local timelines, since they report all the statuses with public visibility (see the previous section) in chronological order. To retrieve the list of the posts in each instance timeline we leverage the Mastodon API and query each instance separately. From each list we extract the users who posted at least one status and put them into the seed set. To respect politeness and not to excessively load the instance servers, we stop to query API when we reach

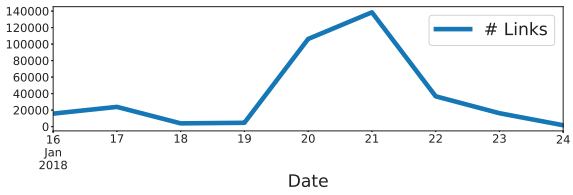

Figure 3: The number of new links extracted by the monitoring tool, day-by-day.

10% of instance users or 30% of statuses [7]. The resulting seed set contains more than 62K users spread over more than 950 instances.

Finally, in our crawler we implement a breadth-first search (BFS) strategy which traverses both out-going and in-going links; where the latter are traversed in the opposite direction, i.e. from destination to source. In the crawler we also add a filter which discard links towards profiles hosted in other social platform supporting the open protocols ActivityPub or OStatus. Indeed, these protocols allow users to interact with users on other platforms, forming what is named "fediverse".

Instance name	Users (MD)	Users (N)	Coverage
pawoo.net	336182	131478	39%
mstdn.jp	156150	92322	59%
mastodon.social	124612	70409	57%
friends.nico	59918	17888	30%
mastodon.cloud	39803	31128	78%
mastodon.xyz	14207	6953	49%
octodon.social	9296	5080	55%
music.pawoo.net	8499	287	3%
mamot.fr	8269	5533	67%
social.tchncs.de	7896	2574	33%

Table 1: Number of users in the most used instances. Users (MD) and Users (N) report the number of users returned by the meta-data and the crawl, respectively. The last column indicates the portion of users who are in the crawled network.

After the end of the crawling process we obtained a network made up of 479,425 nodes and 5,649,762 directed links. Based on the number of users we get from the instance meta-data at the time of the crawl ending, our network covers 46% of users in Mastodon. Specifically, in Table 1 we observe that coverage of the crawl is greater than 50% in most instances and that the three biggest instances are covered on average to an extent of 52%.

Once we have developed the tools for the crawling, we are able to track the evolution of the network, i.e. the third main element released in our dataset. Every day we run a monitoring tool which extracts the new followers and followees of each user in the crawled network and, whereas it detects users not yet in the network, it runs the network crawler again starting from these new profiles[8]. Moreover, the tool is less request-demanding since it retrieves the last connections, due to a chronological ordering in the followee and follower web pages. Our tracking of the evolution of the Mastodon network started on 15 January, 2018 and it is still going. The status of the evolution - numbers of new links - is reported in Figure 3. In 9 days we have gathered more than 370K new links, with a peak during the week-end (20-21 January, 2018). Finally, in the release network we add to each link a timestamp, indicating the day on which we retrieve it.

Findings

Beyond showing the main basic characteristics of the Mastodon network, in this section we aim to give a first answer to two main questions: to which extent does the decentralized and instance-based nature of Mastodon influence its overall structure? More specifically, how does it differ from the most famous microblogging service, Twitter? We leverage the followee and follower data of each users in the first snapshot of the acquired dataset to construct a directed network which enables us to investigate the asymmetric relationships typical of this kind of networks. In addition, we consider also the mutual network - which is built by reciprocated edges (Mastodon reciprocity is 0.35), whose extremes are users following one another - in order to enable comparison with Twitter features as reported in literature (Myers et al. 2014). Considering the global or the instance-based networks, we run a batch of well-known analyses on both the directed and mutual networks and present relevant findings.

Degree Distribution

Considering the asymmetric nature of the "follow" relationship of the Mastodon directed network, nodes have both an in-degree (the number of followers) and an out-degree (number of followees). Figure 4a displays the in-degree distribution as a blue line, the out-degree distribution as a green line and the degree related to the mutual network as a red line. They are plotted as complementary cumulative distribution functions (CCDF) aggregated on all users so as to highlight the heavy-tail shape commonly observed in online social networks. Surprisingly, the in-degree and out-degree distributions are very similar to the extent of exhibiting the same median value of 16; this is opposite to Twitter, where the median value for the out-degree distribution is higher than for the in-degree. This highlights the fact that while the typical Twitter user follows more people than he/she has followed, in Mastodon users have a more balanced behavior. This balanced behavior reflects on how the difference between followers and followees of each node is distributed. In fact, more than 95% of users have a difference in the interval $(-250, 250)$. Thus, in both platforms we can find a small population of celebrity users and the presence of social bots which in Twitter was estimated to be around 15% (Varol et al. 2017); by contrast, Mastodon profiles with spambot traits are more marginal, less than 5%. The degree distribution of

[7] We compute the percentage on the statistics released in the instance meta-data

[8] The duplicate filter implemented in the spider avoids running the crawler over the entire network.

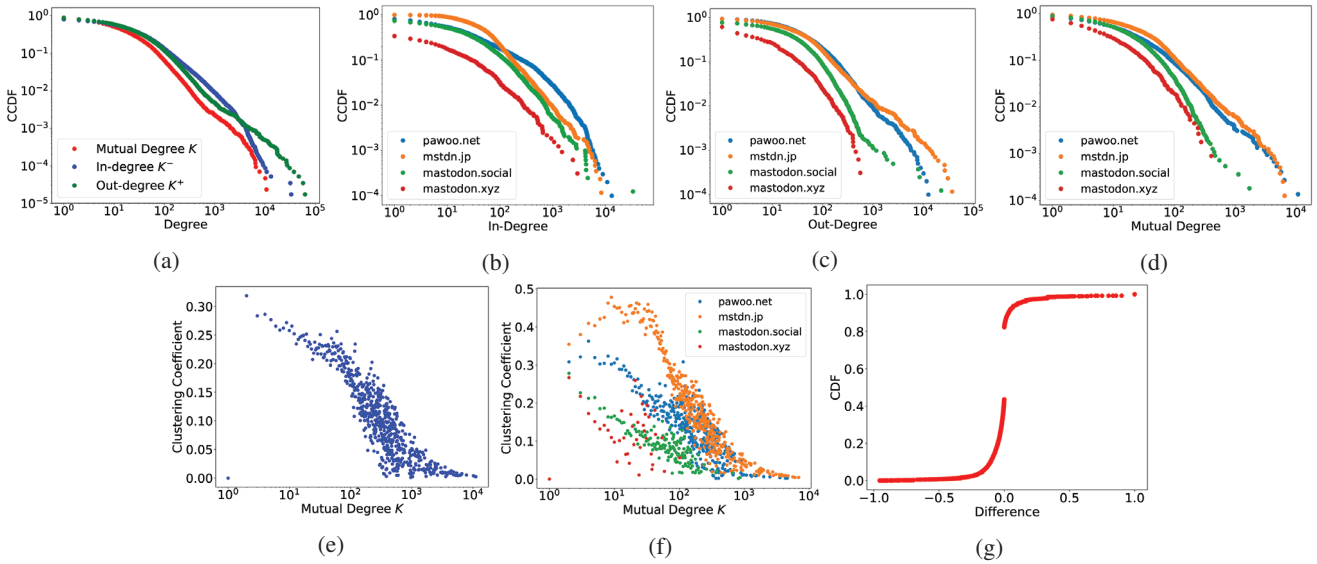

Figure 4: In (a) the in-degree, out-degree and mutual degree distributions of the Mastodon social graph. The in-degree (b), the out-degree (c) and the mutual degree (d) distributions for the largest Mastodon instances, i.e. pawoo.net, mstd.jp, mastodon.social, mastodon.xyz. In (e) and (f) the average clustering coefficient as a function of the degree, in the entire network and in the subnetworks of the four most common instances. In (g) the distribution of the difference between the clustering coefficient measured on the entire network and on the instance subnetwork.

the mutual graph is shown in Figure 4a. Here, we still observe relatively large degrees, although smaller than both the in-degrees and the out-degrees.

Figures 4b, 4c, and 4d show the degree distributions (CCDF) for the four largest instances of Mastodon. It is evident that they largely differ one from one another when considering each of the degree-related metrics; instead, they were found to be comparable in the country-based subgraphs of Twitter. The centralized and group-unaware paradigm of Twitter makes the users' behavior uniform across country, while the decentralized approach of Mastodon makes it possible to build subnetworks of people with different features.

Clustering Coefficient

The clustering coefficient in social networks measures the fraction of users whose friends are friends among one another. As in the Twitter analysis, we focus on the local clustering coefficient (cc) of nodes in the Mastodon mutual network. In Figure 4e we show the average local clustering coefficient as a function of the mutual degree. As in most social networks, the local clustering coefficient decreases while the degree increases. In the comparison of this metric with two of most widespread online social networks, we find that its lies in the middle between Facebook and Twitter. Specifically, if the average clustering coefficient for the degree equal to 5 is about 0.4 and 0.23 in Twitter and Facebook respectively, in Mastodon we get 0.28. The same trend holds for higher degrees, for a degree of about 20 in Facebook cc is 0.3, in Twitter it is 0.19, while in Mastodon it is 0.23. With a degree close to 100 Facebook and Twitter networks are very similar to each other and have a coefficient around 0.14, in Mastodon, rather, the cc is higher, 0.17. In general, the Mastodon network shows a tightly clustered structure. This is a property which makes the released network consistent with a social network.

The above properties of the clustering coefficient results from the combination of the subnetworks supported by the difference instances. But it is also interesting to analyze the differences in the local clustering between the main Mastodon instances, as shown in Figure 4f. In the figure we report the average local clustering coefficient versus the mutual degree for the instances: pawoo.net, mstdn.jp, mastodon.social and mastodon.xyz. Instances are very different from one another. First, mstdn.jp, the second largest instance, has a higher average clustering coefficient (0.35) compared to the other instances (pawoo.net - 0.26, mastodon.social - 0.13 and mastodon.xyz - 0.08), and it is also higher than the clustering coefficient of the entire network. The second and even more interesting fact is that in the mstdn.jp subnetwork the clustering coefficient increases up to a peak ($cc = 0.46$) at degree around 30, then slows down. That indicates the presence of clustered regions around nodes with a small-medium connectivity. The same behavior, at a different magnitude order, has been observed in the Twitter Japanese subgraph [9], where there are quasi-clique subgraphs centered around high degree nodes. The above results highlight that the clustered structure of the network strongly depends on the instances, as also indicated by Figure 4g. Here we plot the cumulative distribution function of the increase/decrease of the clustering coefficient measured on the instance subgraph and on the whole network. The distribution is concentrated in the interval $(-0.1, 0.1)$

[9]Note that mstdn.jp is a Japanese instance.

and reveals that the tendency of neighborhood's nodes of being clustered is limited within the instance boundary.

Degree and Instance Assortativity

Assortativity measures preference for a node to be linked to others that are similar (or dissimilar) w.r.t. a specific property. Typically, we distinguish between nominal and numerical attributes, since the metrics adopted are different. In the first case, we compute the modularity of the network w.r.t. a given category, while in the second case we use the Pearson coefficient to compute how correlated a numerical property of nodes connected by a link is. Here we focus on the degree assortativity and on the nominal assortativity measured on the instance hosting the nodes.

As for the degree assortativity, we consider four cases associated with the directionality of the links: source in-degree (SID), source out-degree (SOD), destination in-degree (DID) and destination out-degree (DOD). Then, we also evaluate the degree assortativity on the mutual graph. In contrast to the findings in Twitter, we do not measure any significant correlation for the pairs (SOD,DOD), (SOD,DID) and (SID,DOD), while we observe a slightly negative correlation (-0.1) on the pair (SID,DID). If we interpret the "follow" relationship as an expression of interest in a user, the lack of correlation for (SOD,DOD) and (SOD,DID) suggests that people who are interested in many people are connected to users whose level of popularity may vary (DID) and who can be interested in many or few other users (DOD). Otherwise, a lack of correlation for (SID,DOD) implies that no matter the level of popularity of a node (SID), it will be linked to people interested in many or few other people. Finally, the negative correlation between SID and DID means that the more popular you are, the less popular the people you follow are. Generally, we find that Mastodon shows degree assortativities that are not consistent with the other online social networks. This is also confirmed by a dis-assortative trait (-0.13) in the mutual network[10].

Also of interest is the nominal assortativity of the property "instance". Indeed, it may reveal whether or not users tend to connect to other people within the same instance. Moreover, if we focus on the hub-users in the main instances, we are able to verify if their popularity is limited to the instance they belong to or if they reach a high popularity because they engage people in other instances. In this case, we compute the instance assortativity on the Mastodon mutual graph. For the whole network the assortativity is 0.56, which holds steady if we limit ourselves to the case of hub-users. This positive instance assortativity confirms that the users' neighborhood is mainly contained in the instance people belong to; and indicates that hubbiness is strongly influenced by and limited to the instance users are located in.

The instance network

Finally, we investigate how the decentralized architecture of Mastodon impact the relationships among users sited in different instances. In fact, instances built around interests may result in well-separated and scarcely interconnected groups. To this aim we analyze the directed network of the instances, as shown in Figure 5, where we draw a weighted link from the instance i_1 to instance i_2 if there is at least one link from i_1's member to i_2's member and the weight is proportional to the number of links connecting i_1 and i_2's members[11]. By visually inspecting the structure of the network we observe that the three major instances are well interconnected, and that there are many instances surrounding them. So, despite the decentralized and fragmented architecture, Mastodon users keep connected to the core of the network and are able to search for friendships in other instances, even if a friendship suggestion mechanism is still lacking.

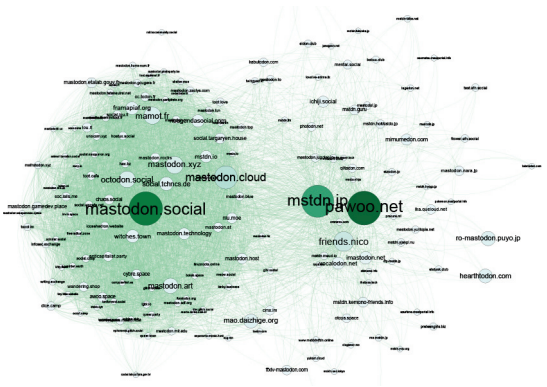

Figure 5: The directed network of the Mastodon instances.

Conclusion

The paper aims to provide the community with a new dataset that is interesting for many aspects. First of all, it concerns one of the most recent and fastest-growing social networks: Mastodon. It is a platform that aims to overturn the centralized and invasive model of the most popular social networks by proposing a decentralized approach, free of sponsored contents and recommendation systems. Given the importance of the debate on these topics, monitoring and understanding this novel approach might be helpful. The results of an initial network analysis reported in our paper already underline its particular features. The absence of interference by the platform will also allow us to understand the intrinsic mechanisms of growth and evolution of a social network when not mediated. For instance, Mastodon data represent a suitable sink of data for the understanding of the processes driving the formation and evolution of ego-networks in online social networks. Besides, its instance-based nature can shed a light on another fundamental topic that still escapes full understanding: communities. Lastly, the wealth of the dataset - which includes not only the structure of the network but also temporal annotation and meta-data about topics and geographical information - makes it suitable for a plethora of different studies from the validation of triadic closure models to a deeper understanding of network forma-

[10]The degree assortativity in Facebook is 0.226.

[11]The percentage of directed link whose extremes are in different instances is 62%.

tion and evolution models based on topics or interests.

References

Ahn, Y.-Y.; Han, S.; Kwak, H.; Moon, S.; and Jeong, H. 2007. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07. ACM.

Backstrom, L.; Huttenlocher, D.; Kleinberg, J.; and Lan, X. 2006. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, 44–54. ACM.

Benevenuto, F.; Rodrigues, T.; Cha, M.; and Almeida, V. 2009. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, IMC' 09, 49–62. ACM.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, P. K. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the 5th International Conference on Web and Social Media*, ICWSM '10.

Cheng, X.; Dale, C.; and Liu, J. 2008. Statistics and social network of youtube videos. In *Proceedings of 16th International Workshop on Quality of Service*, IWQoS '08, 229–238. IEEE.

Coletto, M.; Aiello, L. M.; Lucchese, C.; and Silvestri, F. 2016. On the behaviour of deviant communities in online social networks. In *Proceedings of the 10th International AAAI Conference on Weblogs and Social Media*, ICWSM '16, 72–81.

Gaito, S.; Zignani, M.; Rossi, G. P.; Sala, A.; Zhao, X.; Zheng, H.; and Zhao, B. Y. 2012. On the bursty evolution of online social networks. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, 1–8. ACM.

Gonzalez, R.; Cuevas, R.; Motamedi, R.; Rejaie, R.; and Cuevas, A. 2013. Google+ or google-?: dissecting the evolution of the new osn in its first year. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, 483–494. ACM.

Jiang, J.; Wilson, C.; Wang, X.; Sha, W.; Huang, P.; Dai, Y.; and Zhao, B. Y. 2013. Understanding latent interactions in online social networks. *ACM Transactions on the Web (TWEB)* 7(4):18.

Kairam, S. R.; Wang, D. J.; and Leskovec, J. 2012. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of the fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, 673–682. ACM.

Kumar, R.; Novak, J.; and Tomkins, A. 2010. Structure and evolution of online social networks. In *Link mining: models, algorithms, and applications*. Springer. 337–357.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World wide web*, WWW '10, 591–600. ACM.

Leskovec, J., and Horvitz, E. 2008. Planetary-scale views on a large instant-messaging network. In *the 17th International Conference on World Wide Web*, WWW '08, 915–924. ACM.

Magno, G.; Comarela, G.; Saez-Trumper, D.; Cha, M.; and Almeida, V. 2012. New kid on the block: Exploring the google+ social graph. In *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, IMC '12, 159–170. ACM.

Manikonda, L.; Hu, Y.; and Kambhampati, S. 2014. Analyzing user activities, demographics, social network structure and user-generated content on instagram. *arXiv preprint arXiv:1410.8099*.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, IMC '07, 29–42. ACM.

Mislove, A.; Koppula, H. S.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2008. Growth of the flickr social network. In *Proceedings of the First Workshop on Online social networks*, WOSN '08, 25–30. ACM.

Myers, S. A.; Sharma, A.; Gupta, P.; and Lin, J. 2014. Information network or social network?: the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, 493–498. ACM.

Schneider, F.; Feldmann, A.; Krishnamurthy, B.; and Willinger, W. 2009. Understanding online social network usage from a network perspective. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '09, 35–48. ACM.

Su, J.; Sharma, A.; and Goel, S. 2016. The effect of recommendations on network structure. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16.

Traud, A. L.; Mucha, P. J.; and Porter, M. A. 2012. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications* 391(16):4165–4180.

Ugander, J.; Karrer, B.; Backstrom, L.; and Marlow, C. 2011. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*.

Varol, O.; Ferrara, E.; Davis, C. A.; Menczer, F.; and Flammini, A. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the 11th International AAAI Conference on Weblogs and Social Media*, ICWSM '17.

Viswanath, B.; Mislove, A.; Cha, M.; and Gummadi, K. P. 2009. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, WOSN '09. ACM.

Xie, J.; Kelley, S.; and Szymanski, B. K. 2013. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys* 45(4):43.

Yang, J., and Leskovec, J. 2015. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems* 42(1):181–213.

Zhang, K.; Yu, Q.; Lei, K.; and Xu, K. 2014. Characterizing tweeting behaviors of sina weibo users via public data streaming. In *Web-Age Information Management*. Springer. 294–297.

Zhang, J.; Hamilton, W. L.; Danescu-Niculescu-Mizil, C.; Jurafsky, D.; and Leskovec, J. 2017. Community identity and user engagement in a multi-community landscape. In *Proceedings of the 11th International Conference on Web and Social Media*, ICWSM '17.

Zhao, X.; Sala, A.; Wilson, C.; Wang, X.; Gaito, S.; Zheng, H.; and Zhao, B. Y. 2012. Multi-scale dynamics in a massive online social network. In *Proceedings of the 2012 ACM conference on Internet Measurement Conference*, IMC' 12, 171–184. ACM.

Zignani, M.; Gaito, S.; Rossi, G. P.; Zhao, X.; Zheng, H.; and Zhao, B. Y. 2014. Link and triadic closure delay: Temporal metrics for social network dynamics. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, ICWSM '14.