UNIVERSITÀ DELLA CALABRIA

UNIVERSITA' DELLA CALABRIA

Dipartimento di

Ingegneria Informatica, Modellistica, Elettronica e Sistemistica

Dottorato di Ricerca in

Information and Communication Technologies

CICLO

XXXVI

GRAPH MINING AND MULTIMODAL REPRESENTATION LEARNING FOR EMERGING DECENTRALIZED SOCIO-ECONOMIC DOMAINS

Settore Scientifico Disciplinare ING-INF/05

Coordinatore: Ch.mo Prof. Giancarlo Fortino

GIANCARLO FORTINO
24.02.2024 22:44:01
GMT+00:00

Firma _____

Supervisore: Ch.mo Prof. Sergio Greco

GRECO
SERGIO
24.02.2024
12:45:09
UTC

Firma_____

Co-Supervisore: Prof. Andrea Tagarelli

ANDREA
TAGARELLI
24.02.2024
22:33:14
GMT+01:00

Firma_____

Dottorando: Dott. Lucio La Cava

Lucio La Cava
24.02.2024 21:20:43
GMT+01:00

Firma _____

To my Grandparents,

To my Family.

“Research is to see what everybody else has seen,
and to think what nobody else has thought.”

—Albert Szent-Györgyi

Preface

Complexity is an omnipresent element in our lives, intricately woven into every aspect that defines or surrounds us. Rather than fearing complexity, we must courageously embrace it, as through its intricate nature we uncover the keys to understanding our past, fully experiencing our present, and envisioning a future that perpetually unfolds into greater possibilities. Complexity, far from being a barrier, becomes the very substance that defines the profound journey of our existence.

This work seeks to explore two quintessentially complex systems, namely the social and economic domains, through the lens of computer science. The contemporary social sphere, nowadays experienced and perpetuated through the Internet, stands as the fundamental fabric of our existence. Indeed, grounded in peer interactions and the continuous exchange of knowledge, opinions, and divergent perspectives, it serves as a foundational backbone to our evolution. On the other hand, the economic sphere holds significant sway over the progression of our society, skillfully managing both tangible and virtual assets. This adept management facilitates the provision of goods and services, playing a crucial role in enhancing our lives and steering our evolutionary path.

While these domains have been extensively examined in human history and drawn the focus of researchers across diverse fields, the current complex challenges impacting our society and the increasing inclination towards decentralizing social and economic aspects for enhanced autonomy and control point to a novel journey to be experienced.

Throughout this work, such a journey will be approached through the complementary fusion of two robust yet constantly evolving knowledge realms — namely, *graph mining* and *multimodal deep learning*. The former leverages statistically grounded theories to steer and facilitate the exploration of complex systems, shedding light on the latent phenomena concealed within the contemporary social and economic landscapes. Conversely, the latter capitalizes on learning deep representations of the diverse shapes that information assumes today — e.g., images, text, transactions, and relationships — enabling a boundless and comprehensive exploration of complexity from different perspectives.

Rende (CS), Italy
2023

Lucio La Cava
University of Calabria

Contents

Part III Multimodal Representation Learning for the Web3

Part I

Decentralized Online Social Networks

Overview

This part of the thesis delves into the emerging landscape of *Decentralized Online Social Networks*. It covers five chapters, which are organized as follows:

Chapter 1. This chapter introduces the realm of Decentralized Online Social Networks (DOSNs) through an exploration of the main characteristics of this novel social paradigm and provides an outline of the primary scientific contributions related to DOSNs.

Chapter 2. This chapter presents an in-depth analysis of the network of Mastodon instances achieved through the design and study of the largest and most up-to-date Mastodon dataset existing to date. Besides, this chapter will reveal and discuss the fingerprint of Mastodon, referring to its unique set of traits, and its temporal evolution.

Chapter 3. This chapter delves into user relationships and behaviors within DOSNs, with particular emphasis on Mastodon. The focus will be on investigating the key characteristics of user connections, the impact of the prominent instances, and the main strategic phenomena and user roles that emerge within DOSNs.

Chapter 4. This chapter investigates how decentralization influences user behaviors in terms of information production and consumption. Particular attention will be given to the identification of repeated or alternate behavioral patterns across multiple instances in Mastodon.

Chapter 5. This chapter will conclude the first part of this thesis. The chapter focuses on the migration of Twitter users to Mastodon following Elon Musk's acquisition, which represents to date a unique opportunity for studying collective behavior. The goal of this chapter is to shed light on the drivers of social influence and coordinated behavior that led to this mass migration, constituting one of the largest social migrations in Internet history to date.

1

Background and Related Work

In the last decade, we witnessed an unprecedented proliferation of Online Social Networks (OSNs). Roughly and generally speaking, OSNs aim to shrink timing and distances that characterize inter-personal relationships through the Internet. However, the extreme popularity gained by Facebook and the other worldwide available yet centralized OSN platforms (i.e., hosted and controlled by a single company) has soon led their owners to pursue a collateral social-marketing goal, which is mostly implemented through content personalization mechanisms and advertisement strategies. As it is well-known, side-effects such as the formation of information bubbles and concerns about the protection of data and user privacy normally characterize most existing centralized OSNs.

The above aspects contributed to raise the opportunity for developing new paradigms of OSNs to become "user-centric" rather than "company-centric" platforms. As a major consequence, privacy control, as well as spontaneous and recommendation-free communications among the users, are favored and unbiased as much as possible from the invasiveness of advertisements.

In this context, Decentralized Online Social Networks (DOSNs) are emerging as alternatives to the popular centralized platforms. DOSNs are built upon two key aspects: the availability of open-source software to allow everyone to set up their server hence avoiding centralization, and the existence of specific communication protocols to enable fluid interconnections between servers that embrace the same protocol.

These core components lead to a *federation* model, in which the servers, also called as *instances*, can communicate to each other through the same protocol. This implies that users which are signed up for a particular server can actually interact with users of other servers, analogously to what normally happens with email services. DOSNs hence become part of a massive social network, namely the *Fediverse*. As a consequence of this mechanism, users can use their accounts on a DOSN platform to follow users on other platforms, without needing an account there.

The Fediverse currently provides several services, such as *Mastodon* and *Friendica* for microblogging, *PeerTube* and *Funkwhale* for video hosting, *PixelFed* for image hosting. Among these platforms, Mastodon is the one that has encountered

the greatest attention increase over the years. Mastodon provides a user experience comparable to Twitter (e.g., published contents are called *toots*, whereas the analogous of the retweet functionality is called *boost*) and Reddit (e.g., niche communities and content moderation are emphasized, however Mastodon communities are independent of each other). Moreover, Mastodon affords content warnings, i.e., synopses of toots that can preview disturbing content. Mastodon is built upon the *ActivityPub* protocol,[1] which implements a layer for client-to-server communications and another one for the server-to-server communications. Thanks to this protocol and a subscription-based mechanism (implicitly carried out by the instances), users can interact with each other even if they belong to different instances.

The extended followship mechanism in Mastodon also leads to an original yet remarkable timeline structure, namely *home*-timeline, which provides toots generated by followed users, *local*-timeline, which yields toots created within the instance, and *federated*-timeline, which contains all public toots from all users (either from the same instance or not) that are known to the instance where a user is registered.

Furthermore, Mastodon instances allow their users to apply rules and policies on the generated contents. Administrators can declare both the main topics of their instance and prohibited contents. Users can mark some contents as inappropriate for a given instance by placing a *content warning* on the content. Along with the content warning, a spoiler (i.e., a textual component summarizing the obfuscated content) will be displayed to the user, letting her/him decide whether to view it or not.

Mastodon also allows administrators to close registrations for their instances, e.g., in the case of a "private instance" among friends, or to efficiently moderate contents. Nevertheless, this feature does not affect user interactions which, as outlined above, are guaranteed by specific protocols. Finally, administrators might also decide to moderate relationships with other instances, e.g., by blocking or silencing them, to prevent the spreading of harmful or inappropriate content.

DOSN analysis is a relatively novel research field. Early works mainly investigated motivations, opportunities and challenges related to different solutions for the decentralized paradigm, from distributed systems like peer-to-peer networks to hybrid systems integrating external and private resources for storing user data [95]. Two surveys on these topics as well as on issues related to DOSN infrastructures, data management, privacy and information diffusion, can be found in [95, 60].

To the best of our knowledge, Mastodon is the platform in the Fediverse that has received noticeable attention from the research community [43, 217, 245, 179, 246, 247]. Zulli et al. [247] have recently performed a qualitative analysis based on an interview to a sample of instance moderators. From that study it emerges that the federative structure of Mastodon enables content variety and community autonomy, and also emphasizes horizontal growth between instances rather than growth within instances; however, any analysis of the interactions on the Mastodon instances is missing.

From a network science perspective, the studies by Zignani et al. [245, 246] are particularly relevant, as they were the first to analyze a portion of the Mastodon user-

[1] `https://www.w3.org/TR/activitypub/`

network, focusing on degree distribution, triadic closure, and assortativity aspects, and comparing such characteristics to those in Twitter [245]. From the analysis of the in-degree and out-degree distributions, Mastodon is found to show a more balanced behavior between followers and followees than what observed in Twitter. Also, the 95% of users exhibit a difference between followers and followees bounded in the range (-250, 250). Concerning social bots, the authors reported a low presence (around 5%), which is significantly lower than the 15% observed on Twitter by Varol et al. [221]. Clustering coefficient in Mastodon ranges between those of Facebook and Twitter. The degree assortativity in Mastodon was also inspected, considering source in-degree (SID), source out-degree (SOD), destination in-degree (DID), and destination out-degree (DOD). The authors observed lack of correlation between (SOD, DOD), (SOD, DID) and (SID, DOD), which indicates that users who follow many users are connected to users whose popularity may vary (DID) and who in turn follow few or many users (DOD). Moreover, the observed negative correlation (-0.1) between SID and DID implies that the higher the popularity of a user is, the less popular the users s/he follows will be. Overall, disagreement is observed between the degree assortativity in Mastodon and the ones shown by well-known social networks. Finally, Zignani et al. found that users' hubiness is bounded within its instance and influenced by the latter. Also, in [246], the authors investigate how the decentralization process affects relationships between users, unveiling that instances show individual footprints (based on degree distribution and clustering coefficient statistics observed on the top 10 instances in Mastodon) that influence relationships.

Recently, Nicholson et al. [171] explored the rules declared by the most relevant Mastodon instances, also comparing them with Reddit's content moderation.

Studies on other decentralized social platforms were carried out by Bielenberg et al., who focused on Diaspora providing insights into the implementation, topology, and user growth of this platform [27]. More recently, Hassan et al. studied Pleroma to evaluate the impact of "decentralized" content moderation on users [108] and the effect of historical events on its evolution [107]. Zia et al. [28] explored the propagation of toxic content in Pleroma, also proposing a novel detection model. Finally, Hassan et al. [10] studied the issue of decentralized moderation in the Pleroma platform.

2

Understanding the growth of the Fediverse through the lens of Mastodon

Summary. Mastodon is the most relevant platform in the Fediverse to date, and also the one that has attracted attention from the research community. Existing studies are however limited to an analysis of a relatively outdated sample of Mastodon focusing on few aspects at a user level, while several open questions have not been answered yet, especially at the instance level.

In this work, we aim at pushing forward our understanding of the Fediverse by leveraging the primary role of Mastodon therein. Our first contribution is the building of an up-to-date and highly representative dataset of Mastodon. Upon this new data, we have defined a network model over Mastodon instances and exploited it to investigate three major aspects: the structural features of the Mastodon network of instances from a macroscopic as well as a mesoscopic perspective, to unveil the distinguishing traits of the underlying federative mechanism; the backbone of the network, to discover the essential interrelations between the instances; and the growth of Mastodon, to understand how the shape of the instance network has evolved during the last few years, also when broading the scope to account for instances belonging to other platforms. Our extensive analysis of the above aspects has provided a number of findings that reveal distinguishing features of Mastodon and that can be used as a starting point for the discovery of all the DOSN Fediverse.

2.1 Contributions

Our research stems from a twofold motivation: to provide a fresh view on Mastodon based on recently updated data, and to fill a lack of knowledge on topological features of the Mastodon network focusing at the *instance* level.

As previously discussed, early studies have primarily focused on the analysis of Mastodon users, and they captured a relatively small snapshot of Mastodon dated four years ago. Clearly, this might have overlooked salient traits that can be discovered at the instance level, as well as it raises the need for getting a timely picture of Mastodon which has presumably changed over time. To overcome these limitations, our study builds upon an up-to-date and representative network data over the instances, and utilizes it to provide insights into their relations. The goal is manifold: it includes the opportunity of enhancing our understanding of the *macroscopic* and *mesoscopic* structures of Mastodon to unveil the distinguishing traits of the underlying federative mechanism, and to discover the essential interrelations between the instances; but also we want to understand how the instance network has changed, within Mastodon as well as at the boundary of Mastodon itself.

We elaborate on the above aspects by developing an extensive analysis framework to answer the following research questions:

Q1 – *Network data and models*: How are the Mastodon instances detected and modeled as a network?
Q2 – *Structural features*: What are the salient structural features of the network of Mastodon instances, at *macroscopic* as well as *mesoscopic* level?
Q3 – *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?
Q4 – *Network backbone*: What is the backbone of the network of Mastodon instances, and does it preserve the structural features of the whole network?
Q5 – *Growth*: How has the shape of the network of Mastodon instances evolved during the last few years?

Plan of the Chapter. The remainder of this Chapter is organized so as to address the above stated research questions. Section 2.2 describes our crawling methodology, the data collected and the network models we built upon this data (**Q1**). Section 2.3 contains the structural analysis of the Mastodon instance network, from the macroscopic and mesoscopic perspectives (**Q2**-**Q3**). Section 2.4 describes our methodology of identification of the backbone of the Mastodon instance network (**Q4**). Section 2.5 analyzes the evolution of the Mastodon instance network (**Q5**) from three points of view: comparison with the earlier Mastodon network, emphasis on the online portion of the network, and an analysis of centrality of the instances. Finally, Section 2.6 concludes the work and provides pointers for future research.

2.2 Polite data crawling and network modeling

To answer our first research question (**Q1**), here we describe the crawling methodology adopted to collect public information from Mastodon, the steps carried out to build and validate our Mastodon instance dataset, and the network models we derived from the collected data.

2.2.1 Crawling methodology

The publicly available dataset on Mastodon relationships provided in [245] contains data extracted during the period between 2017 and 2018. This clearly raises concerns about the possibly partial obsolescence of those data, since social networks continuously evolve and there is no reason to assume that Mastodon and the Fediverse would represent an exception to this rule. Therefore, to satisfy the need for up-to-date data, we carried out an extensive crawling phase based on a newly designed crawler.

Crawling requirements and design principles. Our crawler was developed under strict and self-imposed constraints, i.e., following the *privacy by design, privacy by default* approach, and exclusively relying on the publicly-available Mastodon REST APIs [1] — using such APIs, we accessed data through *GET* and *POST* methods of the HTTP protocol, and managed the payload of requested data in a JSON format. Under these constraints, we were able to make our crawling methodology fully compliant with ethical and privacy-related principles.

[1] `https://docs.joinmastodon.org/api/`

Given the decentralized nature of Mastodon, it is not straightforward to detect the myriad of instances available today. Nonetheless, to get updated information on the current landscape of Mastodon instances, the *instances.social* website [2] is commonly used as a de-facto tracker of Mastodon instances. We exploited it to generate a list of seeds (i.e., starting points for the searching process), which correspond to currently online Mastodon instances.

Mastodon instances provide developers with *authentication tokens* to ensure control over the scope of the interactions. Moreover, by leveraging on authenticated requests, developers might achieve better interaction capabilities with instances. These conditions certainly comply with our desired privacy and ethical principles. Therefore, we submitted our seed list (i.e., the instances obtained from *instances.social*) to the authentication process, getting approved from approximately 900 instances out of about 1 100. Also, being able to traverse instances timeline — via authenticated requests — we discovered about 81 000 new users to explore. Then, we carried out a *breadth-first-search* over them, detecting incoming and outgoing links and progressively increasing the number of users to explore, by discovering new ones during the link detection process.

We point out that Mastodon allows redirecting or moving a user's profile. Although notable, this feature could determine some inconsistencies during the crawling, such as redirects to other instances while exploring the user profiles. Therefore, we avoided generating edges for users who presented similar behaviors.

Moreover, two side yet relevant remarks arise regarding our crawler implementation. First, to efficiently handle the collected data, we used a caching mechanism (*Redis*) coupled with a NoSQL database (*MongoDB*). Also, to prevent computational bottlenecks, we avoid repeated checkings over the database during the crawling phase (e.g., for checking duplicated edges), so that we eventually refine the complete network dataset in an "off-line mode", once the crawling process has finished, by exploiting particularly efficient processing functionalities provided by suitable data and network manipulation software libraries.

Our crawling session ended up with 27 989 557 links detected. After performing basic data-cleaning steps, particularly removing duplicate links, we obtained about 1.4M and 18M unique users and links, respectively, managing to cover 16 282 instances.

It should be emphasized that, to respect privacy principles, we firmly avoided using scraping techniques or systems, i.e., we abstained crawling information from instances which did not provide us with an authentication token. Notice also that the detected links were immediately anonymized, and any information that could impact on the users' privacy was replaced with numerical data generated through a proper hashing function — as a consequence, it is not possible to trace back the original information on users from our raw dataset. Finally, we point out that our fetching of descriptive text data (e.g., toots) was *minimal*, i.e., it occurred only during the initialization of our crawling process: indeed, we produced the seed-user set by discovering them through toots available in the timelines of the seed instances, relying only on authenticated requests. Nonetheless, we never stored this data since we processed it in real-time. After this initial phase, the crawling continued via breadth-first-search, thus ignoring textual data.

Spotting Mastodon instances. As previously mentioned, platforms in the Fediverse utilize a shared protocol, allowing for seamless interactions among their users. A related key-aspect is that, when requesting followings or followers of a Mastodon user, the APIs return all of them, regardless of the Fediverse platform. In this regard, one question becomes how we can distinguish between instances that belong to Mastodon from other platforms' instances in the Fediverse. We answered this question through a verification process, as summarized next.

[2] `https://instances.social/`

Table 2.1. Current landscape of Mastodon instances as provided by *instances.social* and *fediverse.party* websites. Symbol $\bigcup$, resp. $\bigcap$, stands for the total of Mastodon instances calculated as the size of the set union, resp. intersection, between the instances set provided by the websites.

	instances.social	*fediverse.party*	$\bigcup$	$\bigcap$
Online	1 193	not available	1 193	997
Online+Offline	7 313	3 396	9 433	1 276

To date, some relevant websites provide up-to-date Mastodon information, namely the aforementioned *instances.social* and *fediverse.party*,[3] so that we exploited them to filter our data through their lists of known instances. Note however that, while *fediverse.party* does not distinguish online from offline instances, *instances.social* provides fine-grained filtering capabilities. In this regard, we focused on the setting of two main parameters provided in the *instances.social* APIs: *include_dead* and *include_down*. As declared in *instances.social*, an instance is considered *dead* if inactive for at least two weeks, and *down* if it is not currently online yet live within a two-week window. We set either both *include_dead* and *include_down* to *true*, or *false*, to obtain all the known Mastodon instances, resp. online-only instances.

As reported in Table 2.1, we merged information retrieved from both websites, regardless of the instances' status (i.e., online or offline), obtaining 9 433 known Mastodon instances. In addition, we requested online-only ones to *instances.social*, getting 1 193 instances to date. Note also that, by restricting our census of the Mastodon instances to the information shared between the two platforms, the total online and the overall total would be decreased of 16.4% (997) and 86.5% (1 276).

Validated datasets. Based on the above information, we analyzed our crawled data to properly detect the status of the instances. Results are summarized in Figure 2.1.

We intercepted 6 960 out of 9 433 Mastodon instances (both online and offline), and 1 116 out of 1 193 currently online instances. We point out the significance of the latter value, given the coverage of most of the online Mastodon instances to date. Moreover, our dataset doubles the earlier state-of-the-art in terms of currently online instances. Clearly, the freshness of our data (November-December 2020) influences this value.

It should be emphasized that our collected data includes a remarkable amount (9 322) of non-Mastodon instances, i.e., belonging to other Fediverse platforms. This clearly strengthens the concept of Fediverse, but also opens to the discovery of the primary role taken by Mastodon within the Fediverse. In fact, although we detected non-Mastodon users and instances through Mastodon ones, and hence this knowledge of the Fediverse might be partial, our collected data offers an unprecedented opportunity for deepening our understanding of the position of Mastodon in the Fediverse (i.e., how Mastodon instances and users interact with the rest of the Fediverse), given the premises of independence yet cooperation among platforms in the Fediverse.

Further important remarks also arise regarding the growth of Mastodon. Although we are not aware of the status (i.e., online or offline) of the instances in the earlier state-of-the-art dataset [245] at the time of their creation, we hypothesize that, after a first boost due to enthusiasm and novelty, Mastodon reached its stability as a DOSN. Indeed, the number of currently online instances is moderate compared to the number of all-time known ones and refers to non-transient yet well-rooted platforms. Overall, our dataset turns out to be

[3] `https://fediverse.party/en/mastodon`

	Total collected	*Mastodon Overall*	*Mastodon Online*	*Time*
This work	16 282	6 960	1 116	Late 2020
Earlier [245]	not available	4 015	548	Mid 2017 - Early 2018

Fig. 2.1. Validated data based on the information reported in Table 2.1, and illustrative comparison between dimensions of the earlier state-of-the-art (in red) and dimensions of our dataset.

significantly larger and more recent than the earlier Mastodon dataset, making it more suitable for novel and further studies.

2.2.2 Network models

Let us denote with $\mathcal{U}$ the set of users and with $\mathcal{I}$ the set of instances available in the extracted Mastodon data. We can define a directed network modeling the Mastodon data as $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, 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 concide with v provided that $i \neq j$. Given $\mathcal{G}$, we derive three Mastodon networks at instance level, which are formally defined as follows.

Instance network. The first network we define is the graph modeling relations between all the instances in $\mathcal{I}$, hereinafter referred to as Instances network, as the directed weighted network $G_{\mathcal{I}} = \langle V, E, w \rangle$, where $V = \mathcal{I}$ is the set of nodes, E is the set of edges such that $(i, j) \in E$ means that 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).

Online instance network. Our second network is induced from the set of instances that are detected as online at the time of the crawling process we carried out. Therefore, 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.

Expanded network. 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,

Table 2.2. Networks created from our collected dataset, and comparison with the earlier state-of-the-art network. All networks but EXPANDED refer to Mastodon-only instances.

Network name	*#Nodes*	*#Edges*
EXPANDED-INSTANCES	16 282	318 218
INSTANCES	6 960	216 504
ONLINE-INSTANCES	1 115	75 046
Earlier [245]	4 015	95 221

we also define 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, we define the EXPANDED-INSTANCES network 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 Mastodon instances network.

All the above defined networks and the one inferred from the earlier dataset are summarized in Table 2.2. Note that the number of nodes in the ONLINE-INSTANCES network is decreased of one w.r.t. the information given in Fig. 2.1, since one online instance is actually disconnected from the network. Moreover, in Figure 2.2, we provide an illustration of the overall network of instances created from our collected dataset, i.e., EXPANDED-INSTANCES.

Remarkably, the *Earlier* network is found to be mostly contained in our INSTANCES network; more precisely, about 80% of the instances in the *Earlier* network are also contained in our INSTANCES. As we shall discuss later in this work, this has important implications in the growth of Mastodon during the last three years.

It should also be noted that our collected data allows us in principle to build networks at the *user* level as well, e.g., we could define the network of the relations between users of each particular instance; nonetheless, this goes beyond the scope of this work, whose focus is the analysis and understanding of the relations among the instances in Mastodon. Therefore, we leave the study of instance-specific networks of users as future work (cf. Section 2.6).

2.3 Structural analysis of the INSTANCES network

In this section, we answer our second research question (**Q2**) by presenting an extensive analysis of the network we built over the Mastodon istances, i.e., the previously introduced INSTANCES network. To unveil the main characteristics that define Mastodon, we will take a macroscopic as well as a mesoscopic perspective, and organize the discussion into the two next subsections.

2.3.1 Macroscopic structural analysis

We begin our investigation of the INSTANCES network at a macroscopic level. We refer to Table 2.3 for a summary of statistics on the main structural characteristics of the INSTANCES network, each of which is analyzed in the following.

Degree distribution. Figure 2.3 shows the boxplot, density function, and Complementary Cumulative Distribution Function (CCDF) of the INSTANCES network in-degrees, with various types of distribution fittings; results obtained for the out-degrees and the total degrees are analogous, and we report them in ***Appendix A.1***.

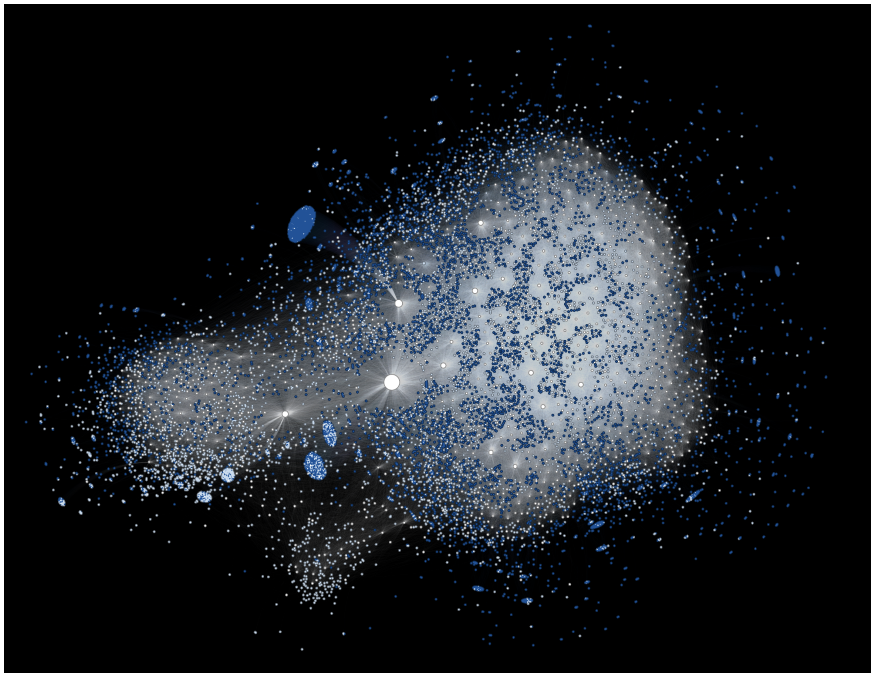


Fig. 2.2. Illustration of the EXPANDED-INSTANCES network. Node colors denote different types of instances: white and light blue indicate online and offline Mastodon instances, respectively, whereas dark blue corresponds to non-Mastodon instances. Node size is proportional to its degree. The displayed layout is based on the force-directed drawing ForceAtlas2 model. (*Produced by using the Graphistry service, available at https://www.graphistry.com/.*)

Looking at the boxplot and the histogram with associated density function, there is evidence of right-skewness of the in-degree distribution, with a small bunch of "outliers" scaling in the regime of thousands. In detail, while the first quartile, median, mean, third quartile, and non-outlier maximum degree are 91.5, 202, 331, 413, and 878, respectively, there are 21 instances having in-degree above 880, up to a maximum degree of 4 685.

We investigated about the outlier instances, focusing on the top-5 by degree, resp. in-degree and out-degree, as shown in Figure 2.4. We found these correspond to *mastodon.social*, *pawoo.net*, *mastodon.xyz*, *octodon.social*, and *mstdn.io*, in all the three cases. These are clearly among the most popular instances in Mastodon and, as expected, they allow their users to discuss a large variety of topics. Interestingly, *mastodon.social* is always the top-1 instance regardless of the type of degree, whereas *pawoo.net* and *mastodon.xyz* alternate each other at the second and third rank.

Such instances are also well-recognized in the CCDF plot, where we observe a probability of 50% of having at least 200 in-degree, which already drops to 20% for an in-degree around 600, and further decreases below 4% for the outliers. The CCDF plot also displays the best fitting of power-law, lognormal, exponential and Poisson distribution to the observed data. The resulting fitting curves appear to provide indications of lognormality and, to a limited extent, of power-law fitting.

Table 2.3. Summary of structural characteristics of the INSTANCES network, including details on community structure and core decomposition.

	INSTANCES	INSTANCES inner-most core		
		degree	*in-degree*	*out-degree*
#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
average degree*	41.966	152.328	157.702	150.98
average in-degree	31.107	136.455	137.933	135.015
% sources	12%	0%	0%	0%
% sinks	6.6%	0%	0.005%	0%
degree assortativity*	-0.274	-0.117	-0.158	-0.135
degree assortativity	-0.253	-0.14	-0.171	-0.151
average path length	2.330	1.270	1.330	1.310
diameter	5	2	2	2
transitivity*	0.128	0.832	0.798	0.807
clustering coefficient*	0.836	0.837	0.810	0.816
clustering coefficient *(full averaging)**	0.687	0.837	0.810	0.816
#strongly connected components	1 305	1	2	1
#weakly connected components*	1	1	1	1
modularity by *Louvain**	0.289	0.032	0.039	0.037
#communities by *Louvain**	5 (5)	3 (3)	3 (3)	3 (3)
modularity by *Louvain***	0.353	0.242	0.246	0.246
#communities by *Louvain***	6 (8)	4 (5)	3 (4)	4 (6)
#communities by *Infomap***	6 (54)	1 (3)	1 (4)	1 (3)

* Statistic calculated by discarding the edge orientation
** Statistic calculated by taking into account the edge weights

The above prompted us to assess the corresponding statistical significance, whereby we resorted to a Kolmogorov-Smirnov test. Results are summarized in Table 2.4. In the first subtable, the high p-values suggest that the null hypothesis that the data are from a power-law distribution cannot be rejected, although this holds on a limited regime ($x_{\min}$) starting from degree values, resp. in-degree and out-degree values, of the order of hundreds. In particular, for the in-degree case, note that $x_{\min}$ is above the mean of the distribution. The remaining subtables in Table 2.4 correspond to four different scenarios we investigated for the lognormality fitting, namely (from top to bottom in the table) full regime (i.e., whole observed data), removal of the outliers, removal of lower-degree (≤ 50) instances, removal of both outliers and lower-degree instances. As it can be noted, the Kolmogorov-Smirnov test yielded high significance values for the lognormality fitting in all cases, except when the outliers only are discarded; particularly, the significance is maximized when the lower-degree instances are removed (i.e., p-values from 0.687 to above 0.9). In such cases, the test informs us that we cannot reject the null hypothesis and so we conclude that the observed data are lognormally distributed.

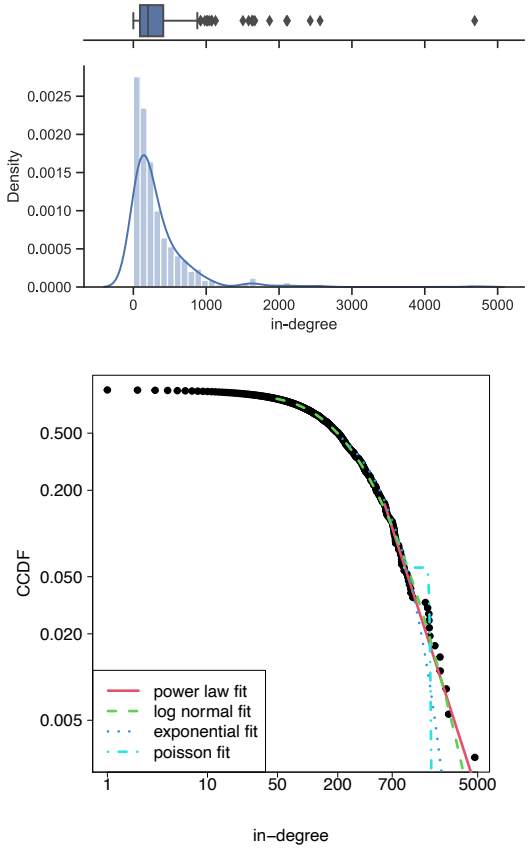


Fig. 2.3. INSTANCES in-degree distribution: boxplot and Probability Density Function (top), and Complementary Cumulative Distribution Function, with various distribution fittings (bottom).

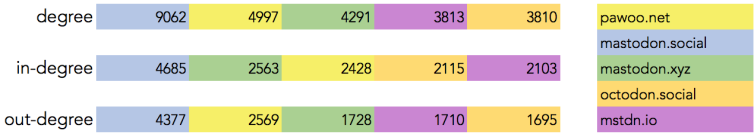


Fig. 2.4. Top-5 instances by degree, in-degree, and out-degree, respectively, in the INSTANCES network.

Sources and sinks. We inspected the presence of instances having no incoming links (i.e., sources) as well as of instances having no outgoing links (i.e., sinks). As reported in Table 2.3, the percentage of both types of instances is not negligible, with the incidence of sources being nearly double than sinks. This might provide clues for the presence of users belonging to small instances (e.g., private ones) interested in contents produced by users located in other instances; indeed, since small instances would host few users, the lack of incoming links is plausible.

Table 2.4. Power-law and lognormal fittings through *Kolmogorov-Smirnov* test performed on the INSTANCES network.

		degree	*in-degree*	*out-degree*
power-law	x_{min}	890	588	457
	α	2.987	3.166	3.057
	p-value	0.939	0.889	0.936
lognormal	interval	[1, 9 062]	[1, 4 685]	[1, 4 377]
	μ, σ	5.54, 1.26	5.18, 1.24	5.11, 1.19
	p-value	0.113	0.187	0.113
	interval	[1, 1 307]	[1, 878]	[1, 742]
	μ, σ	5.37, 1.15	5.05, 1.15	4.96, 1.09
	p-value	0.01	0.054	0.024
	interval	[51, 9 062]	[51, 4 685]	[51, 4 377]
	μ, σ	5.82, 0.95	5.53, 0.87	5.45, 0.81
	p-value	0.963	0.687	0.932
	interval	[51, 1 307]	[51, 878]	[51, 742]
	μ, σ	5.65, 0.80	5.40, 0.74	5.31, 0.67
	p-value	0.47	0.646	0.816

On the opposite side, the percentage of sink instances sheds light on that they might contain well-consolidated user groups, among which there is no need to interact with users belonging to other instances. This peculiarity might suggest sort of self-sufficiency in the federation.

Triadic closure. We analyze how well the triadic closure principle is met in the INSTANCES network, by looking at both transitivity (i.e., the probability that two incident edges are completed by a third one to form a triangle) and local clustering coefficient (i.e., how strongly connected are the neighbors of a node).

We observe a rather low value of transitivity (0.128), which is actually not surprising given the relatively low density of the network. By contrast, local clustering coefficient is very high (0.836), and remains as such even when accounting for sink or source instances (0.687). This evidence is remarkable as it hints at a federative structure among the instances. Note also that the dichotomy between a relatively lower transitivity and a higher local clustering coefficient, and more in general, the low correlation between the two statistics is typical of networks characterized by a skewed degree distribution. In this respect, the Mastodon INSTANCES network also keeps this feature.

As a further remark on length-2 closed loops (i.e., reciprocal edges), we observe a high fraction of reciprocal edges (above 65%). As we shall further observe through our core decomposition analysis, reciprocity tends not to be limited to users within the same instance, but involves instances that can be placed very differently, from the periphery to the internal of the network, and vice versa.

Degree assortativity. One key structural property of a network at macroscopic level refers to degree correlation, or degree assortativity, which measures how the probability of a link between two nodes depends on their degrees [169, 170]. Real-world social networks are often found to have positive degree assortativity, i.e., well-connected individuals are linked to other well-connected ones. A recent study has also shown that this evidence does actually hold for those social networks built upon shared memberships of group [82].

Remarkably, the Mastodon INSTANCES network exhibits a degree assortativity which is significantly negative (-0.253), which means that well-connected instances are connected to many instances with few other connections. This might be ascribed to the heterogeneous degrees characterizing the instances in the network. In this respect, we argue that, since instances can be bounded to specific topics, users belonging to different instances with diversified degrees tend to interact with each other to reach a broader range of contents, thus ultimately improving their experience on the platform and increasing the speed of information transfer.

The degree disassortativity, i.e., negative degree correlation, exhibited by the Mastodon INSTANCES network outlines a novelty w.r.t. well-known centralized social networks. It should be noted that, unlike centralized social networks, Mastodon users' behavior is not impacted by recommendation mechanisms. Therefore, the followships tend to be built upon the topical interests and preferences that users have, which leads to a form of topically-induced link formation rather than a popularity-based attachment. Moreover, this further supports the strong interrelation between instances which, as we shall discuss in the next section, characterizes peripheral and inner-core locations in Mastodon.

2.3.2 Mesoscopic structural analysis

We organize our presentation of the mesoscopic structural analysis of the INSTANCES network into two parts: the first one is devoted to the evaluation of the community structure that is detected over the network, whereas the second part is concerned with the core decomposition of the network.

Community Detection. We resorted to two well-known community detection algorithms for discovering communities in the Mastodon network of instances, i.e., the Louvain method [30] and the Infomap method [190]. Louvain is a two-step, hierarchical greedy optimization method that attempts to maximize the modularity of a partition of the network, whereas Infomap optimizes the Map equation, which exploits the information-theoretic duality between finding community structure in networks and minimizing the description length of a random walker's movements on a network. It should be noted that both methods have been used with success for networks of many different types and sizes, and today, they are the most widely used methods for detecting communities in large networks. For the Louvain algorithm, we exploited both the original, undirected implementation as well as the directed variant.[4] In the latter case, we also took into account the edge weights. As regards Infomap, we exploited its weighted directed implementation.[5]

The number of communities found by the aforementioned algorithms is shown in Table 2.3, for all considered scenarios (cf. notes marked with * and ** below the table). Please note that we report two values for each case: the one within parenthesis corresponding to the total number of communities while the first value refers to the number of communities that contain at least ten instances. Furthermore, since the Louvain algorithm optimizes modularity, we also report the modularity values corresponding to the community structures discovered by Louvain; in this regard, we find evidence of modular structure within the INSTANCES network, with modularity from about 0.29 (undirected network) to 0.35 (weighted directed network).

The number of communities produced by the Louvain method ranges from 5 for the undirected scenario, to 8 when applying the weighted directed variant; in the latter case, only 2 out of 8 communities contain less than ten instances. Infomap, conversely, appears to detect

[4] `https://github.com/nicolasdugue/DirectedLouvain`

[5] `https://www.mapequation.org/infomap/`

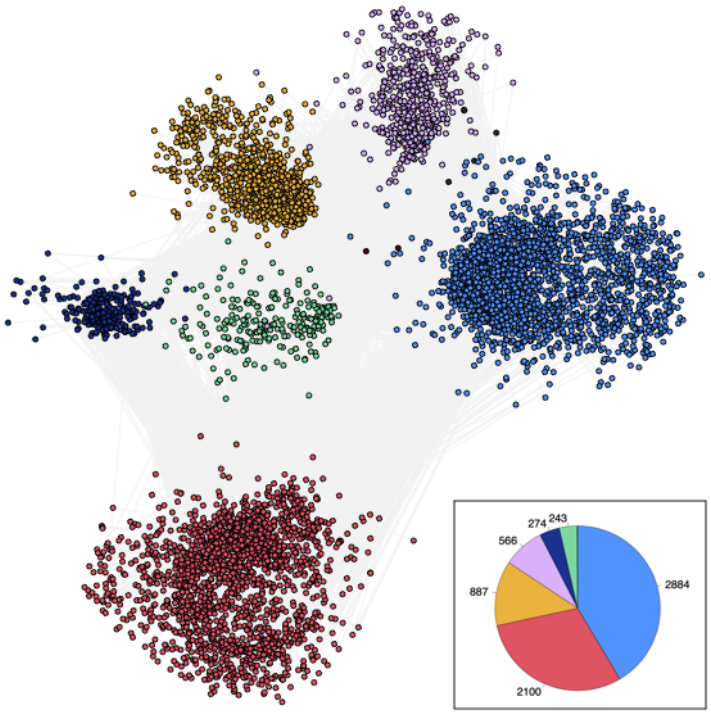


Fig. 2.5. Community structure on the INSTANCES network obtained by the directed Louvain method. The displayed layout is based on the force-directed drawing Fruchterman-Reingold model, with weight 150 for edges incident nodes in the same community and weight 1 for edges incident nodes in different communities. The pie-chart in the bottom-right corner shows the size proportion of the communities, along with the size values for the top six largest communities.

a much higher total number of communities; however, by inspectioning them, we found out that a large majority are poorly significant as they consist of less than ten instances. Indeed, as reported in the table, we point out that the two methods actually behave the same in terms of number of relatively large communities (i.e., 6 communities produced by either method considering weights and orientation of edges).

Figure 2.5 illustrates the communities detected by the directed Louvain method on the INSTANCES network. As it can be observed, two main communities arise in terms of size (displayed at the bottom and on the right in the figure), which together contain about 72% of the instances in the network. Other two communities (on the top in the figure) also stand out, as they contain about 21% of the instances. Remarkably, albeit less evident for the smallest communities, we can observe high connectivity between all of them, which hints at the high interrelation between instances belonging to different communities.

We also delved into the community structure obtained via the Louvain algorithm to get more insights into the community boundaries. The largest community (i.e., the rightmost one in Figure 2.5) contains the most relevant Mastodon instance in the Fediverse, namely

Table 2.5. Unweighted and weighted *conductance* scores for the community structures obtained by Louvain and Infomap methods on the INSTANCES network. The first row refers to the average over all pairwise scores between communities, whereas the other rows refer to comparisons between the three largest communities detected by the methods. Weighted variants of conductance account for edges weight when calculating volumes and cuts.

	Louvain		Infomap	
	unweighted	*weighted*	*unweighted*	*weighted*
INSTANCES	0.285	0.239	0.037	0.035
Top-1 vs. Top-2	0.315	0.191	0.370	0.219
Top-1 vs. Top-3	0.816	0.881	0.561	0.721
Top-2 vs. Top-3	0.100	0.059	0.231	0.335

mastodon.social, which represents the first instance born with the Mastodon project. Consequently, given its role as a reference point in Mastodon, the large constellation of instances observed around it is not surprising. Within the same community, we spotted *mstdn.io* and *octodon.social*. The above three instances are not topically bounded and use English as the primary language, with *mstdn.io* also embracing French and *octodon.social* extending its range of languages to Japanese and Portuguese. We point out that these instances might share the same community given their relevance and longevity (e.g., *mstdn.io* is up since early 2017) in the Fediverse. Moving our focus to the second-largest community in size (i.e., the one at the bottom of the figure), we distinguished two relevant instances, namely *pawoo.net* and *mstdn.jp*. Their co-existence within the same community is justified by the fact that both discuss various topics and share the official language (i.e., Japanese). Furthermore, in the third-largest community, we located the remaining instance, i.e., *mastodon.xyz*, of the previously mentioned top-5 largest ones in Mastodon. This is general-purpose and uses English and French as primary languages. As a final remark, we point out that all the top-5 instances reported in Section 2.3.1 are established in the largest three communities discovered by the Louvain algorithm, and hence their relevance is further strengthened due to their central role within these communities.

We replicated the same explorative analysis also for the community structure detected by Infomap. Remarkably, mostly interesting patterns are found again. Indeed, the largest community still includes the same instances as in the case of Louvain, with the addition of *mastodon.xyz*, which in the Louvain solution is included in the third largest community. Moreover, this similarity aspect also holds for the second largest community detected by Louvain and Infomap, respectively, which comprises the two Japanese instances. These coherent findings from two different community detection methods would support an underlying logics in the organization of the instances and their interrelations.

We further investigated the community structures produced by Louvain and Infomap according to the aspect of *conductance*. This is defined as the ratio between the cut size among two communities (i.e., the sum of the weights of the edges that link two sets of nodes) and the smaller of the volumes (i.e., the sum of the degrees of the nodes in a set) of the two communities; we considered both the weighted and unweighted versions of conductance, where for the latter, the edge weights are equal to one.

As reported in Table 2.5, conductance varies depending on the community detection approach. Both methods induce good separation among communities in the INSTANCES network, which is indicated by the low values of conductance, but in the Infomap solution this is much more evident than in the Louvain community structure. Taking a finer-grain perspective with a focus on the three largest communities produced by the two methods, respectively, the

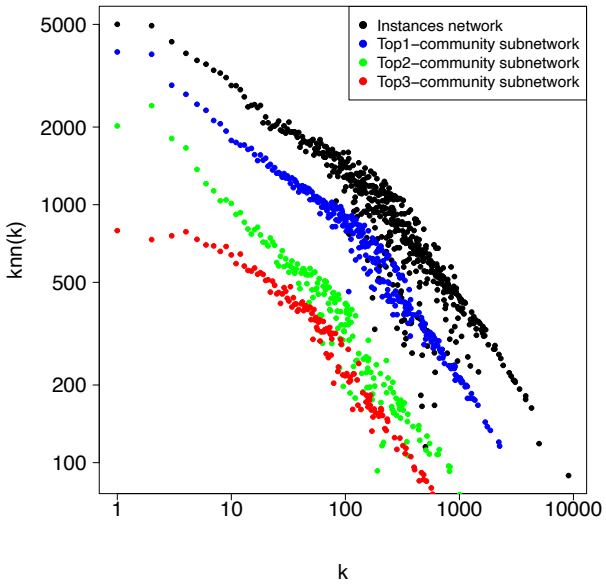


Fig. 2.6. Illustration of the average nearest neighbor degree $knn(k)$ as a function of the degree k. Values on the x-axis and y-axis are log scaled. The three largest community subnetworks are extracted from the structure obtained by the weighted directed Louvain method on the INSTANCES network.

analysis of the pairwise conductance among such communities yields two main outcomes: the conductance between the largest and the third largest communities ranges from about 0.56 (0.72, for the weighted version) w.r.t. Infomap to above 0.8 when using Louvain; by contrast, regardless of the community detection method, both unweighted and weighted conductance is much lower for the pairs involving the second largest community. We tend to ascribe these differences in conductance since, on the one hand, the largest communities have significant inter-community communication flow (and hence, high cut size) due to the involvement of most relevant instances over the network; on the other hand, this inter-community connection may be limited due to different cultures and languages, as it is the case for the second largest community which indeed is centered around the Japanese-language *pawoo.net* and *mstdn.jp* instances.

The above remarks also prompted us to measure dependencies between degrees of neighbor nodes in the communities, by computing the *average nearest neighbor degree* distributions. Figure 2.6 reports the values of the average nearest neighbor degree as a function of the degree k, which we denote by $knn(k)$. As it can be noted, the decreasing trend by k is not only clear for the whole INSTANCES network — which is indeed aligned with the negative degree assortativity previously analyzed — but also for the subnetworks induced from the top three communities exhibiting disassortative traits.

Core Decomposition. The core decomposition of a network graph consists in assigning each node with an integer number (the core index) capturing how well the node is connected with respect to its neighbors. The result is a threshold-based hierarchical decomposition of

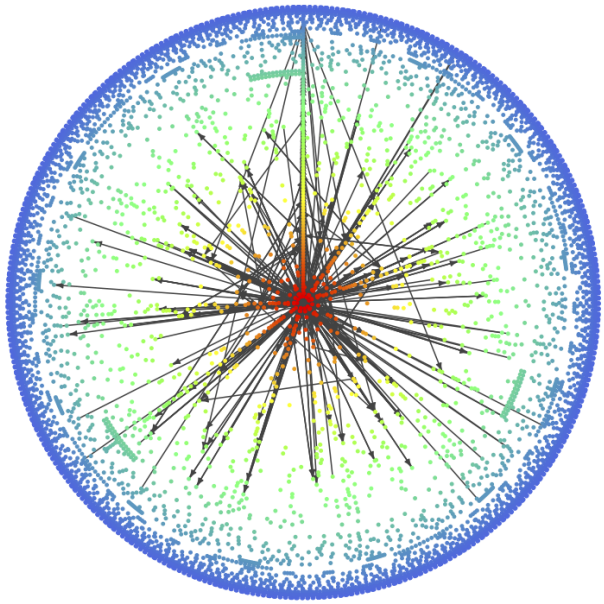


Fig. 2.7. Core decomposition of the Instances network, based on node in-degrees. Nodes having the same core-index are assigned the same color (inner-most, resp. outer-most core correspond to red, resp. blue). To avoid cluttering, only edges having a weight greater than the first quartile of (unique) edge weights are displayed.

the graph into nested subgraphs, based on a threshold (k) which is set on the degree of nodes [199]. The identification of such tightly-knit substructures, or cores, has long been used for understanding mesoscale structural characteristics of a network, with several applications related to the computation of the local importance of nodes [154], including the estimation of the spreading potential of nodes [129, 40]. A key advantage of core decomposition lays on theoretically grounded definition and uniqueness of its solution, which can also be computed efficiently in linear time w.r.t. the number of edges in the input graph.

Given a graph $G = \langle V, E \rangle$ and any subset $S \subset V$, let us denote with $G[S] = \langle S, E[S] \rangle$ the subgraph of G induced by S, where $E[S] = E \cap (S \times S)$. For any choice of an integer value $k \geq 0$, the k-core of a graph is the maximal induced subgraph $G[C_k] = \langle C_k, E[C_k] \rangle$ such that the number of neighbors of every node v in C_k is at least k. The *degeneracy* K of the graph is the highest value of k such that $C_k \neq \emptyset$. The core associated with the graph degeneracy is also called the *inner most core*. The set of all k-cores (i.e., $V = C_0 \supseteq C_1 \supseteq \cdots \supseteq C_K$) represents the core decomposition of the graph. Moreover, the *core-index*, or *coreness*, of a node v is the largest k such that $v \in C_k$ and $v \notin C_{k+1}$. Note also that the above definitions originally apply to undirected graphs, however they are straightforwardly adapted to directed graphs so that the degree of a node may refer to either its in-degree or out-degree.

The core decomposition of the Instances network revealed further hints at the presence of a federative mechanism. As reported in Table 2.3, one major finding is the remarkable number of instances in the inner cores of the network, and even in the inner-most core, which ranges from 189 based on total degree to 208 based on in-degree, and also includes the top-5 instances

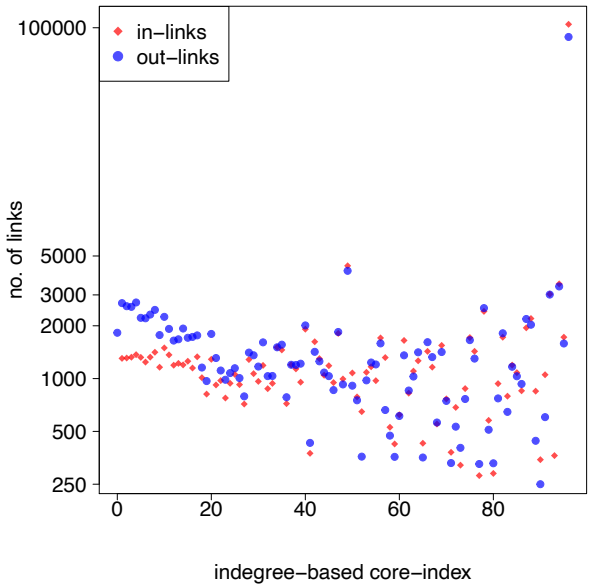


Fig. 2.8. Comparison between no. of in-links and out-links for each indegree-based core-index of the *instances* network. Values on the y-axis are log scaled.

previously discussed in Section 2.3.1. This prompted us to explore the inner-most core of the INSTANCES network, in its three variants, i.e., via a total-degree-based, in-degree-based, and out-degree-based decomposition of the network.

A few remarks arise from the observation of the inner-most cores, such as the high degeneracy under all the considered scenarios: indeed, we found out the values 201, 96, and 97 of degeneracy corresponding to a core decomposition based on total degree, in-degree, and out-degree, respectively. As expected, density is high, and so is also the reciprocity — note that statistics still refer to directed subgraphs, although the core-decomposition may have been generated based on the undirected (i.e., total degree-based) definition. Although less evident, the disassortative trait is still present, denoting the coexistence in the inner-most core of instances exhibiting diversified yet high degrees. Given the high density, average path length and diameter almost halve compared to their original values. According to a more cohesive structure, transitivity significantly increases (nearly seven times the value in the whole network) and aligns with the local clustering coefficient. Also, the number of strongly connected components (i.e., #SCCs) heavily shrinks to 1, resp. 2, for the indegree-based, resp. degree- and out-degree-based decomposition.

We also evaluated the presence of a community structure in the three variants of inner-most core. The (undirected) Louvain method determines an exceptional reduction in modularity compared to the value found for the whole network (i.e., from nearly 0.3 to 0.03-0.04), along with a small decrease in the number of significant communities (from 5 to 3). By contrast, when considering edge orientation and weights, we observe a slight decrease in modularity (from about 0.35 to 0.25) and community structure size (from 8 to 4-6). Interestingly, the number of communities detected by Infomap strongly decreases compared to the value observed for

the whole network, with one giant community and the remaining 3-4 containing less than ten instances.

Our major focus in this analysis was about whether there exists a significant amount of connections between the periphery and the inner parts of the INSTANCES network. To this purpose, we started with a visual exploration of the cores of the network, whose results are shown in Figure 2.7 for the indegree-based decomposition. Looking at the figure, the display of the cores unveils evidence of directed links with a relatively high weight (i.e., solid lines in the chart), which correspond to followships from users of instances in the inner-most core to users of more peripheral instances, but also followships in the opposite direction, even coming from the outer-most cores towards the inner-most core. This certainly relates to the negative degree assortativity previously emphasized, and it represents quite a novel pattern in social networks, which usually do not show direct links between core and peripheral nodes.

We further delved into this trait by examining the number of incoming links and outgoing links for each core-index, as depicted in Figure 2.8. Lower core-index values (up to about 20) correspond to a number of outgoing links that is significantly higher than the number of incoming links, which is expected (i.e., instances within these cores tend to interact with more relevant ones, while the opposite occurs less, or not at all). However, as the core-index value increases, unveiling an interesting feature of bidirectionality in the followships behavior for mid and internal cores. Further interestingly, some instances with high core-index (i.e., from about 70) show a prevalence of outgoing links. A final, remarkable trait of the INSTANCES network is that a major portion of connections (i.e., about 190 000, which is more than 87% of the total edge set size) depart from and arrive to the inner-most core.

2.3.3 Discussion

Our analysis of the structural characteristics of the INSTANCES network has revealed noteworthy features, which might set Mastodon apart from well-known centralized social networks. Here, we summarize such features, thus answering our third research question (**Q3**) on the "fingerprint" of the network of Mastodon instances.

The evaluation at the macroscopic level has unveiled various unique traits of Mastodon. Interesting aspects are already present in the degree distribution of the instances: indeed, as opposed to well-known centralized social networks that tends to fit a power-law distribution, Mastodon reveals a better and extensive fitting with a lognormal distribution, along with the presence of few instances that show a degree up to one order of magnitude higher than the average degree in the network.

Referring to the federation concept as a set of independent yet cooperating instances, we have found a number of aspects (e.g., high clustering coefficient values and percentage of reciprocal edges) indicating that Mastodon adopts a federative mechanism. We mark this mechanism as a *mutual reinforcement* to reduce the sectorization bias that can characterize the individual instances: indeed, they might be topically-bounded due to the decentralized nature of the platform, however, their users generally look for a broader spectrum of topics, and hence they interact across different instances. We point out that, besides distinguishing Mastodon from other social networks, this trait is central to the platform itself, thanks to the shared protocol (i.e., ActivityPub) among instances.

Related to this mutual reinforcement is the observed negative degree correlation (i.e., degree disassortativity), which represents another distinctive Mastodon feature. This indicates that users belonging to different instances with heterogeneous degrees tend to interact with each other, aiming at achieving a better user experience and increasing the speed of information transfer.

The evaluation at the mesoscopic level has also highlighted traits that contribute to determine the fingerprint of Mastodon. Through community detection, we shed light on the modular structures within instances, which provide further clues to the existence of a cohesive and federative framework among them. Further investigations have revealed how these modules might be composed and influenced. We observed topics, languages and temporal processes (e.g., their creation time) as main influence factors. Moreover, through core decomposition, we spotted an unusual and conspicuous number of connections from the inner cores to the peripheral ones (Figure 2.7), with a peculiar bidirectional balance between links observed starting from intermediate core-index values. Additionally, we observed that the majority of links between instances involve the inner-most core.

In conclusion, we can state that Mastodon reveals clear distinctive signs compared to what is commonly observed for other social networks, making it unique in multiple aspects. Among these traits, we detected the logical emergence of a federative mechanism, which allows independent instances to cooperate with each other, thus connecting their respective users.

2.4 Backbone of the INSTANCES network

We focus here on a *network simplification* task which is designed to "prune" the network graph, i.e., to detect and remove irrelevant or spurious edges with the purpose of making it easier unveiling hidden substructures of a network. One simple solution to the above problem is to exploit information on the edge weights so to remove all edges having weight below a pre-determined, global threshold. Unfortunately, besides the difficulty of choosing a proper threshold for any input network, this approach tends to remove all ties that are weak at network level, thus discarding local properties at node level.

By contrast, a theoretically well-founded approach to filtering out noisy edges from a network is based on *generative null models*. The general idea is to define a null model based on node distribution properties, use it to compute a p-value for every edge (i.e., to determine the statistical significance of properties assigned to edges from a given distribution), and finally filter out all edges having p-value above a chosen significance level. In other terms, this allows to maintain only those edges that are least likely to have occurred due to random chance, hereinafter referred to as *backbone* of the network. Clearly, imposing lower significance level will yield to more restrictive substructures, thus giving place to a potential hierarchy of backbones.

Identifying the backbone of a network graph allows us to isolate the latent structure of the network under analysis, based on the removal of statistically insignificant edges. In this regard, statistical models for graph pruning have been conceived to deal with weighted networks, so that the node degree and/or the node strength are used to generate a model that defines a random set of graphs resembling the observed network. One of the earliest methods is the *disparity* filter [201], which evaluates the strength and degree of each node locally. The null hypothesis is that the strength of a node is redistributed uniformly at random over the node's incident edges. Unlike disparity, the null model proposed by Dianati [67] is maximum-entropy based and hence unbiased. Upon it, two models are defined: the *marginal likelihood filter* (*MLF*), which is a linear-cost method that assigns a significance score to each edge based on the marginal distribution of edge weights, and the *global likelihood filter*, which accounts for the correlations among edges. While performing similarly, the latter is more costly than the former, therefore we will consider the MLF model in our analysis.

2.4.1 Details on the disparity and MLF models for weighted directed networks

Both disparity and MLF were originally conceived for undirected networks, but they can easily be extended to weighted directed networks as well. In the following, we provide a brief review of the two models, and refer the interested reader to the original works for further details [201, 67].

In the weighted directed network scenario, let us denote with k_i^{in} and k_i^{out} the in-degree and out-degree, respectively, for any node v_i. Also, the total strength s_i associated to the node has two contributions, namely the incoming strength s_i^{in} and the outgoing strength s_i^{out}, which are obtained by summing up all the weights of the incoming or outgoing links, respectively.

The disparity model aims to preserve the edges carrying a weight that represents a local significant deviation with respect to a statistical null model for the local assignment of weights by using the disparity function. Moreover, since in-degree and out-degree distributions are assumed not to be necessarily correlated, incoming and outgoing links associated to a node are considered separately.

To assess the effect of inhomogeneities in the weights at the local level, the following functions are defined for any node v_i:

$$\Upsilon^{out}(v_i) = k_i^{out} \sum_j (p_{ij}^{out})^2, \Upsilon^{in}(v_i) = k_i^{in} \sum_j (p_{ji}^{in})^2,$$

where the summation term in the first, resp. second, expression characterizes the level of local heterogeneity in the outgoing edge weights, resp. incoming edge weights. The null model for the disparity filter requires for each incoming, resp. outgoing, link of v_i with weight w, the calculation of the p-value α_{ji}^{in}, resp. α_{ij}^{out}, that its normalized weight $p_{ji}^{in} = w/s_j^{in}$, resp. $p_{ij}^{out} = w/s_i^{out}$, is compatible with the null hypothesis. The incoming and outgoing links that carry weights that can be considered not compatible with a random distribution, can be filtered out according to a chosen significance level α; more specifically, all the incoming, resp. outgoing, links with $\alpha_{ji}^{in} < \alpha$, resp. $\alpha_{ij}^{out} < \alpha$ reject the null hypothesis and hence can be considered as significant heterogeneities.

Let T denote the number of unit edges, where each weighted edge is seen as multiple edges of unit weight. In the MLF model, the null model assigns each edge to nodes v_i and v_j which are selected independently and randomly with probabilities proportional to their outgoing and incoming strengths, respectively. More in detail, the probability that an edge from v_i to v_j is associated to them is given by $p_{ij} = \frac{s_i^{out} s_j^{in}}{T^2}$. The probability that a weight w out of T unit edges will choose nodes v_i and v_j as their endpoints is given by a binomial distribution with probability of success w and number of trials T. Therefore, the null model defines for every couple of nodes a probability for their edge weight, σ_{ij}, depending on their strengths. For each directed edge (v_i, v_j) with weight w_{ij}, its p-value, denoted as γ_{ij}, is computed as:

$$\gamma_{ij} = \sum_{w \geq w_{ij}} \Pr[\sigma_{ij} = w | s_i^{out}, s_j^{in}, T] = \sum_{w \geq w_{ij}} \binom{T}{w} p_{ij}^w (1 - p_{ij})^{T-w}.$$

According to this null model, the higher the strengths of two nodes, the higher the weight of an edge connecting them is to be in order to be considered statistically significant. Conversely, the lower the strengths of the two nodes, the lower the weight of a linking edge must be in order to be retained by the filter.

Table 2.6. Summary of structural characteristics of the INSTANCES network before and after graph pruning. For each of the two graph pruning models, we tested under significance levels at 1% and 5%, which are indicated within brackets.

	unpruned	*MLF* (0.05)	*MLF* (0.01)	*Disparity* (0.05)	*Disparity* (0.01)
#nodes	6 960	6 454	6 081	3 949	3 260
#edges	216 504	97 141	68 100	19 208	12 849
reciprocity	65.1%	57.9%	58.4%	67.8%	68.5%
density	0.004	0.002	0.002	0.001	0.001
average degree*	41.966	21.390	15.860	6.429	5.182
average in-degree	31.107	15.051	11.199	4.864	3.941
% sources	12%	13.1%	14.1%	24.1%	26.5%
% sinks	6.6%	6.5%	6.7%	6.9%	7.9%
degree assortativity*	-0.274	-0.249	-0.224	-0.244	-0.272
degree assortativity	-0.253	-0.231	-0.214	-0.253	-0.283
average path length	2.330	3.080	3.270	2.530	2.520
diameter	5	6	7	5	6
transitivity*	0.128	0.070	0.066	0.024	0.018
clustering coefficient*	0.836	0.502	0.493	0.846	0.839
clustering coefficient *(full averaging)**	0.687	0.411	0.390	0.447	0.370
#strongly connected components	1 305	1 276	1 278	1 228	1 121
#weakly connected components*	1	1	1	1	1
modularity by *Louvain**	0.289	0.495	0.540	0.469	0.495
#communities by *Louvain**	5 (5)	7 (7)	7 (7)	8 (8)	8 (9)
modularity by *Louvain***	0.353	0.442	0.442	0.369	0.375
#communities by *Louvain***	6 (8)	6 (8)	7 (9)	7 (9)	7 (11)
#communities by *Infomap***	6 (54)	10 (103)	12 (82)	6 (30)	5 (38)

* Statistic calculated by discarding edge orientation
** Statistic calculated by taking into account the edge weights

2.4.2 Results on the pruned networks

Table 2.6 reports on main structural characteristics of the INSTANCES network after graph pruning processes through the application of the MLF and disparity models (for the sake of presentation, values in the original, i.e., unpruned network are also reported).

Using MLF, 55.1% and 68.5% of the edges are removed with significance level of 0.05 and 0.01, respectively. The pruning effect on the network size is much more evident when using the disparity model, with at least 43.3% of nodes and 91.1% of edges removed. By contrast, the reciprocity percentage in the MLF pruned networks is about 7% less than in the original network, whereas the disparity pruned networks show a small increase of reciprocal edges (about 3%).

Interestingly, while it can be observed a significant decrease in the average in-degree — from around 31 in the original, unpruned network up to 11-15, resp. about 4-5, using MLF and disparity, respectively — the degree assortativity remains negative in all cases, with a small

decrease in module under MLF, and comparable or increased in module for disparity with significance level 0.05 and 0.01, respectively. Average path length is also comparable, yet with some fluctuations, to its original value after applying disparity, while it marginally increases under MLF. The latter determines an increase in the network diameter by one and two units for significance level 0.05 and 0.01, respectively, while disparity raises it by one unit only under 0.01.

We also observe a general reduction of the transitivity, which becomes nearly one-half by MLF and even lower by disparity model, compared to its original value. The local clustering coefficient remains substantially unchanged under disparity, while it is nearly halved under MLF; however, when excluding source and sink nodes, the effect is similar for both pruning models and the gap from the original value is smaller.

Both pruning approaches lead to a relatively small decrease in the number of strongly connected components, which is slightly more evident under the disparity pruning. Looking at the community structures, the impact of the two pruning models varies depending on the specific community detection method. As concerns Louvain, the number of communities has little variations, while the pruning is always beneficial according to an improved modularity. On the other hand, the number of communities produced by Infomap almost doubles in the MLF pruned network, which suggests that the pruning effect due to the MLF model would favor the detection of many small communities by Infomap.

2.4.3 Discussion

To answer our fourth research question (**Q4**), we discussed how the backbone of the Mastodon network of instances is unveiled by leveraging on graph pruning approaches based on generative null models. These approaches allow us to discard noisy edges in the network and shed light on its well-rooted underlying structures.

MLF and disparity models showed to differ in pruning intensity, with the latter being more severe. To explain this, recall that the disparity filter accounts for both the degree and strength of a node locally, whereas MLF relies on a maximum-entropy based, unbiased null model: with this in mind, the disparity filter is likely to be more heavily conditioned on in-degree and out-degree distributions that were observed in the INSTANCES network to best fit with a lognormal distribution and, to a less extent, with a power-law distribution, which is the best scenario of application of disparity [201].

A major remark that stands out is nonetheless a certain consistency with the structural properties of the INSTANCES network; that is, apart from a significant decrease in transitivity as an expected effect of the pruning, the observed structural characteristics in the pruned networks remain comparable to those of the original network, or even more emphasized as for the modularity. In particular, the disassortative trait is still present in the backbone, which certainly strengthens our previous finding so to distinguish Mastodon from many other social networks in terms of negative degree correlation.

It is also worth noticing that the top-5 instances discussed in Section 2.3.1, regardless of the particular community detection method are still found after all the considered pruning scenarios, which indicates that these instances are important constituent of the backbone of the INSTANCES network. Overall, even the pruned scenarios exhibit traits related to a tight structure among instances. We recall that pruning approaches aim to filter out noise from networks, and therefore these traits are further strengthened following a cleaner perspective.

2.5 Evolution of the network of Mastodon instances

In this section we answer our fifth research question (**Q5**), namely "*how has Mastodon evolved during the last few years?*". To this aim, we organize our presentation into four parts: the first one is devoted to a comparison between the main network we introduced into this work, i.e., INSTANCES, and the network of instances previously studied in [245]; the second part focuses on the subnetwork composed of online instances in our updated network; the third part describes our analysis of centrality of the instances based on the PageRank method; finally, we report a summary of the main lessons learned from our study of the Mastodon growth.

2.5.1 Comparison with the earlier Mastodon network

In this section we discuss a comparative evaluation between the Mastodon instances network introduced in this work and the earlier network presented in [245]. By replicating on the latter the same structural analysis as we have carried out on the INSTANCES network, we gain insights into the evolution of Mastodon during the last 3 years, focusing on a set of macroscopic and mesoscopic characteristics that may confirm some traits or highlight novel trends. Table 2.7 shows the structural properties that we observed on earlier Mastodon network along with the percentage increase values obtained by the corresponding characteristics in our network (cf. Table 2.3).

Besides the already noticed larger size of the current network (with a +73% of instances and a +127% of links) compared to the earlier one, a general remark that arises is that the two networks of Mastodon instances show differences in several characteristics, though with two major exceptions: the one relating to the average path length and the diameter, which are both unchanged, and the other one referring to both global and local clustering coefficient, which are slightly lower (i.e., from -1% to 5%) in the current network. This would hint at a consolidated small-world behavior by the Mastodon instances.

As for the remaining properties, we observe in our updated network a small decrease in the percentage of reciprocal edges, while the percentage of sources and sinks is respectively more than doubled and decreased of less than one third. One reasonable explanation for that relates to an increased presence of offline instances in the INSTANCES network whose further neighborhood however cannot be explored via API. It should however be noted that our updated INSTANCES network still outperforms in size the earlier network even when removing their respective sources and sinks, with a percentaage increase of 66% in the number of instances.

Also, in the INSTANCES network, the number of strongly connected components is more than doubled, while the decrease in density (about one third) should be related to a high increase in the number of sources, which is in turn an effect of our deep crawling.

A major aspect of interest we found out in the INSTANCES network, that is, degree disassortativity, holds similarly for the earlier network as well. This is noteworthy as it unveils consistency of this property in Mastodon over time.

Our updated network also shows more communities but a slighly less modular structure than the earlier one. Again, this should be ascribed to the impact due to the increase in the number of source instances. As concerns core decomposition, we notice that a high degeneracy also characterizes the earlier network, and the values according to three degree variants are actually close to those observed in the INSTANCES network in proportion of the respective node-set sizes, with only 9%, resp. 2%, of change when considering degree-based decomposition, resp. in-degree- or out-degree-based decomposition.

Table 2.7. Evolution of the INSTANCES network through a comparison of structural characteristics with the earlier state-of-the-art network. Percentage changes refer to the increase/decrease of a statistic from the value observed in the earlier network. For the community statistics, percentage values refer to the number of communities containing at least ten instances.

	Earlier [245]	*% change*
#nodes	4 015	+73%
#edges	95 221	+127%
reciprocity	70.9%	-8%
density	0.006	-33%
average degree*	30.612	+37%
average in-degree	23.716	+31%
% sources	5.63%	+113%
% sinks	9.39%	-30%
degree assortativity*	-0.287	-5%
degree assortativity	-0.291	-13%
average path length	2.340	0%
diameter	5	0%
transitivity*	0.135	-5%
clustering coefficient*	0.848	-1%
clustering coefficient *(full averaging)**	0.710	-3%
#strongly connected components	604	+116%
#weakly connected components*	1	0%
modularity by *Louvain**	0.356	-19%
#communities by *Louvain**	4 (4)	+25%
modularity by *Louvain***	0.397	-11%
#communities by *Louvain***	3 (5)	+100%
#communities by *Infomap***	5 (63)	+20%
degree-based degeneracy	141	+43%
degree-based inner-most-core #nodes	120	+58%
degree-based inner-most-core #edges	11 227	+130%
in-degree-based degeneracy	69	+39%
in-degree-based inner-most-core #nodes	123	+69%
in-degree-based inner-most-core #edges	11 401	+152%
out-degree-based degeneracy	70	+39%
out-degree-based inner-most-core #nodes	115	+70%
out-degree-based inner-most-core #edges	10 385	+155%

* Statistic calculated by discarding edge orientation
** Statistic calculated by taking into account the edge weights

2.5.2 Narrowing the focus on online Mastodon instances

Our further step towards a comprehensive understanding of the current landscape of Mastodon instances is an analysis of the subnetwork corresponding to the online instances only, i.e., the

Table 2.8. Evolution of the INSTANCES network through a comparison of structural characteristics with the ONLINE-INSTANCES network. Percentage changes refer to the increase/decrease of a statistic from the value observed in the INSTANCES network. For the community statistics, percentage values refer to the number of communities containing at least ten instances.

	ONLINE-INSTANCES	*% change*
#nodes	1 115	-84%
#edges	75 046	-65%
reciprocity	70.6%	+8%
density	0.06	+1400%
average degree*	87.074	+107%
average in-degree	67.306	+116%
% sources	2.1%	-83%
% sinks	1.4%	-79%
degree assortativity*	-0.290	+6%
degree assortativity	-0.284	+12%
average path length	2.020	-13%
diameter	4	-20%
transitivity*	0.369	+188%
clustering coefficient*	0.712	-15%
clustering coefficient *(full averaging)**	0.689	0%
#strongly connected components	41	-97%
#weakly connected components*	1	0%
modularity by *Louvain**	0.229	-21%
#communities by *Louvain**	5 (5)	0%
modularity by *Louvain***	0.336	-5%
#communities by *Louvain***	4 (7)	-33%
#communities by *Infomap***	4 (15)	-33%
degree-based degeneracy	168	-16%
degree-based inner-most-core #nodes	153	-19%
degree-based inner-most-core #edges	17 449	-32%
in-degree-based degeneracy	80	-17%
in-degree-based inner-most-core #nodes	173	-17%
in-degree-based inner-most-core #edges	20 320	-29%
out-degree-based degeneracy	82	-15%
out-degree-based inner-most-core #nodes	158	-19%
out-degree-based inner-most-core #edges	17 987	-32%

* Statistic calculated by discarding edge orientation
** Statistic calculated by taking into account the edge weights

ONLINE-INSTANCES network, which was previously introduced in Section 2.2.2. The goal here is to check whether our findings on the overall INSTANCES network reflect on its subnetwork of online instances as well, or on the contrary there are significant deviations on some traits. Table 2.8 summarizes the comparison between the two networks.

A first remark clearly arises from the reduction in both the number of instances (-84%) and links (-65%) compared to the INSTANCES network. This is clearly expected due to the deep API-based crawling capabilities, yet at the same time it captures the currently active snapshot of Mastodon in the Fediverse. We conjecture that, after a few years of novelty and curiosity that Mastodon attracted, a steady-state has been reached as composed by those instances that might be recognized as most established for the Mastodon users. The subnetwork of online instances also turns out to be more tightly knit than the INSTANCES network, as indicated by the increase in reciprocity (+8%), density (one order of magnitude higher), increase in average degree and in-degree (more than doubled), decrease in average path length and diameter (-13% and -20%, respectively), the almost tripled transitivity, halved number of strongly connected components, and decrease in the percentage of sources and sinks (about 80%). The latter aspect suggests that the high percentages (particularly for sources) found in the INSTANCES network is likely to be ascribed to the different status (i.e., online or offline) of the instances, which affects the search outcomes. Nonetheless, since the ONLINE-INSTANCES network contains currently active and online instances, their users might be more engaged in developing interrelations. Notably, the disassortative trait is still present in the ONLINE-INSTANCES network, thus confirming it as a well-rooted feature in the network of Mastodon instances.

As concerns the community structure, the ONLINE-INSTANCES network exhibits similar characteristics to the overall INSTANCES, with a relatively small reduction in undirected modularity – which becomes negligible for the weighted directed case – and in the number of communities only when accounting for edge weight and orientation. Interestingly, we point out that the number of significant communities (i.e., those containing at least ten instances) in the ONLINE-INSTANCES is the same for both Louvain and Infomap method, which is consistent with the situation already observed in the INSTANCES network. Analogous considerations can be made for the core decomposition results: indeed, not only the degeneracy remains still high in all three variants – with a decrease of about 15% w.r.t. INSTANCES – but also the sizes of the corresponding inner-most cores are proportionally higher than those in INSTANCES according to the respective node-set sizes, with 405%, resp. 419 and 403%, of increase when considering degree-based decomposition, resp. in-degree- and out-degree-based decomposition. This is remarkable as it provides evidence of a well-established trend in Mastodon as to have a conspicuous concentration of instances in the inner-most central region of the network.

2.5.3 Instance centrality

To further investigate the growth of Mastodon, we evaluated how the *importance* of the instances developed during the years. To this aim, we resorted to the well-known *PageRank* method [36], which assigns prestige scores to the nodes in a network. Using this tool, we comparatively evaluated results obtained on all the networks derived from our crawling data, i.e., the EXPANDED-INSTANCES, the INSTANCES, and the ONLINE-INSTANCES network, as well as the earlier state-of-the-art network. This approach allowed us to assess whether and to what extent the centrality of instances was affected by the introduction of new instances and/or the dismission of instances (e.g., instances went to offline mode for a relatively long period, cf. Section 2.2).

We assessed the strength of relatedness between the PageRank solutions obtained on the above networks by means of two standard rank correlation methods, namely *Kendall correlation coefficient* [1] and *Fagin's intersection metric* [76]. The Kendall correlation τ evaluates the similarity between rankings represented through set of ordered pairs assigned to the same set of nodes, relying on the number of inversions of pairs needed to transform one ranking into the

other. Given two rankings $\mathcal{R}'$ and $\mathcal{R}''$ obtained on the same set of N items, this non-parametric correlation is formally expressed as follows:

$$\tau(\mathcal{R}', \mathcal{R}'') = 1 - \frac{2\Delta(\mathcal{P}(\mathcal{R}'), \mathcal{P}(\mathcal{R}''))}{N(N-1)},$$

where $\Delta(\mathcal{P}(\mathcal{R}'), \mathcal{P}(\mathcal{R}''))$ denotes the symmetric difference between $\mathcal{R}'$ and $\mathcal{R}''$, i.e., the number of unshared pairs between the two rankings. The values returned by τ are within the range [-1, 1], where 1 indicates that the two rankings are identical and -1 means that one rank is the reverse order of the other.

The Kendall correlation assigns the same importance to all items, regardless of their position. Therefore, we also employed Fagin's intersection metric F, which considers partial rankings and assigns higher weights to items at the top of the lists. Given two ranking lists $\mathcal{R}'$ and $\mathcal{R}''$, Fagin's metric is formalized as follows:

$$F(\mathcal{R}', \mathcal{R}'', k) = \frac{1}{k} \sum_{q=1}^{k} \frac{|\mathcal{R}'_{:q} \cap \mathcal{R}''_{:q}|}{q},$$

where k is a parameter to determine the number of items to consider from the top of both rankings, and $\mathcal{R}_{:q}$ indicates the set of nodes from the 1*st* to the q*th* position in the ranking. Hence, we refer to F as the average over the sum of the weighted overlaps calculated considering the first k nodes in both rankings. Its values are within the range [0, 1], where the highest the value, the better the score.

The results of our evaluations are reported in Table 2.9. Since Kendall correlation requires a comparison of rankings of nodes from the same set, the results obtained on any pair of networks correspond to the shared set of nodes. Note also that in all cases the Kendall correlation values are high, and associated with p-values equal to zero, which means rejection of the hypothesis of absence of correlation.

Let us first consider the comparison between Expanded-Instances and Instances networks (i.e., first column of Table 2.9). The Fagin's intersection values observed for various k values are very high, as expected. Indeed, as described in Section 2.2, non-Mastodon instances represent the boundary of our Expanded-Instances network, and we have only partial knowledge of them. Hence, the more central role taken by the Mastodon instances is reasonable. However, while these remarkably high values denote that non-Mastodon instances only slightly influence the Mastodon ones, the relatively smaller value obtained for F_{10} indicates that this influence actually affects the PageRank scores of the top-10 instances.

The comparison between Online-Instances and Instances (i.e., second column of the table) allows us to inspect the role of online vs. non-online instances. Paying attention to F_{10}, the noticeable value indicates how the most central Mastodon instances are well-rooted in their role, regardless of potential variations, even evident, in the number of the online ones at any given time. Although a bit less evident, this trend remains valid in the range [50, 100], while we observed a further decrease with $F_{1\,000}$; the latter might indicate the existence of relevant instances among the offline or temporarily inactive ones.

As concerns Expanded-Instances vs. Online-Instances, we observe in general slightly lower values w.r.t. the previous comparison, which might be ascribed to the absence of some offline yet relevant instances as well as to the lack of contributions given by non-Mastodon instances. However, the still high correlation values observed further support the centrality of the currently online Mastodon instances.

The last comparison we consider is between the Instances and the *Earlier* network. It is not surprising to observe a general decrease in the correlation values, given the temporal difference

Table 2.9. Ranking analysis performed via Kendall's tau (τ) and Fagin's intersection metric (F) with various k values (indicated as subscripts) on the PageRank solutions obtained on the Mastodon networks.

	Expanded-Instances vs. Instances	Online-Instances vs. Instances	Expanded-Instances vs. Online-Instances	Instances vs.*Earlier*
τ	0.953	0.899	0.883	0.626
F_{10}	0.860	0.935	0.886	0.569
F_{50}	0.911	0.814	0.813	0.563
F_{100}	0.918	0.832	0.835	0.545
F_{500}	0.938	0.821	0.828	0.487
$F_{1\,000}$	0.915	0.748	0.738	0.473

between the two networks. Nonetheless, the corresponding ranking lists over the shared set of instances show a good Kendall correlation (above 0.6); in addition, the Fagin's intersection metric values are also quite high, for various k values. Overall, despite the introduction of new instances after about three years, the results indicate that a significant portion of those instances that are shared between the two networks have not drastically changed their strategic location in their respective networks.

2.5.4 Discussion

To investigate the evolution of Mastodon from an instance perspective, we compared the various networks presented in Section 2.2, keeping our focus on the Instances network. First, we performed a comparison with the *Earlier* network, which refers to about three years ago, as previously discussed. This allowed us to highlight the structural development of Mastodon. Although the different sizes have some effect on the measures, we noticed some well-established features. In particular, we report the same average path length and diameter in both networks, whose consolidation suggests that Mastodon instances act in a small-world fashion. The *Earlier* network also yields the same disassortative trait spotted for the more recent Instances network. Thus, we can establish the existence of this property over time. This consolidation allows us to consider the negative degree correlation as a distinctive trait of Mastodon w.r.t. centralized social networks and paves the way for further studies concerning interactions among instances through the federative approach. From a mesoscopic perspective, the evaluation of the core decomposition brings us to confirm another important trait. Indeed, as initially shown by the *Earlier* network and subsequently confirmed by the Instances one, Mastodon instances exhibit high degeneracy values, comparable when accounting for the different sizes.

Along with the temporal development, we analyzed the active status of Mastodon, focusing on the subnetwork induced from its online instances (i.e., the Online-Instances network). This narrowing produces an exceptional reduction in size (-84% instances and -65% links among them), which might indicate a stationary state in the number of online Mastodon instances. Indeed, the spotted online ones are found to be no longer transient (e.g., testing instances) or driven by the feeling of novelty, but they are rather well-defined and established. Remarkably, the Online-Instances network still exhibits the disassortative trait. The core decomposition of the Online-Instances network confirms the high degeneracy found in the other Mastodon networks, setting it as a well-established trait of Mastodon, along with a remarkable concentration of instances in the inner-most core.

Finally, the study of the instance centrality based on PageRank allowed us to assert that the most prestigious instances are settled in their roles, also considering the temporal development,

which indicates an actual evolution of the platform towards a stable state. The instance-rankings computed over different pairs of instance networks have generally shown good or very high correlation, with relatively small fluctuations mainly due to the centrality contributions from offline or boundary instances, but also as an effect of the growth of the platform over time.

Overall, and in light of all the above findings, we can state that Mastodon has reached sort of structural stability. This well-established structure couples with a solid federative mechanism among instances, which acts as mutual-reinforcement towards the sectorization bias induced by the decentralized scenario.

2.6 Chapter review

Nowadays, DOSNs express an alternative, user-centric approach to support online interactions among users, with attention on self-hosted networking services, and ownerships in terms of code of conduct and moderation policies for each operating server. Mastodon represents the most widely adopted and recognized platform in the Fediverse of DOSNs. In this work, we provided the first in-depth analysis of the network of Mastodon instances based on the largest and most up-to-date Mastodon relational data existing so far, which was originally built in this work. We analyzed our instance network models from different perspectives (i.e., macroscopic, mesoscopic, backbone) to highlight the main structural characteristics that describe Mastodon. Among these, we spotted remarkable traits that define the fingerprint of Mastodon, setting it apart from well-known OSNs. Finally, we investigated the evolution of Mastodon, by identifying features that have changed over time and those representing supporting pillars.

Our next research stage will be focused on the underlying network of Mastodon *users*. As discussed in this work, our collected data are fine-grained at a user level, although the stored information concern user memberships and relations only, i.e., social contents are discarded. In this regard, by still leveraging the network of links formed by users, we plan to study the behavior of users not only within their home instances, but also their relations across instances. In particular, we believe it would be interesting to investigate the roles that users may take in their instances, at various levels of activity, and possibly alternate or opposite roles that the same users may exhibit when interacting with users of other instances. This would allow us to also analyze the dichotomy between contribution and consumption of information induced by the user behavior relations over the instances of Mastodon and other DOSNs in the Fediverse.

3

Information Consumption and Boundary Spanning in Decentralized Online Social Networks: the case of Mastodon Users

Summary. In this work, we aim to fill a lack of study on user relations and roles in DOSNs, by taking two main actions: understanding the impact of decentralization on how users relate to each other within their membership instance and/or across different instances, and unveiling user roles that can explain two interrelated axes of social behavioral phenomena, namely information consumption and boundary spanning. To this purpose, we build our analysis on user networks from Mastodon, since it represents the most widely used DOSN platform. We believe that the findings drawn from our study on Mastodon users' roles and information flow can pave a way for further development of fascinating research on DOSNs.

3.1 Contributions

Our research work aims to fill a lack of study on user relations and roles in DOSNs, and in this respect we want to pursue two main interrelated goals:

- First, since DOSNs embrace a myriad of instances, we are interested in understanding whether and to what extent interesting, decentralization-driven user behaviors arise within the membership instances, across different instances, or even correspond to mixed behaviors. This might favor the comprehension and the modeling of the information flow within and between multiple instances.
- Second, since the human-centric approach of DOSNs undervalues artificially imposed interactions, such as those deriving from boosting or advertisement mechanisms, we want to assess how users shape their roles in a more spontaneous social networking context. In this respect, our focus is on user roles that are essential to explain two interrelated axes of behavioral phenomena in online social networks, namely information consumption and boundary spanning. The dualism between information consumption and boundary spanning can indeed profoundly affect the scope of the information flow within a network and its fluidity, e.g., whether information flows rapidly and spreads widely or remains confined to specific areas of the network.

To the best of our knowledge, no works have been proposed so far to analyze user roles and behaviors in DOSNs based on the above aspects. Note also that our perspectives on the aforementioned aspects of interest in this work are totally independent from knowledge about textual or media contents produced and exchanged through the Fediverse, thus exploiting only the topological information of the user relation network.

To conduct our research study, we shall focus on the most widely known and representative Fediverse platform, i.e., Mastodon, which is hence the best suited as case in point for investigating on the DOSN landscape. Moreover, this also allows us to capitalize on up-to-date data resources and relating findings from our earlier work [139].

Our roadmap to delve into the understanding of the above discussed aspects will be developed so as to pursue a number of objectives that can be summarized into the following research questions:

Q1 – *The User Network structure*: What are the main structural characteristics of the network of following relations between Mastodon users?
Q2 – *Representative instances*: Are the users belonging to the most relevant Mastodon instances representative of the entire user network?
Q3 – *Boundaries and bridges*: Are Mastodon users involved in inter-instance links and how do they act as local bridges?
Q4 – *Over-consumption*: Are there Mastodon users who tend to over-consume, i.e., lurk, others' information? Is this behavior bounded to the membership instances or it spans across the instance boundaries?
Q5 – *Dual role users*: Are there users who behave as both lurkers and bridges within their own instance?
Q6 – *Alternate role users*: Can user behavior vary according to the observation scale? That is, can a user be a lurker within her/his instance and simultaneously act as a bridge between instances, or vice versa?

Plan of the Chapter. The remainder of the Chapter is organized as follows. Section 3.2 introduces to the Mastodon data and the network models we used in our analysis. Section 3.3 presents our structural analysis of the Mastodon user networks, by first considering the full set of users and their relations, then focusing on a representative subset of the user network corresponding to a selection of the most relevant instances in Mastodon. Section 3.4 provides insights into user roles that are relevant to boundary spanning and information consumption behaviors for the users in Mastodon. Section 3.7 summarizes the main lessons learned from our analysis, while Section 3.8 concludes the Chapter and provides pointers for future research.

3.2 Data Extraction and Network Modeling

In our earlier work [139], we developed a privacy-friendly crawler upon the publicly available Mastodon REST APIs [1] to build an up-to-date and highly representative dataset. It is worth emphasizing that, to preserve privacy requirements, we relied on authenticated requests only, i.e., those towards the instances that allowed accountable requests through their APIs, and we also avoided using any scraping tools.

To account for the decentralized nature of Mastodon, we leveraged on the *instances.social* website,[2] which politely keeps track of the Mastodon instances panorama. In particular, this website enabled us to locate some "seed" instances (i.e., the online ones at the time of crawling) from which we started our exploration of the Mastodon Fediverse. By getting information on the timelines of about 900 instances, we reached more than 80 000 users, who represented the starting point of a *breadth-first-search* to discover new connections and, consequently,

[1] `https://docs.joinmastodon.org/api/`
[2] `https://instances.social/`

more users. In this regard, we point out that although the toots (delivered over the timelines of the seed instances) were inspected to discover the corresponding users, the toot data was never stored, therefore *our study described in this work is totally agnostic of textual contents*. Furthermore, the interactions between users in terms of incoming and outgoing links were anonymized through proper hashing functions at the time of their acquisition.

After processing the fetched data, we came up with about 1.4M unique users and 18M unique links between them, traversing more than 16 000 instances. The protocol underlying Mastodon, i.e., ActivityPub, supports seamless communication between all the Fediverse platforms. This implies that the data obtained by means of the APIs can in principle also concern interactions between Mastodon instances and instances pertaining to other services in the Fediverse. Therefore, using the aforementioned *instances.social* and the *fediverse.party* platforms,[3] we discerned the Mastodon instances within our dataset, splitting them into online and temporarily offline ones, where the latter correspond to instances that keep an inactive status for at most two weeks (according to the instances documentation provided by *instances.social*). As a result, we discovered 6 960 Mastodon instances, among which 1 116 were online.

Upon the extracted data, we built networks whose entities (i.e., nodes) represent users, and we modeled their relationships either at the level of the whole Mastodon network or at the level of individual instances. Let us denote with $\mathcal{U}$ the set of users and with $\mathcal{I}$ the set of instances available in the extracted Mastodon data. We define the *Mastodon user network* as a directed graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, 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 only if $i \neq j$.

Given a target instance $i \in \mathcal{I}$, we define the *instance-specific user network* of i as the directed subgraph $G_i = \langle V_i, E_i \rangle$ induced from $\mathcal{G}$, such that $V_i = \{u \mid (u, i) \in \mathcal{V}\} \subseteq \mathcal{U}$ and E_i is the set of edges (u, v) with u following v in instance i.

Given a set of target instances $\mathcal{M} \subset \mathcal{I}$, we define the network of the relations between users of the instances $\mathcal{M}$, dubbed *merged network*, as the directed subgraph $G_{\mathcal{M}} = \langle V_{\mathcal{M}}, E_{\mathcal{M}} \rangle$ induced from $\mathcal{G}$, such that $V_{\mathcal{M}} = \{u \mid (u, i) \in \mathcal{V} \wedge i \in \mathcal{M}\} \subseteq \mathcal{U}$ and $E_{\mathcal{M}}$ is the set of edges (u, v) with u following v in some instance of $\mathcal{M}$.

3.3 User Network Structure

In this section we begin with answering our first research question (**Q1**), i.e., understanding the main structural traits of the user network in Mastodon (Section 3.3.1). Next, we take the opportunity of investigating on the presence of noisy or irrelevant user relations according to requirements specified for their membership instances (Section 3.3.2). This eventually leads us to answer our second research question (**Q2**) by identifying and analyzing the subset of the user network corresponding to the most relevant instances in Mastodon (Section 3.3.3), which will be used as our workbench for the subsequent user behavioral analysis.

3.3.1 The Mastodon user network

As our first action towards the understanding of behaviors of Mastodon users, we gained a comprehensive view of the main features of the Mastodon user network, both from a macroscopic and mesoscopic perspective.

[3] `https://fediverse.party/en/mastodon`

Table 3.1. Main structural characteristics of the user networks derived from Mastodon.

	User network (full)	*User network (filtered)*	*Top-5 instance merged user network*
#nodes	1 315 739	1 237 985	657 712
#edges	17 252 347	16 239 329	8 227 553
density	1e-05	1e-05	2e-05
% sources	34.2%	34.7%	29.7%
% sinks	7.5%	6.7%	3.1%
average degree*	21.925	22.007	21.445
average in-degree	13.112	13.118	12.509
degree assortativity*	-0.040	-0.042	-0.072
degree assortativity	-0.032	-0.033	-0.048
transitivity*	0.003	0.003	0.002
clustering coefficient*	0.393	0.398	0.401
clustering coefficient *(full averaging)**	0.315	0.323	0.357
reciprocity	32.8%	32.2%	28.6%
average path length	5.326†	5.312†	5.088 (5.162†)
#strongly connected components	565 865	526 349	223 935
#weakly connected components*	327	320	141
modularity by *Louvain*	0.737	0.736	0.688
#communities by *Louvain*	580 (125)	578 (111)	384 (89)
modularity by *Louvain**	0.717	0.717	0.658
#communities by *Louvain**	565 (127)	518 (109)	412 (90)
modularity by *Leiden**	0.743	0.741	0.688
#communities by *Leiden**	629 (138)	629 (127)	417 (97)
#communities by *Infomap*	594 (53)	568 (52)	273 (48)
#communities by *Infomap**	359 (19)	349 (19)	178 (17)

* Statistic calculated by discarding the edge orientation
† Statistic calculated as $ln(N)/(lnln(N))$, where N denotes the number of nodes

In the first data column of Table 3.1, we report the main structural characteristics analyzed. First, it stands out the high numbers of nodes and edges available in our network, which indeed captures a large fraction of the existing Mastodon user base.[4] We also notice a reasonable fraction of source nodes (i.e., users having only outgoing links), which would include newcomers at the time of the data acquisition, and in general users that take a peripheral role in the platform; moreover, the fraction of sink nodes (i.e., users having only incoming links) is quite small, which would indicate a moderate presence of users who do not appear to be interested in establishing or reciprocating connections with other Mastodon users.

[4] At the time of writing of this work, we achieve a coverage that ranges between 62% and 78% of the full audience of Mastodon, according to the *fediverse.party* and *instances.social* websites, respectively.

Another helpful indicator concerning the linkage between users is expressed by the degree correlation or *degree assortativity*, i.e., the probability that a link between two nodes depends on their respective degrees [169, 170]. Given the observed value, which is slightly negative yet close to zero, it happens that the Mastodon user network is an uncorrelated network: this turns out to be explained due to a very interesting trait of Mastodon, since the lack of degree correlation highlights how users relationships within Mastodon are generally driven by genuine interests, rather than by exogenous recommendation mechanisms that may occur within or across instances.

We further delved into the relationships between users by investigating aspects of transitivity (i.e., the likelihood that two incident edges are completed by a third one, thus forming a triangle) and local clustering coefficient (i.e., how strongly connected are the neighbors of a node); the latter was measured by averaging either over all nodes (indicated as "full averaging" in Table 3.1) or only nodes with degree greater than one. The relatively higher local clustering coefficient w.r.t. the transitivity is not surprising, given the low density of the network. We note that coupling this with the quite low average path length hints at the presence of a small-world phenomenon. Moreover, the non-negligible observed reciprocity further strengthens the local connectivity in terms of dyadic relations.

To understand the community structure of the Mastodon user network, we resorted to the widely used *Louvain* [30] and *Infomap* [192] methods, along with the more recent *Leiden* method [216]. Louvain exploits a hierarchical greedy approach based on two phases, modularity optimization and community aggregation, which are repeated until there are no more changes to be made on the communities and a maximum of modularity is achieved. Infomap optimizes the Map equation, which leverages the information-theoretic duality between finding community structures in a network and minimizing the description length of the movements of a random walker in a network. The Leiden method is designed to improve upon the Louvain method, by providing guarantees on the connectivity of the discovered communities through an iterative algorithm that includes local moving and community aggregation stages, with the addition of an intermediate stage of refinement of the community connectivity. We used both the undirected and directed implementations of the Louvain[5] and Infomap[6] algorithms, whereas for the Leiden algorithm, we used the only available undirected implementation.[7]

Table 3.1 reports the number of communities found by each of the above algorithms, along with their modularity, if optimized by the corresponding method. We distinguish between overall and meaningful communities, the latter being regarded as those having at least ten users (enclosed in round brackets), and we indicate with * in the table the cases when we discard edge orientation. As shown by the modularity values above 0.7 for both Louvain and Leiden algorithms, the Mastodon user network appears to be characterized by a modular structure. The number of communities found ranges between 359 (by undirected Infomap) and 629 (by undirected Leiden), which however is significantly reduced when focusing on the meaningful communities, from 127 (by undirected Louvain) and 138 (by undirected Leiden) down to 19, resp. 53, by undirected, resp. directed, Infomap.

3.3.2 Filtering out noisy instances

In our structural analysis of the Mastodon user network, we also considered to measure the effects of simplification of the network in terms of pruning of user relations involving potentially

[5] `https://github.com/nicolasdugue/DirectedLouvain`

[6] `https://www.mapequation.org/infomap/`

[7] `https://github.com/vtraag/leidenalg`

Table 3.2. Size of the top-5 relevant Mastodon instances and their aggregation (*merged network*).

	#nodes	#edges	density
mastodon.social	305 968	3 408 327	4e-05
pawoo.net	306 753	4 329 562	5e-05
mastodon.xyz	16 076	35 631	1e-04
mstdn.io	16 853	112 805	4e-04
octodon.social	7 082	34 493	7e-04
merged network	657 712	8 227 553	2e-05

noisy or irrelevant instances. To this purpose, here we capitalize on related findings from our previous study [139], in which we assessed the relevance of the instances according to the number of links they receive. In this regard, we observed statistical significance (based on a Kolmogorov-Smirnov test) of a lognormal fitting of the in-degree distribution when removing the instances that are pointed by less than 51 other instances.

Performing this pruning step on our Mastodon user network implies the removal of approximately 100k nodes and 1M edges; nonetheless, it should be noticed that the main characteristics of the Mastodon user network structure have remained substantially unchanged. In fact, as shown in the central column of Table 3.1, all statistics are in line with those relating to the original network. Such consistency allows us to argue that the most relevant instances strongly determine the backbone of the whole Mastodon user network, also ensuring its robustness w.r.t. the removal of potentially noisy elements.

3.3.3 Narrowing the focus: the top-5 instances

We further investigated the impact of instance selection on the main traits of the Mastodon user network by focusing on the most important instances. Again following the lead of [139], we selected the top-5 instances by relevance, according to the maximization of a threefold criterion based on number of registered users, number of involved links, and data access permission policy; this resulted in the following selection of instances: *mastodon.social*, *pawoo.net*, *mastodon.xyz*, *mstdn.io* and *octodon.social*.[8] Interestingly, we highlight that all the selected instances are generalistic and not topically-bounded, thus allowing users to discuss and interact through various arguments. Table 3.2 provides a summary of their size information. We then created the user network upon these instances according to the *merged network model* (defined in Section 3.2).

On this merged network, we replicated the structural analysis carried out on the whole and pruned user-networks, in order to unveil whether and to what extent the user network of the top-5 instances can be considered representative of the entire Mastodon user network. As shown in the rightmost column of Table 3.1, the similarity between the statistics on the top-5 instance merged network and the corresponding ones of the whole user network is evident, which supports our above hypothesis of representativeness of the Mastodon user network. Note that, as concerns the average path length, we managed to calculate it exactly on the merged network, while for the much larger full and filtered networks we were forced (due to

[8] Note that, according to the *instances.social* website, the top five ranked instances might partly be changed, since the time of our crawling of the Mastodon network, w.r.t. one or all of the criteria we considered for the instance selection.

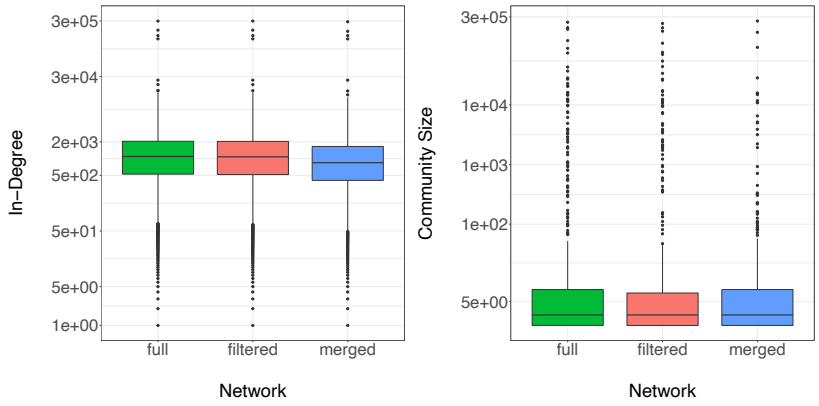


Fig. 3.1. In-degree distributions (on the left) and community size distributions (on the right) of the *full*, *filtered*, and top-5 instance *merged* networks. (Community size distributions refer to the solutions by undirected Louvain)

computational issues) to an approximation generally valid for random scale-free uncorrelated networks (cf. Table 3.1); in this respect, the approximated values of average path length appear to be very close to the exact value computed on the top-5 instance merged network.

To further strengthen our hypothesis, we also examined the in-degree distributions and the community size distributions of the three user networks under evaluation. As shown in Figure 3.1, for both in-degree and community size, the three networks have box plots that are very close to each other. In particular, concerning the in-degree distributions, the medians are equal to 1 092.5, 1 078.5, and 846, for the full, filtered, and merged networks, respectively; moreover, the median of the community size distributions settles on 3 for all the considered networks. This is remarkable, as not only confirms the relatively low relevance of the instances removed from the entire user-network, but more interestingly, it indicates that the network of the top-5 instances can effectively be used as a proxy for the Mastodon user relations; in addition, by focusing on a small number of large instances, this proxy can in principle enhance our interpretability of the behavioral patterns to discover in the Mastodon user relations.

Upon the above findings and remarks, we chose to narrow our focus on the top-5 instance merged network in the subsequent behavioral analysis of Mastodon users.

3.4 Boundaries, bridges, and over-consumption

In this section we delve into the relations between users in top-5 instance merged network in order to discover user roles that are relevant in terms of two particular behavioral phenomena, namely boundary spanning and information consumption. We elaborate on the former in Section 3.4.1 and on the latter in Section 3.4.2, which will lead us to answer our **Q3** and **Q4** research questions, respectively.

3.4.1 Instance boundaries and bridges

Here we will focus on those Mastodon users that are involved in inter-instance links and on how they can be regarded and scored as local bridges in the Mastodon user network (**Q3**).

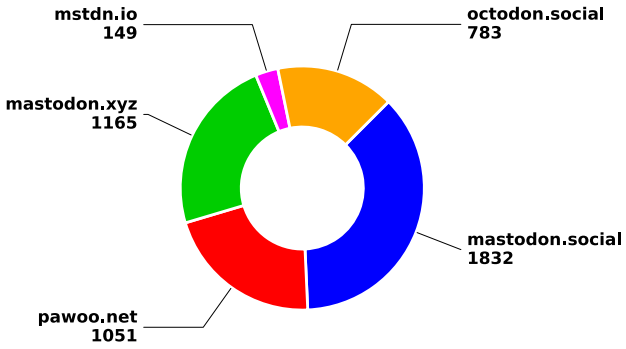


Fig. 3.2. Distribution of shell nodes within the top-5 instances composing the *merged network*.

Shell nodes and inter-instance edges. In Table 3.2, we notice that the sum of the number of nodes and edges over the top-5 instances are 652 732 and 7 920 818, respectively, which differs from the size of the merged network.

We inspected our data looking for the reasons of the above fact, and found that the additional 4 980 nodes and 306 735 edges in the merged network take a specific role therein. Given a merged network $G_{\mathcal{M}} = \langle V, E \rangle$, we define the aforementioned two types of entities as follows:

- *Shell nodes*: a node $(v, i) \in V$ is said a shell node if $\nexists (u, j) \in V : i = j \wedge (((u, j), (v, i)) \in E \vee ((v, i), (u, j)) \in E)$.
- *Inter-instance edges*: $((u, j), (v, i)) \in E$ is said an inter-instance edge if $i \neq j$.

Loosely speaking, a shell node is a user linked to users of other instances only, while an inter-instance edge is a link for users of different instances. The Mastodon top-5 instance merged network has indeed 4 980 shell nodes and 306 735 inter-instance edges; the distribution of such nodes w.r.t. the various top-5 instances is shown in Figure 3.2, whereas details about the distributions of the inter-instance edges will be considered later (cf. Table 3.5).

Visualization of the inter-instance subnetwork. In order to get more insights into the linkage between instances, we visually inspected the inter-instance subnetwork, which models the edges connecting nodes that belong to different instances in the top-5 instance merged network.

As it can be observed from Figure 3.3, several interesting patterns emerge. The first eye-catching aspect is the pervasiveness of *mastodon.social* (colored in *blue*), which appears to be dominant in establishing user relationships across instances. The roots of this phenomenon plausibly lie in the relevance of *mastodon.social* since it is commonly recognized as one of the supporting pillars of the Fediverse and the first instance of the Mastodon project. The central area of Figure 3.3 is characterized by a particularly intense mix of colors, indicating the presence of user connections that involve all the instances in the network; this is clearly consistent with the key principle of the Fediverse as an ecosystem made of independent yet cooperating instances.

Through a detailed inspection of the network in Figure 3.3, we also spotted some regions showing further relevant patterns, such as a strong linkage between users of some pairs of instances or a dense interleaving among multiple instances' users. In this regard, as shown in Figure 3.4 (a), resp. Figure 3.4 (b), there is a tight connectivity among users of *mastodon.social* with users of *mstdn.io*, resp. users of *mastodon.social* with users of *mastodon.xyz*. An analogous situation can be observed in Figure 3.4 (c), where a strong coupling between users of

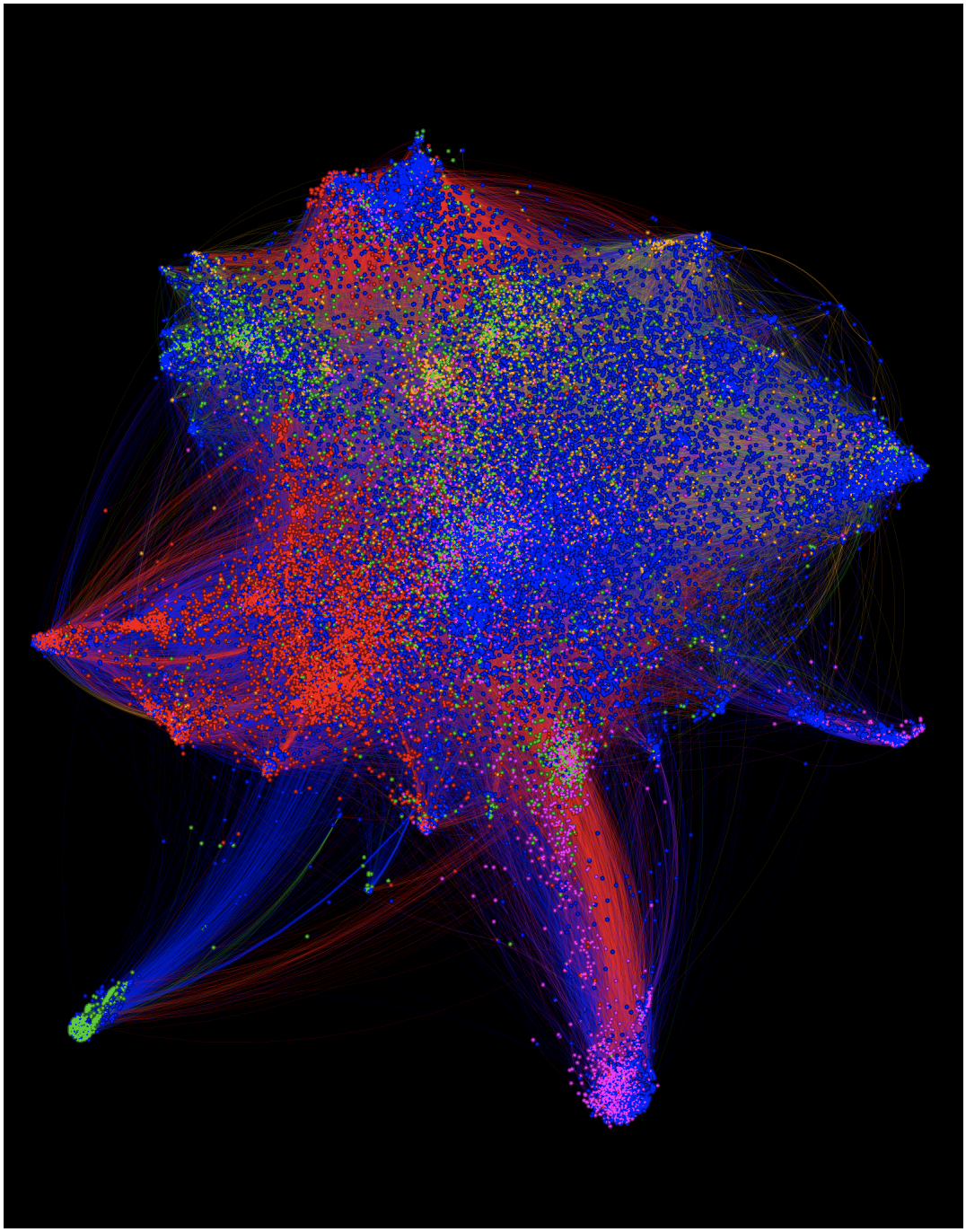


Fig. 3.3. Illustration of the inter-instance subnetwork. Nodes correspond to users belonging to the instances composing the top-5 instance *merged network* and only inter-instance edges are drawn. Nodes are colored according to their membership instances, i.e., *mastodon.social* (*blue*), *pawoo.net* (*red*), *mastodon.xyz* (*green*), *mstdn.io* (*magenta*), and *octodon.social* (*orange*). The color of an edge corresponds to the color of the source instance. The displayed layout is based on the force-directed drawing *ForceAtlas2* model. *(Produced by using the Graphistry service, available at https://www.graphistry.com.)*

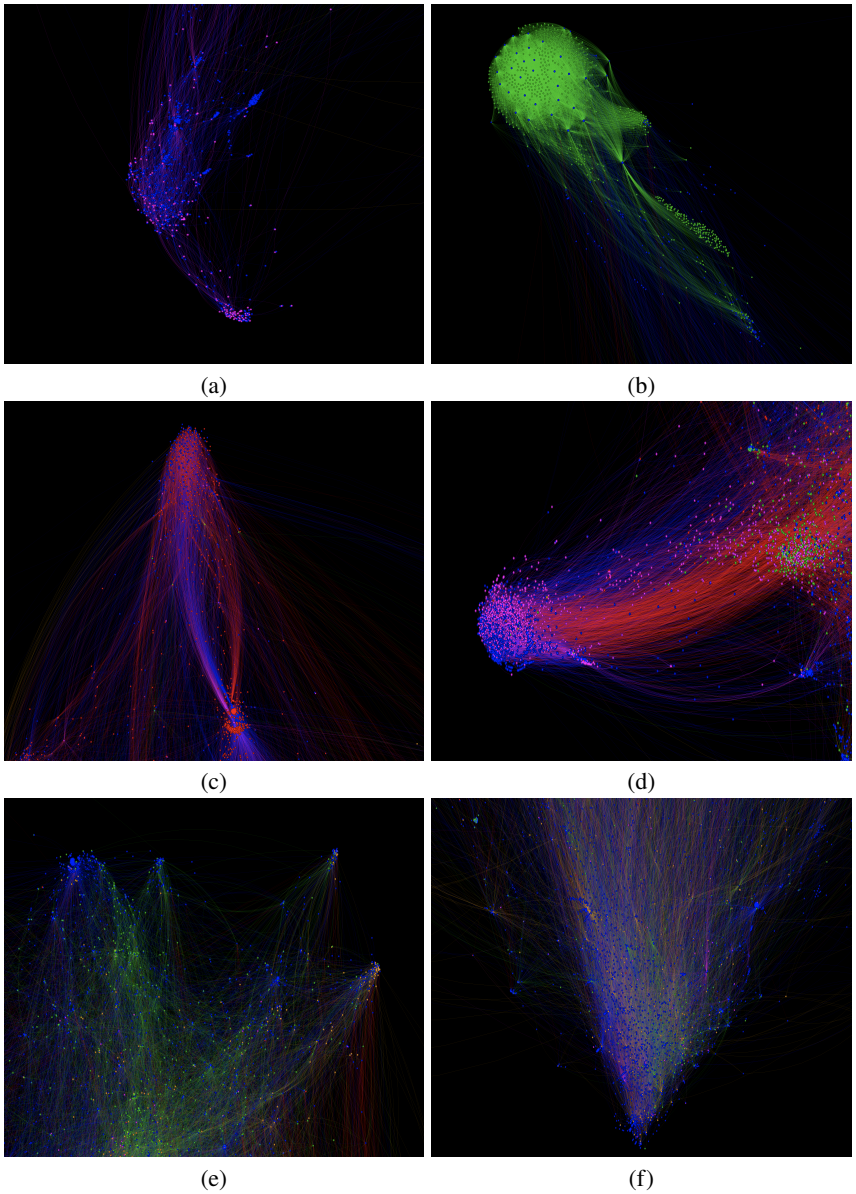


Fig. 3.4. Zoom-in views from Fig. 3.3. Nodes are colored according to their membership instances, i.e., *mastodon.social* (*blue*), *pawoo.net* (*red*), *mastodon.xyz* (*green*), *mstdn.io* (*magenta*), and *octodon.social* (*orange*). The color of an edge corresponds to the color of the source instance.

mastodon.social and users of *pawoo.net* emerges, with the addition of some sporadic users belonging to *mstdn.io*, which are nonetheless well connected with the other two instances.

Moreover, it is worth noticing how the pairwise interactions between (users belonging to) different instances occur with remarkable intensity even among the largest instances, as shown in Figure 3.4 (c). This trait is particularly interesting, since although such instances can definitely be regarded as self-sufficient and represent stand-alone social platforms — given their remarkable size — their users tend to interact outside the boundaries so as to gain a more global user behavioral experience.

Figure 3.4 (d) illustrates two regions of the network characterized by two different patterns: a particularly marked connection between users of *masto- don.social* and users of *mstdn.io* (on the left), which is also massively involved in a linkage with users from *pawoo.net*, and another group of users from different instances (as shown by a mix of colors, on the right). The latter hints at a comprehensive connectivity between all the instances composing the merged network, as also confirmed by the identification of other regions characterized by users belonging to different instances, i.e., Figures 3.4 (e) and (f). These events observed in the merged network reveal that the boundary spanning spotted so far is not limited to pairwise instance links, but it involves multiple instances. As a consequence, we can argue that Mastodon users fully exploit the potential of seamless interaction between independent instances provided by the platform.

It should be noticed that the force-directed layout (*ForceAtlas2*) we used for drawing the network emphasizes the peripheral positioning of portions of the network, such as those corresponding to the above cases, in which there exists a higher connectivity among a bunch of nodes of two or few instances than with nodes of the other instances.

Bridges. The above discussed boundary entities relate to another aspect of interest to our analysis of the merged network, which is the presence of nodes connected by edges acting as local *bridges* at varying degrees.

An effective method to identify such edges is to measure for each pair of linked nodes their *topological overlap*, or normalized embeddedness [172], which is the fraction of common neighbors a pair of connected vertices has. Indeed, edges acting as bridges are expected to share few or no neighbors, and in fact the topological overlap enables smoothing the notion of local bridge, so that the lower the topological overlap of a linked pair of nodes, the higher the strength of their link as local bridge. Originally conceived for undirected networks, the topological overlap has also been adapted to directed networks. Following [211], the *directed topological overlap* (DTO) for an edge (u, v) is defined as:

$$DTO(u,v) = \frac{|\mathcal{N}_u^{out} \cap \mathcal{N}_v^{in}|}{(|\mathcal{N}_u^{out}| - 1) + (|\mathcal{N}_v^{in}| - 1) - |\mathcal{N}_u^{out} \cap \mathcal{N}_v^{in}|}$$

where $\mathcal{N}_u^{out}$ and $\mathcal{N}_v^{in}$ denote the out-neighbors and in-neighbors of the nodes u and v, respectively. Note that the DTO is defined only for edges (u, v) such that $|\mathcal{N}_u^{out}| > 1$ and/or $|\mathcal{N}_v^{in}| > 1$; otherwise, for isolated dyads, the DTO is assumed to be zero.

Since our focus is on users rather than links, we define the *node-centric DTO* ($nDTO$) of a node u as follows:

$$nDTO(u) = \frac{1}{|\mathcal{N}_u^{in} \cup \mathcal{N}_u^{out}|} \left(\sum_{v \in \mathcal{N}_u^{in}} DTO(v,u) + \sum_{v \in \mathcal{N}_u^{out}} DTO(u,v) \right)$$

The application of the node-centric DTO measure produces a ranking of the nodes, whereby higher ranks (i.e., lower scores) correspond to stronger bridge nodes. (Note that, for the sake of

Table 3.3. Percentage of Mastodon users regarded as strong-bridges and bridges for selected cut-off thresholds of the $nDTO$ score percentiles.

network	#nodes	#strong-bridges (%) *(w/o sources and sinks)*	$nDTO$ score		
			$5th$	$10th$	$25th$
mastodon.social	305 968	121 585 (39.7%) 81 112 (26.5%)	61.5%	64.6%	70.9%
pawoo.net	306 753	97 590 (31.8%) 38 842 (12.7%)	50.6%	53.9%	61.8%
mastodon.xyz	16 076	12 521 (77.9%) 1 086 (6.8%)	85.3%	86.2%	88.8%
mstdn.io	16 853	4 414 (26.2%) 290 (1.7%)	75.1%	82.6%	86.2%
octodon.social	7 082	1 436 (20.3%) 626 (8.8%)	64.8%	68.9%	74.4%
merged network	657 712	238 042 (36.2%) 122 927 (18.7%)	56.0%	59.7%	66.8%
full network	1 315 739	521 844 (39.7%) 197 237 (15.0%)	52.3%	57.7%	67.3%

readability, in the DTO and $nDTO$ definitions we have omitted the reference to the membership instances of u and v).

Table 3.3 shows the percentage of nodes corresponding to selected percentiles of $nDTO$ score over each of the top-5 instances as well as the merged and full networks. As it can be noted already for the 5-th percentile, it stands out that more than a half of the node set is identified as bridges, with peaks above 75% in *mastodon.xyz* and *mstdn.io*.

We also quantified the existence of nodes showing a strong status as bridges. We identified such nodes, dubbed *strong-bridges*, as the nodes having a $nDTO$ score equal to zero. In Table 3.3, we report the number of strong bridges for each of the top-5 instances: compared to the total number of nodes, the percentage of strong-bridges appears to be always significant, ranging from about 20% (*octodon.social*) to about 78% (*mastodon.xyz*).

We also evaluated the impact of source and sink nodes on the overall percentage of strong bridges found in our analyzed networks. As reported in Table 3.3, even by filtering out such nodes, the portion of strong bridges remains evident, unveiling a relatively low bias due to source and sink nodes at least in the largest networks, where the percentage of strong bridges ranges from about 13% in *pawoo.net* to above 27% in *mastodon.social*.

3.4.2 Over-consumption

In this section, we answer our fourth research question (**Q4**) regarding the identification of users that tend to over-consume information produced by others. To this purpose, we take a particular perspective on this problem, which relies on the theory of *lurking behavior analysis* [208, 72].

A key concept in this theory is that (online) social networks are characterized by a *participation inequality* principle, whereby the crowd of a social network does not actively contribute, rather it mostly remains hidden or "silent", without taking an active role in the visible participation and interactions with other members. This kind of users should not be trivially regarded as totally inactive users (i.e., registered users who do not use their account to join the online community), rather a silent user can be perceived as someone who gains benefit from information

Table 3.4. Percentage of Mastodon users regarded as lurkers w.r.t. selected cut-off thresholds based on the LurkerRank score percentiles.

network	*LR*		
	$95th$	$90th$	$75th$
mastodon.social	2.8%	7.7%	30.1%
pawoo.net	4.3%	9.3%	28.3%
mastodon.xyz	2.0%	6.0%	73.6%
mstdn.io	4.6%	8.9%	77.6%
octodon.social	38.7%	41.6%	56.3%
merged network	3.7%	9.0%	24.3%
full network	4.4%	8.2%	23.1%

produced by others (e.g., reading posts and comments, watching videos, etc.) without mostly giving back to the online community; within this view, these users are also called *lurkers*. It has been shown in several works (e.g., [208, 72, 204, 89, 227]) that lurking is normal and also an active, participative and valuable form of online behavior, including a form of cognitive apprenticeship that corresponds to legitimate peripheral participation. In this respect, lurkers might have a great potential in terms of social capital, since they acquire knowledge from the community; therefore, when engaged, they become beneficial for the propaganda and development of the community.

Modeling and analyzing lurking behaviors has been formulated as a eigenvector-centrality-based node ranking problem, which is totally content-agnostic, as it does not require other information than the graph topology [211]. The LurkerRank method was designed to assign each user a score expressing her/his lurking status. In ***Appendix A.2***, we report the mathematical details of this method. It should be noted that the LurkerRank method applies to a network graph with *reversed edge-orientation*, therefore hereinafter we shall consider any edge (u, v) as a link from u to v where v is a follower of u.

To answer the research question **Q4**, our main goal is to understand whether and to what extent lurkers of an instance are target nodes of an information flow coming either from the same instance or from a different instance. To this purpose, we compute the LurkerRank method to each of the top-5 instance networks as well as to the merged and full networks. In Table 3.4, we report the percentage of users identified as lurkers for selected percentiles of LR values, where LR symbol is used to denote the scoring function of LurkerRank (cf. ***Appendix A.2***). Looking at the table, we notice that the full and merged networks as well as each of the top-5 instances, but *octodon.social*, show a percentage of lurkers that is below 5% and 10% for the $95th$ and the $90th$ percentile, respectively, while for *octodon.social*, the percentage values at $95th$ and $90th$ percentiles are comparable and set around 40%. However, when extending to the $75th$ percentile, the percentage of users increases to at least approximately 30% (for *mastodon.social* and *pawoo.net*), with a peak above 70% in *mastodon.xyz* and *mstdn.io*. Note also that the increment on *octodon.social* appears to be at a significantly lower rate than for the other instance networks.

Information consumption. Once the lurkers at varying degrees were identified within the merged network, we investigated the links towards lurker nodes of a specific instance w.r.t. the overall incoming links, in order to understand how much the information flow is "consumed" by (i.e., it is directed to) lurkers, and whether this occurs internally or externally to their membership instance.

Table 3.5. Percentage of outgoing edges, resp. incoming edges, between pairs of selected instances that correspond to edges towards, resp. from, lurkers. Percentiles refer to LurkerRank scores.

source instance	target instance	#edges	edges *to lurkers*			edges *from lurkers*		
			95th	*90th*	*75th*	*95th*	*90th*	*75th*
mastodon.social	*mastodon.social*	3 408 327	12.6%	15.1%	25.3%	0.4%	0.6%	2.2%
	pawoo.net	45 713	3.4%	8.3%	19.5%	1.5%	1.9%	5.3%
	mastodon.xyz	46 069	5.8%	7.7%	44.4%	0.8%	1.0%	2.4%
	mstdn.io	29 425	7.0%	9.6%	26.8%	0.8%	1.0%	2.4%
	octodon.social	29 935	5.9%	8.7%	22.1%	0.9%	1.1%	3.0%
pawoo.net	*mastodon.social*	29 572	48.5%	49.4%	54.7%	1.0%	1.4%	3.8%
	pawoo.net	4 329 562	13.4%	16.8%	31.8%	0.1%	0.2%	0.8%
	mastodon.xyz	944	0.7%	12.0%	34.4%	0.1%	0.2%	1.6%
	mstdn.io	3 192	3.3%	6.4%	43.7%	0.5%	0.7%	2.3%
	octodon.social	945	11.7%	19.2%	39.8%	0.1%	0.1%	0.6%
mastodon.xyz	*mastodon.social*	36 328	19.5%	21.8%	34.1%	1.0%	1.9%	10.8%
	pawoo.net	6 684	1.3%	6.6%	17.8%	2.8%	5.9%	30.5%
	mastodon.xyz	35 631	4.5%	8.9%	46.2%	0.0%	0.1%	3.0%
	mstdn.io	1 417	9.5%	12.8%	41.6%	0.4%	0.6%	4.2%
	octodon.social	2 404	3.2%	5.9%	14.6%	0.2%	0.6%	5.2%
mstdn.io	*mastodon.social*	28 523	15.2%	16.7%	23.2%	1.8%	2.7%	12.6%
	pawoo.net	6 691	1.6%	4.9%	13.3%	5.7%	8.3%	35.5%
	mastodon.xyz	803	6.2%	8.1%	24.7%	1.0%	1.9%	15.8%
	mstdn.io	112 805	4.5%	6.6%	25.3%	0.0%	0.0%	0.7%
	octodon.social	626	7.7%	11.8%	22.0%	1.3%	2.7%	18.4%
octodon.social	*mastodon.social*	34 158	17.6%	19.1%	28.0%	2.5%	7.5%	14.5%
	pawoo.net	84	0.0%	0.0%	0.0%	4.8%	4.8%	6.0%
	mastodon.xyz	2 281	4.4%	5.9%	18.8%	1.1%	3.0%	5.8%
	mstdn.io	941	10.1%	13.3%	39.3%	1.0%	1.9%	5.5%
	octodon.social	34 493	20.8%	24.3%	37.2%	0.5%	1.2%	7.7%

We report the results of our analysis in Table 3.5 under the column "edges *to lurkers*", for each pair of the top-5 instances — including self-pairing, i.e., within-instance links — and for various lurking score percentiles. At a first glance, it can be noted a certain variety in the percentage values, which indicates a remarkable differentiation of information consumption by lurkers within and across the various instances.

On the one hand, there is an evidence of information flow directed to lurkers from inside their membership instance, although this happens at different extents; in particular, at the $95th$ percentile, the percentage of links directed to lurkers ranges from 4.5% in *mstdn.io* and *mastodon.xyz* to about 21% in *octodon.social*.

On the other hand, however, there is also a remarkable amount of information flow directed to lurkers from outside their membership instance. In this respect, the *mastodon.social* instance turns out to be the best target for lurkers, given the highest percentages of links coming from the other instances (and *mastodon.social* itself) and directed to lurkers. In particular, we notice a considerable amount of information flow from *pawoo.net* to lurkers in *mastodon.social*, which is about 50% of the connections from *pawoo.net* to *mastodon.social*. Also, *mastodon.xyz* and

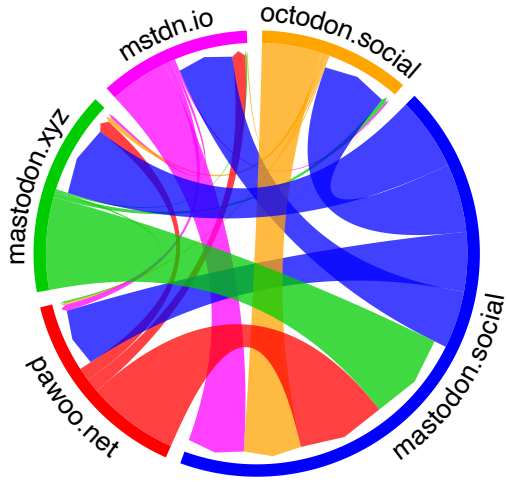


Fig. 3.5. Graphical illustration of the information flow between the top-5 Mastodon instances, modeled from information producers to information consumers. Self-loops and flow values (cf. Table 3.5) are omitted to avoid cluttering. Each flow has the same color as the source instance.

mstdn.io lurkers tend to absorb most information from outside, while *pawoo.net* is particularly relevant as information producer for its own users as well as for users of the other instances.

Overall, the above remarks highlight an important trait of the Mastodon network as a mix of within-instance and across-instance information consumption of its users. Figure 3.5 provides an illustration of the information flow between the top-5 instances, which complements our understanding from the results shown in Table 3.5 by highlighting a sort of mutual reinforcement among such instances, in terms of information production, resp. consumption, behaviors exhibited by their users.

Information spreading. Analogously to the previous analysis, we examined how the outgoing links from an instance might be originated by lurkers.

Clearly, as expected from the definition of lurking behavior, lurkers do not contribute much to the diffusion of the information they consume, which is indicated by the small percentage values reported in Table 3.5, under the column "edges *from lurkers*". However, some exceptions stand out. In particular, already at the $95th$ percentile, lurkers of *mastodon.xyz*, *mstdn.io* and *octodon.social* contribute to spread information towards *pawoo.net* in a non-negligible way. This trend strengthens at the $90th$ percentile and, for *mastodon.xyz* and *mstdn.io*, is boosted at the $75th$ percentile (with peaks above 35% from *mstdn.io* to *pawoo.net*). The above is remarkable as we recall that *pawoo.net* tends to attract lurkers belonging to other instances (e.g., *mastodon.social* and *octodon.social*), and conversely, we also spotted that lurkers from all the other instances (especially from *mastodon.xyz*, *mstdn.io*, *octodon.social*) might contribute to the diffusion of information towards *pawoo.net*.

Table 3.6. Dual role users in the top-5 instance networks, the merged network, and the full network.

network	#nodes	$LR@95th \cap nDTO@5th$	$LR@90th \cap nDTO@10th$	$LR@75th \cap nDTO@25th$
mastodon.social	305 968	0.7%	5.1%	26.1%
pawoo.net	306 753	1.6%	4.9%	20.4%
mastodon.xyz	16 076	1.1%	4.9%	72.3%
mstdn.io	16 853	0.5%	7.3%	76.6%
octodon.social	7 082	37.6%	39.5%	51.5%
merged network	657 712	1.5%	6.0%	18.9%
full network	1 315 739	2.9%	5.2%	17.3%

3.5 Dual role users

Our fifth question (**Q5**) concerns unveiling the existence of Mastodon users that take a twofold role as lurkers and bridges. To this purpose, for each of the top-5 instances, the merged network, and the full network, we analyzed the overlap between a set of lurkers and a set of bridge users, selected from their respective ranking solutions according to the percentile thresholds used in the previous analysis, such that each set pair refers to the same percentile proportion (i.e., $95th$ vs. $5th$, $90th$ vs. $10th$, and $75th$ vs. $25th$).

Table 3.6 reports on the results of our analysis. The *octodon.social* instance shows by far the largest percentage of users exhibiting the dual role, which is already above 37% w.r.t. the toughest overlap (i.e., $LR@95th \cap nDTO@5th$), then settling around 51% for the smoothest overlap (i.e., $LR@75th \cap nDTO@25th$). Note that the latter point corresponds to an increase rate that is much lower than for the other instances, where the percentages of dual roles keep far below 10% w.r.t. the two largest overlaps while increasing up to a minimum of 20% (*pawoo.net*) and a maximum of 77% (*mstdn.io*) w.r.t. the overlap $LR@75th \cap nDTO@25th$. Interestingly, this is in accord with the higher, resp. lower, smoothness in the role identification shown by *octodon.social*, resp. the other instances, for varying scoring percentiles, as we already observed in our previous analysis (cf. Table 3.5). Moreover, the difference in percentages corresponding to the $LR@75th \cap nDTO@25th$ between the two largest instances (i.e., *mastodon.social* and *pawoo.net*) and the other three instances also depends on the size of their respective user-bases: in fact, as the number of users in an instance gets smaller, the volume of information produced becomes more limited, and hence their users tend not only to consume it but also to act as *information flow facilitators*; by doing this, they can contribute keeping the instance sustainable with fresh contents and timely interactions.

Considering the merged and full networks, there is evidence of a certain presence of dual role users — thus indicating that a dual role behavioral phenomenon can also occur in a cross-instance context — with percentages that are in line with some of the instances, particularly *pawoo.net*.

3.6 Alternate role users

Here we consider our sixth research question (**Q6**). The goal is to understand whether by mixing the scales of observation, i.e., either locally within an instance or globally at the level of the merged network, distinct behaviors of users may arise. As for the previously analyzed

Table 3.7. Mastodon users who behave differently according to the observation scale, i.e., locally within their instance (denoted with superscript (L)) or globally at the level of merged network (denoted with superscript (G))

	mastodon.social	*pawoo.net*	*mastodon.xyz*	*mstdn.io*	*octodon.social*
#users	305 968	306 753	16 076	16 853	7 082
$LR@95th^{(L)} \cap nDTO@5th^{(G)}$	0.6%	1.6%	0.8%	0.3%	1.2%
$LR@90th^{(L)} \cap nDTO@10th^{(G)}$	4.8%	5.0%	2.7%	2.4%	36.2%
$LR@75th^{(L)} \cap nDTO@25th^{(G)}$	25.0%	21.2%	66.7%	72.7%	45.1%
$nDTO@5th^{(L)} \cap LR@95th^{(G)}$	0.2%	3.1%	0.3%	0.1%	0.1%
$nDTO@10th^{(L)} \cap LR@90th^{(G)}$	0.7%	12.2%	0.4%	0.2%	0.1%
$nDTO@25th^{(L)} \cap LR@75th^{(G)}$	13.3%	26.6%	2.7%	1.2%	0.5%

research questions, we focus on lurkers and bridge users, thus aiming to identify whether users can be regarded as lurkers inside their membership instance yet as bridges in a cross-instance environment, and vice versa.

In Table 3.7, we report the percentage of users of a given instance that are identified as users showing a lurking role locally and a bridging role globally (upper subtable). As it can be noted, while a few cases (below 2%) are already identified w.r.t. $LR@95th^{(L)} \cap nDTO@5th^{(G)}$, this alternate behavior becomes more evident w.r.t. larger overlaps, on all instances though at different extents. These results allow us to understand how the information flow moves within a decentralized context. An intra-instance (or local) lurker is a user who tends to absorb information, while an inter-instance (or global) bridge is a user who contributes to connect multiple regions of a network of instances. It follows that the users having this dual scale-dependent role are those who, while consuming locally produced information, enable the information coming from their instances to flow into the Fediverse, thus becoming potential information facilitators. Moreover, as already partially unveiled in our previous analysis on dual role users, the higher percentages of alternate role users generally found for instances with a smaller user base suggest a tendency of users in such instances to act as a touch-point and interconnect different regions that cross the instance boundaries.

In Table 3.7, we also report the percentage of users of a given instance that are identified as users showing a bridging role locally and a lurking role globally (bottom subtable). We observe that the percentage values are generally much lower than the previously discussed behavioral case. This should not be surprising since if a user takes a within-istance bridge role, s/he is already committed to broker information and hence will likely be less inclined to absorb information from the outside. Nonetheless, in this scenario, *mastodon.social* and *pawoo.net* represent an exception, showing non-negligible overlaps of alternate role users under less restrictive percentile thresholds. We tend to ascribe this phenomenon to aspects related to the topology of those instances; in particular, the sparsity of connections over a large user base would favor some users to absorb information from other instances while acting as bridges locally.

3.7 Discussion

Here we summarize the main findings that raised from our extensive analysis of the Mastodon user relations.

To answer our first research question (**Q1**), we explored the main structural characteristics of the Mastodon user network. Among the noteworthy facts, we observed a lack of degree correlation, which should be ascribed to a form of spontaneous connectivity between users that relates to the absence of boosting mechanisms for "artificial" interactions, such as those due to the widely used recommendation strategies adopted by the centralized OSNs. From a mesoscopic perspective, based on Louvain, Leiden, and Infomap community detection methods, the user networks exhibit a moderately high modularity (around 0.7) and a high number of communities; this trait, which indicate the existence of small densely connected groups of users tailored to specific shared interests, appears to be consistent with the spontaneous connectivity trend in Mastodon.

Remarkably, all the specific traits discovered on the full user-network remained valid also after our step of graph pruning aimed at removing irrelevant instances. This was further strengthened when we considered a set of instances able to represent the entire Mastodon user network, as outlined by our second research question (**Q2**). To this purpose, supported by some pertinent results from the study in [139], we focused on the five most relevant instances in Mastodon. After evaluating their main structural features and finding high consistency with the results obtained on the full user-network, we concluded that the top-5 instance network can be regarded as representative of the whole Mastodon user network, and indeed we used it in our subsequent tasks of user behavior analysis.

To answer our third research question (**Q3**), we investigated the linkage between Mastodon users accounting for the instance boundaries. We indeed found out a significant fraction of inter-instance links and of shell nodes (i.e., users having connections with other instances' users only), thus unveiling an evident boundary-spanning phenomenon, as also confirmed by our visual inspection in Figs. 3.3 and 3.4. We delved into the boundary-spanning mechanisms in Mastodon through the identification of users acting as bridges at varying degrees. To this purpose, by leveraging the notion of directed topological overlap, we discovered a widespread presence of bridge users, with a non-negligible fraction of what we called strong-bridges, i.e., users having a topological overlap equal to zero. Interestingly, this still holds even by removing the source and sink nodes from the network. Therefore, we can state the existence of structurally strategic nodes holding connections between across-instance regions of the network, which positively impacts on the effectiveness of information flow between the users over all Mastodon.

As for our fourth research question (**Q4**), we modeled the information over-consumption phenomenon through the Mastodon user network in terms of lurking behaviors. As thoroughly discussed in the literature, lurkers are silent users who tend to mostly consume information from the others' actions rather than produce information; but at the same time, by holding a certain social capital and given their pervasiveness in a social network, such users might significantly contribute to boundary spanning and information flow phenomena. In this regard, we built our analysis upon a theoretically well-founded content-agnostic eigenvector-centrality ranking method, LurkerRank. Our goal was twofold: to understand whether and to what extent lurkers of an instance are target nodes of an information flow coming from other users, and whether this involves the membership instance or the other instances. In this regard, we found out that lurkers are present, at varying degrees, over all the selected instances under study, with *mastodon.social* being the preferred instance for information consumption by lurkers. In general, information consumption is not confined to the membership instance, but it extends beyond the boundaries of the instances, so as to further capitalize on the information exchanged through different regions of the Mastodon user network. Furthermore, we unveiled that lurkers are also involved in information spreading processes between instances, even in a non-negligible

way as it happens for users in *pawoo.net* that are linked to (i.e., follow) lurkers of the other instances.

Our last research questions regarded the existence of users who show a dual lurker-bridge role, either simultaneously through the whole user network and the instance-specific subnetworks (**Q5**), or alternately as a function of the observation scale, i.e., inter-instance and intra-instance perspective (**Q6**). We found a relatively small fraction of users acting both as lurkers and bridges within their own instances; since these users normally over-consume but have also the potential of disseminating information, they could be regarded as information flow facilitators. This trait is present through all the merged network, and is particularly evident in the smallest instances, where the produced information is limited to the size of the audience in those instances, and hence an amplification is needed. Concerning the alternate and scale-dependent behavior, we spotted the existence of users acting as local (i.e., on their own instances) lurkers and global (i.e., between instances) bridges, whereas the contrary does not hold. Reasonably, the former trait allows users to disseminate information from their own instances outwards, while the latter is unnecessary as they are already responsible for the intra-instance information spreading. As a final remark, we believe that such dual/alternate-role users can be regarded as highly strategical ones, as their complementary structural functionality makes them ideal candidates to determine the speed and scope by which the information flows within Mastodon, and more generally, in a decentralized social context.

3.8 Chapter review

Decentralized Online Social Networks (DOSNs) aim to bring the social paradigm back to its roots made up of spontaneous interactions and genuine interests, in contrast to the marketing-driven engagement mechanisms typically adopted by the centralized OSNs. To guarantee a user-centric vision, DOSNs support the creation of independent and self-hosted servers seamlessly connected among each other. Nevertheless, this metamorphosis of the online social media environment could change how some of the fundamental components underlying human relationships appear and evolve via the Internet.

In this work, we have provided a number of insights into DOSN user relations and behaviors, using as a case in point Mastodon, the most-known service of the Fediverse. We analyzed the Mastodon user network to answer six research questions encompassing the main structural characteristics of the following user relations, the impact due to the most representative instances on the user network, across-instance boundary spanning and bridges, over-consumption and information flow, dual and alternate role users.

As a natural follow-up of this work, we are currently investigating the impact that decentralization has on user behaviors and how these adapt to enable information flowing quickly across instances. In this regard, we focus on the dichotomy between information consumption and production as a proxy to gain some insights into two interesting behavioral phenomena still unexplored in the decentralized scenario: behavioral *consistency* and *alternation* of users acting as consumers and/or producers across the Mastodon instances.

We believe that our work can pave the way for new and interesting studies concerning the decentralized social landscape, also by broadening the scope to other emerging and innovative decentralized approaches, such as *Blockchain-based Online Social Media* (BOSMs) [94], of which *Steemit* represents the best known case to date in the research community [147, 122, 96, 101, 15].

4

Network Analysis of the Information Consumption-Production Dichotomy in Mastodon User Behaviors

Summary. Decentralized Online Social Networks (DOSNs) are today an established alternative to the popular centralized counterparts. In this work, we push forward research on user behaviors in a decentralized context, by exploring the dichotomy between information consumption and production. Using the Mastodon user network as a proxy for the Fediverse landscape, we address two main research questions: Do the consumers, resp. producers, identified in one instance exhibit the same behavior consistently while interacting with other instances? and, Are there users who behave as consumers in one instance and simultaneously as producers in other instances, or vice versa? In this respect, our results reveal interesting traits of Mastodon users, yet unveil the emergence for further studies that can embrace other services in the Fediverse.

4.1 Contributions

Despite the recent corpus of studies on Mastodon, several questions still remain open particularly about how users behave in the decentralized scenario. We believe one important direction concerns how the seamless interaction between users of different instances — which, in contrast to centralized platforms, does not require a user having multiple accounts or subscriptions — might impact on the users' contribution to the community life, and the role(s) they might take in their home instance as well as in any other instance where they are involved.

Within this view, in this work we aim to fill a gap in understanding the user behaviors in the Fediverse, through the lens of Mastodon. Our focus is on the dichotomy between information-consumption and information-production behaviors of users across Mastodon instances. In this respect, we want to answer the following research questions:

RQ1 – *Do the consumers, resp. producers, identified in one instance exhibit the same behavior consistently while interacting with other instances*?

RQ2 – *Are there users who behave as consumers in one instance and simultaneously as producers in other instances, or vice versa*?

The underlying motivation for the above research questions stems from our interest in understanding whether DOSNs, and Mastodon in particular, may exhibit similar user behavioral patterns as those observed in some groups of centralized OSNs. For instance, as discussed in [175], it is often the case that a user can produce information and actively interact in certain

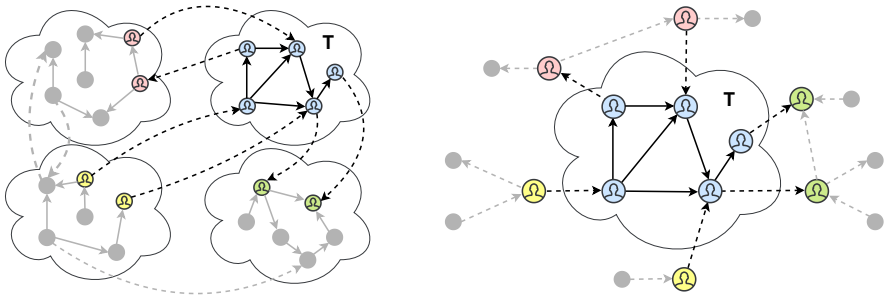


Fig. 4.1. Illustration of our ego-network model applied to a target instance (T) in the Mastodon user network. Node colors indicate home instances. Solid black links are between users belonging to the target instance, whereas dashed black links are from/to other instances. Gray nodes/edges refer to existing entities not involved in the ego-network of T.

platforms where s/he is subscribed, but it may also be the case that the same user can assume a *silent* behavior on other platforms.

To the best of our knowledge, information-consumption vs. information-production has not been studied so far in Mastodon. We also point out that, differently from [140] where information consumption is studied isolatedly, the dichotomous coupling with information-production needs to be analyzed through an unprecedented modeling of the user relations in Mastodon.

4.2 Methodology

Data. We used the data provided in [139], which represents the most complete and up-to-date network dataset of Mastodon. This dataset contains more than 1.3M unique users and about 17M unique links between users. Also, according to [140], it covers about 78% of the Mastodon user base to date, thus enabling a representative study of the Mastodon scenario.

Network model. Let us denote with $\mathcal{U}$ and $\mathcal{I}$ the set of users and instances, respectively, available in the Mastodon dataset. We define the Mastodon user network as a directed graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where the node set $\mathcal{V}$ contains user-instance pairs, i.e., $\mathcal{V} = \{(u, i) \mid u \in \mathcal{U},\ i \in \mathcal{I}\}$, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ where any $(x, y) \in \mathcal{E}$, with $x = (u, i)$ and $y = (v, j)$, means that user v in instance j *receives information produced by* user u in instance i. It should be noted that u and v may coincide only if $i \neq j$.

Given a specific instance $i \in \mathcal{I}$, we define the *extended ego-network* of i as the directed subgraph $\mathcal{G}_i = \langle \mathcal{V}_i, \mathcal{E}_i \rangle$, induced from $\mathcal{G}$, such that $\mathcal{V}_i \subseteq \mathcal{V}$ and $\mathcal{E}_i = \{(x, y) | \ x = (u, j),\ y = (v, k) \wedge (j = i \vee k = i)\} \subseteq \mathcal{V}_i \times \mathcal{V}_i$. Figure 4.1 reports an illustration of our ego-network model.

Identification of consumers and producers. To identify users that tend to over-consume information produced by others, we take a perspective that relies on the theory of *lurking behavior analysis* [208, 72]: the majority of OSN users does not actively contribute, rather it mostly remains hidden or "silent", gaining benefit from information produced by other users. Modeling and analyzing lurking behaviors has been formulated as an eigenvector-centrality-based node ranking problem, which is content-agnostic, and builds upon three key principles

[211]: (*i*) content over-consumption, (*ii*) the authoritativeness of the information received, (*iii*) the non-authoritativeness of the information produced. The first shapes the imbalance between the amount of information a user consumes w.r.t. the amount of information she/he produces, whereas the others refer to the importance as information producer of her/his in-neighbors, and the importance as information consumer of her/his out-neighbors, respectively.

Given a directed graph $\mathcal{G}$, here corresponding to the Mastodon user network or to any instance-specific ego-network, the LurkerRank score $LR(v)$ of any node v according to the *in-out-neighbors-driven lurker ranking* formulation [211] is defined as:

$$LR(v) = \alpha[LR_{in}(v)(1 + LR_{out}(v))] + (1 - \alpha)p(v), \tag{4.1}$$

where LR_{in} (*in-neighbors-driven* lurking function) is:

$$LR_{in}(v) = \frac{1}{|\mathcal{N}_v^{out}|} \sum_{u \in \mathcal{N}_v^{in}} \frac{|\mathcal{N}_u^{out}|}{|\mathcal{N}_u^{in}|} LR(u), \tag{4.2}$$

and LR_{out} (*out-neighbors-driven* lurking function) is:

$$LR_{out}(v) = \frac{|\mathcal{N}_v^{in}|}{\sum_{u \in \mathcal{N}_v^{out}} |\mathcal{N}_u^{in}|} \sum_{u \in \mathcal{N}_v^{out}} \frac{|\mathcal{N}_u^{in}|}{|\mathcal{N}_u^{out}|} LR(u). \tag{4.3}$$

$\mathcal{N}_u^{in}, \mathcal{N}_u^{out}$ are the in-, out-neighbor sets of u, α is a damping factor in $[0, 1]$ (by default 0.85), and $p(v)$ is the value of the PageRank-like *personalization vector* (by default $1/|V|$). To avoid zero or infinite ratios, the values of the in/out-neighborhood size of a node are Laplace add-one smoothed.

According to the LurkerRank, the higher the LR-score of a node, the stronger is the status of the node as consumer in the network. Conversely, as demonstrated in [211, 175], the bottom of a LR ranking can be used to identify the users that act as opposed to consumers, i.e., producers. We hence leverage an analysis of ranking *heads* and *tails* to model the dichotomy between consumers and producers in the ego-network model.

It should be noted that either social and interaction relations can be seen as proxy for information consumption by users; indeed, LurkerRank has been extensively evaluated on both followee-follower and comment/like/mention graphs [211, 212]. Since the available Mastodon user relations are of "following" type, in this work LurkerRank is applied to followship graphs.

Evaluation goals and assessment criteria. Our goal is to answer the previously stated **RQ1**-**RQ2** based on an evaluation of the LR ranking solutions obtained on each instance's ego-network. To this purpose, we first use the *Jaccard similarity* coefficient to measure the matching degree between (portions of) LR rankings L_i, L_j of any two given ego-networks: $Jaccard(L_i, L_j) = (|L_i \cap L_j|)/(|L_i \cup L_j|)$.

To delve into the comparison of two rankings, we also resort to the *binary preference* criterion, which measures how often judged relevant items are retrieved in a list L before judged non-relevant ones [38]:

$$Bpref(L) = \frac{1}{R} \sum_{r} \left(1 - \frac{min(\#n \text{ ranked above } r, R)}{min(N, R)}\right), \tag{4.4}$$

where r and n are *relevant* and *non-relevant* judged items, resp., on a total number R (resp. N) of relevant (resp. non-relevant) items. Note that *not-judged* items may be present in L, although their position is discarded. Bpref ranges within $[0, 1]$, whereby the closer to 1 the better the Bpref.

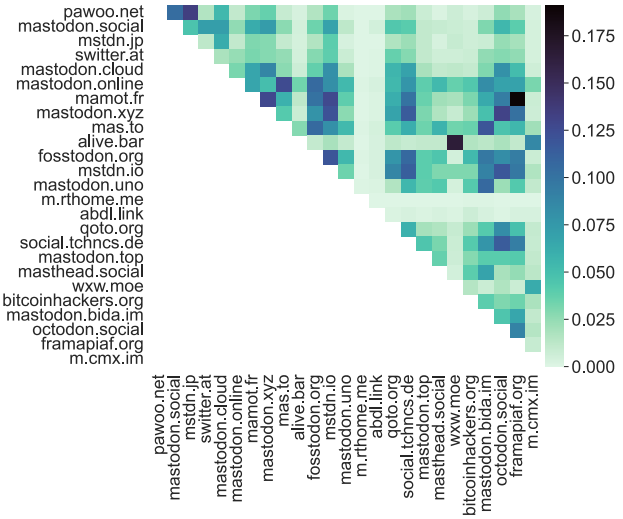


Fig. 4.2. Jaccard similarity between the $tail@25$ from the LR rankings, i.e., top-producers, for each pair of instances in the top-25 instances context.

Given a pair of ego-networks $(\mathcal{G}_i, \mathcal{G}_j)$, and (portions of) their respective LR rankings L_i, L_j, for any choice of a *reference instance*, say L_i, we compute the Bpref of L_j w.r.t. L_i, such that: the users common to the instances i and j correspond to the relevant items, the users in $\mathcal{G}_j$ having j as home instance correspond to the not-judged items, and the users in $\mathcal{G}_j$ not having j as home instance correspond to the non-relevant items. Note that our definition of Bpref is asymmetric as it depends on the choice of the reference instance, i.e., Bpref of L_j w.r.t. L_i is not necessarily equal to Bpref of L_i w.r.t. L_j.

Settings. To ensure significance of our results yet representativeness of the currently active Mastodon landscape, we focused on the top-25 instances by user base according to the *instances.social* platform, which is widely recognized as the de-facto tracker for Mastodon. We organized such instances into three subsets called *contexts*, namely top-5, top-10, and top-25 instances, and for each instance in a given context, we induced its ego-network. Note that the same instance might have less/more external edges, hence a different ego-network, depending on the selected context.

Moreover, given an LR ranking solution, we considered the top-k%, resp. bottom-k%, users by score in LR, dubbed as $head@k$, resp. $tail@k$. We set by default $k = 25$.

4.3 Results

To answer our first research question (**RQ1**), we begin with a pairwise comparison of the ego-network LR heads, resp. tails. Considering the top-5 instances subset, there is some evidence of Jaccard similarity between the head, resp. tail, at $k = 25$, of their corresponding LR rankings, up to 0.04 and 0.14, respectively. Broadening the scope at the top-10 context, we observe analogous Jaccard coefficient values, with peaks of 0.04 resp. 0.14 for pairwise comparisons of $head@25$ resp. $tail@25$, and of 0.09 resp. 0.19, for pairwise comparisons of $head@25$

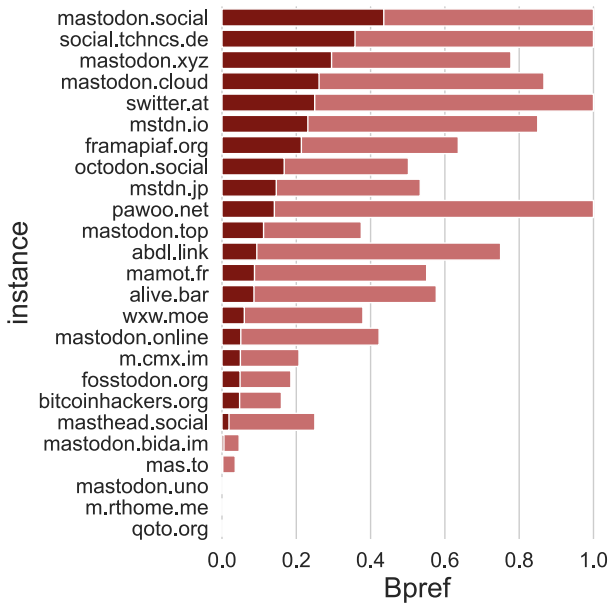


Fig. 4.3. Bpref on the *head*@25 from the *LR* rankings, i.e., top-consumers, for each of the top-25 instances. The average and maximum values for each reference instance over the others are shown in red and light red, respectively. Instances are ordered by decreasing average Bpref values.

resp. *tail*@25 when we account for the largest context (i.e., top-25 instances). These results suggest the existence of a few users that, in at least two different instances, exhibit behavioral consistency, and this holds more for producers than consumers. Moreover, not only this finding is robust to the extent of the context of comparison of instance pairs, but also to the size of the ranking heads and tails (results with $k \in \{5, 10\}$ follow trends analogous to $k = 25$). Due to space limitations, in Fig. 4.2 we report results only for the largest ranking scope ($k = 25$) and context (top-25 instances) of top-producers.

Note that the above results are valid for pairwise comparisons. When extending to triplets of instances, however, we find Jaccard values close or equal to zero in all cases.

We then inspected the interesting cases observed for the pairwise scenario through our Bpref-based evaluation. Figures 4.3-4.4 show results for each instance in the top-25 context. Interestingly, the reference instance with the highest average Bpref turns out to be the first established instance in Mastodon, i.e., *mastodon.social*, for consumers resp. producers, with average values up to 0.436 resp. 0.593, for $k = 25$ in the top-25 context. Moreover, 4 resp. 8 out of 25 instances reach maximum Bpref equal to 1 for the consumer resp. producer evaluation. It should be noted that the maximum Bpref scores relating to producers (Fig. 4.4) are on average greater than those of consumers (Fig. 4.3), as analogously observed for the Jaccard based evaluation.

As concerns our second research question (**RQ2**), we start again by considering a pairwise instance comparison, choosing this time the ranking head (i.e., the top consumers) from one instance and the tail (i.e., the top producers) from the other instance in each pair. Jaccard

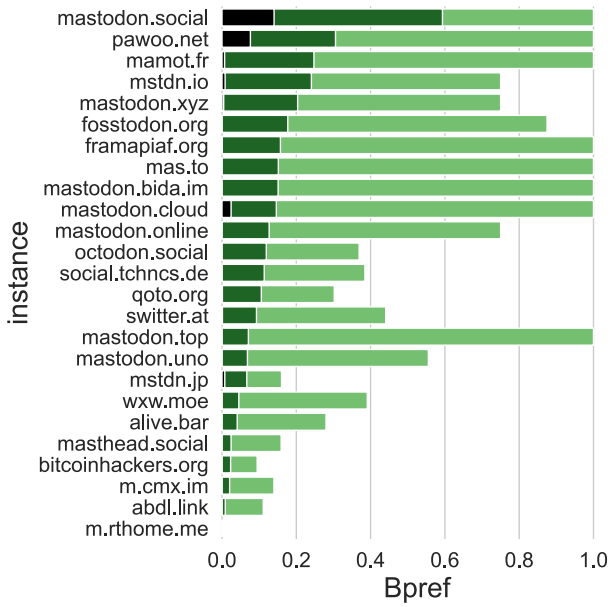


Fig. 4.4. Bpref on the $tail@25$ from the LR rankings, i.e., top-producers, for each of the top-25 instances. The average, maximum, and minimum values for each reference instance over the others are shown in green, light green, and dark green, respectively.

similarity results for $k = 25$ (not shown) reveal maximum values below 0.1 regardless of the instance context. Nonetheless, we also take a finer-grain perspective through Bpref, whereby, for each instance j and reference instance i in a top-N context (with $i \neq j$), we aggregated over both the Bpref of $(tail@k)_j$ w.r.t. $(head@k)_i$ and the Bpref of $(head@k)_j$ w.r.t. $(tail@k)_i$. Figure 4.5 shows results for the top-25 instances and $k = 25$. We observe that 7 out of 25 instances show maximum Bpref above 0.4, with 3 instances reaching Bpref at 1. On average, however, the scores appear to be quite lower than the previous evaluation concerning **RQ1**, with *mastodon.uno* as the instance with the highest average score (0.118).

4.4 Chapter review

We investigated on the impact that the Fediverse decentralization might have on the user behaviors in terms of information production and consumption, either in a repeated or an alternate fashion across two or more instances in Mastodon.

Our analysis of the instance-specific ego-networks' LR ranking heads and tails to capture the consumption-production dichotomy unveiled a few interesting facts, which are summarized as follows.

• There exists a small number of largest Mastodon instances (less than 25) in which either pairwise behavioral consistency and alternation can be observed.

• The fraction of users that are regarded as information consumers, resp. producers, simultaneously in two instances is below 0.1, resp. 0.2, Jaccard similarity. This holds to a less

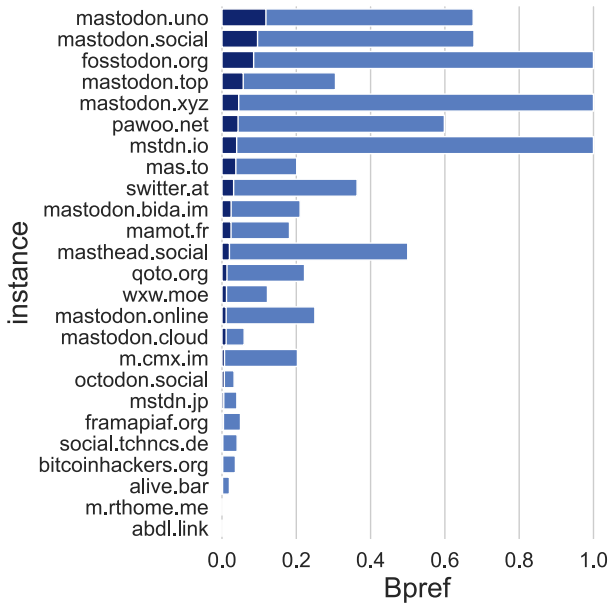


Fig. 4.5. Bpref on the $head@25$ w.r.t. $tail@25$ from the *LR* rankings, and vice versa, for each of the top-25 instances. The average resp. maximum values for each reference instance over the others are shown in blue resp. light blue.

extent when comparing top-consumers of one instance with top-producers of another instance, and vice versa.

• Although statistically not relevant in terms of absolute number of users, the behavioral consistency exhibited by certain consumers resp. producers (**RQ1**) appears to be relatively strong according to our Bpref evaluation of the rankings' heads resp. tails. This is more evident on average for producers, which might be explained by the ease of interaction and content dissemination favored by Mastodon, thus reducing the attitude of having a silent behavior.

• The behavioral alternation (**RQ2**) also turns out to be limited to low fractions of users per instance-pairs, and with lower Bpref strength than for the behavioral consistency.

Overall, **RQ1** and **RQ2** get moderately affirmative answers, and we suspect that this may generally hold in other platforms of the Fediverse. However, our findings need to be taken with a grain of salt, because of two main reasons: (1) our analysis context is content-agnostic and the consumption-production dichotomy is modeled through follower-followee relations, which are in principle a weaker proxy of the user activity in OSNs. (2) Mastodon instances are designed to be interrelated, rather than being perceived as independent and separate OSN platforms; in fact, differently from what happens when the same user has to subscribe to multiple centralized platforms, Mastodon users could feel a limited need of taking a silent behavior on some instances while being producers on other instances. This, however, might change when extending the analysis also to instances of other services in the Fediverse than Mastodon.

5

Drivers of Social Influence in the Twitter Migration to Mastodon

Summary. The migration of Twitter users to Mastodon following Elon Musk's acquisition presents a unique opportunity to study collective behavior and gain insights into the drivers of coordinated behavior in online media. We analyzed the social network and the public conversations of about 75,000 migrated users and observed that the temporal trace of their migrations is compatible with a phenomenon of social influence, as described by a compartmental epidemic model of information diffusion. Drawing from prior research on behavioral change, we delved into the factors that account for variations of the effectiveness of the influence process across different Twitter communities. Communities in which the influence process unfolded more rapidly exhibit lower density of social connections, higher levels of signaled commitment to migrating, and more emphasis on shared identity and exchange of factual knowledge in the community discussion. These factors account collectively for 57% of the variance in the observed data. Our results highlight the joint importance of network structure, commitment, and psycho-linguistic aspects of social interactions in characterizing grassroots collective action, and contribute to deepen our understanding of the mechanisms that drive processes of behavior change of online groups.

5.1 Introduction

After years of steady growth, popular social media are experiencing shifts in user engagement. In Twitter, such a shift has been especially abrupt after business magnate Elon Musk purchased the platform on October 26^{th} 2022. The acquisition itself, as well as several controversial management decisions taken by Musk shortly after [125] (including massive layoffs, the suspension of some journalists' accounts, and the discontinuation of free API access) threw Twitter at the center of a media storm and caused abrupt changes in the typical platform activity, [110] motivating many users to seek substitute services to migrate to.

Decentralized Online Social Networks (DOSNs) have experienced significant growth in recent years, capturing the attention of mainstream social media users. [61] Among these networks, *Mastodon* has emerged as the leading decentralized alternative to Twitter. [245, 139, 140, 141] Similar to email services, Mastodon allows communities to independently manage their own "instances" (i.e., servers) and connect with others through a federated approach facilitated by a common protocol. Following Elon Musk's takeover, Twitter users advocating for a transition to Mastodon promoted the #TwitterMigration movement, resulting in a rapid surge of registration requests on Mastodon instances.

This mass exodus from Twitter represents one of the largest digital migrations in the history of the Social Web and a unique example of collective behavioral change that is documented through large-scale digital traces, that can thus be studied quantitatively and at scale. Moreover, this phenomenon exhibits two uncommon properties that render it especially interesting from the perspective of behavioral change studies. [106] First, despite being prompted by external circumstances, the migration unfolded organically within Twitter, with users engaging in discussions and potentially influencing their peers by signaling their intention to migrate. Second, transitioning to a different social platform entails practical and psychological costs associated with changing habits, [236] as well as a social cost associated with adopting a behavior that deviates from mainstream norms; [88] these characteristics are shared with other grassroots processes of behavioral change that are generally desirable for human societies, [188, 115, 238] which further speaks to the significance of studying this migration.

Early studies have taken initial steps in characterizing the #TwitterMigration phenomenon primarily in terms of user activity, revealing that the majority of migrated users congregated on a few Mastodon instances, [243] while also continuing to post content on Twitter. [119] However, the underlying *drivers* motivating Twitter users to migrate to Mastodon remain largely unexplored, and this work is a first attempt to fill this gap.

We hypothesize that the migration was partly determined by social influence. Drawing inspiration from previous studies, [203, 52] we employed a compartmental model of information diffusion to describe the phenomenon at a macroscopic level, and examined whether the temporal pattern of migration is compatible with the typical dynamics of information contagion. We then shifted our analysis to the mesoscopic level and investigated different communities on Twitter to identify the factors that account for variations in the effectiveness of the influence process. Crucially, the factors we considered are rooted in three branches of interdisciplinary research on behavior change, which led us to formulate three distinct hypotheses regarding the *structure* of the social network, the *commitment* of community members, and their *language use*.

First, we hypothesize that the structural characteristics of the Twitter social network correlate with the rate of behavioral contagion across communities. This hypothesis is motivated by sociological and network science studies highlighting how structural attributes of social graphs, such as group size, cohesiveness, and the presence of influential authorities, can influence the dynamics of information diffusion and behavioral change within social networks. [143] Our second hypothesis draws on controlled experiments demonstrating that a committed minority of individuals can trigger tipping points in opinion formation. [21, 42] Consequently, we examine whether communities exhibiting higher commitment to the #TwitterMigration discussion also displayed faster influence processes. Our third and final hypothesis builds upon prior research in social psychology, which has established connections between psycho-linguistic aspects of social interactions and various outcomes related to consensus, agreement, and coordination. [29, 155, 104, 105, 64] Leveraging recent advances in natural language processing, we extracted high-level language dimensions that convey specific social intents. [50] Here, we hypothesize that communities engaging in conversations rich with positive social intent, such as knowledge exchange, expressions of trust, and identity markers, exhibit a more process of social influence.

Our findings reveal that communities where the peer influence to migrate appears to be more effective exhibit lower density of social connections, higher commitment to the discussion, and frequent emphasis on shared *identity* and exchanges of factual *knowledge*. Collectively, these factors explain more than half of the observed variance in the data.

Our study departs from previous research on online migrations. Existing studies that aimed to characterize user transitions from one digital platform to another have either predominantly

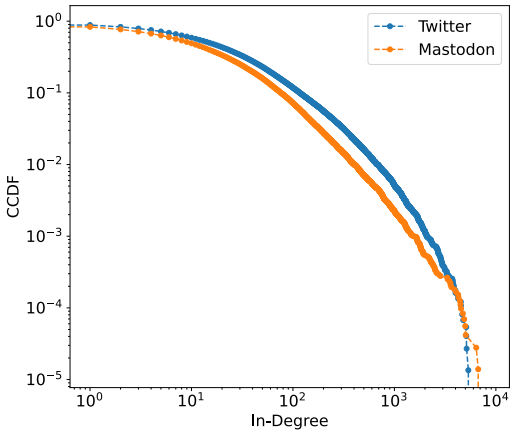
resorted to qualitative approaches [79, 87, 114, 73] or relied on quantitative analyses conducted on small-scale datasets. [26, 25, 152] A few studies have explored user migrations at scale in blockchain-based online social networks, describing the effects of the migration on the structure of the social graph,[14, 86] and finding that network density is a predictor of migration.[16] Recently, a distinct line of research has emerged, focusing on user migrations prompted by *deplatforming*, [187] which refers to the removal of accounts by social media administrators due to user engagement in toxic, offensive, or abusive activities. Studies in this domain have primarily examined the efficacy of deplatforming in disbanding groups of deviant users and reducing the prevalence of toxic interactions online, yet the findings indicate limited success in achieving these objectives. [9, 113, 161, 165] Migration resulting from deplatforming fundamentally differs in nature from the phenomenon under investigation in our study, as it is coerced rather than organic.

5.2 Results

5.2.1 Following the Migration

We collected approximately 2M tweets related to the #TwitterMigration phenomenon, spanning from October 26th, 2022, to January 19th, 2023. Our analysis focused on a subset of over 1.3M tweets written in English, contributed by approximately 500K unique authors. To determine which of these authors had migrated to Mastodon, we leveraged self-disclosed information provided by the authors themselves. During the migration process, numerous users openly shared their Mastodon handles on Twitter to advertise their presence on the new platform. To identify these individuals, we conducted a comprehensive search within the usernames, public profiles, and tweets of all users in our dataset, specifically seeking mentions of Mastodon handles. We found around 75K valid Mastodon handles that were associated with active Mastodon accounts. We also investigated whether there exists any reciprocity in such Twitter-Mastodon profile associations; to this aim, we reverse-searched mentions of Twitter handles within the retrieved Mastodon user profiles, and found about 2.5K Mastodon profiles that declared a Twitter handle in their metadata, of which 98% correspond to a valid match between the mentioned handle and the original corresponding Twitter account (refer to Methods for the detailed procedure).

We analyzed the networks of social links between these 75K users on both Twitter and Mastodon. The Twitter follower network contains approximately 4M links, whereas the Mastodon network exhibits a relatively lower count of 2.5M links, representing a reduction of approximately 38%. The differences observed at the macroscopic structural level between the two networks are primarily due to their contrasting number of edges. In comparison to Twitter, the Mastodon network exhibits relatively lower density, lower average degree, lower transitivity, and an increased presence of small disconnected components. The fewer connections observed in the Mastodon network could potentially indicate that, at the time of data gathering, the process of link formation among newly-migrated users was still underway. The complementary cumulative distribution functions of the in-degrees in the two networks follow a similar trend that however diverges in a relatively large regime (approximately from 10 up to around 10^4, as shown in Figure 5.1, left). Despite the variation in degree distribution and edge density, both networks exhibit a similarly high clustering coefficient, modularity, and percentage of reciprocal follower links. These shared characteristics suggest that both networks foster tightly-knit, reciprocal local neighborhoods that are arranged in well-separated communities.



	Twitter	Mastodon
#Nodes	74 132	71 769
#Edges	4 138 546	2 575 656
Density	8e-04	5e-04
%Reciprocal	48.8	46.9
Transitivity	0.085	0.048
Avg. CC	0.187	0.195
%Sources	5.6	8.4
%Sinks	1.8	2.5
Avg. In-Deg	55.83	35.89
Deg. Assort.	-0.057	-0.085
#SCCs	5646	8038
#WCCs	9	26
# Comm.	292	478
Modularity	0.538	0.499

Fig. 5.1. (Left) Comparison between the CCDF of the in-degree distributions of the Twitter and Mastodon networks. (Right) Main structural traits of the Twitter migration network and Mastodon network; the difference in the number of nodes is due to the presence of isolated nodes, i.e., migrated users having no ties with other migrated users in the Mastodon network.

The connections within the Mastodon network mirror in part the social relationships observed in the Twitter network. Approximately 30% of the links observed in Twitter can be found within Mastodon. When focusing on the *backbone* of the Twitter network, which is a pruned graph whereby spurious connections are filtered out (see Methods), the proportion of Twitter links replicated on Mastodon increases to 41%. This significant presence of shared edges, particularly in the backbone network, suggests that the migration was in part driven by the desire to "replicate" the existing Twitter social network on a new platform, rather than establishing an entirely new social context; indeed, this appears to moderately hold at node-neighborhood level, as hinted by a Spearman correlation of 0.6 between the rankings of local clustering coefficients of shared nodes in Twitter and Mastodon graphs. The structural metrics of both networks and their intersection are summarized in Figure 5.1 (right).

5.2.2 Social Influence of Migrants

The widespread practice of using Twitter to announce one's decision to migrate to Mastodon raises the question of whether social influence played a role in Twitter users' migration choices. To investigate this, we resorted to *compartmental epidemic models* to characterize the "infectiousness" of migration decisions. Epidemic models have been extensively used to simulate information diffusion within social systems, [203, 126] under the principle that the process of influence spreads through social connections, akin to the transmission of communicable diseases through social interactions, and that the population under study is partitioned into predefined compartments expressing epidemiological states. This allows us to focus on understanding global patterns, not on "who-infects-whom", thus in contrast to the approach underlying stochastic information diffusion and maximization models [127] which assume the availability of a network of connections among individuals, possibly with additional information about user-attributes.[39]

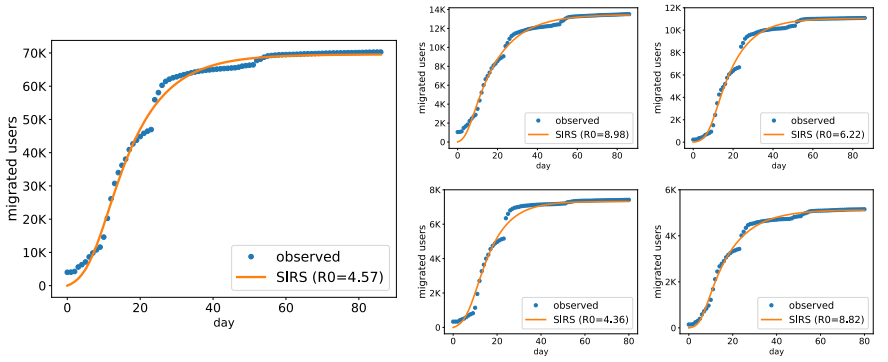


Fig. 5.2. (Left) Cumulative number of Twitter users migrated to Mastodon over the course of 3 months since Elon Musk's acquisition of Twitter. Fit estimated with the SIRS model, along with its R_0. (Right) Fitting with the SIRS model for the top-4 largest communities by Louvain, reported in descending order row-wise.

In this study, we employed the widely known SIRS model, where the entire population is initially susceptible (S) to a disease, subsequently certain individuals become infected (I) and can transmit the disease to others through their social connections, and over time, infected individuals recover (R) and eventually become susceptible to re-infection. We also experimented with the SIR model, a simpler version that does not consider re-infections, obtaining similar results (see ***Appendix A.3***).

We simulated the diffusion process with a daily granularity. On any given day t, the set of susceptible individuals (S) contains Twitter users who have not yet created a Mastodon account by day t. The set of infected individuals (I) comprises Twitter users who registered on Mastodon up until day t. We hypothesize that the triggers of social influence were the public announcements of Mastodon handles on Twitter. Under this assumption, re-infections could correspond to users who announced multiple times their commitment to migrating. However, it was impossible to collect precise information regarding the timing of such announcements for all users. To approximate the day of the announcement, we used the day when their Mastodon account was created. This approximation generally proved to be accurate, considering the tendency of users in advertising their Mastodon profiles in a short time following their registration (0 days as mode, 1 day as median, 5 days as mean) we observed for the subset of users (N = 41K) with both timestamps available (see ***Appendix A.3***). The set I also includes 4K Twitter users who had registered on Mastodon during 2022, prior to our data collection period, who could then have helped to foster the #TwitterMigration movement. These can be regarded as "early-migrant" Mastodon users, with an average registration time of 168 days before October 26^{th} 2022, and median registration time of 183 days, corresponding to the date Musk struck the deal to acquire Twitter. The remaining set of recovered individuals (R) was determined by the model (see Methods). R corresponds to the set of Twitter users who have migrated to Mastodon but whose influence effect has worn out (i.e., their intention to migrate has vanished from their followers' timelines).

To investigate whether the migration phenomenon was compatible with the characteristics of information diffusion, we tested whether a SIRS model could replicate the observed trend of the cumulative number of migrated users. To achieve this, we determined the parameters of the SIRS model that produced the best approximation of the empirical data (see Methods).

The model takes the population size as input, and here we present results based on a population equivalent to the set of all migrated users at the end of our data collection period. In ***Appendix A.3***, we provide additional results considering a larger population, equal to the size of users who have engaged in discussions involving Mastodon, including those who did not migrate in the timeframe of our study. The model with optimal parameters provided a close approximation of the empirical data, with a mean absolute percentage error (MAPE) of 0.077. From the parameters of the fitted model, we derived the *reproduction number* $R_0 = \beta/\gamma$, expressed as the ratio between the rate of infections β and the rate of recovery γ. The β parameter denotes the number of people with whom one infected individual interacts in a unit of time, whereas the parameter γ models the number of people who recover in a unit of time. When $R_0 > 1$, the diffusion process grows as individuals become infected at a higher rate than they recover. Our models estimated an R_0 value of 4.57 (Figure 5.2, left), signifying a highly infectious process that backs our hypothesis of social influence in the migration process.

5.2.3 Is Spreading Community-Driven?

The Twitter follower network is highly modular (Figure 5.1, right), indicating that users participating in the #TwitterMigration discourse belong to distinct and loosely connected communities. This observation aligns with previous research highlighting the highly segregated nature of Twitter communities. [197] The presence of community structure plays a crucial role in information diffusion dynamics. Within communities characterized by dense social ties, information can rapidly spread among members, whereas community boundaries restrict the propagation of information to the rest of the social network. [191, 142] We studied how the social influence process differed across communities.

To this aim, we first used the Louvain method to partition the Twitter follower network connecting migrated users into communities (see Methods), and obtained 292 well-separated communities (modularity of 0.538). The resulting communities are heterogeneous in size, with a long tail of very small isolated groups (see Figure S2 in ***Appendix A.3***). We focused our analysis on the 44 communities containing a sufficiently high number of members (size ≥ 50), that jointly account for 98% of nodes in the migrated network.

We fit the SIRS model on each of these communities, treating them as isolated systems. The estimated reproduction number across communities exhibits a range of values from $R_0 = 1$ to $R_0 = 11.82$, indicating that the process of social influence unfolded at varying rates depending on the social context. Figure 5.2 displays the fits and corresponding R_0 values for the four largest communities (refer to ***Appendix A.3*** for a more exhaustive account on communities).

Notably, despite these fittings involve a different number of users and various rates of social influence, they appear to exhibit a similar trend. This common trait can be attributed to exogenous factors that have contributed to shape the course of migration. A key case in point is the restart of the diffusion process during the third week, prompted by Elon Musk's request to his employees to sign a pledge to work harder at the development of "Twitter 2.0" or leave with three months of severance pay [35].

We also noticed that in most communities, the SIRS model slightly improves on the SIR model in terms of fit with the empirical data, with an average MAPE of 0.124 compared to 0.125 by the SIR; moreover, higher reproduction numbers occur in SIRS, with an average R_0 of 5.08 against 3.92 in SIR (see ***Appendix A.3***), which highlights the role that repeated signals of commitment, which are inherently captured by the SIRS model, may have played in the social influence process.

To gain insights into the factors that contribute to the acceleration of the influence process, we explored the correlations between the parameters of the fitted SIRS models in different

communities and three categories of factors that have previously been associated with the adoption of new opinions and behaviors in social groups, [21] namely *network topology*, *reiterated commitment*, and *language use*. We elaborate on each of these aspects next.

Network features. In terms of network topology, we investigated various community-specific structural aspects, including network size, density, reciprocity, distribution of prestige, and different connectivity metrics (Figure 5.3, left). Notably, we find only a weak correlation between community size and the basic reproduction number (Pearson correlation $r = 0.343$). Communities with higher values of R_0 tend to exhibit sparser connectivity ($r = -0.459$ w.r.t. density), as well as lower levels of reciprocity ($r = -0.425$). Furthermore, communities that experienced faster diffusion were characterized by social ties linking users with similar levels of prestige ($r = 0.298$ w.r.t. assortativity), by a less pronounced hierarchical structure, where the prestige of individuals (measured by their number of followers) was more evenly distributed ($r = -0.474$ w.r.t. standard deviation of in-degree centrality), and by lower levels of clustering ($r = -0.497$ w.r.t. transitivity, and $r = -0.488$ w.r.t. average clustering coefficient).

Reiterated commitment. We examined the influence of iterated commitment on the migration process in terms of the tweets concerning the #TwitterMigration within each of the identified communities. To this aim, we collected a dataset of $\sim$ 8.3M tweets posted by migrated users during the initial month of the #TwitterMigration. We intentionally excluded tweets posted on December 2022, due to their prevalent shifted focus on commentaries related to the 2022 FIFA World Cup.

First, to provide an indicator of the extent to which Mastodon discussions are prevalent within each community, we calculated the *commitment* as the ratio n_M/n_{tot} between the number of tweets about Mastodon (n_M) and the total volume of tweets (n_{tot}) posted by community members over our examination period. We found a positive correlation ($r = 0.54$) between the commitment and the value of R_0 measured at community level, suggesting that a greater and reiterated focus on Mastodon in community discussions might have aided the migration process.

Second, we carried out topic modeling of our tweet corpus, by means of BERTopic, a topic modeling method that has been shown to be more effective in extracting topics from short texts like tweets compared to other traditional topic modeling techniques like Latent Dirichlet Allocation [75]. Based on the resulting users' topic distributions, for each community we selected the distributions of its members, calculated the entropy of such topic distributions, and correlated it with the corresponding R_0 of that community. We noticed a significant variety of entropy values over all communities, which also moderately correlates with R_0 ($r = 0.20$), suggesting that the discussion of a broadest range of topics might lead to higher infectiousness. In addition, we spotted that most communities are characterized by mid-high entropy values, which is surprisingly negatively correlated with the community sizes ($r = -0.23$). An overview of the most prominent topics in the ten largest communities, along with their corresponding entropy scores, is shown in Table 5.1. The above findings hint that *(i)* the migratory process involves users discussing multiple topics, and *(ii)* users migrating might be further persuaded when they perceive broad topical variety.

Language use. Finally, we delved into the pragmatics of language use in relation to the migration process. Specifically, we focused on the *social pragmatics* of language, namely the intended social function of an utterance. Prior research in the social sciences has extensively explored the associations between various forms of social intent and the process of opinion formation. For instance, conveying *trust* has been identified as a crucial factor in aligning divergent points of view. [103] Recent studies have surveyed a range of *dimensions* of social pragmatics commonly observed in everyday language, [63] and have developed language

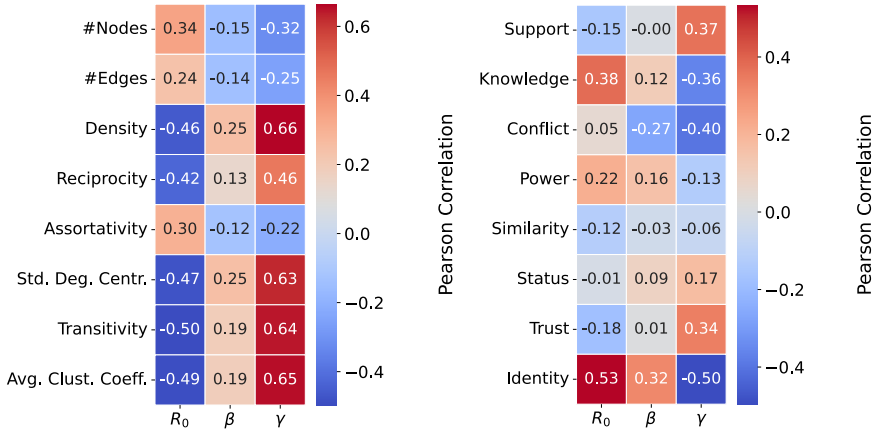


Fig. 5.3. (Left/Right) Pearson correlation heatmap between social and topological aspects characterizing communities and main parameters of the SIRS model.

Table 5.1. Main topic of each of the top-10 largest communities with corresponding community-level topical entropy value. Each topic was manually labeled after excluding the inherent Twitter and Mastodon discourse, which would have hidden the real topics.

Comm.	Topic	Entropy
Top-1	*Society*	0.633
Top-2	*Sports*	0.605
Top-3	*Politics*	0.622
Top-4	*Healthcare*	0.791
Top-5	*Politics*	0.334
Top-6	*Tech*	0.632
Top-7	*Music*	0.773
Top-8	*Politics*	0.821
Top-9	*Writing*	0.634
Top-10	*Research*	0.728

models capable of classifying conversational text based on these dimensions. [50] In the context of Reddit, such a model has been employed to examine the relationship between the use of specific dimensions in public argumentation and an increased likelihood of opinion change. [164] We built a *social dimensions classifier* (see Methods) to our dataset of tweets to identify those that, with high probability, conveyed *Support*, *Knowledge*, *Conflict*, *Power*, *Similarity*, *Status*, *Trust*, and *Identity*. We then computed the ratio $C_{i,d} = N_{i,d}/\mathcal{N}_{i,d}$, which represents the number of tweets marked with dimension d in community i ($N_{i,d}$) divided by the expected number of tweets conveying dimension d in the same community if tweets with dimension d were uniformly distributed at random throughout the entire network ($\mathcal{N}_{i,d}$). Community-level R_0 exhibited the strongest correlation with the ratios of *Knowledge* ($r = 0.38$) and *Identity* ($r = 0.53$) (Figure 5.3, right). These correlations suggest that information exchange and, particularly, the sense of belonging to a shared group or community might have played key roles in the migration phenomenon.

5.2.4 Drivers of Migration

Table 5.2. Ordinary Least Squares regression model fittings for the prediction of R_0 from topological, activity, and social features (top row), and their combinations (bottom row). β coefficients describe the contribution of each feature to the outcome, along with the standard errors (SE) and statistical significance (p-values). Auto-correlation is evaluated via the Durbin-Watson statistic (values closest to 2 indicate no auto-correlation). Regression results are reported via adjusted R^2.

Predicting R_0 from:			
Topological features			
Feature	β	**SE**	p
Density	-0.459	0.137	0.002
Durbin-Watson stat. = 2.639		R^2_{adj} **= 0.192**	

Predicting R_0 from:			
Activity features			
Feature	β	**SE**	p
Commitment	0.542	0.130	0.000
Durbin-Watson stat. = 1.990		R^2_{adj} **= 0.277**	

Predicting R_0 from:			
Social Features			
Feature	β	**SE**	p
Knowledge	0.394	0.117	0.002
Identity	0.543	0.117	0.000
Durbin-Watson stat. = 1.837		R^2_{adj} **= 0.412**	

Predicting R_0 from:			
Topological & Activity Features			
Feature	β	**SE**	p
Density	-0.406	0.116	0.001
Commitment	0.499	0.116	0.000
Durbin-Watson stat. = 2.473		R^2_{adj} **= 0.430**	

Predicting R_0 from:			
Social & Activity Features			
Feature	β	**SE**	p
Knowledge	0.333	0.117	0.007
Identity	0.408	0.131	0.003
Commitment	0.271	0.134	0.050
Durbin-Watson stat. = 1.878		R^2_{adj} **= 0.453**	

Predicting R_0 from:			
Topological & Social & Activity Features			
Feature	β	**SE**	p
Density	-0.356	0.104	0.002
Knowledge	0.340	0.104	0.002
Identity	0.305	0.120	0.015
Commitment	0.282	0.119	0.023
Durbin-Watson stat. = 2.302		R^2_{adj} **= 0.568**	

We observed that factors related to network *topology*, individual *commitment*, and *language use* correlate individually with the speed of the social influence across Twitter communities. To explore the interplay among these three aspects, we conducted experiments using various regression models to predict community-specific R_0 based on combinations of features that displayed the strongest correlations (see Table 5.2). All variables were standardized before using them in the regressions. A more exhaustive set of regressions is presented in ***Appendix A.3***. A model that solely incorporates network *density* yielded the poorest fit ($R^2_{adj} = 0.192$, $\beta = -0.459$). When considering commitment alone, a slightly higher correlation was observed ($R^2_{adj} = 0.277$). When combining density and commitment, the goodness of fit approximately doubled ($R^2_{adj} = 0.430$), indicating that these two factors jointly account for over 40% of the variability in R_0 across communities. Interestingly, a comparable level of fit was achieved by solely considering the prevalence of knowledge and identity messages ($R^2_{adj} = 0.412$). Ultimately, the best fit was obtained by combining all variables ($R^2_{adj} = 0.568$). In this last model, all predictors maintain statistical significance ($p < 0.05$), and the magnitude of their coefficients is comparable, meaning that each variable contributes non-negligible signals in modeling the susceptibility of communities to behavioral changes. Lower density of social links and abundant exchange of factual knowledge exhibit slightly stronger associations with R_0 in this multivariate model, followed by expressions of identity and, last, by iterated expressions of commitment.

5.3 Discussion

In pursuit of experimental evidence of factors underlying behavioral change in social communities, we have studied the #TwitterMigration phenomenon: a rapid and extensive migration of Twitter users to the decentralized social platform Mastodon. Few other studies have touched upon this event, and have done so from angles that are different from our own. They either focused on the decentralization properties of the Mastodon ecosystem from the perspective of the migrated users, [243] or explored the characteristics of Mastodon communities that are associated with higher rates of user retention. [119] Those studies were conducted on data that was either smaller or with fewer dimensions (e.g., no analysis of text) than what we consider in this work.

Our study makes a first attempt at describing the dynamics of this migration from an information diffusion perspective, finding that a simple epidemic model of information spreading closely replicates the temporal trace of migration. Crucially, we observed that the effectiveness of the social influence to migrate (i.e., the value of R_0 estimated by the epidemic model) was community-dependent, and found patterns that help explain why some communities were more successful in migrating more rapidly. Drawing inspiration from prior research on behavioral change, we tested three hypotheses pertaining to the *structure* of the social network, the *commitment* of community members, and their *language use* as potential explanations for these observed differences. By testing each of these hypotheses, we gained interesting insights.

The first finding of our study is that the only structural factor that was consistently associated with an increased R_0 is the sparsity of social connections. This is somewhat unexpected, because it contradicts the conventional understanding that close social proximity, characterized by high clustering and dense social connections, leads to faster and broader adoption of new behaviors. [51] We propose two non-exclusive explanations for this counter-intuitive trend. Firstly, the discussion surrounding the #TwitterMigration phenomenon competed with numerous other topics vying for the limited attention of Twitter users. The prominence of

#TwitterMigration in the overall discourse might have been more diluted in denser follower networks. [232] Secondly, the incentive to migrate may simply be proportional to the fraction of a user's friends who have already migrated, and such fraction increases more rapidly in networks with fewer social connections.

Our second finding is that communities engaging in more frequent discussions on the #TwitterMigration topic exhibited higher R_0. We interpret this result in the light of theories linking the rapid emergence of consensus to the influence exerted by committed individuals, even when they constitute a small minority. [21] While this intuition is grounded in strong theoretical foundations, [42, 117] empirical support has been limited. [150] Our operationalization of commitment, as a simple and straightforward measure to quantify the level of engagement in discussing a specific topic, contributes to validating the underlying theoretical framework.

The third and final finding of our study is that communities in which the influence process unfolded more rapidly are those whose discourse frequently emphasizes a shared *identity* and engages in substantial exchanges of factual *knowledge*. Extensive research in social psychology has established a connection between psycho-linguistic aspects of social interaction and successful, spontaneous coordination. [155, 104, 105, 64] In particular, the Identity Theory posits that cooperation can be facilitated through cognitive mechanisms that foster a sense of belonging to the same social group, [215] suggesting that identity may be pivotal in overcoming social dilemmas involving coordinated behavior that entail inherent risks or a non-zero cost of action. [133] Moreover, the exchange of truthful, factual information has been identified as a prerequisite for constructive debates and, ultimately, persuasion. [102] The combined influence of identity and knowledge explains more variance in our data than the combined influence of density and commitment, highlighting the significant role of psycho-linguistic aspects as key drivers of behavioral change.

Our work comes with limitations that future work can address.

First, our perspective of the social system in which the migration occurred is limited in several ways. Specifically: *(i)* it is plausible that the migration process continued beyond the temporal boundaries captured by our dataset; *(ii)* We focused solely on the network of users who explicitly disclosed their migration, not accounting for users who might have migrated silently and not considering any potential influences from other regions of the Twitter network or exogenous events such as news items ; and *(iii)* we treated the follower network as a comprehensive proxy for all the information channels available to our users on Twitter, which is not true in general. It is hard to gain complete knowledge on the set of Twitter users who migrated to Mastodon without employing more costly information gathering techniques such as extensive surveys. To provide evidence on the robustness of our results to the choice of the set of migrated users, we modeled network influence in two extreme scenarios: one in which we consider only the social graph that connects the 75K users who declared their migration, and one in which the social graph includes all users who were involved in the Mastodon discussion (described in ***Appendix A.3***). Crucially, we observed that the temporal patterns of migration in both scenarios are well described by a process of information diffusion with relatively high contagiousness, represented by R_0 values exceeding one. We argue that a scenario in which one could obtain the full list of migrated users would lie in between these two extremes, and thus still exhibit relatively high values of R_0. We also conjecture that 'silent' users who migrated without explicitly signaling their migration might have not considerably contributed to spreading social influence, effectively making them akin to non-migrated users from the perspective of the information diffusion phenomenon.

Second, our measurements are inherently limited in terms of precision, scope, and validity: *(i)* we adopted only a macroscopic perspective on the migration phenomenon, relying on measurements derived from population-level or community-level aggregates; *(ii)* the network

communities identified through unsupervised techniques may not accurately reflect the communities as subjectively perceived by Twitter users; and *(iii)* both supervised (social dimensions) and unsupervised (topic modeling) methods employed in the analysis of natural language are prone to errors — while they provide valuable insights when interpreted in aggregate, they may fail to accurately categorize specific instances of text.

Last, the factors we explored in relation to the migration rate are not exhaustive, and collectively account for only slightly more than half of the observed variance. The vast literature on behavior change, social influence, and collective action encompasses a broader array of factors than those considered in our study. We encourage future research to investigate additional elements beyond the scope of our work, in order to gain a more comprehensive understanding of the complex dynamics at play.

5.4 Methods

5.4.1 Data Collection

Migration-related Tweets. No public datasets about the #TwitterMigration movement was available at the time of writing. To fill this gap, we carried out an extensive data crawl on both Twitter and Mastodon. We used the full-archive search functionality of Twitter's Academic API v2 to download tweets relevant to the migration that were posted from October 26^{th} 2022 (when Twitter acquisition by Musk was finalized), to January 19^{th} 2023. First, to identify hashtags that were frequently mentioned in the discourse about the migration, we started from the list of hashtags that were featured as *trending* in the days following the public announcement of the acquisition. We then expanded those hashtags by snowball sampling: we collected all the tweets containing those trending hashtags published between October 26^{th} 2022 and November 26^{th} 2022, and counted the frequency of all hashtags mentioned in those tweets. We then manually parsed the top hashtags in the frequency distribution and identified a set of 13 hashtags that unambiguously referred to the migration. To ensure high recall to our data, our dataset is comprised of tweets that *(i)* contain one of the 13 hashtags we found by snowball expansions, or *(ii)* mention Mastodon's official Twitter account (i.e., @joinmastodon), or *(iii)* contain the keyword "mastodon". We excluded retweets. The final query we submitted to the Twitter API was: " (@joinmastodon OR mastodon OR #TwitterMigration OR #RIPTwitter OR #TwitterTakeover OR #TwitterShutdown OR #TwitterIsDead OR #MastodonSocial OR #LeavingTwitter OR #ElonIsDestroyingTwitter OR #MastodonMigration OR #Fediverse OR #Mastodon OR #TwitterAlternative OR #DecentralizedSocialMedia) -is:retweet ". Together with the tweets, we saved the full profile description of the user who posted the tweet, as returned by the APIs.

Matching Twitter Handles with Mastodon Handles. In an effort to recreate their Twitter social network on Mastodon, some users promoted their Mastodon *handles* on Twitter. We used this information to link Twitter users with their corresponding Mastodon profiles. To achieve this, we first devised different regular expressions to identify potential Mastodon-like handles occurring either in the *username*, *description*, or *tweets* of each user. We found $108k$ handles that were compatible with Mastodon profiles. Further refinement of the matching strings was necessary due to the prevalent format of Mastodon handles, which aligns with that of email addresses (i.e., username@domain.tld). To distinguish strings specifically referring to Mastodon accounts, we cross-referenced them with a compiled list of known Mastodon instances that we obtained from the `instances.social` APIs, the most widely-recognized

and comprehensive tracker of Mastodon instances. This step filtered the set of handles down to ∼75K. Last, for each of the Mastodon handles we found, we queried the official Mastodon API (`https://docs.joinmastodon.org/api/`) to collect *(i)* the list of all their followers and followees, and *(ii)* their profile metadata, including the created_at field, which records the precise timestamp of account creation. To comply with general privacy-preserving policies that might be set forth by Mastodon instances, we did not acquire any textual or multimedia content.

5.4.2 Network Modeling

Twitter and Mastodon Social Graphs. We built the social contexts of each migrant user on the two platforms as follows. We first collected the followees of all ∼ 75K migrated Twitter users, obtaining ∼ 111M raw links between ∼ 16.6M users, and analogously we collected ∼ 20M raw links involving the set of followees of migrated users on Mastodon.

We then modeled two distinct social graphs from the Twitter and Mastodon data we collected. The first graph, denoted as $\mathcal{G} = \langle V, E, t \rangle$, is a directed and node-labeled graph, which we also refer to as the *migration network*. In fact, its vertex set V represents the ∼ 75K Twitter users who have a corresponding Mastodon account in our dataset, and the set of edges E models ∼ 4M social ties through the follower relationship, with $(i, j) \in E$ indicating that user $i \in V$ follows user $j \in V$. Each user is also labeled with its migration date through the function $t : V \mapsto T$, which assign a node in V with a timestamp in T denoting the creation date of the corresponding user's Mastodon account. The second graph, denoted as $\mathcal{G}_M = \langle V_M, E_M \rangle$, represents the directed and unweighted Mastodon follower network. It comprises a set V_M of Mastodon users that we collected, and a set E_M of 2.5M edges that represent the follower relationships among these Mastodon users, where $(i, j) \in E_M$ denotes that Mastodon user $i \in V_M$ follows Mastodon user $j \in V_M$.

Graph Backboning. Given the noisiness of online social ties, it is appropriate to perform a network simplification or backboning task aimed at detecting and pruning noisy edges or ties formed due to random chance, thus bringing out the latent structure of a network. In this regard, albeit the simplest solution might be exploiting edge weights to filter out all edges having a weight below a fixed and pre-determined global threshold, the risk of removing social ties locally relevant yet weak at the network level is evident. To this aim, we resorted to the *Disparity Filter* (*DF*) [248] method, which exploits *generative null model* based on node distribution properties to prune networks from statistically irrelevant edges, i.e., via p-value computation w.r.t. specific significance levels. Specifically, the *DF* leverages the null hypothesis that the strength of a node is redistributed uniformly at random over the node's incident edges, thus evaluating the strength and degree of each node locally. By using its publicly available implementation (`https://github.com/malcolmvr/backbone_network`), we applied *DF* to our migration network equipped with an edge weighing function defined as follows. Each edge in $\mathcal{G}$ is assigned a weight through a function $w : E \mapsto \mathbb{R}$, which measures for each $(i, j) \in E$ the similarity between i and j as the Jaccard coefficient over the out-neighbors of i and the in-neighbors of j, i.e., i following j is regarded as similar to j proportionally to as much users followed by i tend to follow j, thus reflecting a simple notion of prestige conferred by i to j.

Community Detection. To unveil the underlying community structure within our *migration network*, we employed the widely-adopted Louvain algorithm for community detection. [30] We used its original (undirected) implementation due to its exceptional scalability that enables accurate community detection in large-scale networks. The Louvain algorithm adopts a

hierarchical greedy approach to maximize the modularity Q, defined as:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{5.1}$$

where A_{ij} represents an entry in the binary adjacency matrix between nodes, k_i and k_j denote the degrees of nodes i and j, respectively, and m corresponds to the total number of edges. The Kronecker delta function $\delta(c_i, c_j)$ takes the value 1 when nodes i and j belong to the same community ($c_i = c_j$), and 0 otherwise. Overall, the modularity Q denotes the quality of a network partitions into communities by weighting the density of intra-community links against inter-community links; it ranges between -0.5 and +1, with higher values indicating better partitioning.

The Louvain algorithm starts by assigning each node to a singleton community and then proceeds by merging communities aiming at maximizing the modularity gain ΔQ, that for a node i is defined as:

$$\Delta Q = \left[\frac{\sum_{in} + 2k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right] \tag{5.2}$$

In the equation, $\sum_{\text{in}}$ is the sum of links within the target community, $\sum_{\text{tot}}$ is the sum of links incident to nodes in the target community, k_i is the sum of links incident to node i, $k_{i,\text{in}}$ is the sum of links from node i to nodes in the target community, and m is the total number of edges in the network. The algorithm stops when no more improvements in modularity can be achieved.

Epidemiological Models. To verify whether the temporal pattern of migration is compatible with an information diffusion phenomenon, we adopted the hypothesis of *infodemic spreading* [203] and fit the SIR and SIRS *compartmental epidemiological models* [18] to our data. The SIR model divides the population into three compartments: *susceptible* (S), *infectious* (I), and *recovered* (R) individuals. The model operates under three assumptions: *(i)* a closed world assumption where individuals cannot enter or leave the population during the spreading phenomenon, *(ii)* equal susceptibility of all individuals to the information, and *(iii)* once recovered, individuals cannot be reinfected. The SIR model is described by the following set of differential equations:

$$\begin{aligned} dS/dt &= -\beta S \cdot I/N \\ dI/dt &= \beta S \cdot I/N - \gamma I \\ dR/dt &= \gamma I \end{aligned} \tag{5.3}$$

In the formula, S, I, and R represent the number of susceptible, infectious, and recovered individuals, respectively. β denotes the infection rate, γ represents the recovery rate, and N indicates the total population size. The basic reproduction number, $R_0 = \beta/\gamma$, serves as a measure of the average number of infections generated by an infected individual, indicating the speed of contagion. In the context of our study, an $R_0 > 1$ corresponds to a scenario of a growing infodemic in which, over time, more and more users decide to migrate.

The SIRS model extends SIR by allowing previously recovered individuals to return to a susceptible state. This model introduces the loss-of-immunity rate parameter, ξ, indicating the rate at which individuals lose their immunity and become susceptible again. The updated set of differential equations is as follows:

$$\begin{aligned} dS/dt &= -\beta S \cdot I/N + \xi R \\ dI/dt &= \beta S \cdot I/N - \gamma I \\ dR/dt &= \gamma I - \xi R \end{aligned} \tag{5.4}$$

In our study, a reinfection event may signify users reaffirming their commitment to migrating. We estimated the parameters of both models using the least square estimation method, [153] which involves minimizing the sum of the squares of the residuals, computed as the difference between the observed data depicting the spreading of the infection and the predictions by the corresponding model. We generated these predictions by solving the systems of Ordinary Differential Equations (ODEs) underlying our compartmental models as an initial value problem, fed with information involving the onset of the outbreak such as the overall population and the initially infected individuals, and leaving the model finding the set of remaining parameters that would minimize the residuals, thus providing the best fit for the observations.

Network Analysis Measures. To characterize the main topological traits underlying the networks in our study, we resorted to classic structural measures, which are summarized as follows. *Transitivity* and *Average Local Clustering Coefficient* provide a global and local perspective of triadic closure, respectively, where the former refers to the probability that two incident edges are completed by a third one to form a triangle, and the latter expresses how much connected are the neighbors of a node, averaged over all nodes. *Source* and *Sink* nodes denote nodes with no incoming, resp. no outgoing, links and provide insights about the origin and termination of information flows within a network. The *Degree Assortativity* measures the tendency of nodes to link with peers having similar degrees, and in this work corresponds to the Pearson correlation between the out-degree of source nodes and the in-degree of target ones. *Strongly* and *Weakly Connected Components* inform about the connectivity of a network in terms of isolated subgraphs; the weakly case refers to subnetworks in which every node can be reached from every other node by ignoring edge directionality, whereas the strongly case requires taking into account the latter. The *Standard Deviation* of the *Degree Centrality* provides insights into the heterogeneity of importance (based on the number of connections nodes have) within a network, where higher values suggest a greater disparity, with some nodes being more relevant than others.

5.4.3 Language Modeling

Topic Modeling. To extract the latent topics that characterize the tweets of a user, we employed BERTopic, [91] a recently-developed unsupervised topic model that leverages Transformer-based pre-trained language models to extract coherent and descriptive topics from unstructured text. The BERTopic pipeline consists of three stages. First, it transforms documents into a high-dimensional embedding space that captures the semantics of the input text. Second, it reduces the embedding to a lower-dimension space that is more suitable for clustering. Lastly, it applies clustering algorithms to identify groups of documents corresponding to distinct topics.

Given a time window of interest, we shape the social discourse of each migrated user by creating a document corresponding to the concatenation of all tweets posted during such a period. We then fed such documents to BERTopic to infer a representative topic for each user based on their tweets. Based on such information, we also computed the topical entropy for each community in our Twitter graph, to quantify the diversity of the collection of topics discussed by the users of that community.

We tried different settings of the main hyperparameters for the various stages of BERTopic. Here we report the best performing configuration. For generating embeddings, we used the default Sentence-BERT, [181] specifically its pre-trained *all-mpnet-base-v2* model. To reduce the dimensionality of the embeddings, we employed Principal Component Analysis (PCA) to project them into a 5-dimensional space. Subsequently, we carried out the K-Means algorithm with $n_{clusters} = 120$ to cluster the reduced embeddings. We used a *CountVectorizer*

and a *c-TF-IDF* modules to extract accurate topic representations. Specifically, to improve topic quality, we set $stop\_words = english$ and $ngram\_range = (1, 3)$ for the former and $reduce\_frequents = True$ for the latter, respectively. Finally, we fine-tuned topics by using the *MaximalMarginalRelevance* module with $diversity = 0.5$ to avoid similar keywords in our topics and the *KeyBERTInspired* module to improve the semantic relationship between keywords and documents in each topic. Other optimizations involve the specification of the minimum size of a topic should have ($min\_topic\_size = 175$) and the automatic topic reduction after training the model ($nr\_topics = auto$).

It should be noted that a systematic exploration of BERTopic hyperparameters goes beyond the objectives of this work; rather, we chose to keep all the components in the BERTopic pipeline, and experimented with a few variations of the pipeline according to the recommended best-practices in using BERTopic, primarily focusing on finding an appropriate setting to achieve a reasonable trade-off between meaningfulness of the induced topics and training/inference time speedup.

Social Dimensions. To infer the social intent that Twitter messages convey, we resort to a theoretical model of ten social dimensions that reflect fundamental social aspects of social interactions. These dimensions have been identified through surveys and an extensive literature review, and they are frequently expressed in social conversations, for example to give emotional support or to convey trust. [63] Previous work developed a model [50] that can accurately detect the presence of these dimensions from conversational text, and that has been successfully employed in multiple studies of online conversations. [242, 164, 6, 19] We used the publicly available LSTM [112] implementation of the model (`http://www.github.com/lajello/tendimensions`) that has been trained on around 9k manually labeled sentences, and demonstrated good performance, achieving an average Area Under the Curve (AUC) of 0.84 across the various dimensions. [50]

Building upon the idea that textual components might be associated with a combination of the ten dimensions [63], instead of being a multi-class classifier, the underlying model comprises a set of independently-trained classifiers, one for each dimension, functioning as a multi-label classifier. Given a message m, each classifier estimates a score $s_d(t) \in [0, 1]$ for each sentence t in m, that encodes the likelihood that the sentence conveys the target dimension; then it returns the maximum score across all sentences $s_d(m) = max_{t \in m} s_d(t)$, thus allowing individual sentences to convey a given dimension and avoiding affecting results by averaging low-score sentences. This intuition matches with theoretical interpretations [63] according to which also brief expressions can unveil dimensions.

To ease the interpretation of the results, we binarize the scores $s_d(m)$ to split messages between those that carry dimension d with high probability and those that do not. We do so through an indicator function that assigns a specific dimension d to a given message m if its corresponding score is above a pre-defined threshold θ_d.

To account for the dimension-dependent empirical distribution of the classifier scores, we avoided relying on a fixed common threshold, and we rather defined dimension-specific thresholds. We adopted a conservative, high-precision approach, and set θ_d to the 90^{th} percentile w.r.t. the empirical distribution of the scores s_d, thus effectively narrowing the set of messages marked with each dimension down to the 10% of the total messages.

Part II

Graph Mining in High Societal Impact Domains

Overview

This part of the thesis concerns graph mining in *high societal impact domains*. It consists of three chapters, which are organized as follows:

Chapter 6. This chapter covers the clustering domain, introducing an adaptation of a correlation clustering method to a fairness-aware context. Extensive experimental evaluations will illustrate to the reader how the proposed approach can produce high-quality clustering solutions while also accounting for fairness aspects.

Chapter 7. This chapter concerns the legal domain, with a particular focus on the Italian Civil Code. The latter will be analyzed through the lens of network science, shedding light on the main patterns underlying the intricacies of law reference networks. Besides, a web-based tool for the modeling, analysis, and visualization of these networks will be introduced.

Chapter 8. This chapter explores the social debate surrounding the topic of climate change. Specifically, through the exploration of the social trace left by users on Twitter, we will investigate the main topics discussed around the Conferences of the Parties, and identify the most relevant actors involved in such a debate.

6

Fairness Constraints in Correlation Clustering

Summary. The study of fairness-related aspects in data analysis is an active field of research, which can be leveraged to understand and control specific types of bias in decision-making systems. A major problem in this context is fair clustering, i.e., grouping data objects that are similar according to a common feature space, while avoiding biasing the clusters against or towards particular types of classes or sensitive features. In this work, we focus on a correlation-clustering method we recently introduced, and experimentally assess its performance in a fairness-aware context. We compare it to state-of-the-art fair-clustering approaches, both in terms of classic clustering quality measures and fairness-related aspects. Experimental evidence on public real datasets has shown that our method yields solutions of higher quality than the competing methods according to classic clustering-validation criteria, without neglecting fairness aspects.

6.1 Introduction

We live in an era where machine learning is increasingly pervasive in our society. Every day we interact with machine learning systems, even without knowing it, and these acquire more and more decision-making power in our lives. For instance, such systems support, or even replace, decision makers in financial [158], medical [157], or legal [130] domains. Given their delicate role, machine learning systems should guarantee correct functioning and not discriminate those who entrust their decisions. In this context, however, a critical aspect emerges: the data used by such systems are often (intrinsically) biased, resulting from incorrect data collection processes. Thus, it is desirable to avoid machine learning algorithms being affected by, or even amplifying, this bias. For instance, in [78], this refers to removing *disparate impact*, according to which no group of individuals should (even indirectly) be discriminated by a decision-making system.

In this respect, and by focusing on an unsupervised machine learning setting, in this work we tackle the problem of *fair clustering*. This corresponds to clustering a set of data objects such that: (*i*) analogously to the classic clustering scenario, similar objects are assigned to the same cluster, whereas dissimilar objects are assigned to different clusters, and (*ii*) the clusters are not dominated by a specific type of sensitive data class (e.g., people having the same sex).

Our key assumption is that the above problem can be addressed under a *correlation clustering* framework [20]. Correlation clustering is a well-established tool for partitioning the set of vertices of an input graph into clusters, so as to maximize the similarity of the

vertices within the same cluster and minimize the similarity of the vertices in different clusters, according to pairwise vertex weights expressing positive and negative types of co-association. Specifically, following our recent work in correlation clustering [156], here we provide insights into its application to the problem of fair clustering, and we compare it to some state-of-the-art approaches in such a context. Furthermore, albeit we do not aim to provide a comprehensive experimental survey on fair clustering, a by-product of our work is that, to the best of our knowledge, it represents a valuable and unprecedented experimental comparison between approaches of fair clustering.

Our contributions in this work are as follows:

- We provide a comparison between state-of-the-art methods in the context of fair clustering, belonging to different approaches;
- We show how, by optimizing aspects of fairness, some methods affect their ability to produce clusters that are qualitatively good according to classic clustering-validation criteria;
- We shed light on the capabilities of our recently proposed algorithm [156] to adapt to a fair clustering scenario. We show that it is able to produce better solutions than the competing methods from a clustering perspective, while still accounting for fairness-related aspects.

The remainder of the Section is organized as follows. Section 6.2 provides related work on fair clustering. Section 6.3 describes how the fair clustering problem can be solved through a correlation clustering framework. Section 6.4 presents our approach to fair correlation clustering. Section 6.5 and Section 6.6 present experimental methodology, while Section 6.7 discusses our main experimental findings. Section 6.8 concludes the chapter, also providing pointers for future work.

6.2 Related Work

Although of relatively recent definition, the problem of fairness in clustering has received considerable attention in the literature [48]. With their seminal work, Chierichetti *et al.* [49] were among the first to formalize the notions around fair clustering and the related problem, following the *disparate-impact doctrine* [78]. Their main contribution is a general pre-processing step, i.e., *fairlets decomposition*, to enable traditional algorithms (e.g., k-center and k-median) meeting fairness principles. Following that forerunner work, fairness has become pervasive in the clustering landscape [23, 24, 189], leading to a fairness-aware declination of numerous traditional clustering formulations, such as k-center [131], k-means [2, 198], k-median [17], spectral clustering [132], and hierarchical clustering [4].

The phenomenon of fairness in clustering has also been extended to alternative approaches, such as correlation clustering. In this regard, Ahmadian *et al.* [5] is the first work to leverage the correlation clustering model for the fair clustering task. More specifically, it takes a complete and undirected graph as input, where vertices are assigned a (single) label representing a given protected class attribute (e.g., sex or ethnicity), and the goal is to provide a fair representation of each considered label in the resulting clusters. Recently, Mandaglio *et al.* [156] proposed to model the fair clustering problem of a relational dataset as a correlation clustering instance. Given a set of objects, defined over a set of features, Mandaglio *et al.* build an associated correlation clustering instance by considering the similarity between the tuples. Although Ahmadian *et al.*'s and Mandaglio *et al.*'s approaches aim to cluster different types of data (graphs and tuples, respectively), both approaches reduce the original problem to a correlation clustering instance. However, Mandaglio *et al.*'s formulation is more general than Ahmadian *et al.*'s one, since the former deals with an arbitrary number of labels (or sensitive attributes), while the latter is limited to a single-label setting.

6.3 Correlation Clustering

6.3.1 Background on Correlation Clustering

The correlation clustering problem, originally introduced by Bansal *et al.* [20], consists of clustering the set of vertices of a graph whose edges are assigned two nonnegative weights, named positive-type and negative-type weights, respectively. Such weights express the advantage of putting any two connected vertices into the same cluster (positive-type weight) or into separate clusters (negative-type weight). The objective is to partition the vertices so as to either minimize the sum of the negative-type weights between vertices within the same cluster plus the sum of the positive-type weights between vertices in separate clusters (Min-CC), or maximize the sum of the positive-type weights between vertices within the same cluster plus the sum of the negative-type weights between vertices in separate clusters (Max-CC). Both the formulations are **NP**-hard [20, 202] and they are equivalent in terms of optimality. However, the available algorithms for Max-CC [45, 209] are inefficient and poorly usable in practice since they are not able to output more than a fixed number of clusters (i.e., six). Conversely, Min-CC admits approximation algorithms [7, 46] that do not suffer from the limitations of the maximization counterpart. For these reasons, in this work we focus on the minimization formulation of correlation clustering:

Problem 6.1 (Min-CC [8]). Given an undirected graph $G = (V, E)$, with vertex set V and edge set $E \subseteq V \times V$, and weights $w^+_{uv}, w^-_{uv} \in \mathbf{R}^+_0$ for all edges $(u, v) \in E$, find a clustering $C : V \longrightarrow \mathbf{N}^+$ that minimizes:

$$\sum_{(u,v)\in E,\, C(u)=C(v)} w^-_{uv} \quad + \sum_{(u,v)\in E,\, C(u)\neq C(v)} w^+_{uv}. \tag{6.1}$$

Min-CC is **APX**-hard [46], but admits approximation algorithms [8, 20, 46, 47, 220] with guarantees depending on the type of input graph. On general graphs and weights, the best known approximation factor is $O(\log |V|)$ [46, 62], provided by a linear programming approach. Conversely, constant-factor approximation algorithms are possible if the graph is complete and edge weights satisfy the *probability constraint*, i.e., $w^+_{uv} + w^-_{uv} = 1$ for all $u, v \in V$. Among these, the one which provides the best trade-off between efficiency and theoretical guarantees is the Pivot algorithm [8], which simply picks a random vertex u, builds a cluster as composed of u and all the vertices v such that an edge with $w^+_{uv} > w^-_{uv}$ exists, and removes that cluster from the graph. The process is repeated until the graph has become empty. This algorithm has $O(|E|)$ time complexity and it achieves a factor-5 expected guarantee for Min-CC under the *probability constraint* or if a *global weight bound* holds on the overall edge weights [156].

Next we discuss how a clustering problem with fairness constraint can be profitably solved through a Min-CC approach.

6.3.2 Problem Statement

Let $\mathcal{X} = \{X_1, \cdots, X_n\}$ be a set $\mathcal{A}$ of n objects defined over a set of attributes. The latter is assumed to be divided into two sets, $\mathcal{A}^F$ and $\mathcal{A}^{\neg F}$. The $\mathcal{A}^F$ set contains *fairness-aware*, or *sensitive*, attributes such as those identifying sex, race, religion, relationship status in a citizen database and any other attribute over which fairness is to be ensured. $\mathcal{A}^{\neg F}$ denotes the attributes that are relevant to the task of interest, and thus can be regarded as *non-sensitive*.

In both cases, we assume that part of the attributes might be numerical, and the others as categorical (binary or multi-value). We use subscripts N and C to distinguish the two types, therefore $\mathcal{A}^F = \mathcal{A}^F_N \cup \mathcal{A}^F_C$ and $\mathcal{A}^{\neg F} = \mathcal{A}^{\neg F}_N \cup \mathcal{A}^{\neg F}_C$.

We consider a clustering task whose goal is to partition the input objects with a twofold objective: (*i*) minimize the inter-cluster similarity according to the non-sensitive attributes $\mathcal{A}^{\neg F}$; (*ii*) minimize the intra-cluster similarity according to the sensitive attributes $\mathcal{A}^F$. The former objective corresponds to the typical clustering objective, since dissimilar objects should belong to different clusters. Pursuing the second objective, instead, would help distribute objects that are similar in terms of sensitive attributes across different clusters, thus fostering the formation of clusters that are equally represented in terms of the sensitive attributes. This is beneficial to ensure that the distribution of groups defined on sensitive attributes within each cluster approximates the distribution across the dataset. Formally, the problem we tackle in this work is:

Problem 6.2 (Fair-CC). Given a set of objects $\mathcal{X}$, two subsets of attributes $\mathcal{A}^F$ and $\mathcal{A}^{\neg F}$, and an object similarity function $sim_S(\cdot)$ defined over the subspace S of the attribute set, find a clustering C^* to minimize:

$$\sum_{u,v \in \mathcal{X},\, C(u)=C(v)} sim_{\mathcal{A}^F}(u,v) \quad + \sum_{u,v \in \mathcal{X},\, C(u)\neq C(v)} sim_{\mathcal{A}^{\neg F}}(u,v) \tag{6.2}$$

The objective in Eq. (6.2) corresponds to solving a complete Min-CC instance where the set of vertices corresponds to the objects in $\mathcal{X}$ and, for each pair of vertices u and v, the positive-type (resp. negative-type) correlation-clustering weight corresponds to the similarity score between the two vertices according to the non-sensitive (resp. sensitive) attributes.

We remark that the Fair-CC problem, as stated above, is introduced here for the first time, while in our previous study in [156] we tackled a different problem: given a set of objects defined over sensitive and not-sensitive attributes, find two attribute subsets that lead to pairwise similarity scores satisfying a certain global condition on the correlation-clustering edge weights. The focus in [156] was to show that the global condition can guide the selection of subsets of features that lead to edge weights expressing the best trade-off between an accurate representation of objects' vectors (i.e., discarding not too many features), and the way how the weights facilitate the downstream correlation-clustering algorithm performing well, i.e., by making it achieve approximation guarantees [156]. Instead, in this work, the set of attributes, over which the similarity scores are computed, are given as input in the Fair-CC problem, and hence they are not needed to be discovered. This is also a more realistic scenario for fair clustering, where the set of sensitive attributes is provided by the specific application scenario.

6.4 Algorithm

The Fair-CC problem requires a function to measure the similarity between two objects with respect to a set of attributes. Following [156], we quantify the degree of similarity between two objects u and v, according to the set of sensitive and non-sensitive attributes, by means of the following $sim_{\mathcal{A}^{\neg F}}(u,v)$ and $sim_{\mathcal{A}^F}(u,v)$ measures, respectively:

$$sim_{\mathcal{A}^{\neg F}}(u,v) := \psi^+\Big(\alpha_N^{\neg F} \cdot sim_{\mathcal{A}_N^{\neg F}}(u,v) + (1-\alpha_N^{\neg F}) \cdot sim_{\mathcal{A}_C^{\neg F}}(u,v)\Big), \tag{6.3}$$

$$sim_{\mathcal{A}^F}(u,v) := \psi^-\Big(\alpha_N^{F} \cdot sim_{\mathcal{A}_N^{F}}(u,v) + (1-\alpha_N^{F}) \cdot sim_{\mathcal{A}_C^{F}}(u,v)\Big), \tag{6.4}$$

Algorithm 1 CCBounds [156]

Input: Set of objects $\mathcal{X}$, sensitive attributes $\mathcal{A}^F$, non-sensitive attributes $\mathcal{A}^{\neg F}$, Min-CC algorithm A

Output: Clustering C of $\mathcal{X}$

1: compute $sim_{\mathcal{A}^{\neg F}}(u,v)$, $sim_{\mathcal{A}^F}(u,v)$, $\forall u,v \in \mathcal{X}$, as in Eqs. (6.3)–(6.4)
2: build the instance $I = \langle G = (\mathcal{X}, \mathcal{X}\times\mathcal{X}), \{sim_{\mathcal{A}^{\neg F}}(u,v), sim_{\mathcal{A}^F}(u,v)\}_{u,v\in\mathcal{X}\times\mathcal{X}}\rangle$
3: $C \leftarrow$ run A on I

where $\alpha_N^F = |\mathcal{A}_N^F|/(|\mathcal{A}_N^F| + |\mathcal{A}_C^F|)$ and $\alpha_N^{\neg F} = |\mathcal{A}_N^{\neg F}|/(|\mathcal{A}_N^{\neg F}| + |\mathcal{A}_C^{\neg F}|)$ are coefficients to weight similarities proportionally to the number of involved attributes, and $\psi^+ = exp(|\mathcal{A}^F|/(|\mathcal{A}^F| + |\mathcal{A}^{\neg F}|) - 1)$ and $\psi^- = exp(|\mathcal{A}^{\neg F}|/(|\mathcal{A}^F| + |\mathcal{A}^{\neg F}|) - 1)$ are smoothing factors to penalize correlation-clustering weights that are computed on a small number of attributes. The latter is reasonable as, in a fair clustering task, we usually have fewer sensitive attributes, and it should be avoided that negative-like weights can dominate the positive-like ones. The exponential function enables a mild smoothing, which is desirable.

As Fair-CC is an instance of Min-CC, it can be solved by Min-CC algorithms. Specifically, although it was originally devised for a slightly different problem (as previously explained in Section 6.3), here we borrow the algorithm proposed in [156] and adapt it to solve the Fair-CC problem. This algorithm, dubbed CCBounds [1] and presented in Algorithm 1, consists of building a Min-CC instance with vertices as the input data objects and edge weights as the similarity scores, and then running a Min-CC algorithm A on such a Min-CC instance.

Theoretical remarks. Let $T_{\mathsf{A}}(\mathcal{X})$ be the running time of the algorithm A on the set of data objects $\mathcal{X}$. CCBounds runs in $O(|\mathcal{X}|^2|\mathcal{A}| + T_{\mathsf{A}}(\mathcal{X}))$ time complexity since it needs to compute a similarity score, over $\mathcal{A}$ attributes, for each pair of objects in $\mathcal{X}$, and then solve the resulting Min-CC instance through algorithm A. Also, the space complexity of CCBounds is $O(|\mathcal{X}|^2)$ for storing the similarity scores in memory. The specific Min-CC algorithm A used in CCBounds is the one proposed in [7], since it provides (under the probability constraint or the global weight bound stated in [156]) constant-factor approximation guarantee in expectation. Also, taking linear time in the size of the input graph, to the best of our knowledge, it is the most efficient algorithm in the Min-CC literature. As a result of this choice, the time complexity of CCBounds becomes $O(|\mathcal{X}|^2|\mathcal{A}|)$.

Another appealing aspect of the fact that Fair-CC is an instance of Min-CC is that Fair-CC inherits the following theoretical result:

Theorem 6.3 ([156]). *If the condition* $\binom{|\mathcal{X}|}{2}^{-1}\sum_{u,v\in\mathcal{X}}(sim_{\mathcal{A}^{\neg F}}(u,v) + sim_{\mathcal{A}^F}(u,v)) \geq \max_{u,v\in\mathcal{X}}|sim_{\mathcal{A}^{\neg F}}(u,v) - sim_{\mathcal{A}^F}(u,v)|$ *holds on the similarity scores and the oracle* A *is an* α*-approximation algorithm for Min-CC,* CCBounds *is an* α*-approximation algorithm for Fair-CC.*

The above theorem provides approximation guarantee on the Fair-CC objective (cf. Eq. (6.2)), which combines the cluster quality measure (first summation) and the fairness-related objective (second summation). It is not known how this quality guarantee translates into the single objective, e.g., the fair objective. This is a challenging open question which we defer to future studies.

[1] https://github.com/Ralyhu/globalCC

6.5 Fairness Evaluation

In this section, we summarize the most-commonly adopted metrics for the evaluation of fairness aspects in clustering. We focus on algorithm-independent measures, i.e., able to generalize across multiple methods, following a *group-level* approach under the *disparate impact doctrine* [78].

Balance. It is one of the most adopted evaluation metrics for fairness in clustering, initially proposed by Chierichetti *et al.* [49] in a context with one sensitive attribute with two protected groups. It has been successively generalized to m protected groups by Bera *et al.* [23]. According to the latter, the balance of a clustering solution can formally be defined as follows [48]:

$$balance(C) = \min_{C \in C, b \in [m]} \min \left\{ R_{C,b}, \frac{1}{R_{C,b}} \right\} \in [0, 1], \tag{6.5}$$

where $R_{C,b}$ is the ratio between the proportion of the objects belonging to a given protected group b in the considered dataset and in a given cluster $C \in C$.

In such a formulation, the lower and upper bounds of a cluster indicate the fully unbalanced and perfectly balanced scenarios, respectively, where the former indicates the case where all the objects in such a cluster pertain to the same protected group, whereas the latter denotes an equal number of objects from each of the protected groups. Therefore, the higher the balance, the better the obtained solution, in terms of equality. Additionally, the considered generalization allows us to obtain a comprehensive evaluation of the balance of our clustering solutions, as it looks at the dataset context, i.e., it will return high scores provided that the balances of the clustering and the input dataset are comparable.

Average Euclidean Fairness. This metric was introduced by Abraham *et al.* [2] to estimate the unfairness by assessing the deviation between the representation of groups obtained focusing on the sensitive attributes in the whole dataset and the given clustering solution. It expresses the cluster-size weighted average of cluster-level deviations (i.e., Euclidean distances) between two frequency (sensitive) attribute vectors, namely $\mathcal{X}_A$, which is computed over the entire set of objects, and C_A, which is computed for each cluster $C \in C$, focusing on a sensitive attribute $A \in \mathcal{A}^F$. Formally, it is defined as:

$$AE_A(C) = \frac{\sum_{C \in C} |C| \times ED(C_A, \mathcal{X}_A)}{\sum_{C \in C} |C|}, \tag{6.6}$$

where ED represents the Euclidean distance between the frequency attribute vectors. Since A can be multi-valued, such a formulation is suited to scenarios where there are multiple protected groups. Also, as this measure is a deviation, smaller values correspond to better solutions.

6.6 Experimental Methodology

6.6.1 Competing Methods

In the following, we briefly overview the competing methods we included in our experiments. For each of those methods, we used publicly available code, which we adopted "as-is", i.e., without making any changes or optimizations.

Fair Clustering Through Fairlets [49]. This method, here dubbed FAIRLETS, is one of the pillars of fair clustering. It is based on the notion of *fairlets decomposition*, that is a grouping

of the input objects into *fairlets*, i.e., minimal subsets of objects that satisfy a given fairness definition, while preserving the clustering objective. Given a good fairlets decomposition, this approach requires traditional clustering algorithms (i.e., k-center or k-median) applied on the centers of the obtained fairlets, to yield the "fair" solutions. FAIRLETS supports two types of fairlets decomposition: an accurate one based on *min cost flow* (MCF), and a more efficient one. We hereinafter refer to those decompositions as *MCF decomposition* and *vanilla decomposition*, respectively. A major limitation of FAIRLETS is that it can handle a single sensitive binary attribute only. We will discuss the impact of such limitations in more detail in Section 6.7.

We involve FAIRLETS in our experimental evaluation by resorting to the unofficial implementation available online.[2]

HST-based Fair Clustering [17]. This approach, here dubbed HST-FC, focuses on the k-median formulation, and employs a quad-tree decomposition to embed the objects in a a tree metric, called *HST*. By leveraging such a tree, HST-FC computes an approximate fairlets decomposition. A fair clustering is ultimately obtained by running k-median algorithms on the produced fairlets. Like FAIRLETS, HST-FC suffers from the limitation that it deals with one binary sensitive attribute only.

In our experiments, we adopt the official implementation made available by the authors of HST-FC.[3]

Fair Correlation Clustering [5]. This method, here dubbed SIGNED, introduces a fairlet-based reduction for the graph clustering scenario with respect to the problem of correlation clustering, leading to the concept of correlation clustering with fairness constraints. Specifically, given a signed graph, i.e., an undirected graph with edges labeled as positive or negative, the algorithm performs a fairlet decomposition (under different fair settings) over the set of vertices. The produced decomposition is used, together with the original graph, to build a reduced (complete and unweighted) correlation clustering instance, where the vertices correspond to the produced fairlets and the sign of the edges between any two fairlets are built according to the majority sign of the edges between vertices within those two fairlets. A clustering on this reduced correlation clustering instance is computed through local-search optimization starting from all singleton clusters, and then expanded into a solution of the original problem. As a fair setting for the fairlets decomposition, we consider the most common case of fair decomposition where clusters are required not to have a sensitive data class. As the SIGNED method requires a signed graph as input, we perform the following preprocessing step to make the relational data compatible with this format. We derive a complete graph whose vertices are the original data objects and an edge (u, v) is labeled as positive with probability $p^+_{uv} = max\{0, sim_{\mathcal{A}^{-F}}(u, v) - sim_{\mathcal{A}^F}(u, v)\}$ and as a negative edge with probability $1 - p^+_{uv}$, where the similarity functions are the ones defined in Eqs. (6.3)–(6.4). We point out that, although we can adapt the same weighting strategy as CCBounds to obtain the edge attributes, we discarded this choice as our experiments showed that it favors the emergence of a degenerated clustering solution (i.e., a single output cluster), due to the strong predominance of positive weights on the edges.

In our evaluation, we use the official implementation made available by the authors of SIGNED.[4]

[2] https://github.com/guptakhil/fair-clustering-fairlets

[3] https://github.com/talwagner/fair_clustering

[4] https://github.com/google-research/google-research/tree/master/correlation_clustering

Table 6.1. Overview of the datasets involved in our experiments.

	#objs.	*sensitive* attribute	*non-sensitive* attributes
Adult	48 842	sex	age, fnlgwt, education_num, capital_gain, hours_per_week
Bank	40 004	marital	age, balance, duration
CreditCard	10 127	sex	customer_age, dependent_count, avg_utilization_ratio, total_relationship_count
Diabetes	101 763	sex	age, time_in_hospital
Student	649	sex	age, study_time, absences

6.6.2 Data

We considered five real-world relational datasets, which have been commonly used in the fair clustering literature. The main characteristics of these datasets are summarized in Table 6.1. As reported in the table, in our evaluation we focused on a smaller subset of the original attributes; note that this is a common practice, which is adopted, among others, by the competing methods outlined above.

Adult.[5] This dataset reports information about the 1994 US Census. For each tuple representing an individual, we considered *age*, *fnlwgt*, *education-num*, *capital-gain* and *hours-per-week* as non-sensitive attributes, and *sex* (i.e., male or female) as a sensitive attribute.

Bank.[5] This provides details on phone calls involving direct marketing campaigns of a Portuguese banking institution to assess whether the bank term deposit will be subscribed or not. We considered attributes *age*, *balance* and *duration* as non-sensitive, and *marital status* (i.e., married or not) as sensitive.

CreditCard.[6] This dataset concerns customer credit card services to estimate customer attrition. We considered attributes *customer_age*, *dependent_count*, *avg_utiliz- ation_ratio* and *total_relation ship_count* as non-sensitive, and *sex* as sensitive.

Diabetes.[5] It reports diabetic patient records, for which we considered *age* and *time_in_hospital* as non-sensitive attributes, and *sex* as a sensitive attribute.

Student.[5] This dataset contains student performances for Mathematics and Portuguese language in secondary education of two Portuguese schools. We considered *age*, *study_time* and *absences* as non-sensitive, and *sex* as sensitive.

6.6.3 Evaluation Goals

Our evaluation objectives concern both fairness and quality aspects of clustering. In the first case, we use the fairness metrics defined in Section 6.5, which allow us to have a group-wide overview of how a method behaves in terms of fair principles. In the second case, we assess the quality of clustering by means of intra- and inter-clustering similarity, considering both the sensitive and non-sensitive attributes, as described below. Finally, we evaluate running times.

[5] `https://archive.ics.uci.edu/ml/datasets/`
[6] `https://www.kaggle.com/sakshigoyal7/credit-card-customers`

Table 6.2. Configurations and hyper-parameters used in our evaluations w.r.t. different experimental setups. k_{avg} is the avg. number of clusters that were obtained over ten runs of CCBounds, and k corresponds to the parameter value provided to FAIRLETS and HST-FC.

	p, q	split ratio	k_{avg}	k
Adult-1k	1,2	650/350	3.12	3
Bank-1k	1,2	650/350	3.48	3
Credit-Card-1k	1,6	800/200	5.6	6
Diabetes-1k	1,2	540/460	5.2	5
Student-1k	1,2	266/383	3.88	4
Adult-10k	1,2	6 500/3 500	2.96	3
Bank-10k	1,2	6 500/3 500	3.28	3
Credit-Card-10k	1,6	4 769/5 358	6.32	6
Diabetes-10k	1,2	5 400/4 600	6.44	6
Adult-Full	2,5	32 650/16 192	3.64	4
Bank-Full	2,5	12 790/27 214	3.64	4
Diabetes-Full	1,2	47 055/54 708	OOM	6

Intra/Inter-cluster similarity. As stated in Section 6.3, we take into account the intra-cluster, resp. inter-cluster, similarity among objects to properly distribute them into clusters, either focusing on their sensitive and non-sensitive attributes (cf. Eqs. (6.3) and (6.4)). We define the following aggregated scores to have an overall measure of goodness of the clusters:

$$inter(\mathcal{A}^{\neg F}) = \frac{1}{|\Theta|} \sum_{u,v \in \Theta} sim_{\mathcal{A}^{\neg F}}(u, v), \qquad inter(\mathcal{A}^{F}) = \frac{1}{|\Theta|} \sum_{u,v \in \Theta} sim_{\mathcal{A}^{F}}(u, v), \quad (6.7)$$

$$intra(\mathcal{A}^{\neg F}) = \frac{1}{|\Omega|} \sum_{u,v \in \Omega} sim_{\mathcal{A}^{\neg F}}(u, v), \qquad intra(\mathcal{A}^{F}) = \frac{1}{|\Omega|} \sum_{u,v \in \Omega} sim_{\mathcal{A}^{F}}(u, v), \quad (6.8)$$

where $\Omega = \{u, v \in \mathcal{X} \mid C(u) = C(v)\}$, and $\Theta = \{u, v \in \mathcal{X} \mid C(u) \neq C(v)\}$. In particular, to obtain fair clusters, we need to maximize (resp. minimize) the $inter(\mathcal{A}^{F})$, resp. $intra(\mathcal{A}^{F})$, scores, so that objects having the same set of *sensitive* attributes will not be clustered together, rather they will be well-distributed across clusters. Conversely, we require to maximize, resp. minimize, the $inter(\mathcal{A}^{\neg F})$, resp. $intra(\mathcal{A}^{\neg F})$, scores, to ensure that objects with the same set of *non-sensitive* attributes will be clustered close with each other and not scattered across different clusters.

Running times. We measure the running times of CCBounds and the competing methods while executing them on the *Cresco6* cluster.[7]

6.6.4 Hyper-parameters and Configurations

Data sampling and attributes selection. To test the selected competing methods under different conditions, and run even the most computationally expensive approaches, we adopt

[7] `https://www.eneagrid.enea.it`

the sampling strategy proposed in [49]. Specifically, by sampling (without replacement) we extracted 1k or 10k tuples from the original full set of tuples, by preserving some desired ratio between the protected classes. The details of the sampling strategy used in our experiments are reported in Table 6.2, where the selected fair attributes and split ratio (i.e., the fraction of tuples pertaining to different sensitive attribute values) are, whenever possible, the same as [49]. Also, both FAIRLETS and HST-FC require two integers p and q as input, whose ratio p/q corresponds to the minimum balance required by each clusters, yielded by these algorithms. The configuration of the aforementioned parameters, inspired by [49, 23], is reported in Table 6.2.

We highlight that, as described so far, we focus on a single and binary sensitive attribute to match the minimum requirements that embrace all competing methods. Nonetheless, some approaches (including our CCBounds) can deal with multiple values assigned to a single sensitive attribute.

Number of clusters. While FAIRLETS and HST-FC require a hyper-parameter k in input, denoting the desired number of output clusters, the same does not apply with the correlation clustering-based approaches. Thus, to create a reasonable comparative environment, we use the (rounded) average number of clusters returned by CCBounds in ten iterations as the k parameter for FAIRLETS and HST-FC. Moreover, we inherit the value k from the nearest subset when the correlation clustering-based approaches run out of memory.

6.7 Results

Table 6.3 summarizes the results achieved by CCBounds and the competing methods. With the exception of very high running times and out of memory errors (indicated with NA and OOM, respectively), all reported measurements correspond to averages over 10 runs of the tested algorithms. The similarity values (Eqs. (6.7)–(6.8)) were obtained by using Euclidean and Jaccard similarities for numerical and categorical attributes, respectively. Moreover, as for the FAIRLETS method, as previously discussed in Section 6.6.1, we report results only for the vanilla fairlets decomposition, since the min-cost-flow (MCF) counterpart has very high running times (more than 7 minutes on the smallest dataset, i.e., *Student-1k*) and produces solutions that are very similar to the vanilla one (results not shown for the sake of brevity).

As for the balance, we notice that, although CCBounds does not match the high scores obtained by "fairness-native" methods (i.e., FAIRLETS and HST-FC), it is still able to score comparably with its direct competing method, i.e., SIGNED. Exceptions arise in the case of *Student-1k* and *Diabetes-1k*, where CCBounds sets up to lower scores, and for some large datasets, where SIGNED does not terminate in reasonable time, while our CCBounds still obtains good results in reasonable time. The paradigm shifts when we consider small yet heavily unbalanced datasets (i.e., *CreditCard-1k*, with an 80:20 ratio); here, although several competing methods struggle to obtain high scores, CCBounds achieves the second-best balance score. Overall, as the balance obtained by CCBounds in all evaluation scenarios ranges from 0.45 to 0.613, we can conclude that it is able of guaranteeing satisfactory balance scores.

In the case of avg. Euclidean fairness, CCBounds obtains very good scores under different scenarios: it is among the best-performer approaches for the *Adult-1k*, *Adult-Full* and *Bank-1k* datasets, and outperforms all the other methods by an order of magnitude on *Bank-10k* and *Bank-Full*. Conversely, CCBounds is unable to match the best scores obtained by some of the competing methods when focusing on the remaining datasets.

Considering the similarity computed on the sensitive attributes, CCBounds does not achieve the best intra-cluster similarity, meaning that it tends to group a few more objects

Table 6.3. Summary of results according to the following criteria (columns from left to right): number of clusters, balance score, avg. Euclidean fairness, avg. intra-cluster and inter-cluster similarities according to either the set of selected sensitive attributes or the set of non-sensitive attributes (cf. Table 6.1), and running time. For each criterion, bold values correspond to the best-performing methods (possibly up to the second decimal point).

		$\#clust.$	balance ↑	AE ↓	$intra(\mathcal{A}^{\neg F})$ ↑	$intra(\mathcal{A}^{F})$ ↓	$inter(\mathcal{A}^{\neg F})$ ↓	$inter(\mathcal{A}^{F})$ ↑	time (s) ↓
Adult-1k	CCBounds	3.12	0.565	**0.007**	**0.685**	0.524	**0.415**	**0.334**	< **1**
	FAIRLETS	3	0.805	**0.004**	0.585	**0.319**	0.596	**0.335**	< **1**
	HST-FC	3	**0.971**	0.01	0.616	0.335	0.599	**0.336**	< **1**
	SIGNED	41	0.66	0.03	0.59	**0.32**	0.60	**0.33**	240
Adult-10k	CCBounds	2.96	0.52	0.03	**0.65**	0.43	**0.43**	**0.33**	3.86
	FAIRLETS	3	0.82	**0.003**	0.60	**0.32**	0.615	**0.33**	< **1**
	HST-FC	3	**0.98**	**0.006**	0.626	0.336	0.618	**0.336**	3.03
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Adult-Full	CCBounds	3.64	0.56	**0.003**	**0.69**	0.47	**0.42**	0.24	75.5
	FAIRLETS	4	0.66	0.02	0.59	**0.32**	0.62	**0.34**	**6.5**
	HST-FC	4	**0.96**	**0.008**	0.63	0.34	0.62	**0.34**	72.86
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Bank-1k	CCBounds	3.48	0.565	**0.006**	**0.727**	0.587	**0.441**	**0.369**	< **1**
	FAIRLETS	3	0.828	**0.002**	0.606	**0.354**	0.613	**0.364**	< **1**
	HST-FC	3	**0.968**	**0.007**	0.621	0.365	0.617	**0.365**	< **1**
	SIGNED	41	0.7	0.03	0.61	**0.35**	0.63	**0.36**	224
Bank-10k	CCBounds	3.28	0.52	**0.0007**	**0.78**	0.63	**0.45**	**0.36**	4.74
	FAIRLETS	3	0.7	0.001	0.59	**0.32**	0.63	**0.36**	< **1**
	HST-FC	3	**0.969**	0.004	0.656	0.365	0.656	**0.365**	3.07
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Bank-Full	CCBounds	3.64	0.55	**0.0004**	**0.72**	0.55	**0.45**	**0.37**	51.1
	FAIRLETS	4	0.68	0.001	0.62	**0.34**	0.65	0.36	**5.3**
	HST-FC	4	**0.94**	0.008	0.66	0.37	0.66	**0.37**	28
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
CreditCard-1k	CCBounds	5.6	0.613	0.127	**0.6**	0.497	**0.46**	0.362	< **1**
	FAIRLETS	6	0.4	0.042	0.485	**0.355**	0.486	**0.375**	< **1**
	HST-FC	6	**0.756**	**0.026**	0.513	0.373	0.481	**0.377**	< **1**
	SIGNED	171	0.56	0.1	0.56	0.41	0.49	**0.38**	173
CreditCard-10k	CCBounds	6.32	0.496	0.17	**0.6**	0.46	**0.46**	0.32	4.1
	FAIRLETS	6	0.94	**0.01**	0.497	**0.34**	0.49	**0.337**	< **1**
	HST-FC	6	**0.955**	**0.013**	0.52	**0.337**	0.491	**0.337**	2.52
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Diabetes-1k	CCBounds	5.2	0.45	0.33	**0.622**	0.519	**0.512**	0.352	< **1**
	FAIRLETS	5	**0.92**	**0.015**	0.537	0.381	0.532	**0.385**	< **1**
	HST-FC	5	0.872	0.05	0.585	0.386	0.529	**0.386**	< **1**
	SIGNED	106	0.85	0.04	0.58	**0.36**	0.54	**0.38**	257
Diabetes-10k	CCBounds	6.44	0.48	0.22	**0.65**	0.54	**0.5**	0.36	4.72
	FAIRLETS	6	**0.92**	**0.01**	0.53	**0.38**	**0.53**	**0.39**	< **1**
	HST-FC	6	0.799	0.065	0.59	**0.388**	**0.53**	**0.386**	2.84
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Diabetes-Full	CCBounds	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	FAIRLETS	6	**0.93**	**0.01**	OOM	OOM	OOM	OOM	**22.2**
	HST-FC	6	0.81	0.06	OOM	OOM	OOM	OOM	761.2
	SIGNED	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
Student-1k	CCBounds	3.88	0.51	0.10	**0.625**	0.463	**0.471**	0.224	< **1**
	FAIRLETS	4	0.82	**0.013**	0.528	**0.339**	0.543	**0.357**	< **1**
	HST-FC	4	**0.93**	0.024	0.563	0.357	0.541	**0.358**	< **1**
	SIGNED	55	0.82	0.04	0.57	**0.34**	0.55	**0.36**	71

with the same sensitive attribute value than the other methods. Nevertheless, the inter-cluster similarities are comparable with the other methods, thus indicating that CCBounds is still able to properly separate the objects into clusters, when accounting for the sensitive attribute.

Instead, when we focus on the similarity computed on the non-sensitive attributes, CCBounds achieves the best performance in all the considered evaluation scenarios, yielding very high-quality clusters.

Finally, we also investigated on running times, spotting FAIRLETS as the best performer, followed by HST-FC and CCBounds, which both guarantee reasonable running times. Although CCBounds has quadratic time complexity due to pairwise similarity calculations (cf. Section 6.4), we managed to perform in parallel such time-consuming steps. On the contrary, SIGNED requires excessively long execution times, often resulting infeasible in practice, along with an abnormal number of clusters produced, which is particularly large even when considering the smallest *1k* datasets. Overall, it should be noted that, albeit the observed running times should be taken with grain of salt due to the (lack of) code optimizations, major remarks are consistent with the time complexities of the corresponding methods.

Discussion. A number of remarks arise from our experimental evaluation. First, although native fairness-aware approaches are able to produce clustering solutions that optimize fairness notions, we found out that such a capability comes with a cost, as the produced clusters are often far from being qualitatively good. On the other hand, CCBounds demonstrated itself to be effective and versatile: it was recognized as the best-in-case approach among the tested ones when it comes to find good-quality clusters, while also being able not to excessively penalize aspects related to fairness.

Second, although we unveiled the weakness in quality shown by the native fair-clustering approaches, we nonetheless shed light on how the approaches based on correlation clustering might suffer from computational issues, by being slower than the other methods, and requiring more memory. This is particularly evident with SIGNED, as it is unable to terminate in all datasets having more than 10k tuples, while it is kept under control in CCBounds, which goes down only in the case of *Diabetes-Full* (containing more than 100k tuples, cf. Section 6.6.2), thanks to the numerous optimization adopted under the hood. However, such a dataset makes it difficult to calculate similarities even for traditional and more efficient approaches, despite the computing capabilities at our disposal.

Finally, by wearing the lens of our proposed approach, we can state that it is able to provide performance in terms of fairness-aware metrics that are comparable to its direct competitor (i.e., SIGNED), but, at the same time, it manages to overcome all the state-of-the-art competing methods considered in our assessment, when it comes to generating qualitatively good clusters, anyway preserving aspects of fairness as much as possible.

6.8 Chapter review

In this chapter, we analyzed how a correlation clustering method, called CCBounds, can profitably be used for the problem of fair clustering. Experimental evidence on real data has shown the meaningfulness of the clustering solutions produced by CCBounds, also revealing its ability of yielding clusters of higher quality than the considered competing methods, according to classic clustering-validation criteria, without discarding aspects of fairness.

In the future, we plan to further evaluate the performance of CCBounds under other conditions, e.g., multiple protected values. Also, we aim to investigate on alternative definitions of the similarity functions and push forward the capabilities of CCBounds towards more challenging scenarios, such as embracing multiple sensitive attributes with many values, allowing us to align with more realistic use cases, and strengthen the versatility of the correlation clustering under fairness constraints.

7

Modeling, Analysis, and Visualization of Law Reference Networks

Summary. The regulation of private law is a focal element in the metamorphosis of society, and by gaining a broader vision of the legal domain we might grasp novel or potentially hidden nuances and characteristics. In this work, we explore the Italian Civil Code (ICC) from an unprecedented perspective based on network analysis. We develop a text processing method to identify and extract article references from the ICC and, upon these, we define network models capturing their relation structures either at a book and corpus scale. The exploitation of the main structural features of these networks leads us to unveil meaningful patterns, holding within and across the books composing the ICC. Furthermore, by leveraging a community detection task, we investigate whether the formation of a community is related to the topic coherence of its assigned articles over the portions of books involved. Our findings reveal useful indicators that may help legal experts and practitioners enhance their knowledge from a novel perspective provided by the network of article references through the ICC.

7.1 Introduction

The Italian Civil Code, hereinafter referred to as ICC, is the legislation source containing norms that regulate private law in Italy. Enacted by Royal decree no. 262 of March 16, 1942, the ICC has been involved in a perpetual process of refinements and enhancements, and subjected to numerous reorganizations to stay updated with respect to legislative needs and social development.

The ICC is compiled as an organic corpus relating to the fundamental and constitutional civil laws. In addition to the organization into various books and their sections, the corpus and its constituent articles have an additional structure that is described by *cross-article references*. These are citations that may occur in the content of an article to refer to one or more articles, from either the same or different books, and hence they are exploited by legislators to clarify the scope and the semantics of specific articles. The ICC article references can naturally be modeled as a network, so to enable the discovery and analysis of relation patterns among the article contents beyond the original, logical organization of the corpus.

Research in artificial intelligence and law has traditional focused on a number of problems that are typically addressed by natural language processing and machine learning models and methods. Despite a relative gap with respect to the network science discipline, a few works have been proposed to investigate the complexity in a legal corpus domain by leveraging

network analysis methods. For instance, Fowler et al. [85] use a network-based representation to characterize the most important precedents at the U.S. Supreme Court. Moser et al. [167] study the structural properties of the citation networks inferred from Austrian Supreme Court decisions. Koniaris et al. [134] model the legislation network of the European Union law, and explore its topological structure and evolution over time, also evaluating its resilience. Mazzega et al. [159] study the network of French Legal codes. Moreover, Zhang et al. [239] propose a software tool to visually explore semantic-based citation networks to ease the analysis of the relations and the evolution of legal issues.

To the best of our knowledge, the Italian Civil Code has not been studied through its article-reference-based structure so far. Therefore, our main goal in this work is to contribute with a study of the networks that can be inferred from the ICC corpus based on article references. This allows us to extend our scope beyond the canonical exploration boundaries of the legal domain, by providing new insights on the interpretation and evaluation of the Italian civil law.

Our main contributions are summarized as follows:

- The ICC articles are not commonly available as hypertexts. Therefore, we develop a text processing method to identify and extract article references that are valid w.r.t. the ICC, hence discarding references to legal sources that are external to the ICC.
- Based on the article-reference relation, we define two network models by differentiating at level of single book or entire ICC corpus.
- We perform an analysis of the book-induced and corpus networks in order to unveil main structural features of such networks and interesting patterns underlying the article references within and across the books of the ICC. Our analysis of the networks is twofold, as we consider macro-scale and meso-scale properties of the networks.
- Upon the outcomes of our performed task of community detection, we also investigate on relations between the formation of communities and the topic coherence within and across books of the ICC. Our findings reveal useful indicators that may help legal experts and practitioners integrate their knowledge from a novel perspective that lays on a unifying, mesoscopic organization of the network of article relations spanning over the whole ICC.

The remainder of the Section is organized as follows. Section 7.2 introduces the ICC corpus, describes our process of extraction of valid article references from the ICC, and presents our defined network models. Section 7.3 contains our analysis of the structural characteristics of the ICC networks. Finally, Section 7.5 concludes the chapter.

7.2 Data Preparation and Models

The Italian Civil Code (ICC) is divided into six, logically coherent books, each in charge of providing rules for a particular civil law theme:

- *Book-1*, on Persons and the Family (articles 1-455) — contains the discipline of the juridical capacity of persons, of the rights of the personality, of collective organizations, of the family;
- *Book-2*, on Successions (articles 456-809) — contains the discipline of succession due to death and the donation contract;
- *Book-3*, on Property (articles 810-1172) — contains the discipline of ownership and other real rights;
- *Book-4*, on Obligations (articles 1173-2059) — contains the discipline of obligations and their sources, that is mainly of contracts and illicit facts (the so-called civil liability);

- *Book-5*, on Labor (articles 2060-2642) — contains the discipline of the company in general, of subordinate and self-employed work, of profit-making companies and of competition;
- *Book-6*, on the Protection of Rights (articles 2643-2969) — contains the discipline of the transcription, of the proofs, of the debtor's financial liability and of the causes of preemption, of the prescription.

The articles of each book are internally organized into a hierarchical structure based on four levels of division, namely (from top to bottom in the hierarchy): "titoli" (i.e., chapters), "capi" (i.e., subchapters), "sezioni" (i.e., sections), and "paragrafi" (i.e., paragraphs). It should however be emphasized that this hierarchical classification was not meant as a crisp, ground-truth organization of the articles' contents: indeed, the topical boundaries of contiguous chapters and subchapters are often quite smooth, as articles in the same group often not only vary in length but can also provide dispositions that are more related to articles in other groups.

The ICC currently in force consists of 2 969 article numbers, which actually corresponds to 3 225 articles considering all variants and subsequent insertions. However, during its history, the ICC was revised several times and subjected to repealings, i.e., per-article partial or total insertions, modifications and removals; to date, 2 294 articles have been repealed.

7.2.1 Extraction of article references

An article in a specific book of the ICC may contain citations of one or more articles, which are from the same or a different book. We will refer to these citations as *article references*.

Unfortunately, identifying and extracting article references from ICC articles is not straightforward, because of two main reasons:

1. The ICC and its books are not designed as hypertexts, neither any index structure containing article references is originally available.
2. An article may also contain references to laws or legal items that do not correspond to ICC articles.

Therefore, since our goal is to infer citation networks from the lists of article references that are contained in each ICC book, we developed an approach to identify and extract *valid* article references, i.e., citations of articles within the ICC only. In the following, we elaborate on the text pre-processing steps that were carried out for the task at hand.

The ICC is obviously publicly available, in various digital formats. From one of such sources, we extracted the contents of each article from all books. Note that we normalized all variants and abbreviations of frequent keywords such as "articolo" (i.e., article), "decreto legislativo" (i.e., legislative decree), "Gazzetta Ufficiale" (i.e., Official Gazette), and finally we lowercased all letters.

An article reference is comprised of three parts:

- *Prefix*, which corresponds to a common root for all lexical variants of article; typically, the abbreviation "*art*" is used.
- *Article id*, which follows the numerical intervals specific to each book (as described earlier in this section).
- *Variant suffix*, which is optional and used to designate a variant of a given article, which is however not alternative, and hence must be treated as a separate article in the book; specifically, an article variant suffix corresponds to a Latin adverbial numeral, to express the multiplicity of occurrence, and hence subsequent versions of a given article id, i.e., "*bis*" (stands for "twice"), "*ter*" ("three times"), "*quater*" ("four times"), and so on.

As mentioned before, in addition to references to other articles in the ICC, an article may contain references to legal items that are external to the ICC corpus. Since they appear in similar textual format, it is not trivial to distinguish between the two types of references. Nonetheless, in the article contents, we can recognize *cue-words* to exploit as hints for the presence or not of references of either type:

- *Relevant* cue-words are used to characterize the presence of valid article references; for instance, "*codice civile*" ("civil code") and its lexical variants.
- *Irrelevant* cue-words are instead used to locate references to external sources. These include "*legge*" ("law"), "*decreto legislativo*" ("legislative decree"), "*sentenza*" ("judgment"), and are usually followed by a date in some format. Moreover, they also include other words, such as "*comma*" ("paragraph"), which may follow a reference inside a portion of the context delimited by round brackets.

It should be noted that relevant cue-words are much fewer and less frequently used than irrelevant cue-words. Moreover, there are cases in which both types of cue-words are found within an article content, also with multiple occurrences of one or both types, while in other cases they are both missing. For example, the following is an excerpt from article 42-bis that contain multiple valid references:

> "`[...] si applicano inoltre art2499, art2500, art2500-bis, art2500-ter, secondo comma art2500-quinquies, art2500-nonies, in quanto compatibili. [...]`"

By contrast, in the next example from article 86, references are to external sources only, and hence must be discarded from our analysis:

> "`[...] la legge 20 maggio 2016, 76 ha disposto (con art1, comma 35) che la presente modifica acquista efficacia dal 5 giugno 2016. [...]`"

To process the article contents for the task of article reference extraction, we first segmented an article's text into shorter passages, dubbed contexts, that are delimited by "." or ";" and preceded by an alphanumerical sequence of at least four characters; the latter constraint is needed to avoid selecting false contexts, e.g., corresponding to one of the many abbreviations that are found in the ICC articles such as "d.p.r.", "c.p.c.", "c.p.p.", "etc.".

Algorithm 2 sketches our designed procedure to identify and extract article references from the text of ICC articles. The procedure takes in input two lists of words, which are used to find indicators of presence of relevant and irrelevant *cue-words* for the task at hand.

In the extraction procedure, the *search* function takes in input a list of cue-words and a context to check whether any cue-word occurs within the context. If irrelevant cue-words are found in the context but relevant cue-words are not, then the next context is processed; otherwise, the context is scanned left-to-right through the *findAll* function, which returns all non-overlapping matches of the article-references pattern in the given context.

By applying Algorithm 2 to the aforementioned examples, the output for the first text will be a list of 6 references to associate with article 42-bis, while in second text the output is an empty list since the occurrence of word "art1" is correctly recognized as an external, hence irrelevant, reference, due to the presence of irrelevant cue-words (i.e., "la legge 20 maggio 2016" and "comma 35").

Algorithm 2 Extraction of article references

Data: Set of articles $\mathcal{A}$ in the ICC; list of relevant cue-words $art\text{-}refs\_words$ and list of irrelevant cue-words $ext\text{-}ref\_words$
Output: Article reference sets AR

```
AR ← {}
foreach a ∈ 𝒜 do
    AR_a ← {}
    contexts ← getContexts(a)
    foreach c ∈ contexts do
        isIrrelevant ← search(ext-refs_words, c) and not search(art-refs_words, c)
        if not isIrrelevant then
            ⟨a, refs⟩ ← findAll("art[0-9]+[a-z]*", c)
            AR_a ← AR_a ∪ {⟨a, refs⟩}
        end
    end
    AR ← AR ∪ AR_a
end
```

7.2.2 Network models

Let us denote with $\mathcal{A}$ the set of articles in the ICC, and with $\mathcal{B}$ a partition of $\mathcal{A}$ into six groups, each corresponding to a book in the ICC, i.e., $\mathcal{B} = \bigcup_{i=1..6} \mathcal{A}_i$, such that $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$, for all $i, j \in \{1, \ldots, 6\}$ with $i \neq j$, where $\mathcal{A}_i \subset \mathcal{A}$ is the subset of articles assigned to book i. Let also $r : \mathcal{A} \mapsto 2^{\mathcal{A}}$ denote a function that associates each article a to a set of articles that are referred to by a, possibly including articles that are from the same book of a or a different one.

Based on the article-reference relation, and by differentiating at level of single book or entire corpus, we define the following network models of interest to the study of the ICC.

Book-induced networks. Our first defined network model focuses on representing the relations between articles of each particular book and their article references, which may cross the boundary of the book itself. Given a book, the induced network is a directed graph built from the article-reference list of all articles of that book. Formally, for each book i, we define the book-induced network for i as the directed graph $G_i = \langle V_i, E_i \rangle$ such that $V_i = \{a \in \mathcal{A}_i \mid r(a) \neq \emptyset\} \cup \{a \in \mathcal{A} \setminus \mathcal{A}_i \mid \exists a' \in \mathcal{A}_i, a \in r(a')\}$ and $E_i = \{(a, a') \mid a, a' \in V_i, a \in r(a')\}$.

Global or **corpus network.** Our second network model encompasses all relations among articles observed in the entire ICC. Therefore, we define the ICC corpus network $G_{\mathcal{B}}$ as the directed graph obtained by merging all book-induced networks, i.e., $G_{\mathcal{B}} = \langle V, E \rangle$ such that $V = \bigcup_{i=1..6} V_i$ and $E = \bigcup_{i=1..6} E_i$.

7.3 Structural Analysis of the ICC Networks

In this section we present our analysis of the book-induced and corpus networks, which is aimed to characterize main structural features of such networks and to unveil interesting patterns underlying the article references within and across books of the ICC. We organize our presentation into two subsections: the first one (Section 7.3.1) is concerned with *macroscopic* structural properties of the networks, whereas the second (Section 7.3.2) is focused on *mesoscopic* structural properties based on the outcome of a community detection task.

Table 7.1. Statistics about the extraction of the article references from the ICC books.

	Book-1	Book-2	Book-3	Book-4	Book-5	Book-6
# articles	395	345	364	891	713	331
# articles w/ references	123	71	45	77	258	88
# article-references	243	132	70	118	551	180
avg. # article-references per-article	0.615	0.383	0.192	0.132	0.773	0.544
avg. # article-references per-article w/ references	1.976	1.859	1.556	1.532	2.136	2.045

Table 7.2. Summary of structural characteristics of the ICC corpus network (first column) and book-induced networks (subsequent columns).

	$G_{\mathcal{B}}$	G_1	G_2	G_3	G_4	G_5	G_6
#nodes	1 147	223	144	95	161	432	171
#edges	1 294	243	132	70	118	551	180
reciprocity	3.4%	4.9%	3%	2.9%	0%	2.2%	7.8%
density	0.001	0.005	0.006	0.008	0.005	0.003	0.006
average degree*	2.218	2.126	1.806	1.453	1.466	2.523	2.023
average in-degree	1.128	1.090	0.917	0.737	0.733	1.275	1.053
% sources	37.2	35	35.4	41.1	43.5	35.4	31
% sinks	42.3	44.8	50.7	52.6	52.2	40.3	48.5
assortativity*	0.016	-0.184	-0.063	-0.058	-0.141	0.012	-0.173
assortativity	-0.035	-0.198	-0.051	-0.072	-0.158	-0.037	-0.196
average path length	2.241	1.639	1.393	1.104	1.064	2.384	1.568
diameter	7	6	3	3	3	7	5
transitivity*	0.099	0.119	0.135	0.160	0.107	0.098	0.109
clustering coefficient*	0.166	0.227	0.225	0.197	0.220	0.129	0.190
clustering coefficient *(full averaging)**	0.081	0.106	0.097	0.039	0.055	0.072	0.091
#strongly connected components	1 128	218	142	93	161	427	166
#weakly connected components *	157	23	29	30	50	47	23
modularity*	0.891	0.866	0.882	0.914	0.946	0.807	0.876
#communities*	174	32	32	30	50	59	30
modularity	0.892	0.868	0.881	0.909	0.946	0.812	0.876
#communities	175	32	32	30	50	60	30

* Statistic calculated by discarding edge orientation

7.3.1 Macroscopic properties

Table 7.2 reports main results of our macroscopic structural analysis if the ICC corpus and book-induced networks. As a first general remark, only a portion of the articles is actually involved in article reference relations; in particular, considering the global network, the number of nodes (1 147) corresponds to about 36% of the articles within the ICC. This is actually not surprising, since it is commonly expected that a law article should be self-contained or self-explanatory. Nonetheless, there are norms regulating specific sections of the law code which, to be completely specified, require one or more references to different articles, possibly crossing

the boundaries of a book. Indeed, this is what happens for a fraction of the ICC, which is not negligible, and hence deserves to be investigated.

In light of the above remark, one trait common to all the networks is the low *density*, which is the actual number of edges divided by the maximum possible number of links in the network, i.e., $|E|/(|V|(|V|-1))$ for any directed graph $G = \langle V, E \rangle$. Partly related is also the low average *degree*, i.e., the average number of references involving an article, as well as low average *in-degree*, i.e., the average number of references to an article. Focusing on the latter, we observe that the ICC corpus network $G_{\mathcal{B}}$, and the book-induced networks for Book-1, Book-5 and Book-6 have average in-degree above 1 (i.e., on average, each article is pointed by at least one other article), whereas the remaining networks exhibit a value lower than 1, indicating broader isolation. Indeed, by looking at an article's in-degree as a measure of its authoritativeness and usefulness to clarify and deepen the semantics of the articles that point to it, we find out that some articles indeed take a central role in the ICC. Figure 7.1 displays a visualization of the ICC corpus network, where articles of the same book are colored the same, and nodes with highest in-degree are made evident with associated label.

By contrast, the periphery of each of the networks represents a significant fraction of nodes, as it can be noted from the percentages of *source* and *sink* nodes reported in Table 7.2. We refer to source and sink nodes as those having only outgoing links and incoming links, respectively, i.e., source articles include others in their definition but are not referred to by other articles, whereas sink articles are used in the definition of one or more articles but they do not need others to complete themselves.

The above duality also impacts on the degree correlation, also known as *assortativity*, which captures how the probability of links between two nodes is influenced by their degree [169, 170]. For all networks, we observe negative correlation, which means that articles with different degrees tend to link each other.

We also analyzed the tendency of articles to form strong ties with closure patterns, specifically dyadic closure or *reciprocity* and *triadic closure*. The former is computed as the fraction of reciprocal edges, and helps us understanding how articles might reinforce each other to enhance and refine their meaning. Reciprocity is found to be low for all networks, with 3.4% for the corpus network and upper bounded by 8% for the book-induced networks. For the latter, it is interesting to observe two contrasting cases: the one corresponding to zero reciprocity, which characterizes the article references found in Book-4, and the other one corresponding to the maximum reciprocity found in Book-6, thus suggesting a more evident dyadic closure for the articles in that book than in others. As concerns the triadic closure, i.e., the probability of closing connected triplets of nodes, we resorted to both global and local measures [32]. In the former case, we evaluated the the probability that two incident edges are completed by a third one to form a triangle, namely *transitivity*. In the latter case, instead, we evaluated the transitivity at node level, i.e., how strongly connected are the neighbors of a node, namely *local clustering coefficient*. Although both measures show low values, it should be noted how the local clustering coefficient shows slightly greater values when ignoring source and sink articles.

We also investigated reachability aspects in the various networks. First, we examined the *average path length*, i.e., the average of the pairwise distances between nodes in a network, where the distance between two nodes corresponds to the length of a shortest path connecting them. As it can be noted from the table, the average path length is slightly above 2 for the corpus network and for the Book-5 network only. Also, the *diameter*, i.e., the shortest-path distance between the two most distant nodes in the network, is maximum in the Book-5 network (7) and determines the diameter for the corpus network as well; Book-1 and Book-6 networks have diameter equal to 6 and 5, respectively, while the remaining networks have diameter 3.

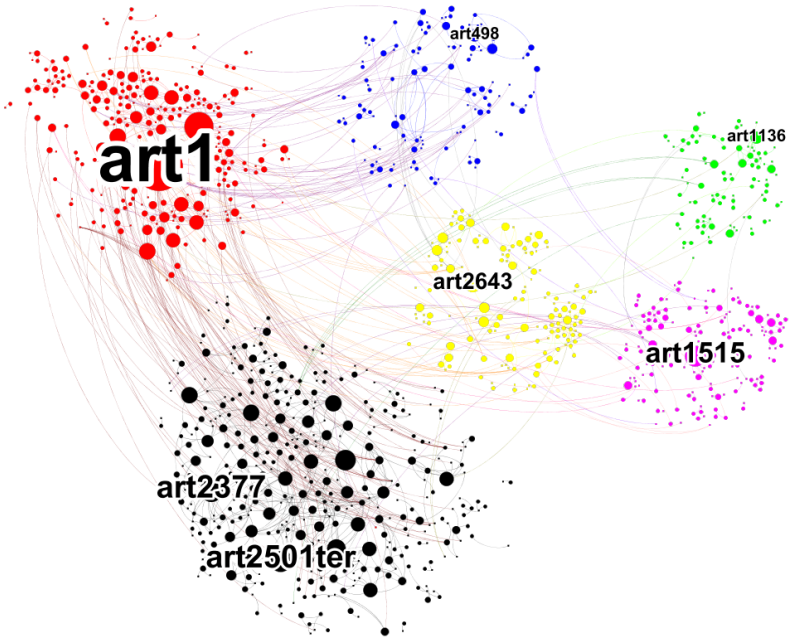


Fig. 7.1. Articles in the ICC corpus network. Node sizes are proportional to the in-degree, and for each book a label representing the article id is associated to the node(s) with highest in-degree. Colors are used to distinguish the six books of the ICC as follows: red (Book-1), blue (Book-2), green (Book-3), magenta (Book-4), black (Book-5), yellow (Book-6).

7.3.2 Mesoscopic properties

Reachability-based approaches to the identification of subnetworks with particular properties of connectivity focus on the existence of paths, regardless of the distance. In this respect, for each of the networks under study, we calculated the *strongly* and *weakly* connected components, which correspond to the maximal subgraphs where every node is reachable from every other node through a directed or undirected path, respectively. Given the outcome of the above discussed analysis steps, it does not come to our surprise that the observed number of strongly connected components is extremely high, which indicates the formation of small groups of articles that are involved in chains of references when accounting for edge orientation, and highlights the lack of *multi-hop references* (i.e., *art. x* refers to *art. y*, which in turn refers to *art. z*). By discarding edge orientation, instead, mutual connectivity clearly increases, and the detected number of weakly connected components is one order of magnitude lower than the

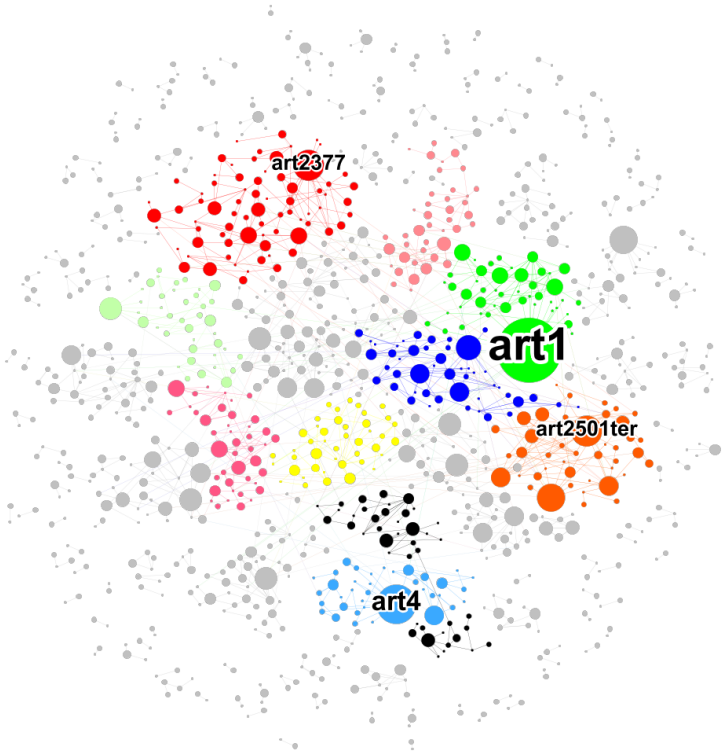


Fig. 7.2. Community structure of the ICC corpus network. Node sizes are proportional to the in-degree, and nodes with in-degree greater than or equal to 10 are labeled with the associated article id. Colors are used to distinguish the top-10 largest communities (nodes assigned to the remaining communities are colored in gray).

node-set size. This trait confirms the existence of strategic articles, namely those referred by other ones and that support (undirected) connectivity between articles.

Connected components can be seen as a raw form of *communities*, based on minimal requirements of connectedness. However, according to a density hypothesis, communities should correspond to locally dense neighborhoods of a network. Therefore, we delved into each of the article reference networks by carrying out a mesoscale-level analysis to shed light on the underlying *community structure*, i.e., a division of a network into regions such that the nodes in each region should be highly linked with each other, whereas few links should exist between the regions. Note however that the amount of edges per se is not an optimal proxy to quantify the community structure: a good community structure is not merely one in which there are few edges between communities, rather it is one in which there are fewer than expected edges between communities. This is the key principle underlying the theory of *modularity* to discover a community structure in a network: intuitively, the modularity of a community is the total difference of the fraction of the edges within the community w.r.t. the expected such fraction if the edges were distributed at random, so that the higher this deviation, the better

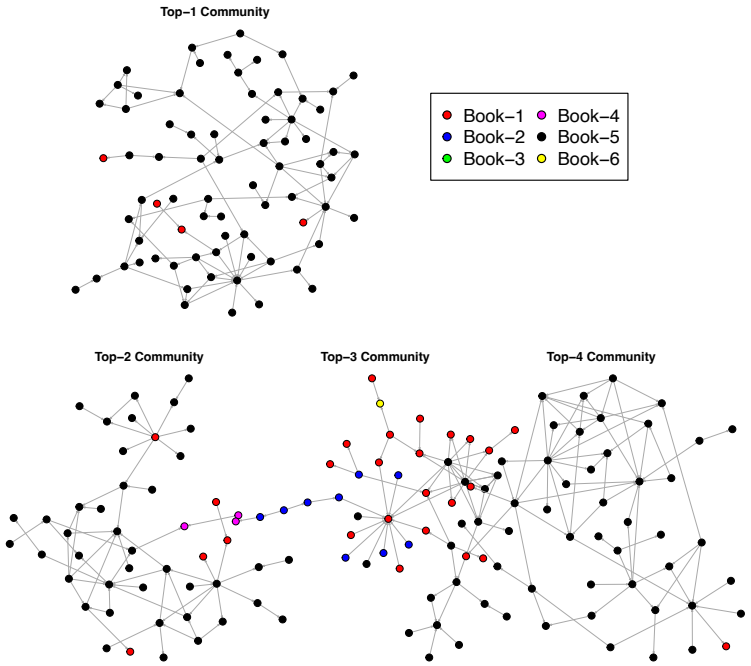


Fig. 7.3. Top-10 largest communities discovered by the Louvain method in the ICC corpus network. Color codes correspond to the ICC books.

the community. In this work, we follow the widely-recognized line of research that resorts to a modularity maximization approach to community detection. In particular, we use the most popular method belonging to this category, namely the *Louvain* method [30], both in its original, undirected version and its variant that accounts for edge orientation while maximizing the modularity.[1]

As reported in Table 7.2, both Louvain and its directed variant lead to the discovery of communities having high modularity (note that modularity is upper bounded by 1) in all networks. This is also consistent with the presence of several connected yet isolated regions within the networks, which are also comparably in size with the connected components previously discussed. Furthermore, it should be noted that accounting for edge orientation does not imply significant differences in modularity as well as number of communities with respect to the outcome of the undirected counterpart. Figure 7.2 illustrates the community structure identified in the ICC corpus network, where the top-10 largest communities are emphasized and distinguished by color. Note that these top-10 communities cover about 40% of the network.

[1] https://github.com/nicolasdugue/DirectedLouvain

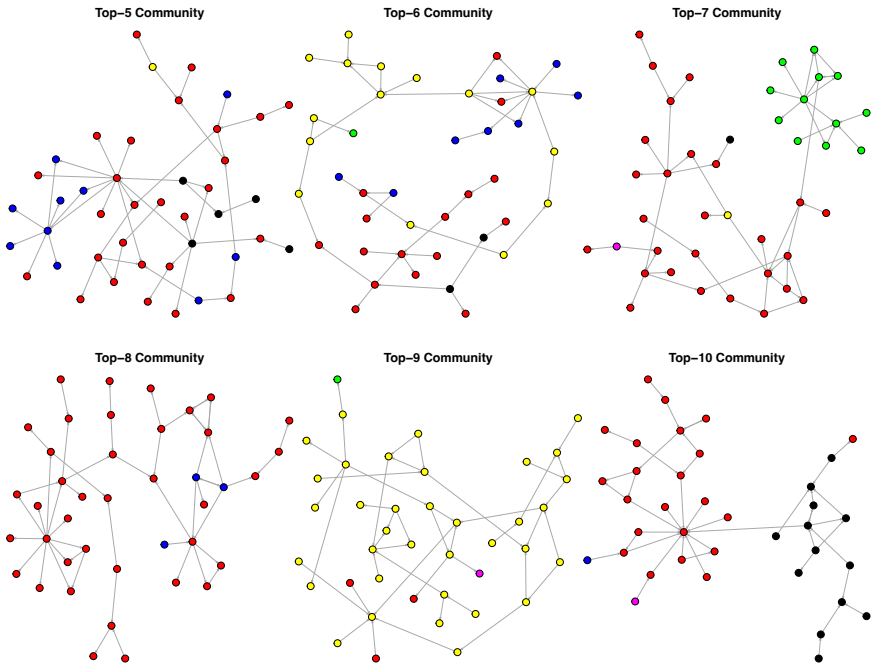


Fig. 7.4. *(Cont.)* Top-10 largest communities discovered by the Louvain method in the ICC corpus network. Color codes correspond to the ICC books.

One interesting question that arises is whether the different communities contain mixed book-memberships, i.e., whether a community may cross the book boundaries through the article reference relations. To this purpose, we explored the top-10 largest communities in the ICC corpus network, which are visualized in Figures 7.3–7.4. As it can be noted from the figures, there are indeed cohesive communities, which are mostly formed by articles from the same book, as well as communities that contain articles from different books. This would suggest that the modular structure in the ICC corpus network can be consistent with either themes of a particular (portion of) book or themes that are shared by (portions of) different books. Clearly, the latter might originate from the opportunity of completing or enhancing the norms provided by a book's article(s) with those provided by other books' article(s).

In this respect, we moved a step forward by exploring the contents of the articles involved in the top-10 communities, in order to gain insights into patterns underlying possible relations between the formation of communities and the topic coherence. From this analysis stage, several remarks stand out, which we try to summarize as follows:

- Community capturing most representative topics of a particular book. This is likely to happen when the book memberships in the community are mostly or fully homogeneous. For instance, the top-1 community contains articles focused on "administration of the capital of a company", which is a representative topic of Book-5.
- Community unveiling fine-grain topical patterns that are mostly discussed in a book yet complemented with references to other book(s). This refers to sort of strong ties that are formed through article references that cross the boundaries of two or more books.

For instance, top-8 and top-10 communities are such a type of community. In the top-8 community, an interesting pattern is found out for "succession" norms (Book-2) in a context of "marital separation" (Book-1).

- Community induced from reinforcement of topic(s) from a book with related topics that differently contribute to the contents of other books. This type of community turns out to be characterized from mixed book-memberships that are distributed over substructures built upon across-book article references. For instance, the top-6 community contains node-articles that belong to five different books; moreover the topic "transcription" that is largely discussed in Book-6 (which is the mostly covered in the top-6 community) is reinforced with the topic "properties", which is shared between Book-1 and Book-2, jointly with the topic "community", which is discussed within Book-1.

In light of the above remarks, we can conclude that a mesoscopic view like that supplied by our discovered community structure can represent a valuable support to enhance the macroscopic (i.e., at book level) or microscopic (i.e., at article level) views that are primarily considered from legal experts and practitioners.

7.4 LawNet-Viz: A Web-based System to Visually Explore Networks of Law Article References

Summary. We present LawNet-Viz, a web-based tool for the modeling, analysis and visualization of law reference networks extracted from a statute law corpus. LawNet-Viz is designed to support legal research tasks and help legal professionals as well as laymen visually exploring the article connections built upon the explicit law references detected in the article contents. To demonstrate LawNet-Viz, we show its application to the Italian Civil Code (ICC), which exploits a recent BERT-based model fine-tuned on the ICC.

7.4.1 Problem, Target Users, and Importance

Promoting access to justice is of utmost importance in the legal domain, which is essential to the regulation of our society and its evolution over time. In this regard, it is a raising opinion to argue for the claim that artificial intelligence offers solutions to ease, speed and secure access to justice. Legal research and analytics systems are nowadays defined as using information retrieval, natural language processing, and machine/deep learning methods to provide enhanced search, understanding and/or predictive capabilities for legal matters, such as statute and case law documents.

In this work, we present LawNet-Viz, a web-based tool for the modeling, analysis and visualization of law reference networks extracted from a statute law corpus. LawNet-Viz is designed to support legal research tasks, hence to help legal professionals searching for referred and semantically related articles, but it also provides a reliable form of legal aid to those who are not familiar with the legal domain. The network of article references can be analyzed at macro-, meso- and microscopic levels, enabling inspection of an article's neighborhood and the associated metadata, along with an extensive set of network statistics including node centrality and ranking measures. Moreover, LawNet-Viz allows the user to refine a network according to the semantic similarity of linked articles, which is captured based on sparse vectorial representations, topic models, word embeddings or deep contextualized embeddings.

The usefulness of LawNet-Viz lies in the capability of providing the user with a responsive and interactive visual exploration of law article reference networks, thus reducing the gap her/his knowledge of legal matters — while guaranteeing a fluid and effective user experience. Indeed, LawNet-Viz has the potential of enhancing the ability of lawyers to identify relevant statutes in a cost- and time-effective manner, thereby increasing access to justice. Also, in civil law countries, it can benefit jurists in drafting new codes or abrogating existing ones, by inferring semantic relations between legal documents. For citizens, LawNet-Viz can be helpful to serve their search and consultation needs. Clearly, LawNet-Viz is not intended as a substitute for a full exploration of the social and ethical considerations related to automating legal research, which is beyond the scope of this work.

LawNet-Viz is a system prototype that is designed and implemented to be easily adapted to any law code system. To demonstrate LawNet-Viz, we show its application to the Italian Civil Code (ICC), which exploits a recent BERT-based model trained on the ICC.

7.4.2 Related Systems

Law information retrieval and related tasks have often benefited from the application of network science methodologies [233, 234, 144, 207, 166]. In [58], law search is modeled as a strategic-predictive process, where the search space follows a network-based structure. In [195], an

information extraction framework is proposed to build a dynamic legislation network from legal documents.

A variety of legal sources and contexts have also been covered. For instance, network-based modeling has been utilized for the study of Courts' decisions, such as in [34] for the extraction of legal indicators from the French court of appeal judgments, in [85] to characterize the most legally relevant precedents at the U.S. Supreme Court, and in [167] to study the structural properties of the citation networks from the Austrian Supreme Court decisions. Further studies of network analysis have focused on the European Union law's legislation [134], the French Legal codes [159], the Italian Constitutional Court [3], and the Italian Civil Code [139].

As concerns visual legal analytics, the research tool proposed in [239] builds and visualizes citation networks to study legal issues in case law documents. EUCaseNet [145] is a toolkit that is comprised of a set of visualization modules and provides different analysis features that can be used by domain experts to explore on the fly the European case law corpus. Also, the Knowlex web application [146] is designed to explore and analyze legal documents from multiple sources (norms, case law, legal literature), by means of two main visual functionalities, namely interactive node graph and zoomable treemap showing the topics, the evolution, and the dimension of the legal literature settled over the years around the norm of interest. Moreover, knowledge graphs represent an important scenario to shape and study legal data [80, 205, 148, 81, 121, 57].

Despite such important contributions in visual legal analytics and knowledge graph research, to the best of our knowledge there is still a lack of systems integrating visual and network analysis with recently developed, advanced methods in natural language processing, particularly the deep contextualized language models such as BERT [65]. The latter is instead a key point in the development of our LawNet-Viz.

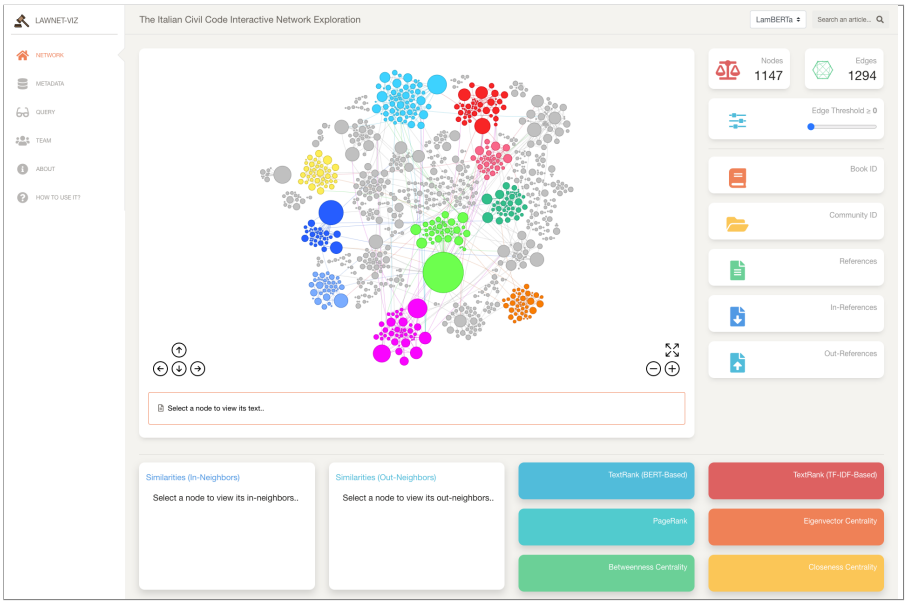


Fig. 7.5. Screenshot of the main interface of LawNet-Viz, with node (i.e., article) colors coding community memberships.

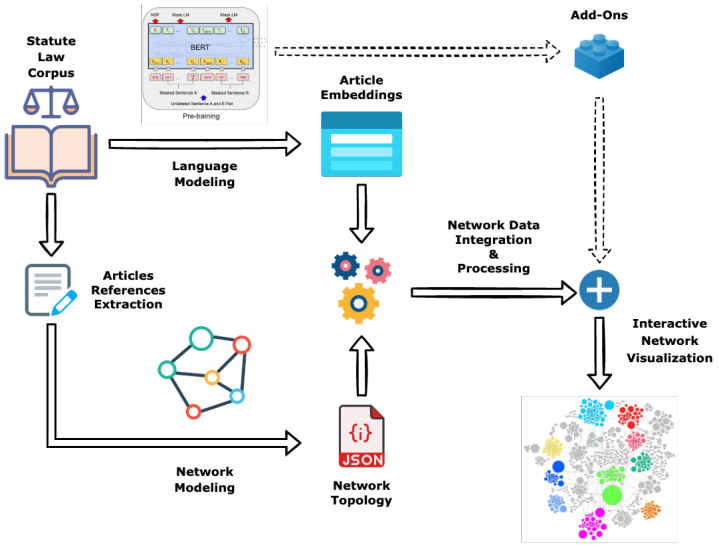


Fig. 7.6. Overview of the architecture of LawNet-Viz.

7.4.3 Design

In this section we provide a conceptual overview of LawNet-Viz and discuss the main functionalities and supported tasks.

LawNet-Viz is logically separated into two independent yet complementary modules that are designed for network and text processing, respectively. The first one is responsible for extracting the references from the articles in the input law corpus to build a law reference network. The resulting graph can be processed through any network analysis software to produce a feature-rich network (i.e., the topology with associated metadata and statistics), which is made available in the form of a JSON object. The second module leverages natural language processing techniques to represent the textual contents of the articles; sparse vectorial representations and deep contextualized embeddings trained on the law corpus are two available but not exclusive options in LawNet-Viz.

The two modules communicate with a third module that is in charge of *(i)* integrating the network topology and metadata with the textual content information, which is used to determine the semantic affinity of any two linked articles, and *(ii)* interactive rendering to enable visual exploration of the integrated data.

LawNet-Viz is designed to be a flexible tool not just in terms of the input law code and relating metadata to build a feature-rich article reference network; it can also play a role as front-end for statute retrieval and question-answering modules. In this regard, we allow extending the core functionalities of LawNet-Viz through additional modules, such as the search and retrieval component which is included in the demonstration of LawNet-Viz (cf. Section 7.4.5).

We sketch the workflow of LawNet-Viz in Figure 7.6. Next, we delve into the interaction *modes* of LawNet-Viz to describe the main functionalities available to the user.

Network mode. The main interface of LawNet-Viz corresponds to a view on the article reference network and provides a number of functionalities to explore and manipulate the network (Figures 7.5 and 7.7). By default, the user is provided with a display of the entire

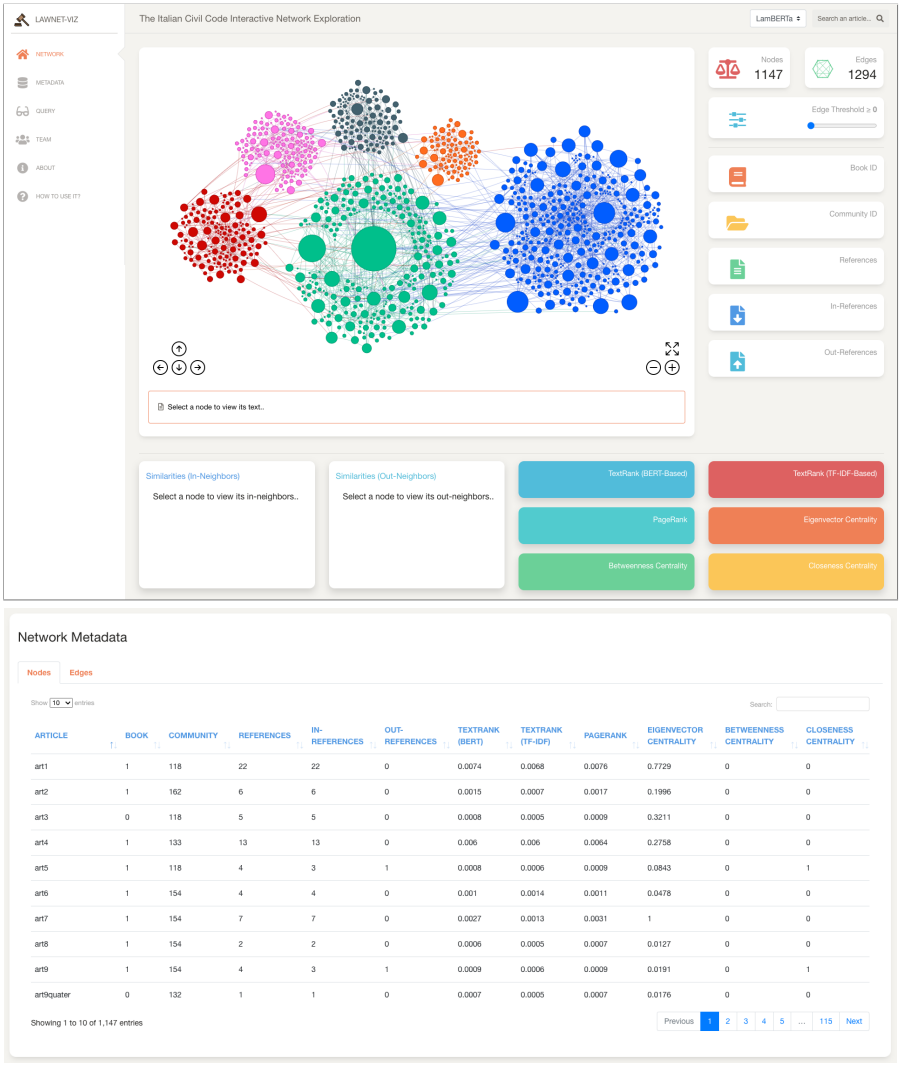


Fig. 7.7. Screenshots of LawNet-Viz network mode, with node (i.e., article) colors coding book memberships (top), and metadata mode (bottom).

network inferred from the analyzed corpus, along with a set of topological statistics. Relations between articles are weighted according to their content similarity; to capture different semantic relation nuances between articles, the user can decide which similarity function to use, and filter edges according to a threshold.

To meet sound user-experience principles, we offer the user a responsive and interactive interface, including but not limited to zooming-in/out and navigating the network through specific controls or using a hand-held pointing device, and selecting articles directly from the network or from a search box. When the user selects an article to delve into details, the article

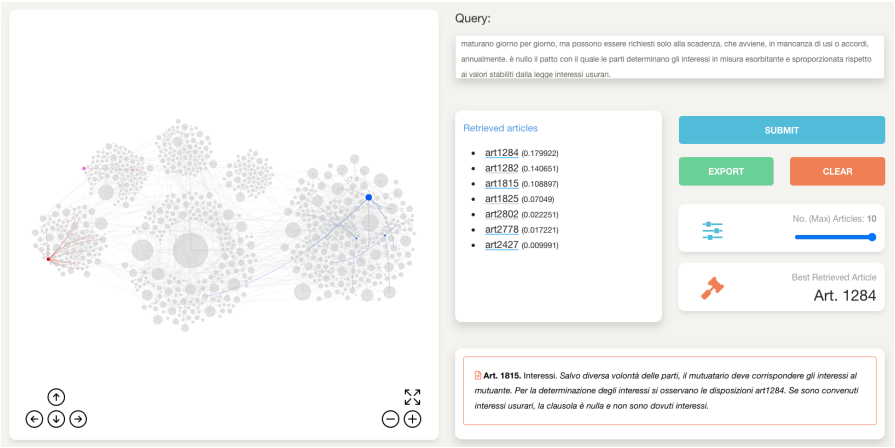


Fig. 7.8. Screenshot of LawNet-Viz query mode.

will be highlighted in the network object container along with its 2-hop expanded neighborhood, thus enabling a visual inspection of the article within its surrounding context. Simultaneously, LawNet-Viz will display all available topological and content related information concerning the article, in particular: *(i)* the article's location in the corpus, which allows the user to notice the high-level subdivision of the corpus (e.g., book) a given article belongs to; *(ii)* the count of incoming and outgoing article's references, and a ranking of these references based on the semantic similarity w.r.t. the article according to the selected similarity model function; *(iii)* statistics on node centrality of the article, to gain an insight into the prominence of the article within the network according to different centrality notions, including standard path-based and eigenvector-based methods, as well as content-biased methods.

Metadata mode. LawNet-Viz provides a table-based interaction mode to explore the metadata associated with nodes (i.e., articles) and edges (i.e., references), which are organized into two tabs (Figure 7.7). Again, the user experience is highly considered, as we enable responsive table exploration through its search, filter, and sort capabilities, based on the article names and/or associated statistics.

Query mode. This is conceived as one add-on of LawNet-Viz, which is designed to fulfill a primary need in exploring a legal corpus, i.e., to retrieve articles relevant to a given input query. Here the user submits a natural language query — which is assumed to be free of references to any article identifier in the law code — and specifies a number k. LawNet-Viz will retrieve the top-k articles from the corpus that satisfy the given query, which are ranked according either to a TF-IDF-like retrieval model or to a relevance probability distribution produced by a contextualized language model specialized for the legal domain. The user can then inspect the retrieved articles, locate their position and context in the reference network, and save the results (Figure 7.8).

7.4.4 Implementation

In this section we elaborate on our development of the constituting components of LawNet-Viz.

Article references extraction. Statute law corpora are not commonly available as hypertexts, therefore a text processing method might be required to identify and extract the article references. We address this task in LawNet-Viz and provide an implementation tailored to the specific syntax (i.e., cue-words, abbreviations, etc.) used to denote article references in the input law code. The reference extractor must also be able to recognize and distinguish article references internal to the code from references to external sources; in our implementation, we discard the external references.

Network modeling. One key component of LawNet-Viz is the graph representation model of the connections between articles to define a *citation* or *reference network*.

We are given a collection $\mathcal{A}$ of articles in the given corpus, and we assume that $\mathcal{A}$ is logically organized into a partition $\mathcal{B}$ of n article groups each corresponding to a high-level subdivision of the input corpus (i.e., a book), that is, $\mathcal{B} = \bigcup_{i=1..n} \mathcal{A}_i$, and $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ for all $\{i, j\} \in \{1..n\}$ with $i \neq j$, where $\mathcal{A}_i \subset \mathcal{A}$ represents the subset of articles of book i. Let us denote with $r : \mathcal{A} \mapsto 2^{\mathcal{A}}$ a *reference function* that associates each article $a \in \mathcal{A}$ to a set of articles referred to by a, either within and across books. For any given book i, we define a directed graph $G_i = \langle V_i, E_i \rangle$ modeling the references carried out by articles in i, where $V_i = \{a \in \mathcal{A}_i \mid r(a) \neq \emptyset\} \cup \{a \in \mathcal{A} \backslash \mathcal{A}_i \mid \exists\, a' \in \mathcal{A}_i \wedge a \in r(a')\}$, and $E = \{(a, a') \mid a, a' \in V_i \wedge a \in r(a')\}$. By merging all book-specific networks, we account for all references observed between articles in the given corpus, so to obtain the corpus-level network as a directed graph $\mathcal{G}_{\mathcal{B}} = \langle V, E \rangle$, such that $V = \bigcup_{i=1..n} V_i$ and $E = \bigcup_{i=1..n} E_i$.

Language modeling. To represent the article contents, we cleaned up the text from non-ASCII characters, removed dates and numbers, including those used as article references, and normalized all variants and abbreviations of terms referring to legislation norms or to the code subdivisions (e.g., chapters, paragraphs, etc.).

We consider different types of text representation models: *(i)* sparse vectorial model (based on TF-IDF weighting function), *(ii)* statistical topic model (Latent Dirichlet Allocation), *(iii)* context-free word embeddings (Word2Vec, FastText), and *(iv)* pre-trained contextualized language model based on BERT [65]. For the implementation of the first three types of models we used the free open-source Python library Gensim [180], whereas for the BERT model(s) we resorted to the HuggingFace Transformers library;[2] in this regard, although any out-of-the-box BERT model can be used, we are aware of the existence of recently developed BERT-based models which are adapted to the legal domain through three main strategies, namely fine-tuning, further pre-training, or from-scratch pre-training of a BERT model on a legal corpus (e.g., [44, 240]). The language modeling component of LawNet-Viz is made flexible to the specific configuration of a legal BERT-based model, allowing for different extensions backed by TensorFlow or PyTorch libraries.

Node and edge features. LawNet-Viz is designed to handle different features associated with articles and article references.

Each node (i.e., article) pair ⟨source, target⟩ is assigned a real-valued weight, which corresponds to the cosine similarity between the vectors of the two articles, according to a selected representation model (cf. Language modeling paragraph).

Attributes at node level include both topological/network-layout information and properties of the corresponding article, such as article ID, article title, article text, book ID, cluster ID, layout coordinates, color, size, number of references (overall, incoming, outgoing), and various centrality scores. The latter include betweenness, closeness, PageRank, Eigenvector centrality, and two variants of TextRank, a well-known weighted extension of the PageRank, where edges

[2] `https://huggingface.co/docs/transformers/index`

are weighted according to some similarity measure; the two TextRank variants implemented in LawNet-Viz utilize the edge features we discussed above. Note that besides the microscopic-level node attributes, the 'book ID' and 'cluster ID' attributes allow for exploring the network at two different mesoscopic levels, i.e., the one corresponding to the logical structure of the law code, and the other one corresponding to the outcome of some graph clustering method (e.g., community memberships). Also, the size and color attributes of any node are defined according to the article's centrality and membership, respectively; by default, the size is set proportionally to the node degree and the color code matches the cluster ID if available, otherwise the book ID.

Data format. All data in LawNet-Viz are structured into a multi-line JSON format enclosing nodes, edges, and their corresponding attributes as discussed above.

We emphasize that our JSON data structure can easily be produced through most of the existing frameworks for network analysis, thus ensuring broad compatibility and interoperability with LawNet-Viz. As a case in point, our JSON data complies with the one that can be exported from Gephi [22] (through its *JSON Exporter* or *SigmaExporter* plugins), which is nowadays widely recognized as the de-facto standard tool for general-purpose network exploration and analysis. Consequently, such a Gephi-compatibility allows users to abstract from programming needs and straightforwardly build a JSON object — along with a set of topological statistics — to be used within our platform; this also includes the capability to transfer the network layout and color coding directly to LawNet-Viz. Nonetheless, we point out that the JSON data structure can also be extended by users to include additional node and edge features.

Web application. Our choice in the implementation of the LawNet-Viz web platform was to leverage exclusively on offline or server-side data processing to minimize potential client-side loads, and make the platform suitable for any desktop or mobile device.

As for the front-end part, we utilize the widely known, free and open-source frameworks *Bootstrap* (v. 5.1.3) [3] and *DataTables* (v. 1.11.4), [4] where the former effectively supports the development of responsive, visually pleasant web user interfaces, and the latter enables the interactive visualization of data tables.

The back-end part is created upon the open-source web-based network visualization *vis.js* library (v. 9.1.0). [5] Note however that, given the visualization and interaction requirements of LawNet-Viz, we extended the core APIs of the "vanilla" version of *vis.js*, and developed specific Python3 scripts, in order to come up with a new set of operations to cover aspects that ensure seamless interplay between topological and textual features, such as dynamic interaction with the articles' node and their texts, responsive filtering of the network based on thresholding of the strength of the article links (i.e., similarity between an article and a referred one), access to different text representation models.

7.4.5 Demonstration

The screen recording of the demo is submitted with this work.[6] The demo shows how to use LawNet-Viz through its three modes.

[3] https://getbootstrap.com/

[4] https://datatables.net/

[5] https://visjs.org/

[6] https://drive.google.com/drive/folders/1csDvMFxQkIDUa2AK6Juc94zGexUBpL4O

As a case in point to demonstrate the functionalities of LawNet-Viz, we present its application to the *Italian Civil Code* (ICC), which contains norms that regulate private law in Italy. The ICC contains more than 3 000 articles which are organized into six, logically coherent books, each in charge of providing rules for a particular civil law theme: *Book-1*, on Persons and the Family, *Book-2*, on Successions, *Book-3*, on Property, *Book-4*, on Obligations, *Book-5*, on Labor, *Book-6*, on the Protection of Rights.

To infer the article reference network from the ICC, we integrate into LawNet-Viz the reference extraction method developed in [139]. The resulting article reference network contains 1 147 nodes and 1 294 edges. Moreover, as BERT-based model, we utilize the *LamBERTa* framework [214], which is designed for law article prediction tasks. Book-specific LamBERTa models are generated by fine-tuning a pre-trained Italian BERT model on a sequence classification task (i.e., BERT with a single linear classification layer on top) given in input a particular book of the ICC. Since each article corresponds to a distinct class, LamBERTa models are designed to face a few-shot learning problem, by exploiting unsupervised learning schemes of labeling of the ICC articles to generate augmented training data.

7.5 Chapter review

Artificial intelligence research in the legal domain is constantly growing and draws on various fields, ranging from natural language processing to machine learning and network science, thus achieving an interdisciplinary imprint. These diversified viewpoints can be valuable to enrich our understanding of the legal domain and to enhance the evolutionary process of law codes, also taking into account social development and needs.

In this work, by taking a network analysis and mining perspective, we presented the first study of citation networks that can be inferred from the ICC articles. Our analysis of the structural features of such networks has shed light on valuable hidden patterns, such as the linkage between community memberships of articles and their topical structure, paving the way for new interpretations and study of the ICC.

It is also worth noticing that our proposed methodology can easily be generalized to other civil law code systems presenting a similar organization as the ICC, i.e., developed upon a logical structure of the corpus into books, and their internal subdivisions.

Besides, it is our opinion that exploring the network underlying the law reference relations as well as semantic affinity among articles in a law code is highly important to support a variety of IR- and AI-based tasks for legal information processing and understanding, including statute law retrieval, entailment and question answering.

In this respect, LawNet-Viz can be regarded as one qualified system for pursuing a twofold ambitious objective: to enhance access to justice for legal professionals as well as for government and administrative agencies, to reduce their workload while responding to the special needs of the legal community, which will increase their productivity and efficiency; to simplify access to intricate regulatory systems for citizens before they will delve into the resolution of their legal problems with the assistance by legal professionals.

8

Evolution of the Social Debate on Climate Crisis: Insights from Twitter During the Conferences of the Parties

Summary. Social media have long been recognized as a valuable proxy for investigating users' opinions by echoing virtual venues where individuals engage in daily discussions on a wide range of topics. Among them, climate change is gaining momentum due to its large-scale impact, tangible consequences for society, and enduring nature. In this work, we investigate the social debate surrounding climate emergency, with particular emphasis on the Conference of the Parties (COP), the foremost global forum for multilateral discussion on climate-related matters. To this aim, we leverage graph mining and text mining techniques to analyze a large corpus of tweets spanning 7 years, aiming to uncover the fundamental patterns underlying the climate debate, thus providing valuable support for strategic and operational decision-making. Our contribution in this work is manifold: (i) we provide insights into the key social actors involved in the climate debate and their relationships, (ii) we unveil the main topics discussed during COPs within the social landscape, (iii) we assess the evolution of users' sentiment and emotions across time. Furthermore, our proposed approach has the potential to scale up to other emergency issues, highlighting its versatility and potential for broader use in analyzing the increasingly debated emergent phenomena.

8.1 Introduction

In recent years, the emergence of social media platforms inexorably changed the way we communicate, share information, and engage with large-scale events. Twitter stands out as one of the most adopted platforms to date, and has been widely exploited as a source of information for researchers seeking to study real-time events. Specifically, the analysis of Twitter data has been instrumental in understanding and responding to extreme events, ranging from natural disasters to social crises. The distance between social and real debates dwindles, and people interacting on social networks are widely recognized as valuable sources of information when it comes to extreme events [194, 186, 162], The vast amount of user-generated social content thus enables us to gain insights into the occurrence, impacts, and response to events, giving rise to the emergent phenomenon commonly dubbed "people as sensors".

Twitter has been proven particularly influential in the context of the annual Conference of the Parties (COP), i.e., the meeting related to the implementation of the United Nations Framework Convention on Climate Change. Such meetings generate unprecedented debates involving a wide plethora of participants, from decision-makers to activists, thus making Twitter

an essential platform for expressing opinions, raising awareness, organizing campaigns, and promoting initiatives, fostering the dynamic exchange of opinions on climate-centered topics. As a result, the platform has become a promising tool for monitoring public opinion and analyzing the main trends, sentiments, and reactions that accompany the social debate around COPs.

By recognizing Twitter as a valuable information source, our goal is to study the evolution of the social debate around the climate crisis, in correspondence with the annual COP meetings, by leveraging social traces left by people discussing climate-related issues as a proxy for real-world debate. Specifically, in this work, we exploit graph mining and text mining techniques seeking to unveil the main patterns and key actors that drive the debate, as well as delve into the sentiment and emotions that distinguish the social response to the climate change narrative on Twitter.

The remainder of this Section is structured as follows. Section 8.2 discusses related studies, Section 8.3 presents methodology and techniques used in our work, Section 8.4 describes experimental results, finally Section 8.5 contains concluding remarks and pointers for future research.

8.2 Related work

Climate change is one of the most studied topics in the social sphere. Kirilenko et al. [128] used Twitter data to investigate the linkage between people's sensory experiences of local temperature and climate change, also assessing the potential influence of mass media on this process. Geo-tagged Twitter data was studied through topic modeling and sentiment analysis by Dahal et al. [59] to characterize the climate change discussion between different countries and over time. Cody et al. [53] assessed how collective sentiment varies in response to climate change news and events, unveiling the role of Twitter as a medium for spreading climate change awareness. By analyzing the tweets about the 2013 IPCC report, Pearce et al. [173] characterized the emerging communities of users around the debate, hinting that contrasting views might lead to greater interactions. In this regard, polarization around the climate social debate has been widely investigated in recent years [118, 71]. Tyagi et al. [218] studied 100 weeks of Twitter discourse about climate change to unveil that deniers of climate change tend to be more hostile towards people who believe in it, and vice versa. Falkenberg et al. [77] studied the Twitter discussion around the COPs, shedding light on increasing ideological polarization, driven by right-wing activity. Network analysis approaches have also been employed to study social media users discussing climate change. For instance, Williams et al. [235] exploited Twitter data to reveal that users tend to segregate within like-minded communities, i.e., echo chambers. Finally, Effrosynidis et al. [74] exploited 15M tweets to provide a comprehensive overview of climate change through several investigation aspects.

Compared to the above works, we focus on the development of the social debate regarding the climate crisis related to the annual COPs through graph mining and text mining tools. Unlike [77] which is the closest study to us, we do not limit to understand polarization as we want to analyze the evolution of the social debate over the years by investigating the main dynamics around user engagement, the key topics discussed by them, and how sentiment and emotions evolve as an effect of the major events that characterize each COP.

8.3 Methods

8.3.1 Problem Statement

Given an input corpus of tweets discussing COPs, we exploit a combination of *Graph Mining* and *Text Mining* approaches to discover valuable information that might help both decision-makers and users improve awareness of and better deal with extreme events such as the climate crisis. Specifically, (i) we infer graph-based models to characterize the main framework underlying the social interactions and shed light on the most influential users for each COP; (ii) we leverage topic modeling to extract the most discussed topics for each COP; (iii) we use lexicon-based frameworks for shaping polarity and emotions expressed by users via textual components, to investigate how they evolve over time.

Our analysis framework is divided into three main modules, which are described next.

8.3.2 Network Analysis Module

We exploit the information concerning retweets to infer graph-based models to analyze main social interaction dynamics around COPs. The inferred *retweet networks* are valuable sources of information for studying social phenomena such as engagement patterns, the spread of information, identifying influential users, and investigating the dynamics of online conversations.

Network modeling. We are given, for each COP $i \in [20..26]$, a corpus $\mathcal{T}_i = \{t_1^i, ..., t_n^i\}$ containing all tweets written during the i-th COP, including one week before and after the conference dates. For each COP i, we infer a directed-weighted *retweet* network $G_i = \langle V_i, E_i, w \rangle$ such that V_i contains all users who *posted* or *retweeted* tweets $\in \mathcal{T}_i$, and $E_i = \{(u, v) \mid u, v \in V_i \wedge u$ retweeted $v\}$. The weighing function $w : E_i \rightarrow \mathbb{R}$ assigns each edge $(u, v) \in E_i$ with a value corresponding to the number of retweets made by user u on tweets posted by user v during the i-th conference period.

8.3.3 Topic Modeling Module

We leverage text analysis of tweets to infer the trending topics and most prominent discussions in the social debate around COPs. Moreover, through this stage of topic modeling, clusters of topically-related tweets can be uncovered. In this respect, we resort to *BERTopic* [92], a powerful method employing BERT (Bidirectional Encoder Representations from Transformers) embeddings and c-TF-IDF (class-based Term Frequency-Inverse Document Frequency) to create dense clusters sharing the same topic. More precisely, BERTopic extracts deep semantic features from the input texts (i.e., tweets) using BERT as encoder, then the BERT-generated embeddings are clustered into similar groups based on the HDBSCAN clustering algorithm. From each of the clusters, topics are finally represented as bags-of-words based on a cluster-level variant of the TF-IDF term relevance weighing function.

8.3.4 Affective Computing Module

Affective computing concerns the analysis of human affects, such as feelings and emotions, based on the computational treatment of subjectivity in a text. Our developed module is designed to detect and measure the sentiments and emotional states expressed by the users in their posted tweets.

Sentiment Analysis. We carried out sentiment analysis through the *Valence Aware Dictionary and sEntiment Reasoner* (VADER) tool [116]. VADER is a versatile, rule-based lexicon sentiment classification method, specifically designed to measure the polarity in a social media text, i.e., to determine if the text expresses a positive, negative or neutral opinion. To this purpose, VADER maps lexical features to emotion intensities, i.e., sentiment scores, being able to understand the basic context of cue words and to understand the emphasis of capitalization and punctuation. The sentiment score of an input text can eventually be obtained by summing up the intensity of each word in the text.

Emotion Recognition. We also resort to *EmoAtlas*[1] for extracting, analyzing, and visualizing emotional information that characterizes an input text expressing the social debate related to COPs. EmoAtlas combines psychological lexicons, network modeling, and artificial intelligence to shape syntactic relationships between words in a text in 18 different languages and detect 8 different categorical emotions. Such emotions are conceived as follows: *joy* is a feeling of happiness, and the opposite of *sadness*, which is a state of sorrow or unhappiness; *fear* is triggered by the perception of danger or threat, and is the opposite of *anger*, the feeling of displeasure or rage; *anticipation* is the act of looking forward to or expecting something, and is the opposite of *surprise*, a feeling of astonishment or disbelief; *disgust* is a feeling of aversion or revulsion, and is the opposite of *trust*, a sense of faith, confidence, or reliance. By means of such a tool, input texts are first processed and enriched, then structured as a semantic network [206], and a score is assigned to each extracted emotion, in order to determine the prevailing sentiment and its magnitude compared to others. The visualization of emotions is carried out using the wheel layout, following the principles of proximity and opposition between pairs of emotions [178]. Emotions are then displayed such that their spatial adjacency or opposition in the wheel respectively indicates semantic proximity and semantic opposition expressed in the text.

8.4 Experimental Results

8.4.1 Data

In order to study a phenomenon of such magnitude as climate emergence, we leveraged the most representative dataset involving climate social debate so far [77]. It consists of a large corpus of tweets spanning 7 years, from COP20 (2014) to COP26 (2021), gathered through the corresponding set of hashtags "cop2x", with $x \in \{0, ..., 6\}$, for a time windows of six months before and after the conference date. We hydrated the corresponding Tweet IDs through the Hydrator tool,[2] following the official Twitter guidelines.

According to the recent changes to the latter, we point out that the most recent COP27 is not part of the dataset, and it was not possible to collect the corresponding data at the time of writing this work due to the new policies enforced for Twitter Academic APIs. [3]

Once obtained all raw data, we processed it as follows. First, we filtered out all non-English tweets to ensure high applicability through major NLP tools to date. Then, to reduce potential noise in the COP social debate, we narrowed our focus to a smaller time window, keeping only tweets published between one week before and after the conference period.

[1] https://github.com/alfonsosemeraro/emoatlas

[2] https://github.com/docnow/hydrator

[3] https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research

Table 8.1. Tweets from 1 week before to 1 week after the COP dates.

	Conference dates	# Tweets	# Users
COP20	12/01/2014 - 12/12/2014	188 085	50 708
COP21	11/30/2015 - 12/12/2015	1 610 105	366 176
COP22	11/07/2016 - 11/18/2016	325 191	90 251
COP23	11/06/2017 - 11/18/2017	357 612	86 899
COP24	12/02/2018 - 12/14/2018	269 906	87 650
COP25	12/02/2019 - 12/13/2019	324 710	118 774
COP26	10/31/2021 - 11/12/2021	2 266 117	644 752

Table 8.2. Main structural traits of the COP retweet networks.

	COP20	COP21	COP22	COP23	COP24	COP25	COP26
# Nodes	44 217	324 186	81 864	77 528	73 951	106 274	565 500
# Edges	95 350	818 433	162 447	158 624	136 895	195 070	1 354 656
Density	4e-05	7e-06	2e-05	2e-05	2e-05	1e-05	4e-06
Average In-Degree	2.156	2.525	1.984	2.046	1.851	1.836	2.396
Degree Assortativity	-0.167	-0.072	-0.101	-0.085	-0.093	-0.100	-0.065
% Sources	90.6	89.6	91.0	89.3	90.0	92.4	90.2
% Sinks	3.2	3.6	3.4	3.7	4.4	3.1	3.9
Average Path Length	5.240	5.797	5.572	5.727	6.077	6.287	7.182
Diameter	19	18	16	19	18	19	23
Reciprocity	0.022	0.020	0.021	0.026	0.019	0.015	0.011
Transitivity	0.005	0.002	0.002	0.003	0.002	0.001	0.001
Clustering Coefficient	0.227	0.197	0.234	0.260	0.248	0.224	0.163
Strongly Conn. Comp.	42 482	310 833	79 236	74 306	71 667	103 752	549 133
Weakly Conn. Comp.	296	2378	731	738	953	998	5218
# Communities	2099	16 310	4275	4655	4579	5328	30 998
Modularity	0.547	0.538	0.610	0.606	0.636	0.636	0.664

We summarize the main details on individual COPs in Table 8.1. Overall, we spotted over 1.5M of tweets involving ∼1.2M unique users across all COPs, with a small subset of ∼3K users appearing in all COPs. We also carried out a number of steps for the categorization of user profiles. We first identified profiles corresponding to international *organizations*, being them politically recognized or not, active in specific fields or generally contributing to the climate debate. We differentiated them from *news spreaders*, which include both reputable news agencies and thematic blogs, and the *official conference accounts* associated with each COP. We also differently labeled individual user accounts, mostly public figures to distinguish between *representatives of* well-known national and international *organizations*, *activists* of any age and gender, *academics*, and *artists*. Finally, we highlighted *super-users* as those highly active in the social debate but not holding authoritative positions.

8.4.2 Retweet Network Analysis

We analyzed the retweet networks corresponding to each COP through multiple perspectives, aiming at unveiling main characteristics in terms of network structure and user roles.

Table 8.2 reports the main structural properties of each retweet network from macroscopic and mesoscopic perspectives, according to commonly used statistics in network analysis — the interested reader can refer to, e.g., [230] for a description of these methods.

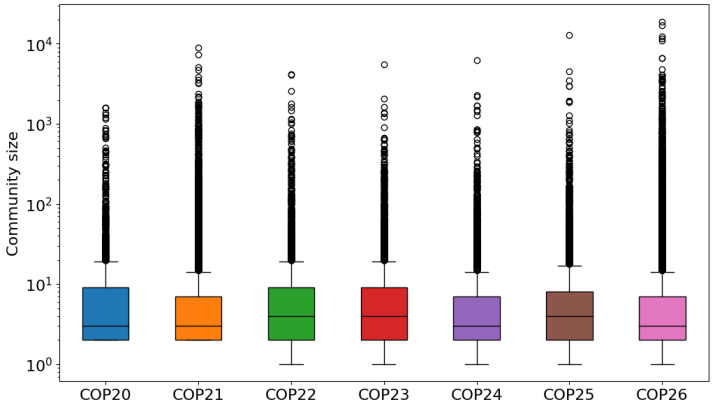


Fig. 8.1. CCDF distribution of community sizes in each COP retweet network.

Interesting traits emerge by considering the number of nodes and edges within each COP. Indeed, COP21 and COP26 stand out in terms of involved users and relationships compared to others, suggesting greater interactions and engagement around the topics discussed in the corresponding conferences. Further investigations unveiled that such a finding is due to temporal and topical factors, e.g., the *Paris Agreement* (COP20) and the *Glasgow Climate Pact* (COP26).

Users' in-degree, i.e., the number of received retweets, might be leveraged as a proxy for their relevance, or interest. While the fraction of relevant users is always quite limited in all scenarios, it turns out that the greater discussion originated from COP21 and COP26 also reflected on users' interest toward some peers. The impact of users' relevance on link formation (i.e., retweets) was also investigated through the degree assortativity, which shapes how the probability of linkage between users depends on their degrees. We observe it remains negative or close to zero in all COPs, indicating a lack of degree correlation among users connecting in the network, i.e., users tend to retweet peers based solely on interest in their content rather than on their relevance within the network.

We then investigated the fraction of users who retweet others yet do not receive any retweet (i.e., sources) and vice versa (i.e., sinks), spotting that around 90% of users within each COP do not receive any retweet, whereas a very small fraction of users never retweeted others' content. Interestingly, such a dichotomy suggests that users generally tend to propagate information produced by a very small fraction of users.

The high values found for path-based measures, such as the average path length and the diameter, indicate the presence of long chains of retweets, denoting potential inefficiency in the network, as it might require several intermediate retweets (and a relatively long time) for a piece of information to propagate from one user to another within the network.

We then delved into link formation (i.e., a proxy for interactions) by investigating how well dyadic and triadic closure principles were met in the networks. As for the former, we spotted a particularly low fraction of reciprocated edges (~2%) for all COPs, hinting at a lack of bi-directionality in interactions. Concerning triadic closure, we delved into both transitivity and local clustering coefficient, spotting in both cases very low values; beyond being expected due to the extremely low density, such values are clues of a reduced engagement and high fragmentation of the networks.

Table 8.3. Top 3 retweeted and retweeters for each COP. Counts indicate the in-degree, resp. out-degree, for each profile.

	Retweeted		Retweeter	
	Category	Count	Category	Count
COP20	Organization	6243	Activist	589
	Organization	3920	News Spreader	572
	Organization	3909	Activist	537
COP21	Organization	50 026	Activist	2429
	Organization	30 185	Activist	2019
	Official Conf. Account	16 681	Activist	1821
COP22	Organization	17 008	Activist	1027
	Organization	7968	Activist	580
	Official Conf. Account	7180	Activist	487
COP23	Organization	16 385	Activist	508
	Official Conf. Account	7123	Activist	489
	Organization	2987	Activist	395
COP24	Organization	17 316	Activist	668
	Repr. of Organization	5106	News Spreader	564
	Activist	4791	Activist	393
COP25	Activist	21 874	Activist	737
	Organization	10 233	Activist	726
	Repr. of Organization	9747	Academic/Repr. of Org.	409
COP26	Official Conf. Account	57 216	Superuser	1955
	Activist	48 373	Artist	1942
	Band	20 443	Activist	1312

Confirmations for the latter emerge from the remarkable number of strongly and weakly connected components across all COPs, i.e., subgraphs where users can reach with each other by following chains of tweets directly, resp. by ignoring the directionality of the links. We found additional hints at such an interesting pattern by looking at the modularity score, i.e., how strongly users are clustered within communities, detected through the widely known Louvain algorithm [31], with internal connectivity way stronger than the external one. Indeed, we spotted that retweet networks are characterized by high modularity, which is also increasing over time, indicating a growing concentration of interactions within specific groups of users, with limited connectivity w.r.t. the remaining users.

We hence deepened the structure of such communities for each COP by inspecting their sizes, as reported in Figure 8.1. We noticed that COP21 and COP26 are again those standing out — along with COP25 —, with some larger communities emerging. Although potentially influenced by the broader set of users and links involved, such a finding might be related to higher specialization in topics characterizing such COPs, as well as a higher engagement around more relevant figures.

We then complemented such analysis by looking for the most retweeted users, i.e., influential ones, and those who retweeted the most, i.e., spreader ones. Both user roles are crucial for the robustness of the networks through which information flows, and are strategic when it comes to maximizing the spreading of useful information, or reducing the diffusion of misinformation and fake news. To this aim, we report in Table 8.3 the top-3 users in descending order of in-degree and out-degree, respectively. As it can be noticed from Table 8.3 (left), the most retweeted accounts in earlier COPs primarily correspond to international organizations involved in climate-related issues, as well as official conference accounts. However, starting

Table 8.4. Main topics discussed in each COP

	Main Topics
COP20	Impact of human activities and efforts to address climate change International climate negotiations, agreements, and commitments Urgent need for action to protect the Earth and save the planet Need for global cooperation and sustainable solutions Impact of deforestation and reforestation on ecosystems and landscapes
COP21	Negotiations and progress surrounding the Paris Agreement Ensuring justice and equity, mobilize financial resources Renewable energy, with focus on solar energy Agriculture, sustainable practices and agricultural policies (Rain)forest preservation, impact of deforestation, and need for reforestation
COP22	Climate hope, sustainability, eco-friendly living, connection to political events Carbon markets and economic aspects of addressing climate change Renewable energy, clean energy initiatives and role of regulatory agencies Participation, activities, and events associated with the conference Sustainable agriculture, climate-smart practices, and ecological aspects
COP23	Promoting climate justice, and mobilizing financial resources Dissemination of news and information related to ecological concerns Agriculture, sustainable practices, and role of farmers Involvement of volunteers, events, partnerships, and a general enthusiasm Challenges faced by Pacific Island nations and their efforts
COP24	International negotiations and the involvement of youth in climate activism Regular updates about sustainability and environmental concerns Clean energy, collective efforts and anticipation for future initiatives Climate change, extinction, and the role of human civilization News, reporting, updates, and events related to the conference.
COP25	Urgency for action, sustainable practices and emissions reduction Media coverage and youth engagement in climate activism Disappointment with the outcomes, and assessment of progress made by countries Role of science, assessments of success or failure, and political influence Climate leaders, and the importance of taking immediate action for a better future
COP26	Need for sustainable practices and collaborative efforts Severity of the climate situation, climate justice and emissions reduction Reducing aviation emissions and promoting sustainable aviation practices Intersection of sustainability efforts with cryptocurrency and related digital assets Inclusion and rights of people with disabilities

from COP23, a greater interest in such issues determined the engagement of individual personalities in the social debate, lending their support to the cause. As concerns top spreaders, Table 8.3 (right) shows the predominance of activists, as expected given their dedication to the cause, followed by accounts devoted to the dissemination of news (e.g., online newspapers or blogs). Besides, we surprisingly spotted that top retweeted and retweeters are likely to share the same communities. Furthermore, recent COPs see some very active emerging profiles on the issue despite not being explicitly stated as activists, an interesting of how these issues are increasingly embedded in the social debate and not the prerogative of a few.

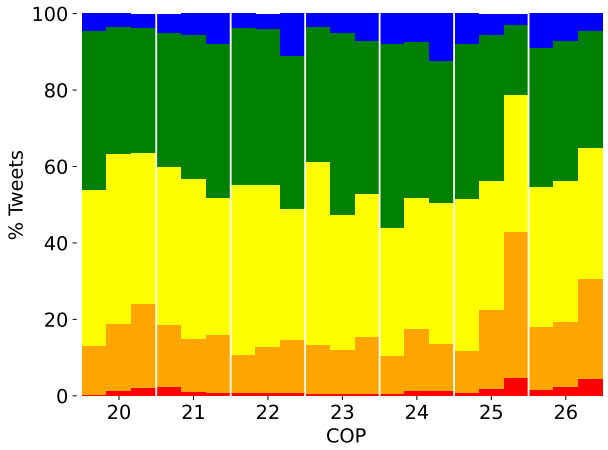


Fig. 8.2. Stacked barchart displaying the percentage of tweets, for each time slice of each COP, belonging to the 5 sentiment classes: extr. positive (blue), positive (green), neutral (yellow), negative (orange) and extr. negative (red).

8.4.3 Topic Modeling

Given the vast amount of data generated while discussing climate change and COPs, we leveraged BERTopic to extract relevant and meaningful topics. Such analysis allowed us to identify and understand the key issues and trends discussed, as well as detect potential shifts in topics that characterized seven years of climate debate. We report the top-5 (by volume of tweets) discussed topics for each COP in Table 8.4.

As expected, some inherent topics are commonly shared by different COPs, reflecting the collective recognition of the urgency to address climate change. We then spotted the topic of deforestation, which consistently emerges from different meetings (COP20, COP21) as an indication of the awareness of their impact on climate change. Similarly, the topic of agriculture spans multiple COPs (i.e., from COP21 to COP23), underlying its essential role in both contributing to and being impacted by climate change. Renewable energies stem from among the most discussed topics in COP21 and COP22, emphasizing the need to transition to more sustainable sources of energy. Similarly, the need for increased funding and financial resources to address climate change characterizes COP21 and COP23, suggesting the need for enhanced efforts. Interestingly, more recent COPs (i.e., COP24 and COP25) are characterized by an emerging discussion around youth engagement and activism in the climate cause, indicating a growing recognition of the importance of youth in driving climate action. Besides, we noticed an increased demand for better information (from COP23 to COP25), as a tool for increasing awareness and bearing to broader climate action. Finally, the latest COPs are characterized by issue-specific discussions, such as the role of science and the influence of politics in climate issues (COP25), or the demand for reducing aviation emissions and making the cryptocurrency ecosystem more sustainable (COP26).

8.4.4 Affective Computing

In order to analyze the variation in public opinion at COPs, we divided our corpus of tweets into three splits: (i) one week before the event, (ii) the dates of the event, (iii) one week after the event. Such a subdivision, in fact, allowed us to assess any pre-COP expectation, as well as immediate reactions due to decisions made during the events.

Using the VADER sentiment analysis tool, we assigned a score $s \in [-1, 1]$ to each tweet of the corresponding time slice. We then used a threshold-based categorization to label each tweet as follows: *extremely negative* ($s \leq -0.75$), *negative* ($s \in (-0.75, -0.15)$), *neutral* ($s \in [-0.15, 0.15]$), *positive* ($s \in (0.15, 0.75)$), and *extremely positive* ($s \geq -0.75$). Figure 8.2 illustrates the percentage values of sentiment, categorized according to the five classes, for each time slice of each COP.

The plot reveals a certain temporal consistency of the neutral component over time, as well as during the time slices that characterize each COP. Extremely negative and positive components account for a reduced fraction of tweets, although they provide valuable insights. As for negative components, we noticed spikes — also in terms of extreme ones — during the latest COPs, which may suggest increasing awareness of the climate crisis and the catastrophic effects it may have (and is already having) on the planet and our society. Besides, negativity tends to increase right after most conferences, suggesting that they actually stir things up. An upward trend is observed for post-conference positive components between COP21 and COP24, which also generated some spikes in extremely positive ones. Overall, we spotted that positive sentiments account for a higher proportion than negative ones, hinting at trust mechanisms w.r.t. events in which the climate crisis is addressed by prominent personalities with decision-making power. Nonetheless, such an observed positive trend reverses starting from COP25, achieving the highest degrees of negativity and the lowest of positivity, including extreme components. This anomaly is all but unexpected, as COP25 represents the most protracted and challenging climate conference, characterized by the inability to achieve a binding agreement and to fulfill the Paris Agreement for the subsequent year, coupled with a new record in levels of greenhouse gases — impeding heat dissipation within the atmosphere.

We gained a deeper understanding of such findings by extracting emotions conveyed by discussions around the climate debate, through EmoAtlas and Plutchik's wheel visualization. Specifically, we want to characterize the extremely negative and positive debate for the three observation periods related to each COP. For instance, Figure 8.3-left illustrates the breakdown of the extremely negative sentiment corresponding to the post-COP25 period, showing *anger*, *disgust* and *sadness* as predominant emotions; by contrast, Figure 8.3-right shows that the pre-COP26 period is characterized by an extremely positive sentiment, given the high scores associated with *joy*, *trust* and *anticipation*. Overall, for the extremely positive components, we point out the high values of *joy* compared to the other scores, as it exhibits peaks during the conference periods when the social debate is most intense, whereas *surprise* decreases in the period following the conferences; also, *anticipation* increases during the conference period compared to the previous week but decreases in the following week, with a significant drop observed after COP23. Consistently higher values of *trust* are reported after the conferences, except for COP26, due to exceptionally high initial values. For the extremely negative component, *anger* and *fear* consistently decrease in the period following the conferences. COP26 is also the only event to exhibit decreasing values of *disgust* across the three periods.

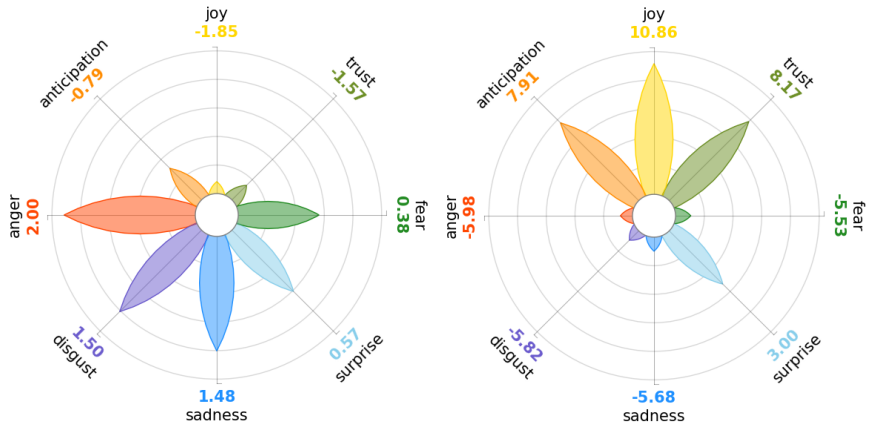


Fig. 8.3. Plutchik's wheel of emotions displaying the 8 basic emotions for the extr. negative component of sentiment after COP25 (left) and for the extr. positive component of sentiment before COP26 (right).

8.5 Chapter review

In this work, we contributed to the understanding of the social debate on climate change by exploiting the discussion surrounding seven years of the Conference of the Parties on Twitter as a proxy for the physical perception of the issue. Graph mining and text mining techniques have been proven helpful tools for exploiting the social traces people left on social platforms, providing valuable insights that might be used to make better decisions, prioritize actions, and foster more effective communication for the urgent challenge of climate change. Future work may include the use of real-time data that further benefit from people's proven ability to act as social sensors in the way they react to emergent phenomena on social, as well as the integration of additional techniques and data sources to provide a comprehensive overview of the climate change perception.

Part III

Multimodal Representation Learning for the Web3

Overview

This part of this thesis work delves into *multimodal representation learning for the Web3*. It comprises five chapters, which are organized as follows:

Chapter 9. This chapter introduces the reader to the landscape of Non-Fungible Tokens (NFTs) and the Web3 through an exploration of the main features associated with this intriguing domain and provides a summary of the key scientific contributions related to it.

Chapter 10. This chapter presents `MERLIN`, an innovative multimodal deep learning framework designed to train Transformer-based language and visual models, as well as graph neural network models, on collections of NFT images and texts. The framework is conceived to address the task of NFT financial performance prediction. Experimental results will demonstrate its ability to achieve remarkable performances according to different financial criteria.

Chapter 11. In this chapter, we leverage Vision Transformers and graph-based modeling to delve into visual inspiration phenomena between NFTs. We unveil the main structural trait shaping visual inspiration networks, and how visual inspiration is interrelated with asset performances. Additionally, we will discuss an explainability module aimed at elucidating our detected inspirations.

Chapter 12. This chapter introduces SONAR, an innovative web-based tool for multimodal exploration of NFT inspiration networks. Our proposed tool provides an interactive visualization of the inspiration-driven connections among NFTs, at both individual level and collection level. The capabilities of SONAR will be assessed through its application on the most widely used NFT dataset to date.

Chapter 13. In this chapter, we introduce our curated dataset of NFT transactions and metadata sourced from OpenSea, the leading NFT marketplace. This dataset stands as the largest and most relevant to date, encompassing more than 70 million transactions. It boasts a wealth of metadata, thus being suitable for various tasks relevant to different research fields.

9

Background and Related Work

Non-Fungible Tokens, commonly shortened as *NFTs*, represent today one of the most fashionable applications of the blockchain technology, as well as the trailblazers for the advent of the *Web3*. Governed by smart contracts, i.e., clauses encoded in programming language that can be deployed using cryptographically signed transactions on the blockchain, NFTs represent information stored on the blockchain that certify the uniqueness of digital assets. Being traceable, NFTs enable inspecting the history of a given asset, from its creation (or *minting*) to all its owners, up to the present, thus aiming at revolutionizing the way we trade any asset that can be digitally represented.

The potential shown by NFTs immediately attracted many digital creators and corporations, interested in certifying assets of various types, from images to in-game objects, video and audio contents. All of that suddenly translated in a fervor of investments that peaked more than $2 billion USD traded in the first quarter of 2021,[1] making NFTs a global-scale phenomenon.

Indeed, such an enthusiasm led to events never seen before in the art landscape,[2] such as cute kittens (i.e., *CryptoKitties*) capable of congesting the Ethereum network,[3] the third highest price of a living artist in an auction (i.e., Beeple's Everydays: The First 5000 Days, for $69.3 Million),[4] automatically generated pixel artworks (i.e., *CryptoPunks*) selling for nearly $24 Million,[5] or the sale of the first Tweet for over $2.9 Million.[6]

The effects of such a disruptive technology are already tangible or under study in several domains [90], such as art and collectibles [136], game development [84, 124], healthcare

[1] https://nonfungible.com/reports/2021/en/q1-quarterly-nft-market-report

[2] https://decrypt.co/62898/most-expensive-nfts-ever-sold

[3] https://qz.com/1145833/cryptokitties-is-causing-ethereum-network-congestion/

[4] https://www.christies.com/about-us/press-archive/details?PressReleaseID=9970

[5] https://decrypt.co/92819/cryptopunks-ethereum-nft-sells-for-nearly-24m-doubling-previous-record

[6] https://www.cnbc.com/2021/03/22/jack-dorsey-sells-his-first-tweet-ever-as-an-nft-for-over-2point9-million.html

[135], preservation of cultural heritage [219].[7] Furthermore, NFTs represent the best-in-case technology today for the development of the *Metaverse* [226, 229, 185, 100], acting as deeds of ownership for virtual lands or as forerunners of digital fashion [120].

The skyrocketing of NFTs in 2021 attracted the attention of numerous research groups intent on unraveling opportunities and challenges manifested by an infant yet impactful technology in the blockchain domain [228, 168, 200, 196, 177, 222, 70, 69, 160, 123, 225, 12, 13, 111, 54, 163, 97, 11, 244, 83, 100, 99, 193, 174, 33, 231, 98, 109, 56, 137], which served as a forerunner for the Web3.

Wang et al. [228] describes technical components, protocols, standards, and desired properties for the state-of-the-art NFT solutions. The seminal work from Nadini et al. [168] represents to date one of the most interesting research contributions on NFTs. Based on about 6M transactions concerning nearly 5M NFTs collected between 2017 and 2021 from the Ethereum and WAX blockchains, the authors unveiled that most traders specialize in particular collections, NFTs in a collection tend to be visually homogeneous, and their visual features can improve the price predictability of NFTs compared to using the transactions history alone. Recently, Guidi et al. [99] have investigated the evolution of the NFT concept toward new features and capabilities, paving the way for the so-called "NFTs 2.0".

Many studies have focused on one major challenge involving the NFT ecosystem, namely identifying useful predictors for NFT financial performance, leveraging either intrinsic or extrinsic features. As for the former, Mekacher et al. [160] analyzed 3.7M transactions collected between 2018 and 2022 for 1.4M NFTs and more than 400 collections to investigate how visual attribute rarity shapes financial performances. The heterogeneous distribution of rarity across most collections is shown to affect selling prices and asset stability, whereas utility-based features (e.g., attack/defense scores of specific traits) improve rarity-based predictions, as found by Ho et al. [111] for the *Axies* play-to-earn gaming NFTs. Recently, He et al. [109] focused on the generation of profitable NFT images from user-input texts.

As concerns external influence factors, although cryptos might provide some hints at the pricing of NFTs [70, 177, 12, 13, 11], social media act as a tangible influence factor for NFT financial performances [244]. In this regard, Kapoor et al. [123] used OpenSea and Twitter data to leverage the linkage between crypto assets and social phenomena for an NFT pricing prediction task, showing that social information allows achieving a boost up to 6% in terms of accuracy w.r.t. models exploiting just NFT-related features.

It should however be noted that, although [160, 111], resp. [123], are the first to deal with intrinsic, resp. extrinsic, factors on the NFT pricing, the visual/utility features exploited by the former approaches are not learned and correspond to simple descriptive metadata stored in publicly available platforms;[8,9] moreover, the latter approach is limited to the OpenSea market and relies on pseudo-financial features, along with information from social media, to address the prediction task according to hand-crafted classes.

Given the highly decentralized and not easy-to-regulate nature of NFTs, moreover, several works have focused on the investigation of the main vulnerabilities [97, 98], frauds [193], and anomalous activities [174], e.g., wash trading [33, 231], in such a scenario.

Network-based approaches represent another growing branch of research within the NFT landscape. Vasan et al. [222] studied the Foundation platform from a network analysis perspective, based on more than 48k NFTs listed by over 15k artists. In that study, the order of

[7] https://cointelegraph.com/news/the-world-s-cultural-heritage-is-being-preserved-one-nft-at-a-time

[8] https://rarity.tools/

[9] https://nonfungible.com/

landing on the platforms of artists and investors was found to affect the chances of earning and spending, respectively. Also, in contrast to what is commonly observed for "traditional" art, fluctuations in asset prices are detected for the same creator; nonetheless, these are found to be in a stable range, which also determines the creator reputation. In addition, ties between artists and collectors are crucial to developing a dense network of investment that endures over time. By analyzing the seller-buyer networks of about 40k sales, Colavizza [54] unveiled that the NFT market is driven by a small set of sellers and (an even smaller) set of buyers. Besides, they reported that preferential buyer-seller ties characterize the growth of the market, and more interestingly, ties persist even during dips, thus becoming a footprint of the NFT landscape growth.

10

Show me your NFT and I tell you how it will perform: Multimodal representation learning for NFT selling price prediction

Summary. In the spotlight after skyrocketing in 2021, NFTs have attracted the attention of crypto enthusiasts and investors intent on placing promising investments in this profitable market. However, the NFT financial performance prediction has not been widely explored to date. In this work, we address the above problem based on the hypothesis that NFT images and their textual descriptions are essential proxies to predict the NFT selling prices. To this purpose, we propose MERLIN, a novel multimodal deep learning framework designed to train Transformer-based language and visual models, along with graph neural network models, on collections of NFTs' images and texts. A key aspect in MERLIN is its independence on financial features, as it exploits only the primary data a user interested in NFT trading would like to deal with, i.e., NFT images and textual descriptions. By learning dense representations of such data, a price-category classification task is performed by MERLIN models, which can also be tuned according to user preferences in the inference phase to mimic different risk-return investment profiles. Experimental evaluation on a publicly available dataset has shown that MERLIN models achieve significant performances according to several financial assessment criteria, fostering profitable investments, and also beating baseline machine-learning classifiers based on financial features.

10.1 Contributions

In this work we address the following problem: based on the key assumption that a (possibly non-expert) user interested in trading NFTs looks at their raw contents as primary source to deal with, given an NFT image and associated textual description, we want to predict the NFT financial performance in terms of selling price.

To this purpose, we propose a novel bimodal deep-learning framework that is designed to train Transformer-based language models and visual models, along with graph neural network models, on collections of NFT images and texts. The objective is to learn dense representations from the NFT images and texts, upon which a price classification task is carried out. To accomplish this, our approach remains independent from any NFT financial feature engineering task, as it just requires to know the categories of selling prices for the NFTs in input to the training phase. Yet, in the inference phase, our framework can make flexible predictions according to different user's preferences reflecting a risk-return investment principle.

We believe that this perspective on the NFT performance prediction problem can offer a number of advantages: (i) it corresponds to a more natural statement of the problem which only

requires the context of the input data, (ii) it avoids depending on feature engineering tasks, i.e., the selection of financial explanatory variables for the prediction task, and (iii) by discarding any requirements in terms of prior knowledge on the financial domain (i.e., NFT market), it would better support newbies in NFT trading.

By means of the proposed framework, we aim at answering the following research questions:

(RQ1) *Catch'em all* — Can we learn a model capable of suggesting profitable investments in NFT trading?

(RQ2) *Take it easy* — Can we answer **RQ1** by just looking at intrinsic contents of NFTs, i.e., images and textual descriptions, thus discarding any requirements in terms of financial indicators?

(RQ3) *Attention is all you want* — Is it more valuable to be a "good artist" or a "good writer"? i.e., What is the impact on the NFT performance prediction by the image, its textual description, and an attentive combination of both?

(RQ4) *Be aware of your neighbors* — What is the effect of introducing contextual awareness learned from similarity-induced aggregate information?

10.2 Scope and limitations

To the best of our knowledge, this work is the first to propose a deep-learning financial-agnostic solution to the problem of NFT price-category prediction. The expected impact of our study is hence twofold: raising the bar to new horizons of NFT financial performance prediction, and gaining insights into influencing factors for the value of digital assets in the Web3.

It should be noted that our experimental findings derive from a publicly available dataset, which is, to date, the most used one for NFT transactions. Although it clearly cannot provide a full picture of the NFT trading realm, we nonetheless point out that our results can reasonably be considered generalizable, due to the high coverage and diversification of the data we used, which (i) include NFTs from heterogeneous platforms (i.e., designed for different scopes), and (ii) cover the NFT trading history up to the steady-state of mid-2021, thus allowing us to catch the main consolidated patterns appeared in the market; in this regard, extreme bull-/bear-driven events of 2022 were left out of our evaluation. Also, our findings are not aware of external sources of influence (e.g., from social media) on the NFT market, neither it considers the history of NFT prices to account for trend analysis; nonetheless, while such aspects are definitely worthy of investigation in future research, the focus of our study is deliberately on information within everyone's reach, thus not requiring any financial domain experience.

10.3 Problem Definition

We formulate the NFT performance prediction problem as follows. Given a collection of N NFT data objects, where each object is a pair consisting of an *image* and a *text*, which corresponds to a description for the image, we assume that each NFT is associated with one or more selling prices, over which we take the mean value. From the distribution of the N NFT average-prices, we derive quantile-based intervals to define a set C of *NFT price categories*.

The NFT performance prediction problem is formulated as a classification problem, where the goal is to predict the price category of a previously unseen NFT. This is accomplished

according to learned dense representations of NFT images and texts, which are agnostic of financial indicators on the NFT market. To learn the prediction model, training instances correspond to the individual data objects, with ground-truth label y_i associated to the i-th training instance corresponding to the category $C \in \mathcal{C}$ that encloses the average-price of the NFT instance. The objective of the task is to minimize the cross-entropy loss function:

$$\mathcal{L} = - \sum_{i=1..N, C \in \mathcal{C}} y_{i,C} \ln \hat{y}_{i,C}, \tag{10.1}$$

where $y_{i,C}$ is 1 if the i-th data object actually belongs to class C, 0 otherwise, and $\hat{y}_{i,C}$ is the prediction for the i-th data object w.r.t. C from a probability distribution matrix $\widehat{\mathbf{Y}} \in \mathbb{R}^{N \times |\mathcal{C}|}$.

10.4 The MERLIN framework

We describe our proposed framework named MERLIN - *Multimodal rEpResentation LearnIng for NFT performance prediction*. We first present design requirements in Section 10.4.1 and an overview of the framework in Section 10.4.2. In Sections 10.4.3 and 10.4.4, we detail the MERLIN stages and its constituting components.

10.4.1 Design requirements

To pursue the goal of predictive modeling in the NFT financial domain (**RQ1**), our architectural choices are naturally fit by deeply contextualized learning models, as we do not want to rely on (financial) feature engineering tasks (**RQ2**). In this respect, and further motivated by the opportunity of learning dense representations of both images and texts for the task at hand, MERLIN exploits the "de-facto" standard in NLP and computer vision, i.e., *Transformer-based pre-trained models*, as well as in graph representation learning, based on *graph neural network* (GNN) models. Indeed, such architectures match our requirement of avoiding manual or domain-driven selection of prominent features and, by relying on the so-called *attention mechanism* [223], they allow lending more significant weights to features detected as more relevant to the task at hand (**RQ3**).

Transformers were originally defined to model language semantics and non-linear relationships between terms, similarly to sophisticated recurrent and convolutional neural networks; however, by employing bidirectional self-supervised training and an attention mechanism that learns contextual relations between (sub-)words in a text, Transformers are much more effective in capturing subtle and complex lexical patterns, including the sequential structure and long-term dependencies, thus obtaining the most comprehensive local and global feature representations of a text sequence [66, 149, 182]. Analogous considerations apply to the visual component of our input (the NFT images), for which we resort to Vision Transformers (ViT) that have recently emerged in computer vision, proposed by Google [68] and Facebook [41]. While adapting the input representation approach used in NLP to images (i.e., image patches are treated as (sub)-words), this type of Transformers has also shown to benefit from the availability of benchmarks, such as ImageNet, to incorporate supervised learning during pre-training.

Moreover, MERLIN utilizes a GNN in order to model the node relations through a message passing scheme to learn the neighborhood importance of each node [241, 237]. Unlike random-walk-based approaches [176, 93], which consider only nodes co-occurring in a random walk and optimize the embeddings to encode random walk statistics, GNN carries out an aggregation scheme by which each node iteratively combines the neighbors and its own features to obtain

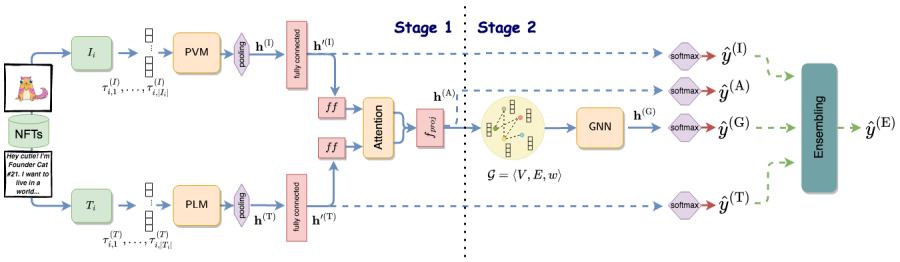


Fig. 10.1. Illustration of the MERLIN framework. Dashed lines refer to pipeline flows alternative to the main flow (solid line)

a new representation. In MERLIN, this allows us to further enhance the learned representations of NFTs in terms of contextual awareness based on similarity search (**RQ4**).

10.4.2 Overview

We propose a flexible two-stage pipeline, as shown in Figure 10.1.

• Stage 1 includes the image and text learning modules, which are aimed to produce dense representations, or embeddings, for the visual and textual components of NFTs, respectively. Note that, being multilayer Transformer encoder models, they have similar architecture (e.g., input representation as a combination of token embeddings, position embeddings, segment embeddings), although they might use different pretraining objectives. As the two learners work on separate modalities, the image and text embeddings are then combined together and fed into an attention module to fuse the visual features and the lexical/semantic features.

• Stage 2 starts from the fused visual/text embeddings to build an NFT-set similarity graph, where each node is a group of NFTs that is cohesive by average-price and each edge expresses similarity according to the visual and text features. By treating an aggregation of the fused visual/text embeddings as initial features of the nodes, a GNN module is in charge of learning NFT-set dense representations. The neighborhood aggregations learned from the GNN act as an underlying similarity search context, whose results are eventually leveraged for the downstream classification task.

Note that MERLIN is designed to provide alternative points of prediction. In particular, the embeddings individually learned from the PLM and the PVM, the embeddings learned from the attention module, and the embeddings learned from the GNN module can be fed to a $softmax$ that yields the prediction probabilities $\hat{\mathbf{y}}$ in an *alternative* or *complementary* fashion; in the latter case, a multi-view aggregation mechanism (i.e., ensemble) is carried out on all the individual predictions to produce the final prediction.

10.4.3 Stage 1

Representing input data. We are given a collection $\mathcal{D}$ of N NFT data, where each data object $D \in \mathcal{D}$ is a pair consisting of an *image*, denoted as I, and a *text*, denoted as T, which corresponds to a description for the image. Both images and texts are initially represented as token sequences, i.e., for any $D_i = \langle I_i, T_i \rangle \in \mathcal{D}$, $I_i = [\tau_{i,1}^{(\mathrm{I})}, \ldots, \tau_{i,|I_i|}^{(\mathrm{I})}]$ and $T_i = [\tau_{i,1}^{(\mathrm{T})}, \ldots, \tau_{i,|T_i|}^{(\mathrm{T})}]$, where $\tau_{i,j}$ symbols denote the j-th tokens of D_i and superscripts (I), (T) are used to denote image and text tokens, respectively.

Clearly, the tokenizers of the two components work on different raw data, and hence the tokens assume different meanings in the two cases: each text is tokenized to yield word pieces (usually to minimize the number of unsegmented words in the text) [66], while each image is split into a sequence of fixed-size non-overlapping patches, which are then linearly embedded [68].

Learning from texts. Given an NFT D_i, MERLIN employs a Transformer-based pre-trained language model, for short PLM, to deeply contextualize the token sequence of the textual component T_i by mapping it onto a space of dimension $d^{(\mathrm{T})}$, i.e., $\mathsf{PLM}(T_i) \in \mathcal{R}^{|T_i| \times d^{(\mathrm{T})}}$. A pooling function $pooling(\cdot)$ is then applied to the token embeddings to output a single embedding $\mathbf{h}_i^{(\mathrm{T})}$ of size $d^{(\mathrm{T})}$: $\mathbf{h}_i^{(\mathrm{T})} = pooling(\mathsf{PLM}(T_i))$. Following BERT and related PLMs, a default choice for the pooling function is to output the embedding of the special token [CLS], which is considered to be representative of the whole input text.

Learning from images. The image learning module shares with the textual one the architectural choices based on Transformer. Indeed, given a pre-trained visual model, for short PVM, the goal is to deeply encode the token sequence of the visual component I_i into a space of dimension $d^{(\mathrm{I})}$, i.e., $\mathsf{PVM}(I_i) \in \mathcal{R}^{|I_i| \times d^{(\mathrm{I})}}$. Analogously to the text learning component, the token embeddings are subject to a pooling function to yield a single embedding $\mathbf{h}_i^{(\mathrm{I})}$ of size $d^{(\mathrm{I})}$: $\mathbf{h}_i^{(\mathrm{I})} = pooling(\mathsf{PVM}(I_i))$. Also for PVM, a [CLS] token is added so that its output embedding can be pooled to serve as representation of an entire image.

Fusing the representations. To obtain a fused fixed-size dense representation of text and image for each NFT (**RQ3**), we need to jointly encode $\mathbf{h}_i^{(\mathrm{T})}$ and $\mathbf{h}_i^{(\mathrm{I})}$ to a common space. First, we use two fully-connected neural networks to map the two embeddings into a smaller space, $d'^{(\mathrm{T})}$ and $d'^{(\mathrm{I})}$, respectively. On the resulting embeddings, $\mathbf{h}_i{}'^{(\mathrm{T})}$ and $\mathbf{h}_i{}'^{(\mathrm{I})}$, a *self-attention* mechanism is carried out. A feed-forward neural network ff computes energy scores for $\mathbf{h}_i{}'^{(\mathrm{T})}$ and $\mathbf{h}_i{}'^{(\mathrm{I})}$:

$$e_i^{(\mathrm{I})} = ff(\mathbf{h}_i{}'^{(\mathrm{I})}), \quad e_i^{(\mathrm{T})} = ff(\mathbf{h}_i{}'^{(\mathrm{T})}), \tag{10.2}$$

on which a $softmax$ operator is applied to obtain a probability distribution over the two attention energies:

$$\alpha_i^{(\mathrm{I})} = \frac{\exp(e_i^{(\mathrm{I})})}{\exp(e_i^{(\mathrm{I})}) + \exp(e_i^{(\mathrm{T})})}, \quad \alpha_i^{(\mathrm{T})} = \frac{\exp(e_i^{(\mathrm{I})})}{\exp(e_i^{(\mathrm{I})}) + \exp(e_i^{(\mathrm{T})})}. \tag{10.3}$$

Eventually, we scale our latent representations $\mathbf{h}_i{}'^{(\mathrm{T})}$ and $\mathbf{h}_i{}'^{(\mathrm{I})}$ by the corresponding energies, to obtain the following contexts:

$$\mathbf{c}_i^{(\mathrm{I})} = \alpha_i^{(\mathrm{I})} \cdot \mathbf{h}_i{}'^{(\mathrm{I})}, \quad \mathbf{c}_i(\mathrm{T}) = \alpha_i^{(\mathrm{T})} \cdot \mathbf{h}._i{}'^{(\mathrm{T})} \tag{10.4}$$

These are then provided to another feed-forward neural network ff_{proj}, which projects them onto another space of dimension $d^{(\mathrm{A})}$ (where superscript (A) stands for "attention"):

$$\mathbf{h}_i^{(\mathrm{A})} = ff_{proj}\left(\mathbf{c}_i^{(\mathrm{I})} \oplus \mathbf{c}_i^{(\mathrm{T})}\right), \tag{10.5}$$

where $\oplus$ denotes the concatenation operator.

10.4.4 Stage 2

Modeling an NFT similarity context. The second stage of `MERLIN` is aimed to enhance the contextual awareness among NFTs by learning new representations based on a similarity-search context built upon the current representations. We define an undirected graph $\mathcal{G} = \langle V, E, w \rangle$, where the node-set V is a partition of $\mathcal{D}$ into n cohesive groups, E is the edge-set expressing similarity relations between nodes (i.e., NFT groups) and $w : E \mapsto \mathbb{R}$ is a weighting function to compute node similarity. The node-set is specified by first ordering the N NFTs by average-price, then forming n groups (i.e., nodes) of almost equal size $l = N/n$ such that each node $v \in V$ contains NFTs having average-prices close within a certain interval of varying length. Moreover, each node is associated with a class label that corresponds to the average-price category.

Given $v \in V$ and an integer $k > 0$, an edge is drawn from v to each of its k most similar nodes, with edge weight defined as:

$$w(v,u) = \cos\left(\textit{n-pooling}(\{\mathbf{h}_i^{(\mathrm{A})}\}_{D_i \in v}), \textit{n-pooling}(\{\mathbf{h}_j^{(\mathrm{A})}\}_{D_j \in u}\right),$$

where $\cos(\cdot)$ denotes the cosine similarity function, and $\textit{n-pooling}(\cdot)$ is a pooling function on the set of NFTs belonging to the same node; by default, we define $\textit{n-pooling}(\cdot)$ as the average over the NFT embeddings learned at the end of Stage 1.

Learning from the graph of NFT-set similarities. The next step of Stage 2 is to provide the graph $\mathcal{G}$ in input to a GNN module. We denote with $\mathbf{X} \in \mathbb{R}^{|V| \times d^{(\mathrm{A})}}$ the initial-feature (or attribute) matrix associated with the nodes in $\mathcal{G}$, i.e., the embeddings produced by the *n-pooling* function. The goal is to learn new node-features $\mathbf{h}^{(\mathrm{G})}$ in a latent space of dimension $d^{(\mathrm{G})}$ (where superscript (G) is for "graph") modeling the relations between cohesive groups of NFTs.

A particularly suitable GNN for our setting is the Graph Attention Network (GAT) [224, 37]. Unlike a graph convolutional network (GCN) that assigns predetermined weights to the neighbors of a node, a GAT learns the weights through a self-attention mechanism in order to capture the importance of different neighbors; more precisely, GAT modifies the aggregation process of GCN by learning the strength of the link between neighboring nodes through self-attention [237]. Formally, the importance of a node v's features w.r.t. node u is computed through the attention coefficients as follows:

$$e_{uv} = a(\mathbf{W}\mathbf{x}_u, \mathbf{W}\mathbf{x}_v) \tag{10.6}$$

where $\mathbf{W} \in \mathbb{R}^{d^{(\mathrm{G})} \times d^{(\mathrm{A})}}$ is a trainable weight matrix. The attention mechanism is performed by a feed-forward neural network exploiting the $LeakyReLU$ non-linearity function:

$$\alpha_{uv} = \frac{\exp(LeakyReLU(e_{uv}))}{\sum_{k \in \mathcal{N}_u} \exp(LeakyReLU(e_{uk}))}, \tag{10.7}$$

where $\mathcal{N}_u$ denotes the set of neighbors of node u. At each step, each node u updates its hidden state, denoted as $\mathbf{z}_u$, by aggregating the features of its neighbors as follows:

$$\mathbf{z}_u = \sigma\left(\frac{1}{Q}\sum_{q=1}^{Q}\sum_{v \in \mathcal{N}_u} \alpha_{uv,q}\mathbf{W}_q\mathbf{x}_v\right), \tag{10.8}$$

where Q denotes the number of independent attention mechanisms (heads) [223], $\alpha_{uv,q}$ and $\mathbf{W}_q$ are the normalized attention coefficients and weight matrix corresponding to the q-th head [224], and $\sigma(\cdot)$ is the $ReLU$ activation function.

10.5 Experimental methodology

Data. We resorted to the most used and publicly available dataset on NFT purchase transactions [168], which includes sales from the *CryptoKitties*, *Gods-Unchained*, *Decentraland*, *OpenSea*, and *Atomic* markets. It contains 6.1M transactions involving 4.7M NFTs, spanning across more than 4k collections, which were grouped by the creators into six main categories (in parenthesis, we report the coverage percentage): *Art* (18.46%), *Collectible* (28.85%), *Games* (47.21%), *Metaverse* (0.1%), *Utility* (0.17%), and *Other* (5.21%).

Each NFT is associated with an image, a text description, and the selling prices, which were used to build our training instances. Besides ensuring to filter out NFTs with missing image or text, we also chose to select NFTs having at least a *secondary sell*. This would avoid us incurring such latent patterns as price boosting mechanisms between authors, whose investigation is beyond the objectives of this work. We thus came up with 202,257 NFTs having images (downloaded from the corresponding URLs in the dataset) and descriptions to be used in our experimental evaluation.

As concerns the price categories to be used as class labels for our training data (C), we examined the distribution of average-prices and decided to define 3 classes corresponding to the first quarter (i.e., average-prices up to the first quartile), the union of second and third quarter (i.e., average-prices between the first quartile and the third quartile), and the fourth quarter (i.e., average-prices above the third quartile). For short, we hereinafter refer to the 3 classes as Low, Mid, and $High$, respectively.

We split the dataset into training and validation sets of size 90% and 10%, respectively, by keeping equal per-class distribution w.r.t. the whole dataset (cf. ***Appendix A.4***). Also, note that reproducibility of data samples and results is ensured since we fixed the seed-sets for our randomness-related operations.

Assessment criteria and goals. To evaluate our models, we used a number of statistical criteria derived from each confusion matrix outputted by the models (Table 10.1). These criteria include both global performance measures and more specific performance measures associated with the financial task at hand. The former group contains *accuracy* (A), and weighted macro-averaged *precision* (P), *recall* (R), and *F-score* ($F1$). (We used `scikit-learn` implementations.) The latter group focuses on criteria that have clear meanings in terms of trading losses and gains, particularly relating to the most *profitable* (i.e., with a positive-return) investments. This led us to derive the following criteria from the portion of the confusion matrix involving class $High$:

- *win rate* (WR), which is defined as in Eq. 10.9, i.e., the fraction of predicted most-profitable trades that are actually most-profitable (which is also equivalent to the precision for class $High$);
- *win-loss ratio* (WLR), which is defined as in Eq. 10.10, i.e., the ratio between most-profitable wins and losses;
- *loss rate* (LR), which is defined as in Eq. 10.11, i.e., the fraction of wrongly predicted most-profitable trades (which is also equivalent to the false discovery rate for class $High$);
- *missed opportunity rate* (MR), which is defined as in Eq. 10.13, i.e., the fraction of true most-profitable trades that are incorrectly predicted (which is also equivalent to the false negative rate for class $High$);
- *cautiousness* (Cn), which is defined as in Eq. 10.12, i.e., the ratio of missed opportunities to the total number of errors for class $High$;
- *riskiness* (Rn), which is defined as in Eq. 10.14, i.e., the ratio of wrong predictions to the total number of errors for class $High$ (which is also equivalent to $1 - Cn$).

Table 10.1. Confusion matrix for our 3-class classification task (top) and domain-specific criteria (bottom)

		True Label		
		Low	*Mid*	*High*
Pred. Label	*Low*	$C_{L,L}$	$C_{L,M}$	$C_{L,H}$
	Mid	$C_{M,L}$	$C_{M,M}$	$C_{M,H}$
	High	$C_{H,L}$	$C_{H,M}$	$C_{H,H}$

criterion	definition	criterion	definition
win rate	$WR = \frac{C_{H,H}}{C_{H,L} + C_{H,M} + C_{H,H}}$ (10.9)	**win/ loss ratio**	$WLR = WR/LR$ (10.10)
loss rate	$LR = \frac{C_{H,L} + C_{H,M}}{C_{H,L} + C_{H,M} + C_{H,H}}$ (10.11)	**cautious-ness**	$Cn = \frac{MR}{MR + LR}$ (10.12)
missed opport. rate	$MR = \frac{C_{L,H} + C_{M,H}}{C_{L,H} + C_{M,H} + C_{H,H}}$ (10.13)	**risk-ness**	$Rn = \frac{LR}{MR + LR}$ (10.14)

The above group of statistics, hereinafter referred to as *High-driven performance criteria*, represents a proxy for real-life financial performances. In this regard, note that the win and loss rates are complementary to each other, and determine the win-loss rate, which is strictly related to the *return on investment* (ROI), i.e., the effectiveness of investment choices. Furthermore, we can shape investments on different risk-profiles by looking at the missed opportunity rate; indeed, by identifying "untaken" chances (i.e., $C_{L,H}$, $C_{M,H}$), it can be used to understand whether the model was risky or cautious on its predictions.

Models and Settings. We tested different pre-trained Transformer-based models for both image and text modules. Specifically, we included in our evaluation BERT-base-uncased [66], XML-RoBERTa [149], and S-BERT (all-MiniLM-L6-v2) [182] as PLM, and ViT-base (patch-16-224) [68] and DINO (vitb8) [41] as PVM, which are available in the *HuggingFace* model repository. In all cases, but S-BERT, we set 12 Transformer encoder layers, with hidden size $d^{(\mathrm{I})} = d^{(\mathrm{T})}$ equal to the model default of 768, and 12 attention-heads; as for S-BERT, we used the default 6-layer model. Moreover, we used the [CLS] token embedding as pooling function. Note that PLM and PVM were subject to a fine-tuning stage of training in order to adapt them to our NFT prediction task. Hereinafter, unless otherwise specified, PLM and PVM correspond to BERT and ViT, respectively, which revealed to be our preferred models (cf. Section 10.6.3).

As concerns the NFT-set similarity graph settings, we varied both its parameters, i.e., the node size l and the node neighborhood size k; as we shall discuss later, our best choices for l and k are 50 and 10, respectively. The GNN was implemented through the GATv2Conv [37] module available in PyTorch Geometric, it uses two convolutional layers with 4 concatenated attention heads (Q), hidden dimensionality of 16, and dropout probability equal to 0.5. *Please note that investigating the best available PLMs, PVMs, and GNNs for our tasks is beyond the objectives of this work.*

The three linear-projection layers used in `MERLIN` were equipped with $ReLU$ activation function, batch normalization and dropout probability equal to 0.2. The output embedding size was set as $d'^{(\mathrm{I})} = d'^{(\mathrm{T})} = d^{(A)} = 256$.

All models were trained with the *Adam* optimizer and learning rates $1.0E$-5 for both PLM and PVM, and $1.0E$-3 for GAT; the attention module was trained with learning rate $1.0E$-5 when considering PLM and PVM and with $1.0E$-4 with the GAT. We set the number of training

Table 10.2. Competing baseline methods

Baseline	P	R	$F1$	A	$WR\uparrow$	$LR\downarrow$	$WLR\uparrow$	$MR\downarrow$	$Rn\downarrow$	$Cn\uparrow$
ZeroR	0.169	0.333	0.224	**0.506**	na	na	na	na	na	na
Prior	0.336	0.335	0.335	0.381	0.252	0.748	0.338	0.749	**0.500**	**0.500**
SVM	0.426	0.416	0.407	0.419	0.247	0.753	0.327	0.633	0.543	0.457
Logistic	**0.499**	**0.500**	**0.467**	0.461	**0.393**	**0.607**	**0.648**	**0.236**	0.720	0.280

epochs to 10, which is relatively large considering that both PLM and PVM are pre-trained models, and that the attention and GAT modules are initialized with contextualized embeddings. Furthermore, early-stopping was applied, saving the model at a maximum validation win-rate, so as to maximize the expected number of profitable predictions. All reported results correspond to averages over ten runs (i.e., different seeds); we noticed very small standard deviation (e.g., order of $1.0E$-4, for accuracy), hinting at high stability of PLM and PVM and, as a consequence due to its weight initialization based on them, of the GAT in cascade.

Our experiments were carried out on a 56-core Intel(R) Xeon(R) Gold 6258R CPU, with 256GB RAM and two NVIDIA GeForce RTX3090s, OS Ubuntu Linux 22.04 LTS.

10.6 Results

10.6.1 Competing baseline methods

We considered two types of baseline methods to be comparatively evaluated w.r.t. our MERLIN models. The first type includes simple yet feature-agnostic models: ZeroR, which always returns the most-frequent price-category as predicted class (i.e., Mid), and Prior, which samples the predicted class from the true-class distribution over the input data. The second type of competitors refers to machine-learning classifiers, i.e., SVM and logistic regressor, that are trained over data objects represented by a predetermined set of features. This contains two subsets: the one including min, max, avg and std of the selling prices of the collection an NFT belongs to, and the other one including one-hot encodings of the top-25 most-frequently used terms in the NFT descriptions across $\mathcal{D}$.

As reported in Table 10.2, ZeroR sets the worst global performance, while being not applicable for the $High$-driven evaluation. The other feature-agnostic model, Prior, has significantly lower global performance than the machine-learning classifiers, although it behaves comparably to SVM. The logistic regressor is the best competitor according to all criteria but Rn and Cn. In any case, however, all baselines are significantly outperformed by our MERLIN models, as we shall describe in the next sections. In particular, it should be noted that all baselines have a WLR much lower than 1, thus being unable to yield profitable predictions.

10.6.2 Evaluating MERLIN models w.r.t. RQs

Here we discuss the main results achieved by MERLIN models in order to unveil the capability of our proposed framework to answer the RQs stated in the Introduction. Note that the results presented in this section refer to the best performance obtained by the various constituting modules of MERLIN, which are reported in Table 10.3; sensitivity analysis is postponed to Section 10.6.3.

Table 10.3. Summary of results by the best-performing models in MERLIN. First four rows, resp. GAT-based row, refer to training mode, resp. evaluation mode, of the PLM and PVM models. Best values per criterion are in boldface

Model	*Att.*	*P*	*R*	*F*1	*A*	$WR\uparrow$	$LR\downarrow$	$WLR\uparrow$	$MR\downarrow$	$Rn\downarrow$	$Cn\uparrow$
PLM	✗	0.767	0.738	0.720	0.738	0.891	0.109	8.136	0.622	0.150	0.850
PVM	✗	**0.803**	0.800	**0.800**	0.800	0.798	0.202	3.960	0.264	0.433	0.567
PVM+PLM	✗	0.802	**0.801**	**0.800**	**0.801**	0.807	0.193	4.179	0.284	0.405	0.596
PVM+PLM	✓	0.802	0.799	0.798	0.799	0.814	0.186	4.368	0.309	0.376	0.624
GAT ($l=50$) on PVM+PLM	✓	0.773	0.727	0.701	0.727	0.926	0.074	12.571	0.688	0.097	0.903
ensemble	-	0.779	0.726	0.697	0.726	**0.959**	**0.041**	**23.531**	0.704	**0.055**	**0.945**

Answering RQ2. We begin with assessing the individual performances of PLM and PVM. PLM achieves precision, recall, F1, accuracy equal to 0.767, 0.738, 0.720 and 0.738, respectively. Coupling this with the good performance in terms of win-rate (0.891), loss-rate (0.109), and win-loss ratio 8.136, we unveil that the textual descriptions provided by the NFT creators can already serve as valuable information to predict how an NFT will perform. Moreover, the high missed-opportunity rate and the particularly high cautiousness reveal that our fine-tuned PLM tends to be cautious, as it avoids jumping into very risky trading opportunities.

The visual counterpart PVM achieves even higher performance than PLM according to global criteria, which are all not less than 0.800. However, as concerns the $High$-driven performance criteria, we notice a particularly risky model (i.e., minimum MR and Cn), which doubles, resp. halves, the loss rate ($LR = 0.202$), resp. win/loss ratio ($WLR = 3.960$), w.r.t. PLM. This might partly be ascribed to the fact that some visual features could not be exclusive of certain collections. As a result, although visual features are certainly valuable to predict the NFT financial performances, they also might lead to a more hazardous and loss-prone behavior.

Answering RQ3. To understand the beneficial effects from the combination of both visual and text embeddings, we distinguish two cases, depending on whether the prediction was made on top of the attention module or just on top of the concatenation of the (compressed) visual and text embeddings (i.e., $\mathbf{h}'^{(\mathrm{T})}$ and $\mathbf{h}'^{(\mathrm{I})}$), as reported in the 3rd and 4th rows of Table 10.3.

While achieving no particular advantages in both cases compared to the best-performing PVM according to global criteria, we notice improvements in terms of the $High$-driven performance criteria, which are particularly evident when the attention module was trained and used for prediction. Overall, the attention-based combination of visual and textual features leads to more cautious predictive behavior than the PVM-only model ($Cn = 0.624$), by fixing some hazardous predictions ($Rn = 0.433$), presumably due to the role of the attention on highly informative textual patterns; also, the attention-based combination has beneficial effects in terms of missed-opportunity rate, which halves w.r.t. the PLM-only model.

Answering RQ4. Learning from the similarity relations between price-cohesive groups of NFTs relies on two key parameters, namely the node size l and the node neighborhood size k. We focus here on their best-performing settings, whereas the impact of different settings is discussed in Section 10.6.3. Nonetheless, one important remark that stands out by looking at Figure 10.2 is that, regardless of a particular setting of l, a partition of the graph naturally emerges in terms of the three price-categories. Indeed, as shown in the figure, each of the three classes is well-represented by a densely connected subgraph and, at the same time, the three subgraphs are connected by a few yet non-negligible number of links. This implies that the

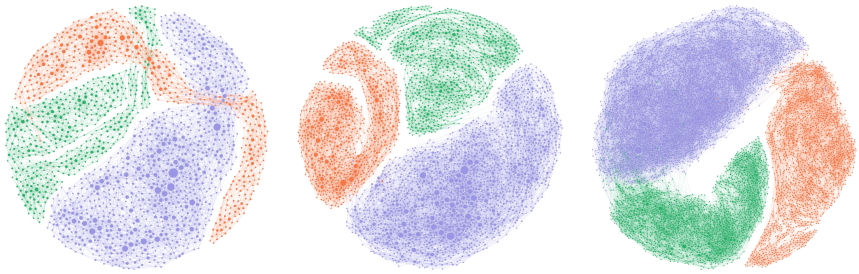


Fig. 10.2. NFT-set similarity graph by varying node size l: 100 (left), 50 (center), 25 (right). Colors correspond to the price categories: orange for Low, purple for Mid, green for $High$. Fruchterman-Reingold layout is used to display the graphs

process of node's neighbor aggregation carried out by the GNN module also involves nodes of different classes.

The best-performing GNN results in Table 10.3 show a striking improvement on financial criteria, by paying just a little in terms of global measures. We notice a remarkably cautious model ($Cn = 0.903$), which peaks up to $WR = 0.926$ and down to $LR = 0.074$, respectively. Accordingly, the score $WLR = 12.571$ is worthy of attention as it increases by 55% versus the best-performing model so far (i.e., the best PLM) with a remarkably better risk-profile.

Ensembling multi-view predictions. A further stage of evaluation concerns the opportunity of leveraging the predictions performed by the constituting components of `MERLIN`, with the goal of providing a more robust prediction according to the $High$-driven criteria. More specifically, rather than aggregating by majority voting or similar criterion, we define a *priority rule*, i.e., a precedence relation among predictors, according to their cautiousness and win-rate: (i) we first consider the most cautious and win-prone model, i.e., GAT, and look at its prediction: if it is not $High$, `MERLIN` returns the GAT outcome as the predicted class; (ii) if it does not hold, we consider the prediction of the second-most cautious model, i.e., the PLM: again, if it is not $High$, `MERLIN` returns the PLM outcome as the predicted class; (iii) otherwise, `MERLIN` entrusts PVM, regardless of the predicted class.

Looking at Table 10.3, we find evidence that supports our initial hypothesis: the devised ensemble predictor is particularly cautious ($Cn = 0.945$) while achieving the best overall results on financial criteria ($WLR = 23.531$), up to doubling, resp. triplicating, the scores of GAT, resp. PLM; an ever sharper improvement when we consider the PVM, where it results seven times better.

10.6.3 Sensitivity analysis

To complement our discussion in Section 10.6.2, we delve into the effect of different choices of the main modules in `MERLIN`. Table 10.4 shows results achieved by various settings for PLM, PVM, and GAT (cf. Section 10.5). Note that, for PLMs and PVMs, their respective prediction results were obtained at inference time.

We notice that XLM-RoBERTa and BERT behave quite similarly to each other, especially in terms of global criteria, and better than S-BERT. Our choice fell on BERT since, although it achieves slightly lower performance in terms of WR and LR, BERT has 40% faster training time w.r.t. RoBERTa according to our early-stopping criterion, while also having less parameters (110M vs. 125M).

Table 10.4. Results achieved by MERLIN models with various settings. Symbol $*$ is used to mark the selected model in each subtable. PLM and PVM models are used in training mode, but evaluation mode when supporting the GAT models

	Model	*Att*	*P*	*R*	*F*1	*A*	*WR* ↑	*LR* ↓	*WLR* ↑	*MR* ↓	*Rn* ↓	*Cn* ↑
PLM	bert-base-uncased *	✗	0.767	0.738	0.720	0.738	0.891	0.109	8.136	0.622	0.150	0.850
	xlm-roberta-base-cased	✗	0.769	0.738	0.719	0.738	0.900	0.100	9.010	0.628	0.137	0.863
	sbert-default	✗	0.740	0.719	0.700	0.719	0.831	0.169	4.909	0.641	0.209	0.791
PVM	vit-base-patch-16-224 *	✗	0.803	0.800	0.800	0.800	0.798	0.202	3.960	0.264	0.433	0.567
	dino-vit	✗	0.791	0.791	0.790	0.791	0.797	0.203	3.919	0.281	0.420	0.580
PLM+ PVM	best PLM, best PVM	✗	0.802	0.801	0.800	0.801	0.807	0.193	4.179	0.284	0.405	0.596
	best PLM, best PVM *	✓	0.802	0.799	0.798	0.799	0.814	0.186	4.368	0.309	0.376	0.624
GAT	$l = 100$, best PLM+PVM	✓	0.771	0.731	0.708	0.731	0.914	0.090	10.560	0.660	0.115	0.885
	$l = 50$, best PLM+PVM *	✓	0.773	0.727	0.701	0.727	0.926	0.074	12.571	0.688	0.097	0.903
	$l = 25$, best PLM+PVM	✓	0.756	0.707	0.669	0.707	0.926	0.074	12.489	0.769	0.088	0.912
	$l = 50$, best PLM+PVM	✗	0.761	0.679	0.621	0.679	0.952	0.048	19.688	0.876	0.052	0.948

We also notice that the PVMs were pre-trained using different approaches: supervised image classification for ViT-base (patch-16-224) [68] and knowledge distillation for DINO (vitb8) [41], albeit they shared the dataset (i.e., ImageNet). The best performer revealed to be ViT, with an increase in performances of 2% w.r.t. DINO.

Regarding the GAT module, the evaluation based on different node-size values (i.e., $l = \{100, 50, 25\}$) evidenced roughly comparable performances, with $l = 50$ and $l = 25$ ensuring the best global and $High$-driven scores. We narrowed our attention to $l = 50$ as the best configuration, since it is better in terms of global criteria and slightly improves in terms of WLR w.r.t. the $l = 25$, while having an inference time noticeably better than the latter (cf. ***Appendix A.4***). We then assessed the impact of the attention component on GAT with $l = 50$, as shown in Table 10.4 (last row). In this regard, although the WLR appears to increase, skipping the attention module negatively affects both types of performance criteria, along with a tendency of the resulting model to become too much cautious ($MR = 0.876$), thus overlooking more promising opportunities.

Also, by increasing k while keeping l fixed to 50, MR tends to decrease along with an increase in Rn, which sets on 0.05, 0.097, and 0.75 for $k = 5$, 10, and 20, respectively. Also, $MR = 0.99$ obtained for $k = 5$ hints at a degenerate model.

10.7 Lessons Learned

Here we summarize the results obtained addressing our initially stated research questions, thus providing the reader with a memorandum for the main lessons learned in this work.

• **RQ1**: Deeply contextualized pre-trained learning models undoubtedly represent the most suited choice when we need to address complex tasks in new domains while abstracting from feature engineering. By adapting pre-trained language and visual models to the NFT domain and relating selling-price prediction task, our MERLIN can leverage such models' capabilities to effectively learn meaningful representations of the raw NFT data, i.e., images and their descriptions, to be exploited for financial prediction.

• **RQ2**: MERLIN learned models have shown ability to provide valuable suggestions in the NFT trading landscape without learning from financial features. By contrast, the visual as well as textual features learned by our models from the NFT images and descriptions can serve as helpful financial proxies. Yet, being trained on NFT data which only require knowledge

on their average-price category as their class label, our models outperform machine-learning classifiers that were trained over financial features.
• **RQ3**: NFT images and texts convey different yet complementary rich contents for the task at hand. When textual descriptions are really informative and discriminative of NFTs within the same collection, they can suggest profitable investments being fairly cautious w.r.t. risky trading moves. Conversely, visual features might lead to more risky yet still profitable plays in the market, presumably due to their non-uniqueness w.r.t. specific collections and/or creators (e.g., certain visual features are used in multiple collections). Therefore, we would say that being a "good writer" pays more than being a "good artist", unless one wants to try the thrill of risk. Furthermore, an attentive combination of textual and visual embeddings refines the capabilities of `MERLIN` in a complementary way. Indeed, we spotted that attention shifts towards visual features when the textual component is not informative (e.g., the same description is used for all NFTs in the *Sorare* collection) and, conversely, it leverages the textual component when visual models result excessively risky in predictions.
• **RQ4**: While consistently improving the prediction performances w.r.t. the textual/visual models and their combination, learning from the graph of NFT-set similarities unveils some unexpected yet remarkable patterns. In particular, the embeddings generated by the GAT are highly effective in detecting price categories also at finer resolution (percentiles), and even collections (cf. ***Appendix A.4***) despite never having seen them during the training. It is also worth emphasizing that a multi-view ensemble approach can complement the best skills of the individual modules in `MERLIN`: a $High$-driven strategy giving precedence among predictors based on a mixture of win-rate and cautiousness, enables a particularly profitable yet reasonably cautious model, which doubles the best individual predictor in terms of financial metrics.

As a final note, remarkable aspects arise from the *explanation* of our `MERLIN` models. In ***Appendix A.4***, we provide interpretation of the predictions yielded by `MERLIN` on different examples.

10.8 Chapter review

We presented `MERLIN`, a deep-learning-based framework for the task of predicting NFT performance (selling average-price) by solely relying on images and descriptions of NFTs. To the best of our knowledge, this is the first work to address the above task.

We can outline a few directions for future research. Besides investigating architectural alternatives to the learning modules for improving the `MERLIN` performance, it would be interesting to incorporate the time dimension (e.g., (re)selling times) into the NFT representation learning. Another line of investigation corresponds to detecting anomalous or adversarial patterns, such as those related to price boosting mechanisms between the NFT creators and traders. Yet, accounting for external sources of influence, such as social media, would introduce new perspectives for the analysis.

11

Exploring the Role of Inspiration in Non-Fungible Tokens

Summary. The fervor for Non-Fungible Tokens (NFTs) attracted countless creators, leading to a Big Bang of digital assets driven by latent or explicit forms of inspiration, as in many creative processes. This work exploits Vision Transformers and graph-based modeling to delve into visual inspiration phenomena between NFTs over the years. Our goals include unveiling the main structural traits that shape visual inspiration networks, exploring the interrelation between visual inspiration and asset performances, investigating crypto influence on inspiration processes, and explaining the inspiration relationships among NFTs. Our findings unveil how the pervasiveness of inspiration led to a temporary saturation of the visual feature space, the impact of the dichotomy between inspiring and inspired NFTs on their financial performance, and an intrinsic self-regulatory mechanism between markets and inspiration waves. Our work can serve as a starting point for gaining a broader view of the evolution of Web3.

11.1 Contributions

To date, several questions still remain unanswered about the visual nature of NFTs and their implications on the market. No work has, in particular, investigated the underlying patterns that can be induced from the raw visual features of the NFTs — especially across the boundary of collections owned by different creators — and how such patterns can have effect on the creation of new NFTs, and hence on the NFT market. By contrast, our proposed study aims to fill a gap in the literature by addressing the above aspects.

A major goal of our work is to leverage visual features learned from NFTs in order to build a suitable network model for capturing latent visual influences that some NFTs can exert on others. In particular, in this work, we originally provide an in-depth structural investigation of the visual influence that can be detected *whenever an NFT appears to be visually close to another that was published earlier in the market*. We call this phenomenon **visual inspiration**, which is conceptualized to be as general enough to also include potential extreme cases, such as copying or even plagiarism. To the best of our knowledge, we are the first to deal with this challenging topic.

Throughout this work, we shall investigate the visual inspiration mechanisms underlying the NFT landscape to date, through answering the following main research questions:

(RQ1) *Caught in the net* — Can we effectively model NFT images and their pairwise similarity relations to unveil possible patterns of NFT visual inspiration? (Sect. *Data Extraction and Network Modeling*)

(RQ2) *Just the way they are* — What are the main structural traits exhibited by the inferred NFT inspiration networks? (Sect. *Analysis of the NFT and Collection Networks*)

(RQ3) *To be or not to be... inspired?* — Are inspired and inspiring NFTs performing differently on the market? (Sect. *Market-based Characterization of the NFT Visual Inspiration Phenomenon*)

(RQ4) *Crypto-Flu* — What is the correlation between crypto markets and inspiration processes in the NFT landscape? (Sect. *Crypto Influence Dynamics*)

(RQ5) *Tell me why!* — What are the most relevant visual features to explain the inspiring relation between two NFTs? (Sect. *Explainability Aspects of the NFT Visual Learning Model*)

11.2 Scope and Limitations.

The data used in this study are largely representative in terms of diversification and coverage of the NFT landscape. In fact, they contain NFTs from heterogeneous platforms (i.e., markets), thus with different scopes, and cover the market history up to the steady-state of mid-2021, enabling us to capture the main consolidated fluctuations that appeared in the NFT selling scenario. Nonetheless, being out of the temporal range of our data, extreme events that occurred in the NFT market in 2022 are not considered in this study.

We also point out that our data contain timestamped information on the NFTs' selling prices, however they lack the assets' minting times, i.e., release times. We do not treat this missing point as a major inconvenience, since our focus is on those NFTs that have actually drawn attention in the market rather than those published but that might not be noticed by traders.

Also, we are aware of the risk that some visual inspiration processes might be altered or even "boosted" from multiple (fake) accounts belonging to a single person; however, delving into such phenomena would shift the focus to users, while the scope of this work involves visual features of NFTs. This issue is worthy of attention and we consider it part of our future work.

11.3 Data Extraction and Network Modeling

This section describes the data used in our study and our defined NFT and graph representation models.

Data. We use the well-recognized dataset provided in [168], which significantly covers the main NFT markets to date. It contains 6.1M purchase transactions involving 4.7M NFTs and more than 4k collections observed between 2017 and 2021. NFTs are grouped by the authors into six main categories, which indicate their scope, namely Art (18.46%), Collectible (28.85%), Games (47.21%), Metaverse (0.1%), Utility (0.17%), and Other (5.21%) — values within parenthesis are percent proportions. Each transaction in the dataset is associated with metadata that allows keeping track of the actors involved in the transaction and the main information associated with each NFT, such as its selling price, collection and category, and image URLs. After filtering out inaccessible NFTs or NFTs with less than one sell, our final

dataset consisted in more than 180k NFT images of various formats (.png, .jpeg and .webp), and corresponding metadata.

NFT visual representation model. To represent NFT images, we produce dense vectorial representations (embeddings), whose feature space is learned via *Transformer-based pre-trained visual models* (PVMs). These models, which allow avoiding manual or domain-driven selection of prominent features, are indeed the "de-facto" standards in computer vision. More specifically, we exploit a PVM based methodology for NFT visual feature extraction that has shown to be effective in [55]; note that, however, the purpose of exploitation of visual feature extraction in this work is different from the one in [55], which focused on a task of NFT financial performance prediction.

Let us denote with $\mathcal{I}$ the set of NFT images available in our dataset. Any NFT image $n_i \in \mathcal{I}$ can conveniently be represented as a *token sequence* $\mathcal{T}_i = [\tau_{i,0}, \tau_{i,1}, ..., \tau_{i,|\mathcal{T}_i|}]$, where $\tau_{i,j}$ denotes the j-th token of image n_i. Given a pre-trained visual model PVM, image tokens are densely encoded into a d-dimensional latent space, that is, $\mathsf{PVM}(\mathcal{T}_i) \in \mathcal{R}^{|\mathcal{T}_i| \times d}$. The resulting token embeddings are then fed through a pooling function $pooling(\cdot)$ in charge of producing, for each NFT image n_i, a single d-dimensional embedding $\mathbf{h}_i = pooling(\mathsf{PVM}(\mathcal{T}_i))$. A common approach, widely used in state-of-the-art PVMs is to get the output embedding of the special [CLS] token, whose pooling is generally considered a comprehensive representation of the entire (tokenized) input image.

Our choice for PVM in this study falls on *Vision Transformer* (ViT) [68, 41]. ViT is the first vision representation learner capable of achieving very competitive results against state-of-the-art CNN based architectures (e.g., Inception-V3). Being a BERT-like architecture, ViT works by tokenizing an image into patches of a given size (e.g., 16×16 pixels, as in our used implementation)[1] so that a token sequence can be attended to by the *self-attention* mechanism, which is able to draw relationships between patches. ViT models have shown to excel in image classification tasks reaching up to 90% accuracy on ImageNet; in particular, our used implementation of ViT was pre-trained on ImageNet-21k [184].

It should be noted that our design choice for PVM is to use it as is, without any adaptation to the NFT domain; in other terms, our PVM was used in inference on NFTs, but not further trained on NFTs. Although this may appear as a limitation, we instead defined such a setting for the following reason: a visual model that has never seen digital assets, like NFTs, is more likely to mimic the newbie user who is approaching the NFT market, and as such s/he does not have domain-specific knowledge.

Graph representation models for NFT visual inspiration. Once computed the NFT image embeddings $\mathbf{h}$, we use them to model proximity graphs, both at NFT level and collection level, which will lead us to answer our **RQ1**; in the following, we provide the corresponding definitions.

NFT Graph. We define the *NFT visual-inspiration graph* as a directed weighted graph $\mathcal{G} = \langle V, E, w, ts, T_s, T_e \rangle$, where V is the set of NFTs as nodes, T_s, T_e are start-time and end-time of observation, $ts : V \rightarrow [T_s, T_e]$ is a function assigning each NFT with a timestamp corresponding to its *primary sell* (or first sell), E is the set of edges such that, for any $v_i, v_j \in V$ *from different collections*, an edge is drawn from v_i to v_j if v_i follows v_j, i.e., $ts(v_i) > ts(v_j)$, and $w : E \rightarrow [0, 1]$ is an edge weighing function such that, for any $(v_i, v_j) \in E$, $w(v_i, v_j) = sim(\mathbf{h}_i, \mathbf{h}_j)$, where $sim(\cdot)$ denotes a similarity function for numerical vectors; our default choice is the cosine similarity.

[1] https://huggingface.co/google/vit-base-patch16-224

Table 11.1. Main structural characteristics of the NFT visual-inspiration graphs (left subtable) and similarity-linkage variants of the latest Collection graph (right subtable). Symbol * refers to statistics calculated by discarding edge orientation.

	NFT-GRAPH			COLL-GRAPH (*mid* '21)		
	mid '21	*mid* '20	*mid* '19	max	avg	min
#Nodes	52 198	19 011	3564	488	190	158
#Edges	481 873	122 046	15 501	3461	612	324
Density	2e-04	3e-04	0.001	0.015	0.017	0.013
Avg. In-Degree	9.232	6.420	4.349	7.092	3.221	2.051
Degree Assortativity	-0.14	-0.192	-0.300	-0.304	0.162	-0.105
%Sources	8.17	11.66	17.23	43.03	43.16	53.80
%Sinks	72.78	75.49	74.44	12.5	22.63	17.72
Diameter	17	11	5	9	5	3
Avg. Path Length	5.059	3.315	1.642	3.345	1.695	1.53
Transitivity*	0.054	0.025	0.002	0.18	0.334	0.207
Clust. Coeff. *	0.088	0.042	0.011	0.412	0.437	0.432
Clust. Coeff. *(full avg)**	0.062	0.029	0.007	0.357	0.336	0.315
#SCCs	52198	19011	3564	452	182	157
#WCCs	97	36	18	1	8	7
#Comm. by *Louvain**	138	69	37	6	14	14
Modularity by *Louvain* *	0.753	0.832	0.818	0.44	0.571	0.635

Note that our model definition has two key aspects relating to how NFTs are (i) timestamped and (ii) linked to each other in the graph. As for the first aspect, the availability of the primary-selling timestamps allows us to focus on those NFTs that were traded in the market, thus producing a tangible effect on it. Concerning the second aspect, our rationale for edge drawing and orientation is to capture a notion of *inspiration* triggered by some NFTs for others subsequently appeared in the market. It should be noted that some platforms (e.g., OpenSea) envisage warning mechanisms to alert users about potential "copies" of existing NFTs, however such warnings would be raised at collection level only, thus without actually flagging the involved NFTs. Therefore, identifying issues related to NFT copying, or even plagiarism, remains not trivial to cope with. In this regard, we take a more conservative perspective, as our goal is more generally aimed to detect influential effect of some NFTs towards others by measuring its extent in terms of similarity between the NFT images. Note also that we discard trivial relations of visual inspiration that are likely to occur between pairs of NFTs that belong to the same collection.

COLLECTION GRAPH. Analogously, we define the *NFT Collection visual-inspiration graph* as a directed weighted graph $\mathcal{G}_C = \langle V_C, E_C, ts, w_C, T_s, T_e \rangle$, where V_C is the set of NFT collections as nodes, E_C is the set of edges such that, for any $c_i, c_j \in V_C$, c_i points to (or visually inspired by) c_j (i.e., $(c_i, c_j) \in E_C$) if there exists $n \in c_i$ such that $ts(n) > \min\{t | t = ts(n'), n' \in c_j\}$ (with ts denoting the NFT timestamp assignment function as defined in $\mathcal{G}$), and $w_C : E_C \rightarrow [0, 1]$ assigns each pair of adjacent collection-nodes with a proximity score driven by a principle of NFT nearest-neighbor detection of c_i w.r.t. c_j, for any $(c_i, c_j) \in E_C$, according to some pre-defined *collection-linkage criterion*; formally, $w_C(c_i, c_j) = \sigma(\{\max_{n_a \in c_i} sim(\mathbf{h}_a, \mathbf{h}_b), \forall n_b \in c_j \mid ts(n_a) > ts(n_b)\})$, where σ denotes a set-function implementing the linkage criterion. To completely specify the above function,

we devise three linkage criteria, which correspond to as many ways of capturing different sensitivity levels w.r.t. the NFT inspiration processes across collections:

- σ = min, which represents the most tolerant scenario, as it requires that all the involved NFTs reach a certain similarity threshold to hypothesize that the source collection was inspired by the target one;
- σ = max, which represents the most eager case, as it is enough to have just one NFT reaching the similarity threshold to hypothesize that the source collection was inspired by the target one;
- σ = avg, which represents a balanced case, as it requires that, on average, the pairwise similarity must exceed a given threshold to hypothesize that the source collection was inspired by the target one.

In line with the NFT graph definition, $sim(\cdot)$ would be the cosine similarity between two NFT-collections, ranging between 0 and 1; however, to properly account for the amount of NFTs belonging to an inspired collection c_i that do not contribute to the similarity to another collection c_j (the inspiring one), we introduce a penalization factor of the cosine similarity defined as the sigmoid $1/(1 + \exp(-p_{i\to j}/np_{i\to j}))$, where $p_{i\to j}$ is the number of NFTs in c_i that are regarded as similar to others in c_j (i.e., that are involved in the application of σ), and $np_{i\to j}$ is the difference between the total number of NFTs in c_i and $p_{i\to j}$. Note that the final value of $sim(\cdot)$ still ranges within [0, 1], therefore edges on both the NFT and collection graphs are required to have a similarity score ≥ 0.5.

11.4 Analysis of the NFT and Collection Networks

In this section, we answer our **RQ2** through a structural analysis of the NFT and Collection networks, respectively, shedding light on their main macroscopic and mesoscopic traits. To this purpose, we focus on *cumulative yearly* observations, so that given our available data, a graph at time t models similarity relations between NFTs, resp. collections, from the beginning (i.e., 2017) up to year $T_e \equiv t$.

NFT graphs. Table 11.1 summarizes statistics on the main structural characteristics of the time-cumulative NFT graphs at mid 2019, mid 2020, and mid 2021. Our choice of starting with a cumulative observation at 2019 has a twofold justification in that more than 80% of the NFTs in our dataset were first sold starting from 2019, while for earlier periods, the similarity relations, and hence the visual inspiration, would be much less significant because of a certain immaturity of the NFT landscape before 2019. Indeed, the rapid growth of the NFT landscape is tangible on the visual-inspiration graphs, with an almost 15x increase in the number of nodes from 2019 to 2021, and an even sharper proliferation of edges in the same observation period (about 31x), hinting at an increasing presence of inspirational mechanisms among the NFT artworks. This is also strengthened by the increased average in-degree over the three-year observation period. Moreover, the negative (directed) degree assortativity, or disassortativity, reveals the tendency to draw inspiration from, or to inspire, NFTs having different inspiration degrees.

We also studied the power-law fitting of the in-degrees of the NFT graphs, as reported in the upper rows of Table 11.2. In this regard, albeit with non-negligible x_{min} values and with a particularly steep curve (see α), we spotted hints of likely fitting with the power-law distribution, as also confirmed by the high p-values obtained on the *Kolmogorov-Smirnov*'s test (KS-test). These values hold for all the considered timestamps (i.e., from 2019 to 2021) and

might shed light on the existence of latent yet perceivable preferential attachment mechanisms on the inspirational process between NFTs.

Interesting remarks arise from the percentages of source and sink nodes, i.e., nodes having no in-links and out-links, respectively. The fraction of sink nodes remains relatively constant and high over the years, denoting that a large part of the NFTs in our graphs are inspiring. The same does not hold for the source nodes, where a progressive decrease in the number of such nodes during the last three years denotes that the remaining majority of NFTs in the network (~91% of nodes at mid 2021) have inspired other NFTs. This could reasonably be ascribed to the reaching of a sort of *temporary saturation* threshold of NFT artistic development, as most of the visual traits of NFTs could have been used at least once in the market. Such insights are reinforced by the diameter and average path length statistics, which, doubling year by year, reveal how the chain of inspiration was getting longer.

As concerns the triadic closure, both its expressions in terms of transitivity and local clustering coefficient show low values, suggesting that the NFT inspiration (i.e., linkage) process according to visual features could be highly targeted. This is also emphasized by the large and growing number of connected components and communities. In particular, coupled with the observed high modularity, the latter supports the finding on specialized and well-separated inspiration mechanisms among NFTs. In this regard, we notice how the observed communities appear to be induced by specific "visual topics" shared by their corresponding NFTs, such as virtual pets, vehicles, wearable, metaverse lands, gaming cards, and naming services.

Table 11.2. Outcomes of KS-tests on the power-law fitting of the in-degree distributions of NFT and Collection graphs.

	graph	α	x_{min}	p-value
NFT-Graph	*mid* 2021	4.94	168	0.81
	mid 2020	4.11	74	0.96
	mid 2019	5.06	45	0.99
Collection-Graph (*mid* 2021)	max	2.31	23	0.99
	avg	2.70	12	0.99
	min	2.29	7	0.99

In Figure 11.1, we provide an illustration of the NFT graph evolving from 2019 to 2021. Looking at the plots, a first remark that stands out concerns the asset outbreak in the transition between years, especially between 2020 and 2021. The "Big Bang"-like visual effect we can notice in the figure gives evidence that inspiration processes have suddenly occurred in the NFT landscape. In particular, the inspiration process have recently involved almost all NFT categories (as evidenced by the colored out-links), eventually coming up in 2021 with *Art*, *Metaverse*, *Games*, and some *Collectibles*, being the most inspiring ones. Therefore, we can reasonably state that the NFT inspiration phenomenon has allowed the market to undergo a rapid expansion.

Collection graphs. To analyze the NFT visual-inspiration relations at collection level we focus on the latest time-cumulative models (i.e., mid 2021) by varying the linkage criterion function. A summary of structural statistics is reported in the right-most subtable of Table 11.1.

As expected, the linkage criteria have a clear impact on the network size, with the average-linkage balancing between the expansion and contraction effects of the max- and min-linkage, respectively. In contrast to what observed for the NFT graphs, the percentage of reciprocated

edges is not zero, under all linkage criteria (1.1%, 2.6%, and 0.6% for max, avg, min, respectively), revealing that during different timestamps, cross-collection inspiration events among NFTs of different collections are possible. As previously observed in the NFT graphs, the low density indicates a certain "specialization" of the inspiration process, which is again confirmed by quite a small average in-degree, and conversely, and extremely high number of strongly connected components (almost matching the number of nodes in a graph).

Remarkably, testing the power-law fitting of the in-degree distribution in the Collection graphs, we found clues of preferential attachment, as indicated by the p-values resulting from the corresponding KS-tests reported in Table 11.2. Since the observed goodness-of-fit consistently holds for all linkage criteria, this would suggest that the inspiration process can be pulled from some particularly inspiring collections and creators.

The fraction of source nodes, resp. sink nodes follow relatively similar trends under all the linkage strategies, although in an opposite way w.r.t. the NFT graphs: indeed, we spotted that a large fraction of the set of collections actually draw inspiration from others, while the inspiring ones (i.e., those having zero out-degree) represent a minority of the total collections involved in inspiration processes.

The choice of linkage criterion also impacts on the degree assortativity. According to the graph definition, the disassortativity corresponding to the *max* case indicates that nodes having a high out-degree (i.e., collections inspiring to many others) tend to get also inspired by nodes with small in-degree (i.e., collections that inspire little); this is clearly in accord with the high sensitivity of the aforementioned linkage criterion, which is, therefore, able to spot even the subtlest semblances of inspiration. On the contrary, the disassortativity corresponding to the *min* case allows us to reasonably state that collections drawing little inspiration from others (i.e., small out-degree) do this by targeting the most inspiring collections (i.e., high in-degree); again, this fits the behavior due to the linkage criterion, which is here the most cautious in terms of similarity recognition. The two above contrasting behaviors are finally balanced by the *average*-linkage criterion, which leads to a positive degree correlation, which means that, on average, collections that are very inspired do it from those that are very inspiring, thus hinting again at a preferential attachment mechanism, consistently with the previously discussed finding (cf. Table 11.2).

In addition, the values of diameter and average path length, regardless of the linkage criterion, further shed light on the close-knit process of inspiration among collections. The linkage of collections hence appears to shape a tight network system, albeit sectorial, as indicated by triadic closure and modularity-based statistics. In light of such remarks and the previous findings about the NFT graphs, we again report a notable and even more evident matching between community structures and specific visual traits.

In Figure 11.2, we show the effect of the linkage criteria on the 2021's Collection graph. At first glance, it appears that the min linkage criterion tends to trigger scattered links among collections, thus leaving apart potential cases of mild or fairly moderate inspirational scenarios. Conversely, the eagerness of the max linkage criterion yields a proliferation of similarity links, forming two close-knit and dense "clouds" of visual inspiration. Besides, the avg linkage approach balances between the aforementioned strategies, overcoming the extreme sparseness of the min case, thus effectively being able to capture actual similarities, yet without degenerating into potential noise of the max case. Moreover, regardless of the chosen linkage criterion, some shared traits emerge from Figure 11.2: *Art* and *Collectibles* appear to be the most inspiring categories, whereas *Other* results to be the one corresponding to the most inspired NFTs (red out-links), by also being the most generic ones — thus including any potential asset. Interestingly, in mid-2021, *Games* and *Metaverse* appear to have carved out a portion of the market

Table 11.3. Market-related statistics for the inspiring and inspired NFTs computed from the 2021 NFT visual-inspiration graph.

Financial Indicator	Inspiring NFTs	Inspired NFTs	Inspiring/ Inspired
average volume	$231531.69	$146192.15	1.584
average #transactions	151.92	100.00	1.519
average price	$692.91	$899.09	0.771
maximum price	$6661.95	$4605.24	1.447
minimum price	$102.24	$318.89	0.321
st. dev. price	$977.22	$725.69	1.347

away from other visual features, or with an original redefinition of the same, as their presence is almost latent in the plots.

11.5 Market-based Characterization of the NFT Visual Inspiration Phenomenon

Our previous analysis has shed light on the structural traits of the visual-inspiration networks built on the NFTs and their collections. Here we want to provide a simple characterization of the dichotomy between *inspiring* and *inspired* NFTs in terms of main aggregated financial indicators (**RQ3**).

Based on the latest update of the NFT visual-inspiration graph (i.e., 2021), Table 11.3 reports, for various market-related statistics, the ratio between inspiring NFTs (target nodes in a link) and inspired NFTs (source nodes in a link).

We first notice that inspiring NFTs are able to achieve more transactions and higher volumes[2] than inspired ones, with peaks of +58% and +52%, respectively. Surprisingly, we find lower average, resp. minimum, prices for inspiring NFTs, since inspired NFTs obtain a +23%, resp. +68% advantage over the former. We shed light on such a counter-intuitive trait by looking at the ratio of standard deviations, which shows a +35% for inspiring NFTs. As a result, we can state that the inspiring NFTs face an initial struggle in the market with lower minimum prices due to the novelty of the proposed visual traits, with a more evident price instability, whereas inspired ones tend to enter a market "ready" to appreciate the proposed features, thus starting from higher minimum prices. Nonetheless, we also spotted that the market inherently regulates such a phenomenon, with inspiring NFTs able to achieve higher maximum prices w.r.t. inspired ones (+45%), thus conferring a sort of competitive advantage to the former.

11.6 Crypto Influence Dynamics

Answering our **RQ4** requires to analyze the temporal relation between trends of market-related indicators and trends of NFT-image similarities. To this purpose, we resort to the Time Lagged Cross Correlation (TLCC) technique which, given two time series s and s' of the same length,

[2] In the NFT market, volume of an NFT is meant as the total amount sold for all transactions involving that NFT.

determines how well s' is related to past lags of s, thus identifying lags of s that might be predictors of s'. Formally, the TLCC of s, s' is the set of sample Pearson-correlations between $s_{t+\ell}$ and s'_t, for all $\ell \in [-T..T]$ ($T > 0$), at any given time-step t within a chosen time period T. The lag refers to how far the series are offset, and its sign indicates which series is shifted in time, therefore $\ell < 0$ corresponds to a correlation between s at a time before t and s' at time t, i.e., s leads s', whereas $\ell > 0$ corresponds to a correlation between s at a time after t and s' at time t, i.e., s lags s'.

Within this view, we analyze different types of TLCC to assess whether a leading-lagging relation exists between "artistic" and "market-related" indicators. In particular, using either a *monthly* or a *weekly* sampling — i.e., time-steps correspond to either months or weeks — we modeled five types of time series over the full period of observation (i.e., $T_e - T_s$), measuring for each time-step t: (i) the average of the pairwise similarities (including both within- and across-*category* pairs) of NFTs from the NFT graph observed up to t (i.e., $T_e = t$), (ii) the average of the NFTs' mean selling prices up to t, (iii) the closing price of the Bitcoin market at t, (iv) the number of first-sold NFTs at t, and (v) the number of collections containing at least one first-sold NFT, at t.

NFT and market-related time series. Before moving to the analysis of the TLCCs, let us first provide insights into the shapes of the aforementioned time series. (For the sake of brevity, we discard a discussion about the series of collections containing first-sold NFTs; however, main remarks for the latter hold similarly to those for the plots of first-sold NFTs). As shown in Figure 11.3-(a), the average pairwise similarity series exhibit a similar and generally increasing trend, although interesting differences at their starting point arise depending on the time sampling of the evolving NFT visual-inspiration graph. In fact, the weekly series shows non-zero similarities since the first week of 2018, whereas the monthly one "delays" the raising of visual inspiration events to August 2018. For both series, we also notice an abrupt growth in similarities around the 10th month, before the trend becoming slower or even stopping around the 30th month. As a further remark (not drawn from Figure 11.3-(a)), although the average growth rate turns out to be more evident for the within-category similarity, the trend of the average similarities follows that of the across-category similarities.

Looking at the average NFT-selling price series, shown in Figure 11.3-(b), we notice almost identical shapes from the monthly and weekly samplings. Interestingly, the peaks at the beginning of the series correspond to the dawn of the NFT advent. This should be ascribed to the scarcity of assets yet the novelty of this technology, which dictated those high entry-prices. Then, a large time window follows as characterized by low prices, which can be explained by the strong proliferation of NFTs in the market. Eventually, a rising trend is observed during 2021, when the market achieved its maturity, and the *fomo* (i.e., fear-of-missing-out) begins to spread over more people. It should also be noted how the two discussed plots are inter-related, hinting at self-regulation of the market, where similarities and average prices appear to be (inversely) affected by each other.

Considering Figure 11.3-(c), the initial upward trend is rapidly stopped by the bear market where a minimum of $3500 is reached at the beginning of January 2019 (14th month, resp. 58th week). A new trend inversion then begins, thus increasing the BTC price to around $9000 for a time span that goes from the 20th month to the 35th month, then again the bull-run drives the BTC price upward reaching about $60000 at the end of the observation interval.

Finally, the upward trend characterizing both series in Figure 11.3-(d), appears to be more prominent for the monthly sampling. Two distinct peaks are noticed at around the 40th month, resp. 100th week, corresponding with the trend inversion of the BTC bear/bull-run and the start of the bull-run of 2020/2021.

Time-lagged cross correlations. We are now ready to analyze the TLCCs between the time series previously discussed, where we specify T equal to one year. Figure 11.4 shows the TLCC between average NFT selling prices and average NFT similarities, and the TLCC between average NFT-selling prices and Bitcoin market closing prices, where the latter quantity in either comparison corresponds to s' in the definition of TLCC reported at the beginning of this section.

Considering the first comparison (top correlogram in Figure 11.4), according to a monthly sampling we find a strong negative correlation (-0.918) between the average NFT selling prices and the average NFT similarities, with a negative offset ($t = -7$ month). This means that in the mid term, selling prices lead the NFT inspiration process in an inverse way, i.e., high prices reduce similarities, and hence lower the tendency of reusing visual features (i.e., temporary saturation of artistic development), and vice versa. Conversely, the weekly perspective sheds light on a short-term strategy with peak negative correlation (-0.658) at $t = 1$ week, thus indicating that the selling prices lag the NFT similarities, still in an inverse way. Remarkably, this finding unveils that the temporary saturation of artistic development can impact on asset performances by reducing the NFT prices. Besides, as the leading quantity changes depending on the sample resolution (i.e., weekly or monthly), we can spot different latent strategies that would be triggered by the market in response to the variation of the aforementioned quantities.

The bottom correlogram of Figure 11.4 shows the TLCC between the average NFT selling prices and the closing prices of the crypto markets, represented via the BTC-USD trends obtained through the Yahoo! Finance APIs.[3] Looking at the plots, two main remarks stand out: (i) BTC-USD is the leading quantity in both medium- and short-term scenarios, with peaks of correlation up to 0.93 and 0.898 for the monthly and weekly perspective, respectively; and (ii) there exists a plateau of correlation according to both weekly sampling (from the 5th to the 17th week) and monthly sampling (from the 2nd to the 5th month). The coherence of such observations across different samplings and the corresponding strong correlation allow us to confirm that BTC is the main influence factor in determining the NFT selling performances in line with other studies [12].

Furthermore, in Figure 11.5, we show the TLCC between prices and counts of newly created NFTs/collections on both weekly and monthly offsets. Besides spotting an almost equivalent offset between months or weeks for either NFTs and collections, we find strong evidence that their birth impacts on the market. More precisely, collections appear to be a better financial predictor due to the greater volume they generate, and such an indicator is stronger when we consider a monthly perspective, with peaks of correlation up to 0.92 and 0.58 for collections and NFTs, respectively.

11.7 Explainability Aspects of the NFT Visual Learning Model

In this section we aim at answering our **RQ5**, which allows for gaining qualitative insights into the ability of our proposed approach to shape similarities between NFTs, thus providing an explanation underlying a relation of visual inspiration between any two NFT images.

To this purpose, we resort to the SHAP (SHapley Additive exPlanation) framework [151],[4] which leverages the game-theoretic Shapley value to locally explain the outcome of a machine/deep learning model, like our NFT visual-inspiration learner, whereby input features are represented through players of the game. SHAP is a model-agnostic technique, since it just

[3] Available at `https://finance.yahoo.com/`

[4] Code available at `https://github.com/slundberg/shap`

requires class probabilities generated by the learner and, by selectively perturbing the learner's input features, it calculates the contribution that each of the features gets w.r.t. the output probabilities. Details on SHAP are given in ***Appendix A.5***.

In Figure 11.6, we report three exemplary cases, sampled by linked NFTs in our 2021 NFT-graph, and ordered by left to right with increasing level of difficulty for our NFT visual-inspiration learner. Let's begin with two NFTs representing the Van Gogh's *A Pair of Shoes* painting, shown in Figure 11.6-(a). As expected, since the two images only differ in terms of color saturation and contrast, our model was able to identify a strong similarity between them. This similarity is well explained by SHAP as almost all portions of the shoes are detected as relevant features (red cells).

The case depicted in Figure 11.6-(b) is an example of "extreme" inspiration captured by our NFT visual-inspiration learner, since the two NFTs display a common concept, i.e., the handle pickaxe, by an almost identical drawing design, even though with different color choices for their heads. Indeed, the SHAP model correctly detects the "handle" portions and the "Vs" symbols of the images as relevant features for the similarity between the two NFTs (red cells), whereas the "heads" are detected as irrelevant features (blue cells). Also, the textual components of the two NFTs ("Gold" and "Kryptonite") are correctly discarded by SHAP in explaining their similarity. It should be emphasized that, like in other cases, the two NFTs are from different collections.

Figure 11.6-(c) shows two apparently different NFTs, which nonetheless share a common visual concept concerning a warrior figure. The inspiration exerted by the rightmost NFT to the leftmost NFT is well captured, as indicated by the higher Shapley values (i.e., red cells) corresponding to the upper portion of the warrior on the left figure and the torso of the second one. Furthermore, the model is able to distinguish the different weapons (i.e., a katana on the left, and a sword on the right), as indicated by the blue cells located upon their respective locations. Therefore, not only our NFT visual-inspiration learner effectively captures similarity (0.74) underlying the concept of "warrior", but does so by ignoring pointless features (e.g., text over the left image) and, most importantly, regardless of the observation perspective (front with zoom inset on the left, or in profile on the right).

11.8 Chapter review

We presented the first work leveraging visual features learned from NFT images and used to build a network model for capturing and analyzing visual inspiration relations between NFTs.

By answering a number of research questions relating to the visual inspiration phenomenon, our results have shown that (i) the inspirational mechanisms underlying the artistic development of NFTs are pervasive and they have progressively led to a temporary saturation of the visual feature space, thus affecting originality of NFTs; (ii) the dichotomy between inspiring and inspired NFTs has effect on related financial indicators; and (iii) there exist latent and enduring mechanisms of self-regulation between markets and inspiration waves.

Several directions can be outlined as future research. One aspect that is part of our ongoing work is the development of a visual learning model specific for the NFT domain, by re-training or further training a PVM like our used ViT, or alternative models, on NFT image data collections. A related opportunity comes from optimizing an NFT visual model to downstream tasks, such as classification and object recognition, possibly in multi-modal learning scenarios.

We would also like to point out that our definition of NFT visual inspiration graph models can straightforwardly be generalized to other datasets of NFTs, which might also include more

information than the one used in this study, such as the NFT minting timestamps. Also, we acknowledge the availability of a very recently released dataset which however, unlike the dataset we used in this study, is narrowed to the OpenSea market only [56].

We expect our study on NFT visual inspiration can pave the way for understanding important mechanisms arising during the evolution of Web3.

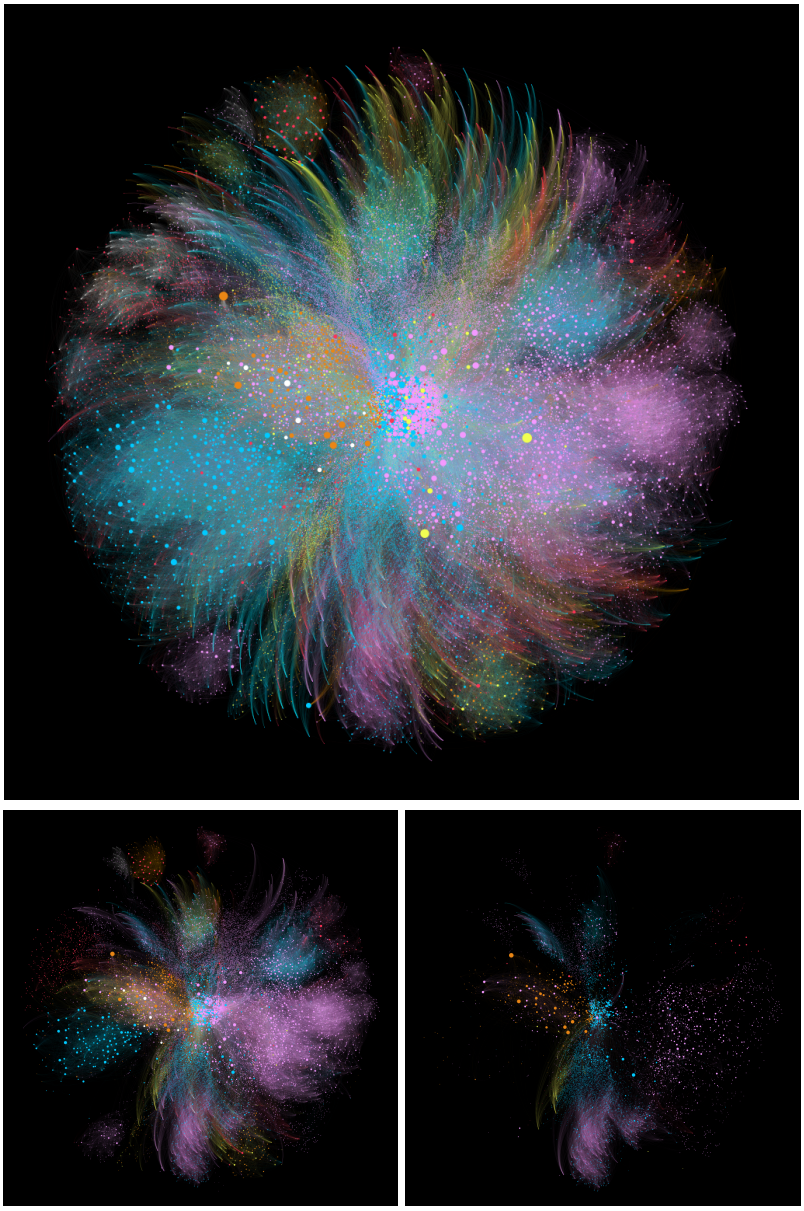


Fig. 11.1. NFT visual-inspiration graphs in 2021 (top), 2020 (bottom-left) and 2019 (bottom-right). Node colors correspond to categories (pink for Games, yellow for Collectibles, blue for Art, orange for Metaverse, white for Utility, red for Other). Links are colored according to the color of the source nodes. Node size is proportional to the sum of weights on the incoming links to the node; to ease observing the graph evolution, the size of each node is kept fixed to the node's status at 2021. *(Best viewed in color)*

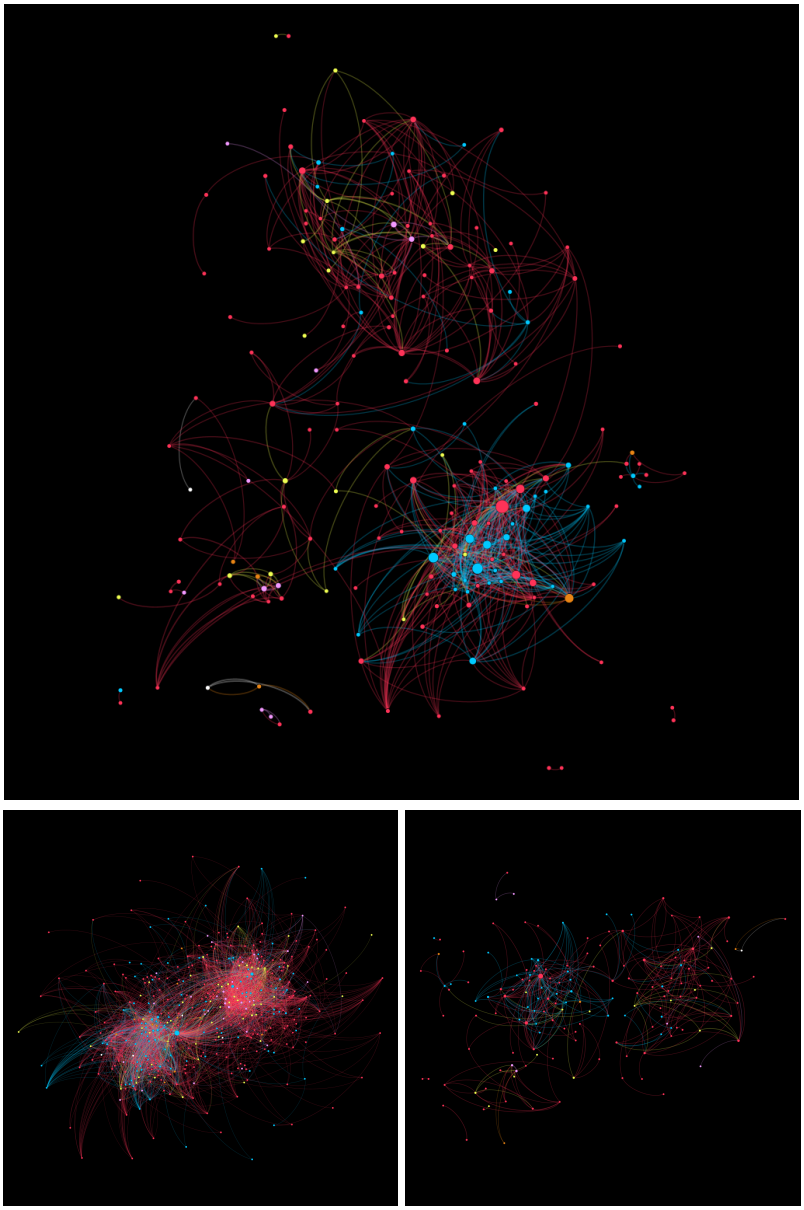


Fig. 11.2. Collection graphs at 2021 by varying linkage criterion: avg (top), max (bottom-left), min (bottom-right). Node size is proportional to the sum of weights on the incoming links to the node. Color setting is the same as for Figure 11.1.

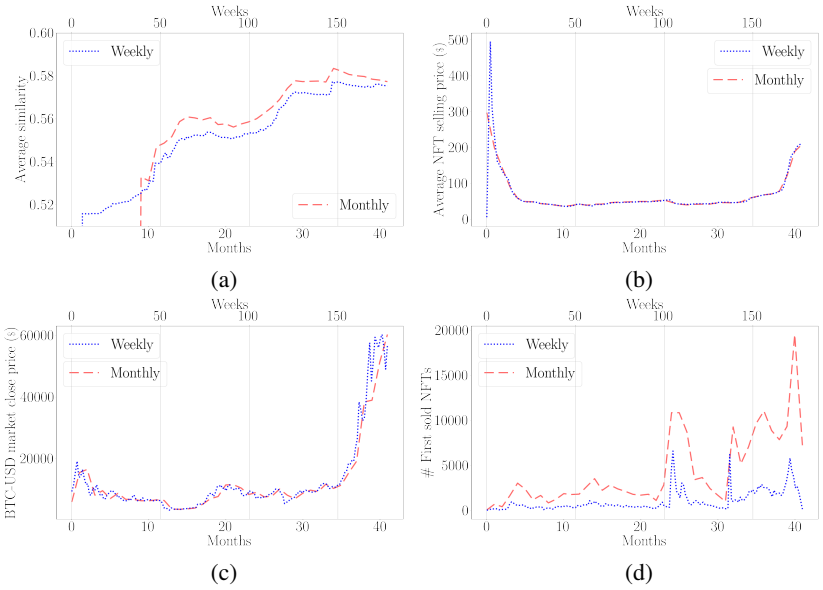


Fig. 11.3. Average NFT similarities over time (a), average NFT selling prices (b), Bitcoin market closing prices (c) and number of first-sold NFTs (d).

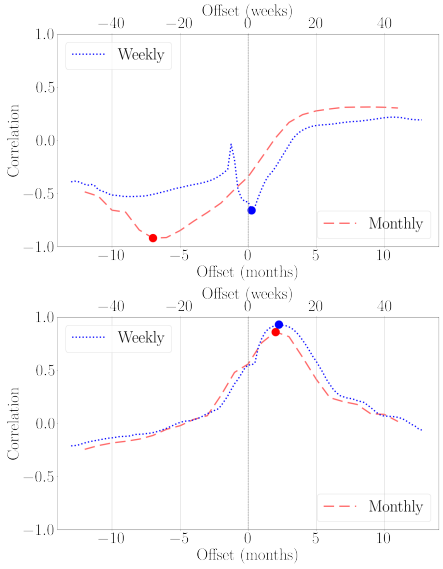


Fig. 11.4. Time lagged cross-correlation of average NFT-selling prices vs. average NFT similarities (top), and of average NFT-selling prices vs Bitcoin market closing prices (bottom). In each plot, the bottom x-axis corresponds to monthly samples (red dashed line), whereas the top x-axis corresponds to weekly samples (blue dotted line).

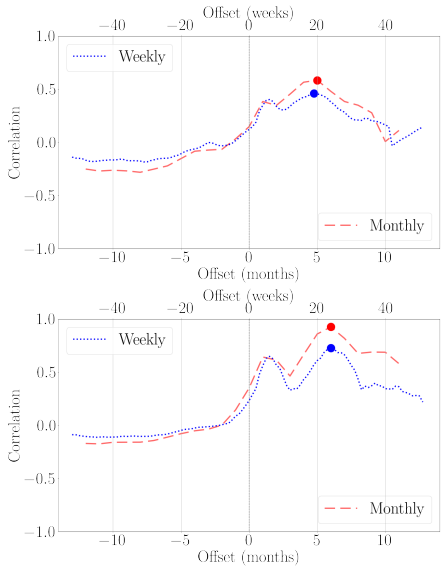
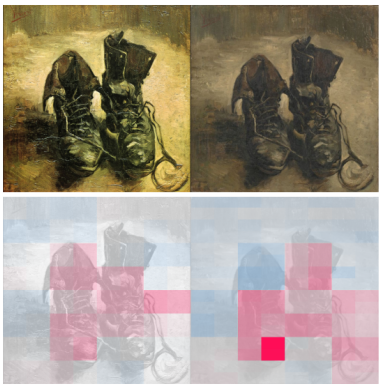


Fig. 11.5. Time lagged cross-correlation of average NFT-selling prices vs. number of first-sold NFTs (top), and of average NFT-selling prices vs. number of collections containing at least one first-sold NFT. In each plot, the bottom x-axis corresponds to monthly samples (red dashed line), whereas the top x-axis corresponds to weekly samples (blue dotted line).



(a)

(b)

(c)

Fig. 11.6. Use cases for explanation of our PVM learner for NFT images. Each case consists of two images placed side by side that correspond to NFTs linked in our inspiration-networks, and hence from different collections (top), along with their SHAP explanations (bottom) in the form of heatmap layers. Blue, resp. red, cells denote negative, resp. positive, impact of features on the model's similarity prediction.

12

SONAR: Web-based Tool for Multimodal Exploration of Non-Fungible Token Inspiration Networks

Summary. In this work, we present SONAR, a web-based tool for multimodal exploration of Non-Fungible Token (NFT) inspiration networks. SONAR is conceived to support both creators and traders in the emerging Web3 by providing an interactive visualization of the inspiration-driven connections between NFTs, at both individual level and collection level. SONAR can hence be useful to identify new investment opportunities as well as anomalous inspirations. To demonstrate SONAR's capabilities, we present an application to the largest and most widely used dataset concerning the NFT landscape to date, showing how our proposed tool can scale and ensure high-level user experience up to millions of edges.

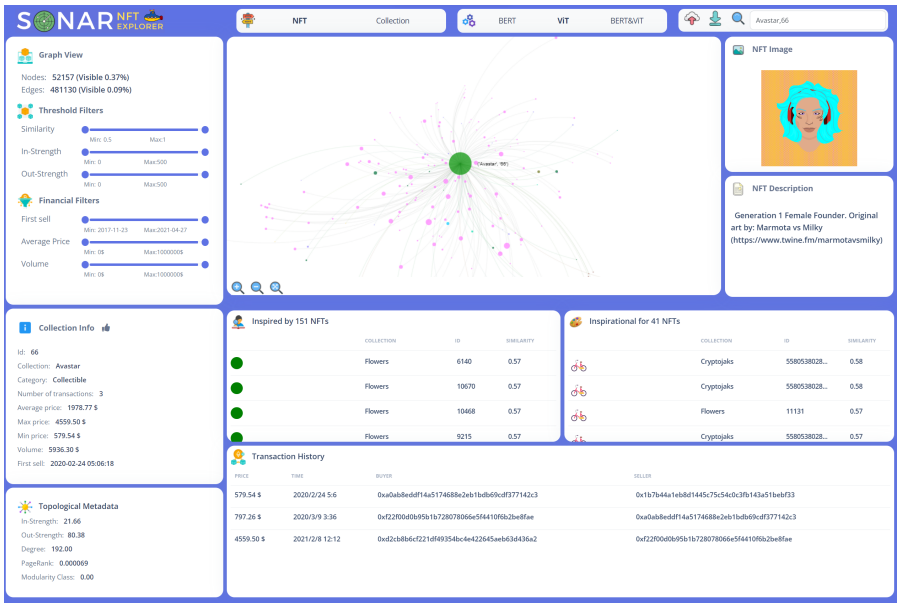


Fig. 12.1. Screenshot of the interface of SONAR

12.1 Contributions

The fervor behind NFTs has resulted in an inestimable number of NFTs, easily becoming potentially noisy for creators and investors, in which pitfalls can be confused with opportunities; therefore, there is a need to shed light on them.

In this regard, it is desirable to have efficient and effective tools that can leverage the pillars of NFTs, i.e., images and descriptions, to allow end-users to explore them dynamically. It should be noted that these markets are heavily dependent on the creator's creativity; indeed, by generating assets, they contribute to the continuous evolution and growth of the domain. Besides that, inspirational mechanisms emerge, allowing creators to delve into novel forms of expression or take cues from others' ideas. Such an originality-inspirational dichotomy is hence the fulcrum for a proper understanding of the NFT landscape, and hence it is today demanding to have tools that can provide NFT creators as well as traders with the capability of efficiently traversing NFT inspirational effects.

To this aim, we propose SONAR, a web-based, user-friendly interactive tool for exploring NFT inspiration networks inferred through the multimodal representation of NFT images and descriptions, enriched with topological and financial knowledge.

SONAR is conceived to enable users to seize the myriad of opportunities offered by such an exciting landscape by posing search queries on NFTs both at individual and NFT-collection level. For instance, investors can use SONAR to buy tokens similar to well-performing ones, whereas creators might spot unexplored lands, giving space to their creativity. Also, SONAR would allow end-users to protect themselves from risks by making it easier to identify potential copyright infringements or plagiarism due to extreme inspirational phenomena. In this work, we demonstrate SONAR's capabilities by exploring the most recognized dataset involving NFTs existing to date [168].

To the best of our knowledge, no software tool has been proposed so far for exploring NFTs through the discovery of inspiration patterns, according to the NFT visual appearance as well as textual description. Here, we aim at filling this gap, with the ultimate goal of contributing to improve our understanding of the continuous evolution of the NFT landscape.

12.2 Design

Figure 12.2 shows the main modules and data flows of SONAR. To favor usability and speed of interaction, we decided to develop SONAR as a single-interface tool, encompassing various functionalities and interaction modes, as described next.

Multimodal exploration. The first point of interaction with SONAR concerns the choice of the granularity level for the exploration of NFTs, namely *NFT-level* and *Collection-level*. This allows end-users to switch from fine- to coarse-grained views depending on their exploration needs. Here, it is also possible to select a proper representation model (cf. Section 12.4) for shaping the embedding modality, i.e., *unimodal* or *bimodal*, upon which network(s) to be explored are inferred.

Interaction modes. SONAR offers three main interaction and responsive modes for exploring the NFT inspiration network(s). The first concerns using hand-held devices or on-screen controls to traverse the network and select nodes (i.e., NFTs or Collections) by clicking on them. The latter can also be obtained by means of a search bar, which interprets user-provided information to highlight in the network the NFT corresponding to the input collection and/or token identifier. In all cases, once a node is selected, the user views its image and text or a

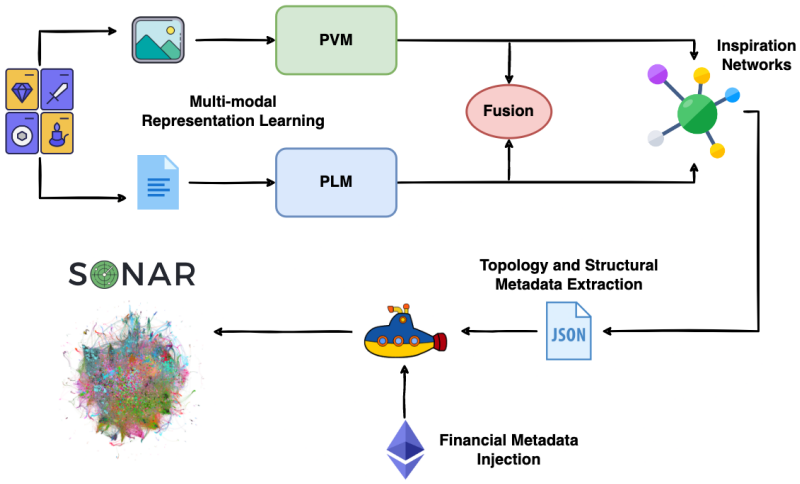


Fig. 12.2. Overview of the conceptual architecture of SONAR

collection-wise representation, depending on whether it is an individual NFT or not. For all the interaction modes, users can take advantage of specific thresholds concerning topological and financial properties of NFTs — more on this in Section 12.3 — to filter out unwanted tokens/collections and resample the network according to the users' exploration objectives.

Topology and Inspiration views. While the user is interacting with the network by selecting a node, the interface of SONAR is populated with the main topological attributes that characterize the node, in order to provide insights into the structural role of NFTs or Collections in the market. Contextually, the one-hop neighborhood of the selected node is highlighted to attend potential incoming or outgoing inspirational mechanisms among these and the focal node. In this regard, specific interactive tables containing the inspiration levels among nodes (being them NFTs or Collections) are filled in, along with visual indicators advising how influential they are, where the influence flow is modeled according to the method described later in Section 12.4. Such a peculiarity allows users to estimate the degree of influence exerted on, resp. received from, other nodes, thus unveiling highly influential tokens, resp. raise warnings for potential plagiarism, up to the level of individual NFTs.

Data import/export. We equipped SONAR with data export and upload capabilities. As concerns the former, the key idea is to allow users to bookmark nodes and export them with a few clicks to save the state of the exploration. This is particularly relevant for the purposes of SONAR, since, while traversing the network, anomalies (e.g., plagiarism) or promising tokens in which to invest might be identified, and users would like to store them for further investigations, or as an input for other systems. Conversely, users can load previously exported snapshots to continue exploring the NFT landscape or deepen at previously spotted phenomena.

12.3 Implementation

In this section, we elaborate on our implementation choices regarding the development of the main components of SONAR.

Node and edge features. Besides NFT image and text data, SONAR is designed to handle a variety of metadata, including those useful to characterize NFT inspirational mechanisms. Particularly, NFT nodes are associated with topological metadata, such as in-/out-strength, full-degree, PageRank score, and the identifier of the corresponding modularity class (i.e., community in the network), NFT category, and collection of membership. Moreover, we enrich NFT metadata with financial information, such as the time of the first-sell, the list of transactions involving them, the overall traded volume, and statistics involving the mean and min/max selling prices. Note that, in the Collection-level graph, each node is labeled with the NFT having the best selling volume.

Also, for network layout setting purposes, layout coordinates, node size, and color code are handled for each node, and width and color for edges. In this regard, the default choice is to plot nodes with color coding collections, resp. categories, in the NFT-level, resp. Collection-level; in both cases, node size refers to the in-strength.

Data representation. We implemented SONAR paying attention to interoperability, flexibility, and compatibility aspects. To this aim, our chosen reference format is the multi-line JSON, which encompasses nodes, edges, and their corresponding metadata, as discussed above. The rationale of this choice is twofold: (i) the JSON format is among the most diffused ones, and it is easily extensible through any programming language, thus allowing the manipulation of the networks or a definition of new node/edge properties, and (ii) it ensures high compatibility between SONAR and the vast majority of network analysis tools available to date. In particular, our JSON structure matches the one exportable via Gephi [22] (and its JSONExporter or SigmaExporter plugins), which is recognized as the de-facto standard in general-purpose graph analysis and permits even non-expert end-users to generate basic input data (i.e., containing just topological and layout rendering information).

Web application. SONAR is developed to provide end-users with the most user-friendly and smoothest interaction capabilities, thus minimizing client-side workloads. To match device-wise (i.e., computers, tablets, smartphones) scalability desiderata, we implemented the front-end by exploiting one of the most-adopted frameworks to date, namely Bootstrap (v.5.2.0).[1] As for the back-end, we had to deal with the enormous amount of data that characterizes the NFT landscape and the different data types involved (i.e., text, transactions, images). To this aim, we leveraged the SigmaJS (v.0.1) framework[2] for graph rendering and manipulation purposes. Furthermore, to fulfill all interaction needs for SONAR, we extended the base set of APIs with customized ones, thus achieving interactive thresholding via queries, dynamic search functionalities, graph uploading/exporting, or event-driven graph navigation. We assessed SONAR's capabilities by using it to handle more than 1.5M edges and almost 100K nodes.

12.4 Data and Network Modeling

SONAR is designed to be versatile w.r.t. the models used for representing the NFT and the networks being inferred from NFT relations, respectively. In the current version of SONAR, our chosen modeling approach leverages state-of-the-art *language and visual models*, as well as a notion of *NFT visual inspiration* network we originally introduced in [138].

NFT visual and text representation models. To represent NFT images and their textual descriptions, our goal is to avoid manual or domain-driven selection of prominent features, by

[1] `https://getbootstrap.com/`
[2] `https://www.sigmajs.org/`

learning dense vectorial representations (embeddings) via a *pre-trained visual model* (PVM) and a *pre-trained language model* (PLM), respectively. Both models are based on the well-known Transformer model architecture [223], which has become ubiquitous on a wide range of tasks in NLP as well as computer vision. The key point in the Transformer architecture is the use of the *attention* mechanism as a more effective alternative to recurrent neural networks and convolutional neural networks [223]. The PLM and PVM are fed with an initial representation of each text and image as a sequence of tokens, which correspond to (sub-)words and fixed-size non-overlapping patches, respectively. Each of such tokens is embedded into the PLM/PVM latent space, and a final low-dimensional representation of each text or image is obtained by feeding the token embeddings through a pooling function.

Our choice for PLM is BERT [66], which has represented a breakthrough in several NLP benchmarks and is still considered a must-have baseline. It essentially consists of a stack of Transformer encoder layers (i.e., 12 attention heads and 12 hidden layers) and its key advantages include the bidirectional pre-training and the unified architecture across different tasks. As concerns PVM, SONAR utilizes ViT [41],[3] which is the first vision representation learner capable of achieving very competitive results against state-of-the-art CNN based architectures (e.g., Inception-V3). Particularly, we use the ViT model that was pre-trained on ImageNet-21k [184].

NFT images and descriptions clearly bring different information, but equally useful to determine the essential nature of the asset. Following [55], we consider a bimodal representation model as well, where the image and text embeddings are eventually combined together and fed into an attention module to fuse the visual features and the lexical/semantic features.

It should be noted that both PVM and PVM are used without further pre-training or fine-tuning to the NFT domain. This choice could be regarded as a limitation, but we believe using pre-trained models that have never seen NFT images or descriptions are more likely to mimic the newbie users.

Graph representation models for NFT inspiration. The learned NFT representations are used to model special proximity graphs, both at NFT level and collection level. Here we summarize key aspects of our defined graphs, while a complete, formal definition can be found in [138].

Chosen an NFT representation modality (i.e., image, text, or bimodal fusion), we define the *NFT-level inspiration graph* as a directed, timestamped, weighted graph, where nodes represent NFTs, each associated with a timestamp corresponding to its *minting* (i.e., creation) time, or alternatively, its *primary sell* (or first sell) time. Edges are drawn between NFTs of different collections in such a way that a node points to any other node having earlier timestamp and the strength (weight) of this relation is measured in terms of cosine similarity of the two NFTs' embeddings.

Analogously, we define the *Collection-level inspiration graph* as a directed, timestamped, weighted graph, where each node is an NFT collection, and an edge is drawn from a collection to another if there exists at least one NFT in the pointing collection that is more recent than all the NFTs in the pointed collection. A real-valued weight is assigned to each edge according to a principle of NFT nearest-neighbor detection of the pointing collection w.r.t. the pointed one: more precisely, for each NFT in the pointed collection, the most similar later NFT in the pointing collection is identified, and the similarity scores are aggregated according to some *collection-linkage criterion*; the aggregation operator is the mean.

Note also that, in order to address the problem concerning the number of NFTs belonging to a collection that have actually taken part in the edge weight computation, we introduce a

[3] `https://huggingface.co/google/vit-base-patch16-224`

penalization factor of the cosine similarity. Given an edge from collection i to collection j, we define the penalization as $1/(1 + \exp(-p_{i\to j}/np_{i\to j}))$, where $p_{i\to j}$ is the number of NFTs in the source node which are regarded as similar, while $np_{i\to j}$ is the total number of NFTs contained within the source collection.

12.5 Demonstration

We use the largest and most recognized dataset concerning the NFT landscape to date [168, 55]. It contains 6.1M purchase transactions involving 4.7M NFTs and more than 4K collections observed between 2017 and 2021. NFTs are grouped by the authors into six main categories, namely Art (18.46%), Collectible (28.85%), Games (47.21%), Metaverse (0.1%), Utility (0.17%), and Other (5.21%) — values within parenthesis are percent proportions. Each transaction in the dataset is associated with metadata that allows keeping track of the actors involved in the transaction and the main information associated with each NFT, such as its selling price, collection and category, and image URLs. After filtering out inaccessible NFTs or NFTs with less than one sell, our final dataset consisted in more than 180K NFT images of various formats (.png, .jpeg and .webp), and corresponding metadata.

Note also that the original dataset lacks information on minting times, but primary-selling timestamps are nonetheless available. We notice that resorting to the primary-selling times for our NFT visual inspiration networks ensures that the assets got at least one transaction, thus actually impacting on the NFT market.

A commented screen recording of the demo of SONAR applied on the data mentioned above is provided with this work.[4]

12.6 Path to Impact

Our work finds its motivations in the opportunity of interactively exploring inspiration flows in the NFT landscape, which heavily depend on creativity, and are paramount to both creators and investors. However, it is tough to discern such phenomena due to the amount of (noisy) data involved.

In this regard, SONAR contributes to raise the bar by filling up a key gap for the evolution of Web3 and pursuing two complementary ambitious goals involving both creators and investors: (i) to detect potential phenomena of plagiarism on different of granularity (unlike OpenSea,[5] which limits this capability to collections), and (ii) to spot potential profitable investments to be made. Coupled with the multimodal exploration capabilities, SONAR can lead to unprecedented opportunities in the NFT landscape, which is now ready to expand its horizons to new frontiers.[6]

[4] https://drive.google.com/drive/folders/1_UfonabB_Siq0umP5JQ7VpeabGarvYj7

[5] https://opensea.io/

[6] https://cointelegraph.com/news/the-world-s-cultural-heritage-is-being-preserved-one-nft-at-a-time

13

A comprehensive collection of Non-Fungible Token transactions and metadata

Summary. Non-Fungible Tokens (NFTs) have emerged as the most representative application of blockchain technology in recent years, fostering the development of the Web3. Nonetheless, while the interest in NFTs rapidly boomed, creating unprecedented fervor in traders and creators, the demand for highly representative and up-to-date data to shed light on such an intriguing yet complex domain mostly remained unmet. To pursue this objective, we introduce a large collection of NFT transactions and associated metadata that correspond to trading operations between 2021 and 2023. Our developed dataset is the most extensive and representative in the NFT landscape to date, as it contains more than 70M transactions performed by more than 6M users across 36.3M NFTs and 281K collections. Moreover, this dataset boasts a wealth of metadata, including textual descriptions and URLs to multimedia content, thus being suitable for a plethora of tasks relevant to database systems, AI, data science, Web and network science fields. This dataset represents a unique resource for researchers and industry practitioners to delve into the inner workings of NFTs through a multitude of perspectives, paving the way for unprecedented opportunities across multiple research fields.

13.1 Value of the data

- To the best of our knowledge, the dataset provided in this work [56] represents to date the largest and most up-to-date publicly available source of information involving NFT transactions collected from the OpenSea platform, the leading and most extensively adopted trading platform in the Web3 ecosystem. By encompassing both bull and bear markets (i.e., the period of significant growth and prosperity, and the subsequent challenging phase) that characterized the NFT landscape in recent years, this valuable collection of data provides a holistic view of the market dynamics.
- This dataset represents a valuable contribution to the scientific community interested in Web3 and Non-Fungible Tokens, thus saving time in granting access to APIs for gathering "from-scratch" NFT transactions from trading platforms, which is typically an intricate and noisy task. Indeed, the APIs provided to date by the major NFT trading platforms are particularly resource and time-consuming, due to the small amount of data retrievable with a single call. Also, such APIs exhibit a recurrent lack of stability, further compounding the challenges of efficient data collection.

- The dataset's uniqueness stems from its inherent feature of encompassing a wide array of information describing the NFT landscape. Given the availability of millions of transactions rich in metadata, the dataset is designed to support tasks of sequential/transactional data processing and analysis (e.g., time-series similarity search) but also to empower graph-based modeling of the complex relationships among traders, thereby fostering a thorough understanding of the domain's intricacies. Additionally, the availability of textual descriptions and URLs containing multimedia resources, such as the artworks of NFTs, further amplifies the modeling capabilities, serving as a unique and comprehensive multimodal resource for delving deeper into the NFT realm through multiple perspectives.
- This dataset is conceived to serve as a catalyst for the work of scholars and practitioners pertaining to a wide variety of research disciplines and application domains, ranging from economy and sociology to cybersecurity.
- This dataset can be used as a benchmark for several innovative and impactful tasks unraveling the complexity of the crypto economy, ranging from NFT price projection to fraudulent and wash trading activities in Web3. Also, the multimodal nature of the dataset enables the development of textual and visual generative models and provides the foundations to design frameworks for forecasting the financial trajectories and performances of NFTs.

13.2 Logical organization of the data

Non-Fungible Tokens (NFTs) represent today a prominent application of blockchain technology, pioneering the emergence of the so-called Web3. NFTs define deeds of ownership, based on blockchain technology and smart contracts, of unique crypto assets on digital art forms, such as artworks or collectibles. These assets are regulated through smart contracts, i.e., pieces of code (typically coded in Solidity) that leverage the blockchain to ensure their trading adheres to predefined rules in the crypto market. The potential of this technology immediately attracted many digital creators and investors, as in the case of the Cryptokitties collection, a 2017 project based on the Ethereum blockchain, which caused significant congestion of the underlying network for days. To date, the NFT market boasts a remarkable market capitalization of more than $6 billion USD, making it an important case study in the novel and evolving crypto economy. It is hence crucial to manage and exploit the abundance of transactions and associated metadata to gain deeper insights into this ecosystem, and foster research works aimed at shedding light on the main characteristics of such an intriguing domain [55, 137, 138].

Within this view, in this work, we introduce the most representative and largest dataset involving NFT transactions to date, encompassing more than 70M NFT unique transactions from the OpenSea market [1], executed by more than 6M users across 36.3M NFTs and 281K collections. Moreover, our dataset includes valuable metadata associated with these transactions, comprising more than 281K images and descriptions. While we also considered other marketplaces, such as Blur and LooksRare, we opted for OpenSea due to its position as the platform with the highest sales volume and the largest number of traders within the entire NFT ecosystem (cf. Table 13.2), making it particularly suitable for our work. Furthermore, our choice was also influenced by the availability of well-documented APIs, which is paramount for gathering consistent and reliable data.

Furthermore, our work seeks to address the gap in publicly available data that accurately represent that domain. Indeed, to the best of our knowledge, the only publicly available dataset

[1] `https://opensea.io/`

to date involving NFTs is the one from Nadini et al. [168], which covers 6.1M transactions and more than 4.7M NFTs (coming from 4 different marketplaces) traded between June 2017 and April 2021. It also includes transactional metadata, like NFT images and descriptions, making it a valuable source for multi-modal information. However, our newly proposed dataset significantly differs from the one in [168]. First, we exceed the latter up to an order of magnitude both in terms of transactions and users involved. Furthermore, we offer a more comprehensive perspective on the NFT domain by including a broader range of metadata and handling various token types. Finally and remarkably, the uniqueness of our proposed dataset lies in the temporal window of 3 years (from January 2021 to August 2023), which makes it the only one to date that can cover all oscillatory phenomena (e.g., bull and bear runs) and major patterns that have characterized the NFT market from its early days to the present.

Logical organization of the data. All data are locally stored in multiple gzip files such that each of the files is formatted as a list of multiline JSON, where each line corresponds to an API call. Each first-level entry in a line corresponds to a record or transaction. The variables describing the transactions are reported in Table 13.1.

Anonymization of the data. In order to comply with the requirements of preserving privacy and to prevent any breaches of data protection regulations and property rights, we have undertaken a series of measures, as detailed in Table 13.1:

- We have retained the values of all variables that pertain to non-sensitive information in their original, unaltered form;
- The values of variables associated with sensitive information have been subjected to anonymization, employing a one-way, irreversible method to ensure that the data cannot be reverted to its original state;
- URLs referencing image data (NFT images) and textual contents (NFT image descriptions) have been substituted with identifiers that point to numerical vectors corresponding to an encoded representation of the images or texts. These representations, or embeddings, have been generated using neural network models and are provided in lieu of the original image and text data.

Table 13.3 summarizes the main characteristics of the collected data, whereas Figures 13.1, 13.2 and 13.3 show relations occurring for various features over our 3-year collecting period. Moreover, Figure 13.4 shows the distribution of the overall volume traded across the different NFT standards and blockchains.

13.3 Experimental design, materials and methods

We obtained data by fetching them from the OpenSea platform using their official APIs. To achieve this, we were granted an API key to perform authenticated calls to the specific endpoints of interest. Our dataset construction followed a multi-phase pipeline, as shown in Figure 6.

Phase 1. We started collecting data by using the events endpoint available in the OpenSea APIs (version 1), which provides details on events occurring on NFTs tracked by the platform. The events endpoint is accessible through authenticated HTTP GET requests, and therefore we set the API Key as the header of each request. We used the requests Python library to perform such API calls. For each call, we exploited the set of parameters reported in Table 13.4. Due to the presence of intricate nested data structures in each retrieved event, we generated raw data chunks, each of which consisted of a maximum of 500 payloads sourced from API calls.

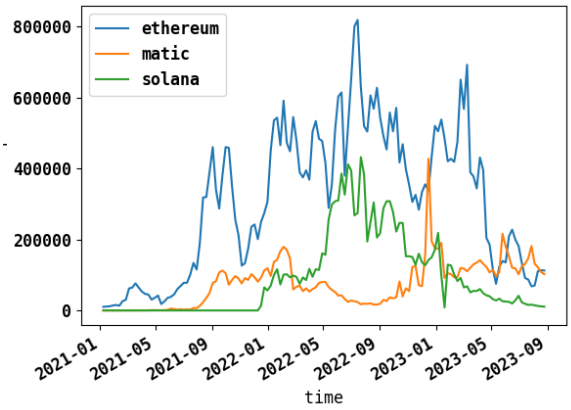


Fig. 13.1. Weekly number of transactions per chain during the three-year collecting period

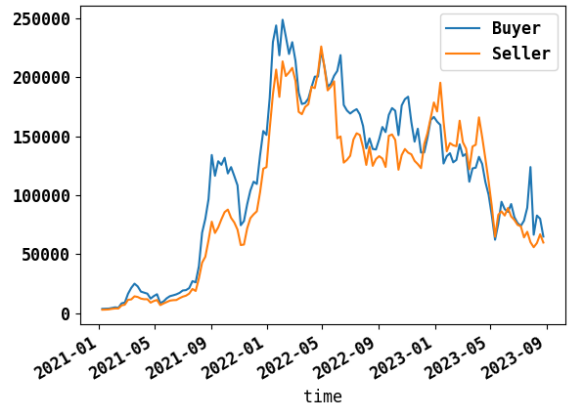


Fig. 13.2. Number of weekly active users by role during the three-year collecting period

We hence stored each chunk as a JSON file. Specifically, during this process, we resorted to the gzip compression format to efficiently condense each chunk, thus achieving a remarkable 12x reduction in memory usage. As a byproduct, the subdivision into chunks of our raw data allowed us to capitalize on soft restart capabilities in the case of network issues or unexpected errors. All these operations were executed through Python programs.

Phase 2. After retrieving all transactions relevant to our query and gathering the corresponding raw data chunks, we performed specific data processing steps aimed at filtering out noisy, irrelevant, and inconsistent metadata. Eventually, we integrated the processed yet unmodified information into a MongoDB database instance. Throughout this loading stage, checks for duplicate insertions were conducted to ensure the integrity of our dataset. The choice of

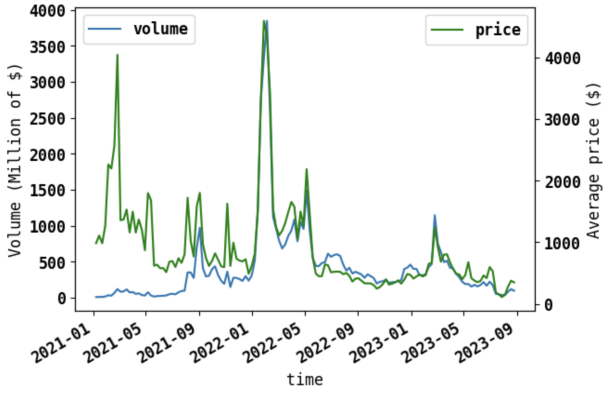


Fig. 13.3. Weekly volume traded (blue) and average price per token (green) during the three-year collecting period

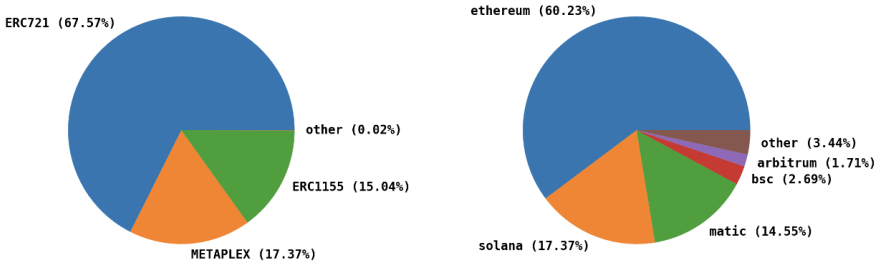


Fig. 13.4. Distribution of the overall volume traded across the different NFT standards (on the left) and blockchains (on the right)

employing a NoSQL database in this phase was driven by the advantages it offers in terms of querying capabilities, as well as efficient handling and manipulation of large-scale and semi-structured data.

Phase 3. We hence prepared our data for being shared with the researchers' community. To this aim, we first exported the data stored in MongoDB into a CSV file for ease of processing. We then leveraged our neural data anonymization module to remove any potentially sensitive data from the original dataset (cf. Anonymization of the data in Section "Data Description"). During this step, we also created textual and visual embeddings of the corresponding raw textual content and image URLs. Our final dataset and the companion embeddings, overall weighing more than 100GB, were eventually released on the HuggingFace platform through the Datasets library, which allows us to efficiently split large datasets in chunks for ease of sharing. In the last few years, being a major reference to freely share neural language/vision models, HuggingFace has attracted remarkable attention and popularity from researchers and practitioners in machine/deep learning related fields, thus ensuring high visibility and ease of access to any data stored in its repositories.

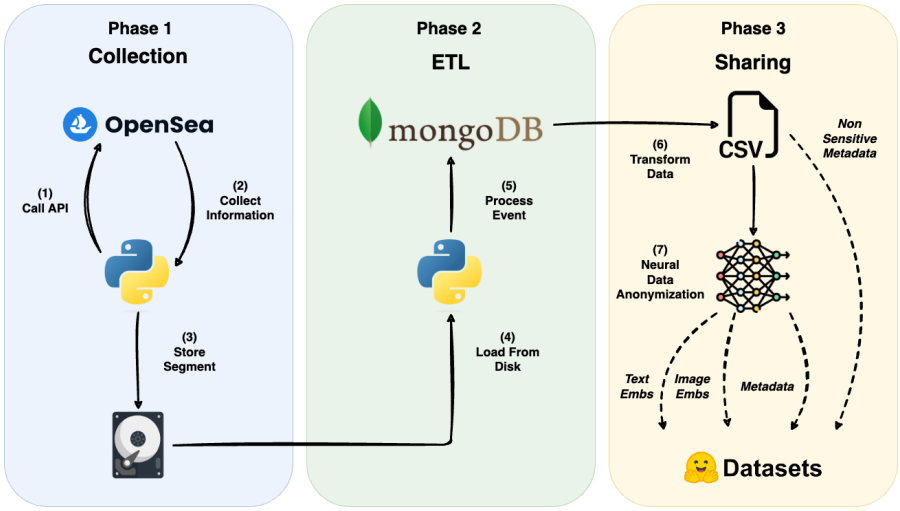


Fig. 13.5. Overview of the development pipeline for the proposed dataset

13.4 Limitations

While the OpenSea platform monitors various events related to NFT dynamics, including completing sales, creating auctions, entering or withdrawing bids, as well as transferring tokens, this work focuses on successful sales (i.e., event type successful). This choice is all but random, as this type of event always entails a transfer of funds between wallets; however, other event types, such as transfer (i.e., transfer of token without currency exchange), can in principle offer valuable insights into the NFT ecosystem. Therefore, gathering data associated with other event types might be considered as a further, complementary development of our presented dataset. Furthermore, our work is biased toward the OpenSea platform, which is currently recognized as the most relevant and adopted NFT trading platform; however, the evolving nature of the NFT landscape fosters the continual emergence of new platforms, such as Blur and LooksRare. Considering such platforms in future data collections would help maintain a comprehensive perspective on the NFT market.

Table 13.1. Detailed description of the variables contained in each transaction of the dataset

Variable	Type	Description	Processing Notes
token_id	String	The id of the NFT — this value is unique within the same collection	Anonymized (hash-codes)
num_sales	Integer	A progressive integer indicating the number of successful transactions involving the NFT up to the current timestamp (cf. tx_timestamp)	Original (not sensitive variable)
nft_name	Vector ID	The name of the NFT	Anonymized (textual embedding)
nft_description	Vector ID	The description of the NFT as provided by the creator	Anonymized (textual embedding)
nft_image	Vector ID	The ID for accessing the NFT image vector	Anonymized (visual embedding)
collection_name	Vector ID	The ID for accessing the Collection name vector	Anonymized (textual embedding)
collection_description	Vector ID	The ID for accessing the Collection description vector	Anonymized (textual embedding)
collection_image	Vector ID	The ID for accessing the Collection image vector	Anonymized (visual embedding)
fees_seller	Float	The absolute amount of fees the seller has gained from this transaction expressed in token	Original
fees_opensea	Float	The absolute amount of fees OpenSea has gained from this transaction expressed in token	Original
fees_seller_usd	Float	The absolute amount of fees the seller has gained from this transaction expressed in US dollars (USD)	Original
fees_opensea_usd	Float	The absolute amount of fees OpenSea has gained from this transaction expressed in US dollars (USD)	Original
payout_collection_address	String	The wallet address where seller fees are deposited	Anonymized (hash-codes)
tx_timestamp	String	Timestamp of the transaction expressed in yyyy-mm-ddTHH:MM:SS	Original
price	Float	The price of the transaction expressed in token	Original
gain	Float	The gain after fees (i.e., $gain = price - fees\_opensea * price - fees\_seller * price$)	Original
usd_price	Float	The price of the transaction expressed in US dollars (USD)	Original
usd_gain	Float	The difference between the price and the fees expressed in US dollars (USD)	Original
token	Categorical	The token type used to pay the transaction	Original
to_eth	Float	The conversion rate to convert tokens into Ethereum at the current timestamp, such that $eth = price * to\_eth$	Original
to_usd	Float	The conversion rate to convert tokens into US dollars (USD) at the current timestamp, such that $usd = price * to\_usd$	Original
from_account	String	The address that sends the payment (i.e., winner/buyer)	Anonymized (hash-codes)
to_account	String	The address that receives the payment (it often corresponds to the contract linked to the asset)	Anonymized (hash-codes)
seller_account	String	The address of the NFT seller	Anonymized (hash-codes)
winner_account	String	The address of the NFT buyer	Anonymized (hash-codes)
contract_address	String	The contract address on the blockchain	Anonymized (hash-codes)
created_date	Timestamp	The date of creation of the contract	Original
chain	Categorical	The blockchain where the transaction occurs	Original
token_type	Categorical	The schema of the token, i.e., ERC721 or ERC1155	Original
asset_contract_type	Categorical	The asset typology, i.e., non-fungible or semi-fungible	Original
asset_type	Categorical	Whether the asset was involved in a simple or bundle transaction	Original

Table 13.2. Statistics of the main NFT marketplaces according to DappRadar. Bold values correspond to highest scores.

Marketplace	Avg. Price	Traders	Sales	Volume
Opensea	$556	**4.90M**	**64.95M**	**$36.10B**
Blur	$1490	312.18K	4.01M	$6.43B
LooksRare	**$8920**	153K	396.10K	$4.85B
Axie Marketplace	$161	2.24M	20.53M	$4.28B

Table 13.3. Main descriptive statistics of the data

Number of transactions	70,972,143
Volume	$58,684,058,679
Most priced NFT	$223,101,250
Unique sellers	3,471,834
Unique buyers	4,793,818
Unique Actors	6,306,461
Number of collections	281,152
Number of NFTs	36,277,260
Average transaction price	$826.86

Table 13.4. Set of parameters used to setup the OpenSea API calls

Query Param	Description	Value Set
cursor	Unique identifier used to keep track of and link multiple API calls within the same session, thus enabling a chronologically ordered download of all events matching the API request	Each API call returns a new cursor pointing to the next page to retrieve during pagination
event_type	The type of event to retrieve during the API call	successful, since we want to retrieve only successful transactions, i.e., those leading to currency exchange
occurred_before	Timestamp variable to retrieve only events listed before it; it is expressed in seconds since the Unix epoch	31/08/2023
occurred_after	Timestamp variable to retrieve only events listed after it; it is expressed in seconds since the Unix epoch	01/01/2021
limit	The number of transactions to retrieve after each API call	150; to avoid incurring in rate-limit penalizations and to be polite, we set a 2-second sleep after each API call

Conclusions and Future Work

"The important thing is not to stop questioning.
Curiosity has its own reason for existing."
– Albert Einstein

Concluding remarks. This thesis has focused on unraveling the intrinsic complexity of emerging decentralized socio-economic and high societal impact domains. This goal was pursued through the lens of *graph mining* and *multimodal representation learning*, which allowed shedding light on latent and fascinating phenomena that unfold within these domains.

The first part of this work delved into the emerging landscape of *Decentralized Online Social Networks*. It provided the first yet comprehensive analysis of the network of Mastodon instances through the design and study of the largest and most up-to-date dataset existing to date. We unveiled the fingerprint of Mastodon, i.e., the set of distinctive traits that differentiate it from traditional Online Social Networks. Besides, we analyzed the Mastodon user network to delve into user relationships and behaviors within this innovative decentralized context, revealing the key characteristics of user connections and the influence of prominent instances. The investigation extended to strategic phenomena, including across-instance boundary spanning, over-consumption, and information flow, as well as user roles in DOSNs, such as bridges, dual, and alternate role users. Finally, we characterized the drivers of social influence that led to the mass migration towards Mastodon in late 2022, marking one of the largest social migrations in Internet history to date.

In the second part of this work, our focus shifted to graph mining for *high societal impact domains*, whose study and understanding have a remarkable influence on our lives. We started by adapting a correlation clustering method to a *fairness-aware* context, demonstrating that our proposed approach produces high-quality clustering solutions while also accounting for fairness aspects. Subsequently, we took an unprecedented perspective on the Italian Civil Code through network analysis. Specifically, besides shedding light on the main patterns underlying the intricacies of *law reference networks*, we developed a web-based tool designed for modeling, analyzing, and visualizing such networks. Our goal is to support legal research tasks and help legal professionals and citizens in navigating law corpora. Finally, we exploited the social traces people left on social media platforms to investigate the *social debate on climate change*. This involved uncovering and characterizing the main topics discussed around the Conferences of the Parties, as well as identifying the most relevant actors involved in such a debate.

The third and last part of this thesis involved the realm of *Non-Fungible Tokens (NFTs)*. We introduced `MERLIN`, an innovative bimodal deep-learning framework designed to train Transformer-based language and visual models, along with graph neural network models, on

collections of NFT images and texts. Through `MERLIN` models, we then carried out financially-agnostic price-category classification tasks, achieving remarkable performances. Besides, we explored the role of inspiration in Non-Fungible Tokens, unveiling that the widespread occurrence of inspiration resulted in a temporary saturation of the visual feature space. Our investigation extended to the dichotomy between inspiring and inspired NFTs and its impact on financial performance. Moreover, we developed SONAR, a web-based tool facilitating the multimodal exploration of NFT inspiration networks. This tool allows interactive visualization of inspiration-driven connections between NFTs, aiming at supporting investors and creators in uncovering new investment opportunities and anomalous inspirations. Finally, we curated the most extensive and representative feature-rich and multimodal dataset related to NFTs to date, making it available to the research community.

The synergy between graph mining and multimodal deep learning demonstrated its effectiveness and adequacy for the goals of this thesis. This combination facilitated the identification and characterization of latent phenomena, enabling the development of tools capable of improving the evolution of the considered domains.

Future research. Despite the body of contributions presented in this thesis, the journey into the complexities of such emerging domains has just begun. Given the continuous metamorphosis of these domains, future research entails addressing many open challenges, discovering new phenomena, and defining new frameworks and approaches.

In the context of Decentralized Online Social Networks, content moderation faces new challenges due to the distributed nature of instances, introducing additional complexity in controlling information flows. Besides, the growing influx of users into DOSNs requires investigating phenomena such as polarization and segregation, to define and implement countermeasures aimed at mitigating these issues.

Concerning the high societal impact domains investigated in the second part of this work, ongoing investigation into critical issues is essential. In this regard, identifying the gaps that require attention, and proposing new approaches that can improve their development is crucial, considering the tangible effects they have on our society. Key aspects of future work also involve instilling the concept of fairness in machine and deep learning approaches, fostering a better symbiosis between artificial intelligence and the legal domain, and raising awareness of climate crisis issues.

Finally, although the so-called Web3 is now an integral part of our society, numerous challenges persist. These include detecting anomalies in the exchange of digital assets (e.g., wash trading and other frauds), which remains a difficult task given the unregulated nature of this new economy. Besides, it is important to develop new modeling forms able to uncover hidden features that have remained latent so far and design new frameworks capable of harnessing all the opportunities given by the establishment of Web3, while also ensuring a proper and rapid evolution.

References

1. H. Abdi. The Kendall Rank Correlation Coefficient. In *Encyclopedia of Measurement and Statistics*. 2007.
2. Savitha Sam Abraham, Deepak P, and Sowmya S. Sundaram. Fairness in clustering with multiple sensitive attributes. In *Proc. EDBT Conf.*, pages 287–298, 2020.
3. Tommaso Agnoloni and Ugo Pagallo. The Case Law of the Italian Constitutional Court, Its Power Laws, and the Web of Scholarly Opinions. In *Proc. of the 15th International Conference on Artificial Intelligence and Law*, pages 151–155. Association for Computing Machinery, 2015.
4. Sara Ahmadian, Alessandro Epasto, Marina Knittel, Ravi Kumar, Mohammad Mahdian, Benjamin Moseley, Philip Pham, Sergei Vassilvitskii, and Yuyan Wang. Fair hierarchical clustering. In *Proc. NIPS Conf.*, 2020.
5. Sara Ahmadian, Alessandro Epasto, Ravi Kumar, and Mohammad Mahdian. Fair correlation clustering. In *Proc. AISTATS Conf.*, pages 4195–4205, 2020.
6. Luca Maria Aiello, Sagar Joglekar, and Daniele Quercia. Multidimensional tie strength and economic development. *Scientific Reports*, 12(1):22081, 2022.
7. N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. In *Proc. ACM STOC Symp.*, pages 684–693, 2005.
8. N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: Ranking and clustering. *JACM*, 55(5):23:1–23:27, 2008.
9. Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Understanding the effect of deplatforming on social networks. In *13th acm web science conference 2021*, pages 187–195, 2021.
10. Ishaku Hassan Anaobi, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Damilola Ibosiola, and Gareth Tyson. Will admins cope? decentralized moderation in the fediverse. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 3109–3120, New York, NY, USA, 2023. Association for Computing Machinery.
11. Giulio Anselmi and Giovanni Petrella. Non-fungible token artworks: More crypto than art? *Finance Research Letters*, 51:103473, 2023.
12. Lennart Ante. The non-fungible token (nft) market and its relationship with bitcoin and ethereum. *FinTech*, 1(3):216–224, 2022.
13. Simona Andreea Apostu, Mirela Panait, Làszló Vasa, Constanta Mihaescu, and Zbyslaw Dobrowolski. Nfts and cryptocurrencies — the metamorphosis of the economy under the sign of blockchain: A time series approach. *Mathematics*, 10(17), 2022.

14. Cheick Tidiane Ba, Andrea Michienzi, Barbara Guidi, Matteo Zignani, Laura Ricci, and Sabrina Gaito. Fork-based user migration in blockchain online social media. In *14th ACM Web Science Conference 2022*, pages 174–184, 2022.
15. Cheick Tidiane Ba, Matteo Zignani, and Sabrina Gaito. Social and Rewarding Microscopical Dynamics in Blockchain-Based Online Social Networks. In *Proc. International Conference on Information Technology for Social Good (GoodIT)*, page 127–132, 2021.
16. Cheick Tidiane Ba, Matteo Zignani, and Sabrina Gaito. The role of groups in a user migration across blockchain-based online social media. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 291–296. IEEE, 2022.
17. Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable fair clustering. In *Proc. ICML Conf.*, pages 405–413, 2019.
18. Norman TJ Bailey et al. *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company, 1975.
19. Duilio Balsamo, Paolo Bajardi, Gianmarco De Francisci Morales, Corrado Monti, and Rossano Schifanella. The pursuit of peer support for opioid use recovery on reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 2023.
20. N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Mach. Learn.*, 56(1):89–113, 2004.
21. Andrea Baronchelli. The emergence of consensus: a primer. *Royal Society open science*, 5(2):172189, 2018.
22. Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An Open Source Software for Exploring and Manipulating Networks. In *Proc. of the Third International Conference on Weblogs and Social Media (ICWSM)*. The AAAI Press, 2009.
23. Suman Kalyan Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. In *Proc. NIPS Conf.*, pages 4955–4966, 2019.
24. Ioana Oriana Bercea, Martin Groß, Samir Khuller, Aounon Kumar, Clemens Rösner, Daniel R. Schmidt, and Melanie Schmidt. On the cost of essentially fair clusterings. In *Proc. APPROX/RANDOM Conf.*, pages 18:1–18:22, 2019.
25. Kelly Bergstrom and Nathaniel Poor. Reddit gaming communities during times of transition. *Social Media+ Society*, 7(2):20563051211010167, 2021.
26. Kelly Bergstrom and Nathaniel Poor. Signaling the intent to change online communities: A case from a reddit gaming community. *Social Media+ Society*, 8(2):20563051221096817, 2022.
27. Ames Bielenberg, Lara Helm, Anthony Gentilucci, Dan Stefanescu, and Honggang Zhang. The growth of Diaspora - A decentralized online social network in the wild. In *Proc. IEEE INFOCOM Workshops*, pages 13–18, 2012.
28. Haris Bin Zia, Aravindh Raman, Ignacio Castro, Ishaku Hassan Anaobi, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. Toxicity in the decentralized web and the potential for model sharing. *Proc. ACM Meas. Anal. Comput. Syst.*, 6(2), jun 2022.
29. Peter M Blau. Social exchange. *International encyclopedia of the social sciences*, 7(4):452–457, 1968.
30. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10:P10008, 2008.
31. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

32. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Report*, 424:175–308, 2006.
33. Gianluca Bonifazi, Francesco Cauteruccio, Enrico Corradini, Michele Marchetti, Daniele Montella, Simone Scarponi, Domenico Ursino, and Luca Virgili. Performing wash trading on nfts: Is the game worth the candle? *Big Data and Cognitive Computing*, 7(1), 2023.
34. Paul Boniol, George Panagopoulos, Christos Xypolopoulos, Rajaa El Hamdani, David Restrepo Amariles, and Michalis Vazirgiannis. Performance in the Courtroom: Automated Processing and Visualization of Appeal Court Decisions in France. In *Proc. of the Natural Legal Language Processing Workshop 2020 co-located with the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, volume 2645 of *CEUR Workshop Proceedings*, pages 11–17. CEUR-WS.org, 2020.
35. Thomas Brewster. Elon musk gives twitter staff 2 days to decide if they want to stay. *Forbes*, 2023.
36. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.
37. Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *Proc. 10th Int. Conf. on Learning Representations (ICLR)*. OpenReview.net, 2022.
38. Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. of the 27th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '04, pages 25–32, 2004.
39. Antonio Caliò and Andrea Tagarelli. Attribute based diversification of seeds for targeted influence maximization. *Inf. Sci.*, 546:1273–1305, 2021.
40. Antonio Caliò, Andrea Tagarelli, and Francesco Bonchi. Cores matter? an analysis of graph decomposition effects on influence maximization problems. In *12th ACM Conference on Web Science*, WebSci '20, page 184–193, New York, NY, USA, 2020. Association for Computing Machinery.
41. Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 9630–9640. IEEE, 2021.
42. Damon Centola, Joshua Becker, Devon Brackbill, and Andrea Baronchelli. Experimental evidence for tipping points in social convention. *Science*, 360(6393), 2018.
43. Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. Multi-task dialog act and sentiment recognition on Mastodon. In *Proc. COLING Conf.*, pages 745–754, 2018.
44. Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: the muppets straight out of law school. *CoRR*, abs/2010.02559, 2020.
45. M. Charikar, V. Guruswami, and A. Wirth. Clustering with qualitative information. In *Proc. IEEE FOCS Symp.*, pages 524–533, 2003.
46. M. Charikar, V. Guruswami, and A. Wirth. Clustering with qualitative information. *JCSS*, 71(3):360–383, 2005.
47. Shuchi Chawla, Konstantin Makarychev, Tselil Schramm, and Grigory Yaroslavtsev. Near optimal LP rounding algorithm for correlation clustering on complete and complete k-partite graphs. In *Proc. ACM STOC Symp.*, pages 219–228, 2015.
48. Anshuman Chhabra, Karina Masalkovaitė, and Prasant Mohapatra. An overview of fairness in clustering. *IEEE Access*, 9:130698–130720, 2021.
49. Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *Proc. NIPS Conf.*, pages 5029–5037, 2017.

50. Minje Choi, Luca Maria Aiello, Krisztián Zsolt Varga, and Daniele Quercia. Ten social dimensions of conversations and relationships. In *Proceedings of The Web Conference 2020*, WWW '20, page 1514–1525, New York, NY, USA, 2020. Association for Computing Machinery.
51. Nicholas A Christakis and James H Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine*, 32(4):556–577, 2013.
52. Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *Scientific reports*, 10(1), 2020.
53. Emily M Cody, Andrew J Reagan, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M Danforth. Climate change sentiment on twitter: An unsolicited public opinion poll. *PloS one*, 10(8):e0136092, 2015.
54. Giovanni Colavizza. Seller-buyer networks in NFT art are driven by preferential ties. *CoRR*, abs/2210.04339, 2022.
55. Davide Costa, Lucio La Cava, and Andrea Tagarelli. Show me your NFT and I tell you how it will perform: Multimodal representation learning for NFT selling price prediction. In *Proc. of the ACM Web Conference 2023 (WWW '23), May 1–5, 2023, Austin, TX, USA*. Association for Computing Machinery, 2023.
56. Davide Costa, Lucio La Cava, and Andrea Tagarelli. Unraveling the nft economy: A comprehensive collection of non-fungible token transactions and metadata. *Data in Brief*, page 109749, 2023.
57. Ademar Crotti Junior, Fabrizio Orlandi, Damien Graux, Murhaf Hossari, Declan O'Sullivan, Christian Hartz, and Christian Dirschl. Knowledge Graph-Based Legal Search over German Court Cases. In *Proc. of the Semantic Web Conference: ESWC 2020 Satellite Events*, pages 293–297, 2020.
58. Faraz Dadgostari, Mauricio Guim, P. Beling, Michael A. Livermore, and D. Rockmore. Modeling law search as prediction. *Artificial Intelligence and Law*, 29(1):3–34, 2021.
59. Biraj Dahal, Sathish AP Kumar, and Zhenlong Li. Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining*, 9:1–20, 2019.
60. Anwitaman Datta, Sonja Buchegger, Le-Hung Vu, Thorsten Strufe, and Krzysztof Rzadca. *Decentralized Online Social Networks*, pages 349–378. Springer US, Boston, MA, 2010.
61. Anwitaman Datta, Sonja Buchegger, Le-Hung Vu, Thorsten Strufe, and Krzysztof Rzadca. Decentralized online social networks. In Borko Furht, editor, *Handbook of Social Network Technologies and Applications*. Springer, 2010.
62. E. D. Demaine, D. Emanuel, A. Fiat, and N. Immorlica. Correlation clustering in general weighted graphs. *TCS*, 361(2-3):172–187, 2006.
63. Sebastian Deri, Jeremie Rappaz, Luca Maria Aiello, and Daniele Quercia. Coloring in the links: Capturing social ties as they are perceived. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), nov 2018.
64. Valerian J Derlega and Janusz Grzelak. *Cooperation and helping behavior: Theories and research*. Academic press, 2013.
65. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019.
66. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics, 2019.
67. Navid Dianati. Unwinding the hairball graph: Pruning algorithms for weighted complex networks. *Physical Review E*, 93:012304, 2016.
68. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. 9th Int. Conf. on Learning Representations (ICLR)*. OpenReview.net, 2021.
69. Michael Dowling. Fertile land: Pricing non-fungible tokens. *Finance Research Letters*, 44:102096, 2022.
70. Michael Dowling. Is non-fungible token pricing driven by cryptocurrencies? *Finance Research Letters*, 44:102097, 2022.
71. Riley E Dunlap, Aaron M McCright, and Jerrod H Yarosh. The political divide on climate change: Partisan polarization widens in the us. *Environment: Science and Policy for Sustainable Development*, 58(5):4–23, 2016.
72. N. Edelmann. Reviewing the definitions of "lurkers" and some implications for online research. *Cyberpsychology, Behavior, and Social Networking*, 16(9):645–649, 2013.
73. Emory James Edwards and Tom Boellstorff. Migration, non-use, and the 'tumblrpocalypse': Towards a unified theory of digital exodus. *Media, Culture & Society*, 43(3):582–592, 2021.
74. Dimitrios Effrosynidis, Georgios Sylaios, and Avi Arampatzis. Exploring climate change on twitter using seven aspects: Stance, sentiment, aggressiveness, temperature, gender, topics, and disasters. *Plos one*, 17(9):e0274213, 2022.
75. Roman Egger and Joanne Yu. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7:886498, 2022.
76. R. Fagin, R. Kumar, and D. Sivakumar. Comparing Top k Lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003.
77. Max Falkenberg, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociocchi, and Andrea Baronchelli. Growing polarization around climate change on social media. *Nature Climate Change*, 12(12):1114–1121, 2022.
78. Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proc. ACM KDD Conf.*, pages 259–268, 2015.
79. Casey Fiesler and Brianna Dym. Moving across lands: Online platform migration in fandom communities. *Proceedings of the ACM on human-computer interaction*, 4(CSCW1):1–25, 2020.
80. Erwin Filtz. Building and processing a knowledge-graph for legal data. In *Proc. of the 16th International Semantic Web Conference*, pages 184–194. Springer International Publishing, 2017.
81. Erwin Filtz, Sabrina Kirrane, and Axel Polleres. The linked legal data landscape: linking legal data across different countries. *Artificial Intelligence and Law*, 29(4):485–539, 2021.
82. David N. Fisher, Matthew J. Silk, and Daniel W. Franks. The perceived assortativity of social networks: Methodological problems and solutions. *CoRR*, abs/1701.08671, 2017.
83. Marius Arved Fortagne and Bettina Lis. Determinants of the purchase intention of non-fungible token collectibles. *Journal of Consumer Behaviour*, 2023.

84. Allan Fowler and Johanna Pirker. Tokenification - the potential of non-fungible tokens (nft) for game development. In *Proc. Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '21, page 152–157, 2021.
85. James H. Fowler, Timothy R. Johnson, James F. Spriggs, Sangick Jeon, and Paul J. Wahlbeck. Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. *Political Analysis*, 15(3):324–346, 2007.
86. Alessia Galdeman, Matteo Zignani, and Sabrina Gaito. User migration across web3 online social networks: behaviors and influence of hubs. In *In Proceedings of IEEE International Conference on Communications*, 2023.
87. Natalie Gerhart and Mehrdad Koohikamali. Social network migration and anonymity expectations: What anonymous social network apps offer. *Computers in Human Behavior*, 95:101–113, 2019.
88. Carroll J Glynn and Michael E Huge. Opinions as norms: Applying a return potential model to the study of communication behaviors. *Communication Research*, 34(5):548–568, 2007.
89. Wei Gong, Ee-Peng Lim, and Feida Zhu. Characterizing silent users in social media communities. In *Proc. of the Int. Conf. on Web and Social Media (ICWSM)*, pages 140–149. AAAI Press, 2015.
90. P. Gonserkewitz, E. Karger, and M. Jagals. Non-fungible tokens: Use cases of nfts and future research agenda. *Risk Governance and Control: Financial Markets & Institutions*, 12(3):8–18, 2022.
91. Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.
92. Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR*, abs/2203.05794, 2022.
93. Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 855–864. ACM, 2016.
94. Barbara Guidi. When Blockchain meets Online Social Networks. *Pervasive and Mobile Computing*, 62:101131, 2020.
95. Barbara Guidi, Marco Conti, Andrea Passarella, and Laura Ricci. Managing social contents in decentralized online social networks: A survey. *Online Soc. Networks Media*, 7:12–29, 2018.
96. Barbara Guidi and Andrea Michienzi. Users and Bots behaviour analysis in Blockchain Social Media. In *Proc. International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–8, 2020.
97. Barbara Guidi and Andrea Michienzi. Sleepminting, the brand new frontier of non fungible tokens fraud. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, GoodIT '22, page 75–81, New York, NY, USA, 2022. Association for Computing Machinery.
98. Barbara Guidi and Andrea Michienzi. Delving nft vulnerabilities, a sleepminting prevention system. *Multimedia Tools and Applications*, 2023.
99. Barbara Guidi and Andrea Michienzi. From nft 1.0 to nft 2.0: A review of the evolution of non-fungible tokens. *Future Internet*, 15(6), 2023.
100. Barbara Guidi and Andrea Michienzi. The social impact of nfts in the metaverse economy. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, GoodIT '23, page 428–436, New York, NY, USA, 2023. Association for Computing Machinery.

101. Barbara Guidi, Andrea Michienzi, and Laura Ricci. A Graph-Based Socioeconomic Analysis of Steemit. *IEEE Transactions on Computational Social Systems*, 8(2):365–376, 2021.
102. Jürgen Habermas. *Communication and the Evolution of Society*. Beacon Press, 1979.
103. Jurgen Habermas. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT press, 1991.
104. Jürgen Habermas. *On the pragmatics of communication*. MIT press, 1998.
105. Jürgen Habermas. *On the pragmatics of social interaction: Preliminary studies in the theory of communicative action*. mit Press, 2001.
106. Martin S Hagger, Linda D Cameron, Kyra Hamilton, Nelli Hankonen, and Taru Lintunen. Changing behavior: A theory-and evidence-based approach. *Cambridge Handbooks in Psychology*, 2020.
107. Anaobi Ishaku Hassan, Aravindh Raman, Ignacio Castro, and Gareth Tyson. *The Impact of Capitol Hill on Pleroma: The Case for Decentralised Moderation*, page 1–2. 2021.
108. Anaobi Ishaku Hassan, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. Exploring Content Moderation in the Decentralised Web: The Pleroma Case. In *Proc. International Conference on Emerging Networking EXperiments and Technologies (CoNEXT)*, page 328–335, 2021.
109. Huiguo He, Tianfu Wang, Huan Yang, Jianlong Fu, Nicholas Jing Yuan, Jian Yin, Hongyang Chao, and Qi Zhang. Learning profitable nft image diffusions via multiple visual-policy guided reinforcement learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 6831–6840, New York, NY, USA, 2023. Association for Computing Machinery.
110. Daniel Hickey, Matheus Schmitz, Daniel Fessler, Paul Smaldino, Goran Muric, and Keith Burghardt. Auditing elon musk's impact on hate speech and bots. *Proceedings of the International AAAI Conference on Web and Social Media*, 2023.
111. Kin-Hon Ho, Yun Hou, Tse-Tin Chan, and Haoyuan Pan. Analysis of non-fungible token pricing factors with machine learning. In *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*, 10 2022.
112. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997.
113. Manoel Horta Ribeiro, Homa Hosseinmardi, Robert West, and Duncan J Watts. Deplatforming did not decrease parler users' activity on fringe social media. *PNAS nexus*, 2(3):pgad035, 2023.
114. Avus CY Hou and Wen-Lung Shiau. Understanding facebook to instagram migration: a push-pull migration model perspective. *Information Technology & People*, 33(1):272–295, 2020.
115. Candice Howarth, Peter Bryant, Adam Corner, Sam Fankhauser, Andy Gouldson, Lorraine Whitmarsh, and Rebecca Willis. Building a social mandate for climate action: Lessons from covid-19. *Environmental and Resource Economics*, 76:1107–1115, 2020.
116. Clayton J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proc. Int. Conf. on Weblogs and Social Media*, 2014.
117. Iacopo Iacopini, Giovanni Petri, Andrea Baronchelli, and Alain Barrat. Group interactions modulate critical mass dynamics in social convention. *Communications Physics*, 5(1):64, 2022.
118. S Mo Jang and P Sol Hart. Polarized frames on "climate change" and "global warming" across countries and states: Evidence from twitter big data. *Global environmental change*, 32:11–17, 2015.

119. Ujun Jeong, Paras Sheth, Anique Tahir, Faisal Alatawi, H Russell Bernard, and Huan Liu. Exploring platform migration patterns between twitter and mastodon: A user behavior study. *arXiv preprint arXiv:2305.09196*, 2023.
120. Annamma Joy, Ying Zhu, Camilo Peña, and Myriam Brouard. Digital future of luxury brands: Metaverse, digital fashion, and non-fungible tokens. *Strategic Change*, 31(3):337–343, 2022.
121. Ademar Crotti Junior, Fabrizio Orlandi, Declan O'Sullivan, Christian Dirschl, and Quentin Reul. Using Mapping Languages for Building Legal Knowledge Graphs from XML files. In *Proc. of the Blockchain enabled Semantic Web Workshop (BlockSW) and Contextualized Knowledge Graphs (CKG) Workshop co-located with the 18th International Semantic Web Conference*, volume 2599 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.
122. Kristina Kapanova, Barbara Guidi, Andrea Michienzi, and Kevin Koidl. Evaluating Posts on the Steemit Blockchain: Analysis on Topics Based on Textual Cues. In *Proc. International Conference on Smart Objects and Technologies for Social Good (GoodTechs)*, page 163–168, 2020.
123. Arnav Kapoor, Dipanwita Guhathakurta, Mehul Mathur, Rupanshu Yadav, Manish Gupta, and Ponnurangam Kumaraguru. Tweetboost: Influence of social media on nft valuation. In *Proc. the Web Conference 2022*, page 621–629. Association for Computing Machinery, 2022.
124. Christos Karapapas, Iakovos Pittaras, and George C. Polyzos. Fully decentralized trading games with evolvable characters using nfts and ipfs. In *2021 IFIP Networking Conference (IFIP Networking)*, pages 1–2, 2021.
125. Gurumurthy Kasinathan. Musk's twitter acquisition. *Economic & Political Weekly*, 58(2):21, 2023.
126. Julian Kauk, Helene Kreysa, and Stefan R. Schweinberger. Understanding and countering the spread of conspiracy theories in social networks: Evidence from epidemiological models of twitter data. *PLOS ONE*, 16(8):1–20, 08 2021.
127. David Kempe, Jon M. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proc. ACM SIGKDD*, pages 137–146, 2003.
128. Andrei P Kirilenko, Tatiana Molodtsova, and Svetlana O Stepchenkova. People as sensors: Mass media and local temperature influence climate change discussion on twitter. *Global Environmental Change*, 30:92–100, 2015.
129. M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, 2010.
130. Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2017.
131. Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. Fair k-center clustering for data summarization. In *Proc. ICML Conf.*, pages 3448–3457, 2019.
132. Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. Guarantees for spectral clustering with fairness constraints. In *Proc. ICML Conf.*, pages 3458–3467, 2019.
133. Peter Kollock. Transforming social dilemmas: group identity and co-operation. *Modeling rationality, morality, and evolution*, 7:185–209, 1998.
134. Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. Network analysis in the legal domain: a complex model for European Union legal sources. *Journal of Complex Networks*, 6(2):243–268, 08 2017.

135. Kristin Kostick-Quenet, Kenneth D. Mandl, Timo Minssen, I. Glenn Cohen, Urs Gasser, Isaac Kohane, and Amy L. McGuire. How nfts could transform health information exchange. *Science*, 375(6580):500–502, 2022.
136. Logan Kugler. Non-fungible tokens and the future of art. *Commun. ACM*, 64(9):19–20, aug 2021.
137. Lucio La Cava, Davide Costa, and Andrea Tagarelli. Sonar: Web-based tool for multimodal exploration of non-fungible token inspiration networks. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3200–3204, New York, NY, USA, 2023. Association for Computing Machinery.
138. Lucio La Cava, Davide Costa, and Andrea Tagarelli. Visually wired nfts: Exploring the role of inspiration in non-fungible tokens. *arXiv preprint arXiv:2303.17031*, 2023.
139. Lucio La Cava, Sergio Greco, and Andrea Tagarelli. Understanding the growth of the Fediverse through the lens of Mastodon. *Appl. Netw. Sci.*, 6(1):64, 2021.
140. Lucio La Cava, Sergio Greco, and Andrea Tagarelli. Information consumption and boundary spanning in decentralized online social networks: The case of mastodon users. *Online Social Networks and Media*, 30:100220, 2022.
141. Lucio La Cava, Sergio Greco, and Andrea Tagarelli. Network analysis of the information consumption-production dichotomy in mastodon user behaviors. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1378–1382, May 2022.
142. Renaud Lambiotte and Pietro Panzarasa. Communities, knowledge creation, and information diffusion. *Journal of informetrics*, 3(3):180–190, 2009.
143. Bibb Latané. The psychology of social impact. *American psychologist*, 36(4):343, 1981.
144. Bokwon Lee, Kyu-Min Lee, and Jae-Suk Yang. Network structure reveals patterns of legal complexity in human society: The case of the constitutional legal network. *PLOS ONE*, 14(1):1–15, 01 2019.
145. Nicola Lettieri, Antonio Altamura, Armando Faggiano, and Delfina Malandrino. A computational approach for the experimental study of eu case law: analysis and implementation. *Social Network Analysis and Mining*, 6(1):56, 2016.
146. Nicola Lettieri, Antonio Altamura, and Delfina Malandrino. The legal macroscope: Experimenting with visual legal analytics. *Information Visualization*, 16(4):332–345, 2017.
147. Chao Li and Balaji Palanisamy. Incentivized Blockchain-Based Social Media Platforms: A Case Study of Steemit. In *Proc. of the 10th ACM Conference on Web Science*, page 145–154, 2019.
148. Luoqiu Li, Zhen Bi, Hongbin Ye, Shumin Deng, Hui Chen, and Huaixiao Tou. Text-guided legal knowledge graph reasoning. In *Proc. of the China Conference on Knowledge Graph and Semantic Computing*, pages 27–39, 2021.
149. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.
150. Lorenzo Lucchini, Luca Maria Aiello, Laura Alessandretti, Gianmarco De Francisci Morales, Michele Starnini, and Andrea Baronchelli. From reddit to wall street: the role of committed minorities in financial collective action. *Royal Society Open Science*, 9(4):211488, 2022.
151. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. 2017.
152. Jacqueline M. Otala, Gillian Kurtic, Isabella Grasso, Yu Liu, Jeanna Matthews, and Golshan Madraki. Political polarization and platform migration: A study of parler and

twitter usage by united states of america congress members. In *Companion Proceedings of the Web Conference 2021*, pages 224–231, 2021.
153. Junling Ma. Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling*, 5, 2020.
154. Fragkiskos D. Malliaros, Christos Giatsidis, Apostolos N. Papadopoulos, and Michalis Vazirgiannis. The core decomposition of networks: theory, algorithms and applications. *VLDB J.*, 29(1):61–92, 2020.
155. Thomas W Malone and Kevin Crowston. The interdisciplinary study of coordination. *ACM Computing Surveys (CSUR)*, 26(1):87–119, 1994.
156. Domenico Mandaglio, Andrea Tagarelli, and Francesco Gullo. Correlation clustering with global weight bounds. In *Proc. ECML-PKDD Conf.*, pages 499–515, 2021.
157. Massimo Martorelli, Senerath Mudalige Don Alexis Chinthaka Jayatilake, and Gamage Upeksha Ganegoda. Involvement of machine learning tools in healthcare decision making. *J. Healthc. Eng.*, 2021.
158. Akib Mashrur, Wei Luo, Nayyar A Zaidi, and Antonio Robles-Kelly. Machine learning for financial risk management: A survey. *IEEE Access*, 8:203203–203223, 2020.
159. Pierre Mazzega, Danièle Bourcier, and Romain Boulet. The Network of French Legal Codes. In *Proc. of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 236–237, 2009.
160. Amin Mekacher, Alberto Bracci, Matthieu Nadini, Mauro Martino, Laura Alessandretti, Luca Maria Aiello, and Andrea Baronchelli. Heterogeneous rarity patterns drive price dynamics in nft collections. *Scientific Reports*, 12(1):13890, 2022.
161. Amin Mekacher, Max Falkenberg, and Andrea Baronchelli. The systemic impact of deplatforming on social media. *arXiv preprint arXiv:2303.11147*, 2023.
162. Marcelo Mendoza, Bárbara Poblete, and Ignacio Valderrama. Nowcasting earthquake damages with twitter. *EPJ Data Science*, 8(1):1–23, 2019.
163. Sarah C. Meyns and Fisnik Dalipi. What users tweet on nfts: Mining twitter to understand nft-related concerns using a topic modeling approach. *IEEE Access*, 10:117658–117680, 2022.
164. Corrado Monti, Luca Maria Aiello, Gianmarco De Francisci Morales, and Francesco Bonchi. The language of opinion change on social media under the lens of communicative action. *Scientific Reports*, 12(1):17920, 2022.
165. Corrado Monti, Matteo Cinelli, Carlo Valensise, Walter Quattrociocchi, and Michele Starnini. Online conspiracy communities are more resilient to deplatforming. *arXiv preprint arXiv:2303.12115*, 2023.
166. Kody Moodley, Pedro V Hernandez-Serrano, Amrapali J Zaveri, Marcel GH Schaper, Michel Dumontier, and Gijs Van Dijck. The case for a linked data research engine for legal scholars. *European Journal of Risk Regulation*, 11(1):70–93, 2020.
167. Markus Moser. and Mark Strembeck. An Analysis of Three Legal Citation Networks Derived from Austrian Supreme Court Decisions. In *Proc. of the 4th International Conference on Complexity, Future Information Systems and Risk*, pages 85–92, 2019.
168. Matthieu Nadini, Laura Alessandretti, Flavio Di Giacinto, Mauro Martino, Luca Maria Aiello, and Andrea Baronchelli. Mapping the nft revolution: market trends, trade networks, and visual features. *Scientific Reports*, 11(1):20902, 2021.
169. M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20), Oct 2002.
170. M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2), Feb 2003.
171. Matthew N. Nicholson, Brian C Keegan, and Casey Fiesler. Mastodon rules: Characterizing formal rules on popular mastodon instances. In *Companion Publication of the 2023*

Conference on Computer Supported Cooperative Work and Social Computing, CSCW '23 Companion, page 86–90, New York, NY, USA, 2023. Association for Computing Machinery.

172. Jukka-Pekka Onnela, Jari Saramäki, Jörkki Hyvönen, Gábor Szabó, M Argollo de Menezes, Kimmo Kaski, Albert-László Barabási, and János Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9(6):179–179, jun 2007.
173. Warren Pearce, Kim Holmberg, Iina Hellsten, and Brigitte Nerlich. Climate change on twitter: Topics, communities and conversations about the 2013 ipcc working group 1 report. *PloS one*, 9(4):e94785, 2014.
174. Konstantinos Pelechrinis, Xin Liu, Prashant Krishnamurthy, and Amy Babay. Spotting anomalous trades in nft markets: The case of nba topshot. *PLOS ONE*, 18(6):1–17, 06 2023.
175. Diego Perna, Roberto Interdonato, and Andrea Tagarelli. Identifying users with alternate behaviors of lurking and active participation in multilayer social networks. *IEEE Trans. Comput. Soc. Syst.*, 5(1):46–63, 2018.
176. Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: online learning of social representations. In *Proc. 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 701–710. ACM, 2014.
177. Christian Pinto-Gutiérrez, Sandra Gaitán, Diego Jaramillo, and Simón Velasquez. The nft hype: What draws attention to non-fungible tokens? *Mathematics*, 10(3), 2022.
178. Robert Plutchik. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984(197-219):2–4, 1984.
179. Aravindh Raman, Sagar Joglekar, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. Challenges in the Decentralised Web: The Mastodon Case. In *Proc. ACM IMC Conf.*, pages 217–229, 2019.
180. Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.
181. Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.
182. Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proc. 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990. Association for Computational Linguistics, 2019.
183. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.
184. Tal Ridnik, Emanuel Ben Baruch, Asaf Noy, and Lihi Zelnik. Imagenet-21k pretraining for the masses. In *Proc. Neural Information Processing Systems - Track on Datasets and Benchmarks*, 2021.
185. Georg David Ritterbusch and Malte Rolf Teichmann. Defining the metaverse: A systematic literature review. *IEEE Access*, 11:12368–12377, 2023.
186. Bella Robinson, Robert Power, and Mark Cameron. A sensitive twitter earthquake detector. In *Proc. ACM Conf. on World Wide Web*, WWW '13 Companion, page 999–1002, 2013.
187. Richard Rogers. Deplatforming: Following extreme internet celebrities to telegram and alternative social media. *European Journal of Communication*, 35(3):213–229, 2020.

188. Björn Rönnerstrand and Karolina Andersson Sundell. Trust, reciprocity and collective action to fight antibiotic resistance. an experimental approach. *Social science & medicine*, 142:249–255, 2015.
189. Clemens Rösner and Melanie Schmidt. Privacy preserving clustering with constraints. In *Proc. ICALP Colloq.*, pages 96:1–96:14, 2018.
190. M. Rosvall and C. T. Bergstrom. Maps of information flow reveal community structure in complex networks. *Proc. Natl. Acad. Sci. (PNAS)*, 105(1118), 2008.
191. Martin Rosvall and Carl T Bergstrom. Maps of information flow reveal community structure in complex networks. *arXiv preprint physics.soc-ph/0707.0609*, 3, 2007.
192. Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.
193. Sayak Saha Roy, Dipanjan Das, Priyanka Bose, Christopher Kruegel, Giovanni Vigna, and Shirin Nilizadeh. Demystifying nft promotion and phishing scams. *arXiv preprint arXiv:2301.09806*, 2023.
194. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proc. ACM Conf. on World Wide Web*, WWW '10, page 851–860, 2010.
195. Neda Sakhaee and Mark C. Wilson. Information extraction framework to build legislation network. *Artificial Intelligence and Law*, 29(1):35–58, 2021.
196. Kentaro Sako, Shin'ichiro Matsuo, and Sachin Meier. Fairness in ERC token markets: A case study of cryptokitties. In *Proc. Financial Cryptography and Data Security Workshops - CoDecFin, DeFi, VOTING, and WTSC*, volume 12676 of *Lecture Notes in Computer Science*, pages 595–610. Springer, 2021.
197. Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, 4(1):381–402, 2021.
198. Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. Fair coresets and streaming algorithms for fair k-means. In *Proc. WAOA Work.*, pages 232–251, 2019.
199. Stephen Seidman. Network structure and minimum degree. *Social Networks*, 5:269–287, 09 1983.
200. Alesja Serada, Tanja Sihvonen, and J. Tuomas Harviainen. Cryptokitties and the new ludic economy: How blockchain introduces value, ownership, and scarcity in digital gaming. *Games and Culture*, 16(4):457–480, 2021.
201. M. Ángeles Serrano, Marián Boguñá, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488, 2009.
202. R. Shamir, R. Sharan, and D. Tsur. Cluster graph modification problems. *Discret. Appl. Math.*, 144(1-2):173–182, 2004.
203. Jonathan Skaza and Brian Blais. Modeling the infectiousness of twitter hashtags. *Physica A: Statistical Mechanics and its Applications*, 465, 2017.
204. V. Soroka and S. Rafaeli. Invisible participants: how cultural capital relates to lurking behavior. In *Proc. of the World Wide Web conference (WWW)*, pages 163–172, 2006.
205. Francesco Sovrano, Monica Palmirani, and Fabio Vitali. Legal knowledge extraction for knowledge graph based question-answering. In *Legal Knowledge and Information Systems*, pages 143–153. IOS Press, 2020.
206. Massimo Stella, Sarah De Nigris, Aleksandra Aloric, and Cynthia SQ Siew. Forma mentis networks quantify crucial differences in stem perception between students and experts. *PloS one*, 14(10):e0222870, 2019.

207. Emilio Sulis, Llio Humphreys, Fabiana Vernero, Ilaria Angela Amantea, Luigi Di Caro, Davide Audrito, and Stefano Montaldo. Exploring network analysis in a corpus-based approach to legal texts: A case study. In *Proc. of the CAiSE for Legal Documents (COUrT) Workshop co-located with the the 33rd International Conference on Advanced Information Systems*, pages 27–38, 2020.
208. N. Sun, P. P.-L. Rau, and L. Ma. Understanding lurkers in online communities: a literature review. *Computers in Human Behavior*, 38:110–117, 2014.
209. C. Swamy. Correlation clustering: maximizing agreements via semidefinite programming. In *Proc. ACM-SIAM SODA Conf.*, pages 526–527, 2004.
210. Andrea Tagarelli and Roberto Interdonato. "Who's out there?": identifying and ranking lurkers in social networks. In *Proc. Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 215–222. ACM, 2013.
211. Andrea Tagarelli and Roberto Interdonato. Lurking in social networks: topology-based analysis and ranking methods. *Soc. Netw. Anal. Min.*, 4(1):230, 2014.
212. Andrea Tagarelli and Roberto Interdonato. Time-aware analysis and ranking of lurkers in social networks. *Soc. Netw. Anal. Min.*, 5(1):46:1–46:23, 2015.
213. Andrea Tagarelli and Roberto Interdonato. Time-aware analysis and ranking of lurkers in social networks. *Social Network Analysis and Mining*, 5:1–23, 2015.
214. Andrea Tagarelli and Andrea Simeri. Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code. *Artificial Intelligence and Law*, pages 1–57, 2021.
215. Henri Tajfel. Co-operation between human groups. *The Eugenics Review*, 58(2):77, 1966.
216. V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), Mar 2019.
217. Jan Trienes, Andrés Torres Cano, and Djoerd Hiemstra. Recommending users: Whom to follow on federated social networks. *CoRR*, abs/1811.09292, 2018.
218. Aman Tyagi, Joshua Uyheng, and Kathleen M. Carley. Affective polarization in online climate change discourse on twitter. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 443–447, 2020.
219. Foteini Valeonti, Antonis Bikakis, Melissa Terras, Chris Speed, Andrew Hudson-Smith, and Konstantinos Chalkias. Crypto collectibles, museum funding and openglam: Challenges, opportunities and the potential of non-fungible tokens (nfts). *Applied Sciences*, 11(21), 2021.
220. Anke van Zuylen and David P. Williamson. Deterministic algorithms for rank aggregation and other ranking and clustering problems. In *Proc. WAOA Work.*, pages 260–273, 2007.
221. Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proc. of the Eleventh Int. Conf. on Web and Social Media (ICWSM)*, pages 280–289, 2017.
222. Kishore Vasan, Milán Janosov, and Albert-László Barabási. Quantifying nft-driven networks in crypto art. *Scientific Reports*, 12(1):2769, 2022.
223. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Conf. on Neural Information Processing Systems*, pages 5998–6008, 2017.
224. Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *Proc. 6th Int. Conf. on Learning Representations (ICLR)*. OpenReview.net, 2018.

225. Victor von Wachter, Johannes Rude Jensen, Ferdinand Regner, and Omri Ross. NFT wash trading: Quantifying suspicious behaviour in NFT markets. *CoRR*, abs/2202.03866, 2022.
226. Fei-Yue Wang. Parallel intelligence in metaverses: Welcome to hanoi! *IEEE Intelligent Systems*, 37(1):16–20, 2022.
227. Lei Wang, Jianwei Niu, Xuefeng Liu, and Kaili Mao. The silent majority speaks: Inferring silent users' opinions in online social networks. In *Proc. of the World Wide Web conference (WWW)*, pages 3321–3327. ACM, 2019.
228. Qin Wang, Rujia Li, Qi Wang, and Shiping Chen. Non-fungible token (NFT): overview, evaluation, opportunities and challenges. *CoRR*, abs/2105.07447, 2021.
229. Yuntao Wang, Zhou Su, Ning Zhang, Rui Xing, Dongxiao Liu, Tom H. Luan, and Xuemin Shen. A survey on metaverse: Fundamentals, security, and privacy. *IEEE Communications Surveys & Tutorials*, 25(1):319–352, 2023.
230. Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*. Cambridge university press, 1994.
231. Xiaolin Wen, Yong Wang, Xuanwu Yue, Feida Zhu, and Min Zhu. Nftdisk: Visual detection of wash trading in nft markets. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.
232. Lilian Weng, Alessandro Flammini, Alessandro Vespignani, and Fillipo Menczer. Competition among memes in a world with limited attention. *Scientific reports*, 2(1):335, 2012.
233. Ryan Whalen. Legal networks: The promises and challenges of legal network analysis. *Michigan State Law Review*, page 539, 2016.
234. Gineke Wiggers and Suzan Verberne. Citation metrics for legal information retrieval systems. In *Proc. of the 8th International Workshop on Bibliometric-enhanced Information Retrieval co-located with the 41st European Conference on Information Retrieval*, volume 2345 of *CEUR Workshop Proceedings*, pages 39–50. CEUR-WS.org, 2019.
235. Hywel TP Williams, James R McMurray, Tim Kurz, and F Hugo Lambert. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global environmental change*, 32:126–138, 2015.
236. Wendy Wood and Dennis Rünger. Psychology of habit. *Annual review of psychology*, 67:289–314, 2016.
237. Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021.
238. Liu Yang and Yang Ren. Moral obligation, public leadership, and collective action for epidemic prevention and control: Evidence from the corona virus disease 2019 (covid-19) emergency. *International Journal of Environmental Research and Public Health*, 17(8):2731, 2020.
239. Paul Zhang and Lavanya Koppaka. Semantics-based legal citation network. In *Proc. of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 123–130, 2007.
240. Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When does pretraining help?: assessing self-supervised learning for law and the casehold dataset of 53, 000+ legal holdings. In Juliano Maranhão and Adam Zachary Wyner, editors, *Proc. of the Eighteenth International Conference for Artificial Intelligence and Law (ICAIL)*, pages 159–168. ACM, 2021.

241. Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *CoRR*, abs/1812.08434, 2018.
242. Ke Zhou, Marios Constantinides, Luca Maria Aiello, Sagar Joglekar, and Daniele Quercia. The role of different types of conversations for meeting success. *IEEE Pervasive Computing*, 20(4):35–42, 2021.
243. Haris Bin Zia, Jiahui He, Aravindh Raman, Ignacio Castro, Nishanth Sastry, and Gareth Tyson. Flocking to mastodon: Tracking the great twitter migration. *arXiv preprint arXiv:2302.14294*, 2023.
244. Vivian Ziemke, Benjamin Estermann, Roger Wattenhofer, and Ye Wang. What determines the price of nfts? *CoRR*, abs/2310.01815, 2023.
245. Matteo Zignani, Sabrina Gaito, and Gian Paolo Rossi. Follow the "mastodon": Structure and evolution of a decentralized online social network. In *Proc. ICWSM Conf*, pages 541–551, 2018.
246. Matteo Zignani, Christian Quadri, Sabrina Gaito, Hocine Cherifi, and Gian Paolo Rossi. The Footprints of a "Mastodon": How a Decentralized Architecture Influences Online Social Relationships. In *Proc. IEEE INFOCOM Workshops*, pages 472–477, 2019.
247. D. Zulli, M. Liu, and R. Gehl. Rethinking the 'Social' in 'Social Media': Insights into Topology, Abstraction, and Scale on the Mastodon Social Network. *New Media & Society*, 22(7):1188–1205, 2020.
248. M. Ángeles Serrano, Marián Boguñá, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488, 2009.

A

Appendix

A.1 Understanding the growth of the Fediverse through the lens of Mastodon

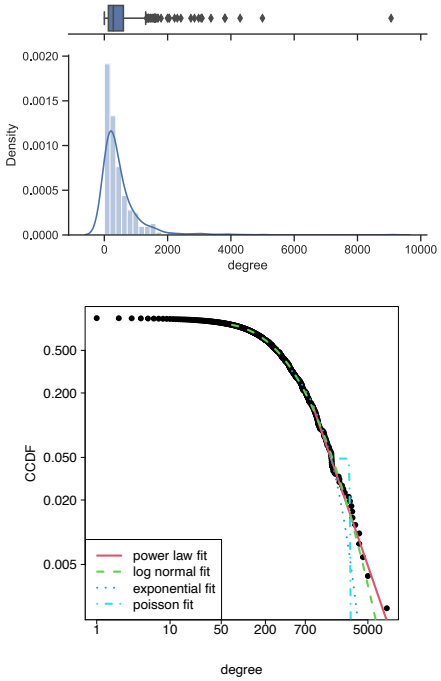


Fig. A.1. INSTANCES full-degree distribution: boxplot and Probability Density Function (top), and Complementary Cumulative Distribution Function, with various distribution fittings (bottom).

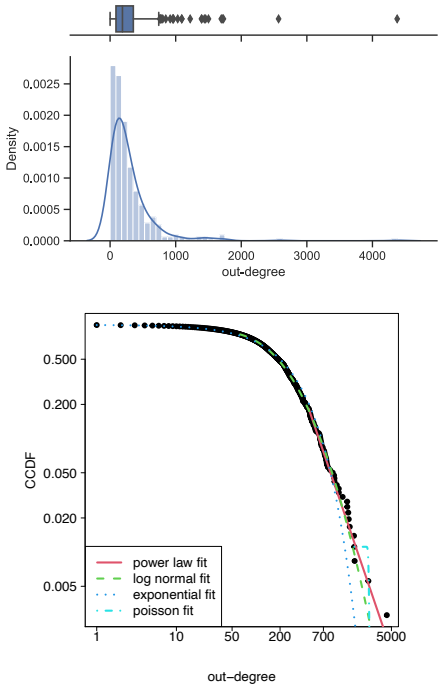


Fig. A.2. INSTANCES out-degree distribution: boxplot and Probability Density Function (top), and Complementary Cumulative Distribution Function, with various distribution fittings (bottom).

A.2 Information Consumption and Boundary Spanning in Decentralized Online Social Networks: the case of Mastodon Users

Lurker ranking methods, originally proposed in [210, 211], are designed to mine silent user behaviors in the network, and hence to associate each user with a score indicating her/his lurking status. Since the basic assumption of lurking behaviors is related to the *amount of information a user receives*, the key idea in the definition of lurker ranking methods is that the strength of a user's lurking status can be determined based on three main principles, namely over-consumption, authoritativeness of the information received, non-authoritativeness of the information produced.

The first principle corresponds to the evidence of a disproportion between information-consumption over information-production exhibited by a user. The second principle relates to the importance as information producers of a user's followees (i.e., in-neighbors), while the third principle related to the low importance as information producer of a user with respect to her/his followers (i.e., out-neighbors).

These principles are implemented in a ranking model so as to differently weighing the contributions of a node's in-neighborhood and out-neighborhood. For the sake of brevity here,

we will refer to only one of the formulations described in [210, 211], which is that based on the full *in-out-neighbors-driven lurker ranking*, hereinafter named as LurkerRank (LR).

Given a directed social graph $G = \langle V, E \rangle$, where any edge (u, v) means that v is is "consuming" or "receiving" information from u, the LurkerRank $LR(v)$ score of node v is defined as:

$$LR(v) = \alpha[\mathcal{L}_{\text{in}}(v)\ (1 + \mathcal{L}_{\text{out}}(v))] + (1 - \alpha)p(v)$$

where $\mathcal{L}_{\text{in}}(v)$ is the in-neighbors-driven lurking function:

$$\mathcal{L}_{\text{in}}(v) = \frac{1}{|\mathcal{N}_v^{out}|} \sum_{u \in \mathcal{N}_v^{in}} \frac{|\mathcal{N}_u^{out}|}{|\mathcal{N}_u^{in}|} LR(u)$$

and $\mathcal{L}_{\text{out}}(v)$ is the out-neighbors-driven lurking function:

$$\mathcal{L}_{\text{out}}(v) = \frac{|\mathcal{N}_v^{in}|}{\sum_{u \in \mathcal{N}_v^{out}} |\mathcal{N}_u^{in}|} \sum_{u \in \mathcal{N}_v^{out}} \frac{|\mathcal{N}_u^{in}|}{|\mathcal{N}_u^{out}|} LR(u)$$

where α is a damping factor ranging within [0,1] (usually set to 0.85), and $p(v)$ is the value of the personalization vector, which is set to $1/|V|$ by default. To prevent zero or infinite ratios, the values of the in/out-neighborhood size of a node are Laplace add-one smoothed. As a result, the higher the LR score of a node, the higher its likelihood to be regarded as a lurker in the network under study.

It should be noted that the actual meaning of "received information" modeled by the links in the LurkerRank input graph can depend on the specific context of network analysis; in practice, it refers to either a social graph (i.e., a linked pair (u, v) means that v *is follower* of u) or an interaction graph (e.g., v *likes* or *comments* u's posts). LurkerRank has been extensively evaluated on both scenarios [211, 212]. Nonetheless, although both social and interaction relations are indicators of information consumption by users, the information on interaction data that can be acquired from a real social network might be significantly sparse, and our context of study does not make an exception to this. Therefore, in this work LurkerRank is applied to a followship graph, which corresponds to the Mastodon user networks defined in Section 3.3 (with reversed edge-orientation).

A.3 Drivers of Social Influence in the Twitter Migration to Mastodon

A.3.1 Development and signaling of the migration process

We analyzed how the #TwitterMigration movement developed as a result of users discussing on Twitter their intention to switch to Mastodon. We hypothesizes that users announcing their Mastodon profiles on Twitter acted as a social trigger that might have persuaded other users to migrate. Given that the date of such announcement is not available for all users, we used the timestamp of Mastodon account creation as a proxy. Figure A.3 (left) reports the distribution of registration times after the purchase of Twitter by Elon Musk. Interestingly, some spikes in registrations emerge in response to controversial choices by the new management (e.g., mass layoffs and policy changes). Nonetheless, for our proxy hypothesis to be valid, it is important that the registration time is close to the time in which the Mastodon account was announced on Twitter. We could measure the temporal gap between these two events for a subset of about

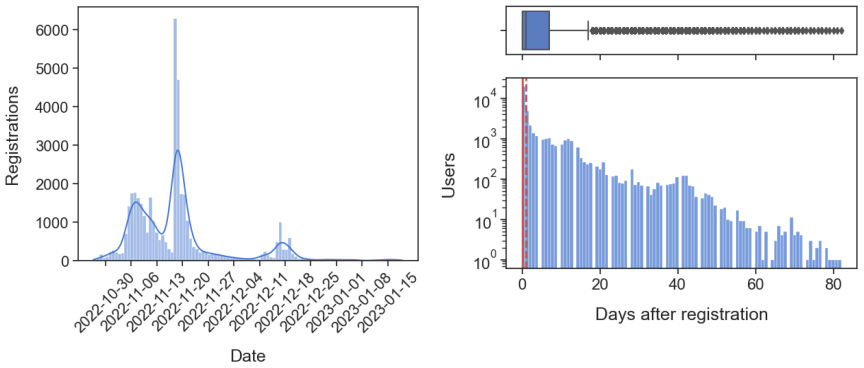


Fig. A.3. (Left) Registration dates of users migrated to Mastodon between October 26^{th} 2022 and January 19^{th} 2023. (Right) Time delta between migration to Mastodon and signaling of the new Mastodon handles on Twitter; the solid and dashed red lines represent the mode and median of the distribution, respectively.

$41K$ users for whom we have both the registration time on the new platform and the signaling time of the corresponding profile on Twitter. As shown in Figure A.3 (right), the time gap is rather short for most users, with mode of 0 days (i.e., less than 24 hours), a median of 1 day, and an average of 5 days.

A.3.2 Fitting with the SIR model and with an extended set of users related to the #TwitterMigration

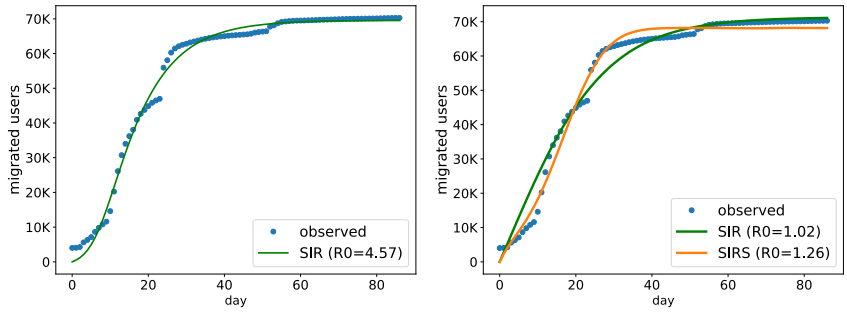


Fig. A.4. Cumulative number of Twitter users migrated to Mastodon over the course of 3 months since Elon Musk's acquisition of Twitter. (Left) Fit estimated with the SIR compartmental epidemiological model. (Right) Fit estimated with the SIR and SIRS compartmental model with the set of users corresponding to the population including all users who have discussed Mastodon on Twitter.

Fitting the SIR model. In addition to the SIRS model, we experimented with the simpler SIR model. We observed that the SIR model could replicate the observed trend of the cumulative

number of migrated users equally (same R_0 and MAPE values) with the SIRS model, as reported in Figure A.4 (left). However, in this regard we point out that although this may seemingly suggest that from a macroscopic point of view both models perform identically, this turns out to be a borderline case, as in general the SIRS model has been shown to be superior in terms of goodness of fit (cf. Results) due to its ability to model reiterated commitment in our infectiousness scenario.

Expanding the population. Our primary focus in this work is on the analysis of the main drivers of social influence behind the #TwitterMigration movement, and accordingly we considered as the population underlying our epidemic models the set of users eventually migrated to Mastodon. Nonetheless, to conduct a comprehensive study, we extended our evaluation by broadening the set of users representing our population. Specifically, this expanded set includes $\sim$ 540K users who have engaged in discussions involving Mastodon between October 26^{th} 2022 and January 19^{th} 2023, regardless of their final decision to migrate or not. Although this choice provides an underestimation of the actual process due to the inclusion of potential noise in the population, our SIR and SIRS models estimated an R_0 value of 1.02 and 1.26 (see Figure A.4, right), respectively. Since R_0 is greater than 1, this confirms growth of the infection process underlying social influence.

A.3.3 Further details on communities

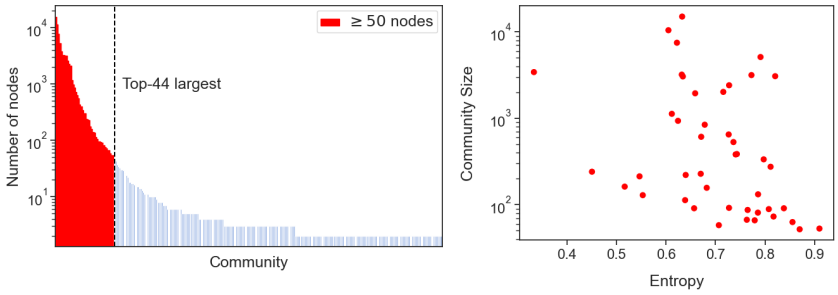


Fig. A.5. (Left) Distribution of users across the communities found via the Louvain method; red-colored communities contain at least 50 users. (Right) Scatterplot showing the relationship between community topical entropy and size for the top-44 largest communities.

To investigate the spreading of the #TwitterMigration's infectiousness across communities, we focused our analysis on a specific group of nodes containing a sufficiently representative number of users. Specifically, by referring to the distribution reported in Figure A.5 (left), it is possible to observe that most users turn out to be clustered in a few communities, with a long tail of communities containing a few members, or even singleton nodes. Consequently, for our analysis, we selected the set of communities having at least 50 users, which gives us coverage of 98% of the set of migrated users, corresponding to the top-44 largest communities. In this regard, we also complement our findings on the relationship between community size and topical entropy by resorting to Figure A.5 (right), which shows the (moderate) negative correlation we spotted between these two quantities across the top-44 largest communities.

A.3.4 Fitting compartmental models with the largest communities

Figure A.6 shows the how the SIR and SIRS model fit the migration data for the top-15 largest communities by number of users. In line with our findings (cf. Results), both the SIRS and SIR models are able to accurately shape the migration process, albeit with different R_0 (with mean values of 5.08 for SIRS and 3.92 for SIR for the top-44 largest communities).

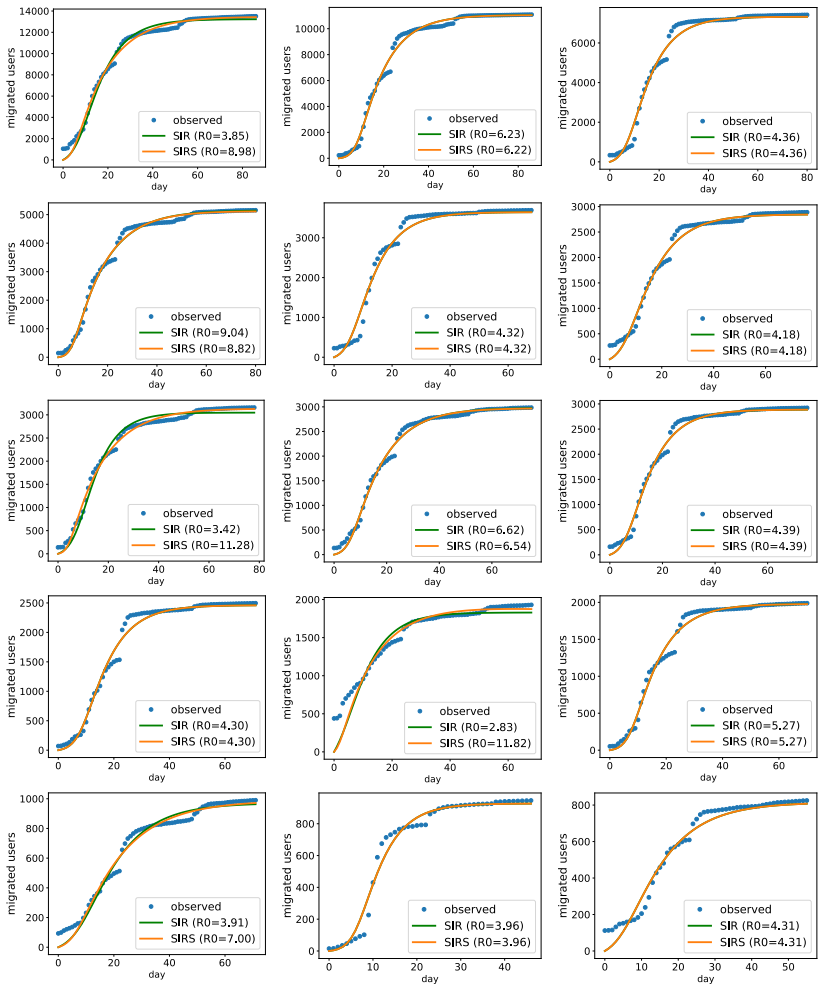


Fig. A.6. Fitting with the SIR and SIRS compartmental epidemiological models and associated R_0 values for the top-15 largest communities by Louvain, reported in descending order row-wise, of the cumulative number of Twitter users migrated to Mastodon over the course of 3 months since Elon Musk's acquisition of Twitter.

Table A.1. Ordinary Least Squares regression model fittings for the prediction of R_0 from the full set of topological (left) and social (right) features. β coefficients describe the contribution of each feature to the outcome, along with the standard errors (SE) and statistical significance (p-values). Auto-correlation is evaluated via the Durbin-Watson statistic (values closest to 2 indicate no auto-correlation). Regression results are reported via adjusted R^2.

Predicting R_0 from:				**Predicting R_0 from:**			
Topological Features				**Social Features**			
Feature	β	**SE**	p	**Feature**	β	**SE**	p
Density	-0.055	0.967	0.955	Support	0.040	0.188	0.834
Reciprocity	-0.123	0.235	0.604	Knowledge	0.507	0.161	0.003
Assortativity	-0.015	0.270	0.955	Conflict	-0.043	0.171	0.802
Std. Deg. Centr.	0.230	1.345	0.865	Power	0.248	0.157	0.124
Transitivity	-0.426	0.604	0.485	Similarity	-0.103	0.220	0.641
Avg. Clust. Coeff.	-0.175	0.500	0.729	Status	-0.126	0.194	0.519
				Trust	0.098	0.153	0.526
				Identity	0.511	0.141	0.001
Durbin-Watson stat. = 2.661		R^2_{adj} **= 0.140**		Durbin-Watson stat. = 2.038		R^2_{adj} **= 0.395**	

A.3.5 Experimenting with regression models to predict community-specific R_0

Table A.1 reports the regression model fittings for predicting community-specific R_0 based on the full set of topological and social features we leveraged in our analyses. Table A.2 reports the fitting of a Least Absolute Shrinkage and Selection Operator (LASSO) regression model on the combination of the best-performing features for predicting R_0 (cf. Results).

A.3.6 User activity in the months following the migration

To assess the effectiveness of the migration from Twitter to Mastodon almost a year after the Musk's buyout, we examined the activity levels of migrated users. To this aim, we utilized Mastodon's official APIs to retrieve the date of the most recent post made by migrated users. This check was performed in mid-October 2023. The presence of a valid payload not only confirmed the user's ongoing activity (i.e., post creation) on Mastodon, yet also allowed us to determine its extent. Figure A.7 (left) summarizes the key findings from our analysis. Remarkably, more than 95% of the migrated users maintained their account on Mastodon. Furthermore, while we observed a decline in activity due to the initial surge following the collective migration, we found that almost half of the migrated users remained active six months

Table A.2. Least Absolute Shrinkage and Selection Operator regression model fitting for the prediction of R_0 from a combination of topological, activity, and social features. β coefficients describe the contribution of each feature to the outcome. Penalty term is set to 0.1. Regression results are reported via adjusted R^2.

Predicting R_0 from:	
Topological & Social & Activity Features	
Feature	β
Density	-0.274
Knowledge	0.244
Identity	0.246
Commitment	0.240
R^2_{adj} **= 0.525**	

after the migration, suggesting interest in well-rooted settlement from a considerable fraction of migrated users. We took a closer look at users who made particular efforts to remain active even after a substantial period following the surge in migrations. To this aim, we calculated the fraction of users who continued to contribute by creating content in the last few months. Notably, as reported in Figure A.7 (left), more than a quarter of migrated users were observed to be actively posting during the month leading up to our mid-October 2023 check. These users constitute a resilient segment of the migrated population that was able to endure the initial noisy growth due to the curiosity to try out a new platform, thus effectively managing to cave out a new social space and maintaining their engagement and participation in the community. We further delved into such an investigation by computing, for each migrated user i, the corresponding *persistency* and *freshness* levels. The former is defined as $p_l(i) = t_p(i)/t_m(i)$ and indicates the duration of which the user i has been active (considering the creation of posts) on the platform $t_p(i)$, compared to the time has passed since migrating $t_m(i)$. The latter has been defined in previous work [213] as $f_l(i) = 1/\log_2(2 + t_\Delta)$, where t_Δ indicates the number of days elapsed between the date of checking the user's activity (i.e., mid-October 2023) and the date of the last post created by the user. These two complementary scores were normalized in $[0, 1]$ and serve as proxies for the migrated users' degree of persistency and propensity to contribute, respectively. As reported in Figure A.7 (right), migrated users almost split in a bipartite fashion. Indeed, while we observed a moderate fraction of migrated users not actively engaging in the new platform, most of them were found to be clustered at high persistency levels, with observable peaks around the maximum value. Similarly, despite observing that a large fraction of migrated users are not particularly keen to contribute, we found two noticeable groups exhibiting mid and high values of freshness, respectively, thus actively contributing with fresh content on the new platform. We further evaluated the extent of this propensity to contribute, unveiling very intriguing traits. Indeed, as reported in Figure A.7 (left), these "fresher" users were responsible for generating nearly the entire volume (90.13%)

Check	Perc.
Account existence	95.13%
Active after 1 month	69.26%
Active after 3 months	54.60%
Active after 6 months	44.73%
Active in the last 3 months	37.61%
Active in the last 2 months	32.78%
Active in the last month	27.50%
Volume ≥ Q3 freshness	73.56%
Volume ≥ Q2 freshness	90.13 %

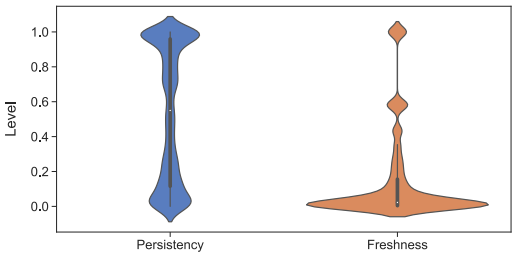


Fig. A.7. Percentage of accounts still existing and active w.r.t. different temporal checkpoints. (Right) Violin plots for the persistency and freshness levels of migrated users.

of posts by migrated users on the new platform. This is particularly evident among those in the last quartile of freshness value (73.56%), which hence act as "super-users". These intriguing findings underscore the pivotal role played by these super-users in sustaining the post-migration ecosystem and pave the way for further investigations. Finally, we explored the potential connection between the level of persistency/freshness and the underlying social influence in the migration process. Remarkably, we observed a non-negligible correlation between the R_0 value of the SIRS model and the corresponding persistency ($p = 0.491$) and freshness ($p = 0.394$) levels of the top-44 communities on which we narrowed our focus within our study. This finding, coupled with the replication of more than 40% of the social ties after the migration (cf. Results), poses a stepping point for the proper understanding of the growing migratory phenomena between social platforms, thereby warranting further investigations.

A.4 Multimodal representation learning for NFT selling price prediction

A.4.1 Running Times at Inference

Figure A.8 shows the running times at inference of various MERLIN models compared to their obtained *WLR* performance. It stands out that, while the non-GAT modules are the most efficient ones, the GAT module leads to a significant improvement in *WLR* at only a moderate inference-time cost (as long as the node size is not below the default of 50) w.r.t. the other modules. Also, the ensemble of the models' predictions allows for boosting the *WLR* performance by doubling the GAT's one, and without any execution overhead thanks to parallelization of the visual and text modules.

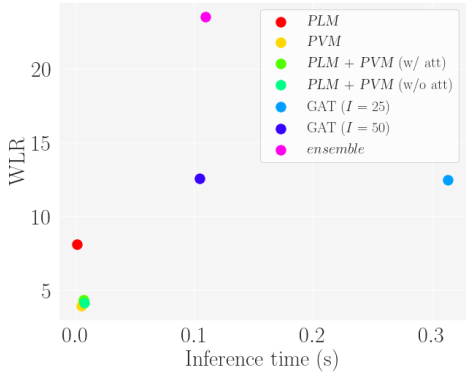


Fig. A.8. Inference times w.r.t. *WLR* scores for various models

A.4.2 Additional Remarks on Evaluation

Data distributions. Figure A.9 shows the distributions of the NFT categories and the price categories (i.e., classes).

NFT-set similarity graph. We additionally evaluated the effect on the structural traits of the resulting NFT-set similarity graph by different values for the node size l (Table A.3). One remarkable fact is that the number of communities, computed by the Louvain method [30], is roughly constant (set around 10) regardless of the choices for l (note that the number of communities is not an input to the Louvain method). By delving into the community structures, Figure A.10 (a-c) unveils that on each graph, with a more marked tendency for lower l, the discovered ten or eleven communities mostly capture the ten percentiles of the average selling price for the NFTs in our dataset. This further supports our initial intuition that there is an important impact of the visual and textual features on the selling price, i.e., *similar NFTs might sell similarly*. Moreover, we observed that the communities appear to be topologically contiguous according to the ordered percentiles: intuitively, this will have a positive impact on the effectiveness of the neighbor aggregation steps performed by the GNN module in MERLIN.

Table A.3. Main characteristics of the NFT-set similarity graph by varying node size l: average degree (Deg), density (Den), diameter (D), average path length (APL), clustering coefficient (CC), modularity (M), no. of communities (#C)

$\mathcal{G}$	|V|	|E|	Deg	Den	D	APL	CC	M	#C
$l = 100$	1808	13 939	15.419	0.009	42	12.756	0.457	0.795	11
$l = 50$	3616	29 959	16.570	0.005	33	9.572	0.400	0.777	10
$l = 25$	7232	62 322	17.235	0.002	27	8.051	0.356	0.768	11

Interesting aspects also emerge about the graph that can be built from the similarities between the embeddings $\mathbf{h}^{(G)}$, i.e., by applying the GAT at inference. Indeed, Figure A.10 (d) shows almost perfect separation of the NFT-sets w.r.t. their source collections, which is information never provided during the training phase. This is an outstanding evidence of the effectiveness of MERLIN to model contextual awareness in latent NFT representations, thus supporting our research hypotheses.

Structural characteristics of NFT-set similarity graphs. As reported in Table A.3, we notice some interesting aspects underlying the NFT-group similarities. For instance, quite large values for diameter and average path length suggest that less similar nodes in our graphs tend to be relatively far apart. Moreover, relatively high clustering coefficient and modularity indicate triadic closure and modular structure, respectively.

Early-stopping. We compared the effects of our defined early-stopping criterion based on win-rate w.r.t. the one based on accuracy. We find that (results not shown) the win-rate-based criterion is clearly more beneficial in terms of $High$-driven performance measures, with an improvement up to 48.7% of WLR, and on average 22.7% of riskiness and 9.2% of cautiousness over all models.

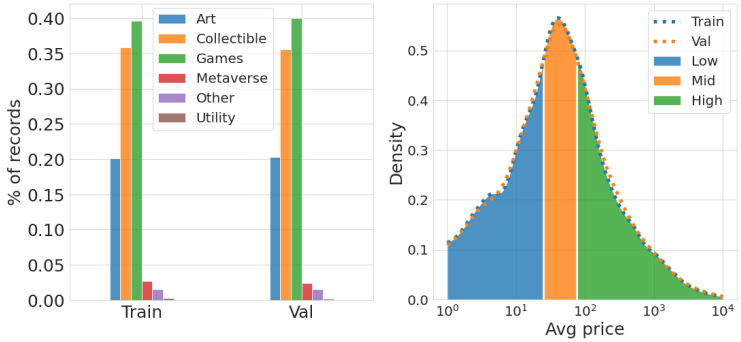
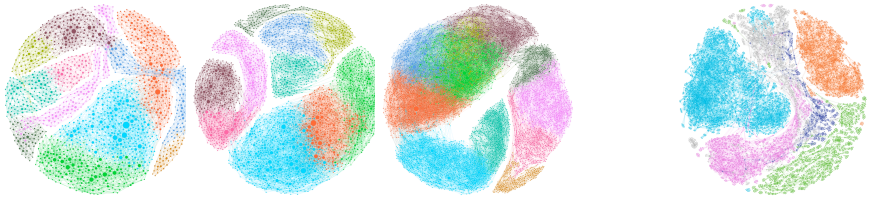


Fig. A.9. NFT category distributions for the training and validation sets (left) and density of the average-price distributions (overall, training, and validation sets), with corresponding areas for the three price-categories (right)



(a) (b) (c) (d)

Fig. A.10. NFT-set similarity graphs, with colors corresponding to communities detected by Louvain method (a-c) and NFT-set similarity graph on validation set, with colors corresponding to top-5 collections (d). Plots (a-c) refer to different values for node size l: 100 (a), 50 (b), 25 (c), whereas plot (d) refers to $l = 50$

A.4.3 Interpretations of MERLIN predictions

Here we discuss the interpretation of MERLIN predictions and their meaningfulness w.r.t. the task at hand, by resorting to *LIME* (Local Interpretable Model-Agnostic Explanations) [183]. Being model-agnostic, *LIME* just requires the class probabilities outputted by the model that is to be interpreted. It learns a linear model that approximates the target one in the neighborhood of the instance that needs to be explained, by perturbing the latter in order to learn feature importance scores w.r.t. the outcome of the model.

In this respect, in Figure A.11 we present different examples of NFTs in our test set, and show the importance of features of images and descriptions w.r.t. class $High$, denoting with the green, resp. red, features that positively, resp. negatively, contributed to the class prediction performed by our best PVM and PLM, respectively.

Figure A.11 (a) shows an NFT of the *Gods-Unchained* collection (ID = 9493839), i.e., a collectible card game, correctly classified by both PVM and PLM. Among the visual features positively impacting on the prediction capabilities, we report (i) the number 7 on the top-left side, which indicates the "mana cost" of a card, i.e., the amount of tokens needed for playing with it; (ii) the top-right symbol indicating the "set" of the card, i.e., its grouping w.r.t. particular events or themes; (iii) the wings of the portrayed creature and its name, along with the descriptive part of the character. Conversely, we spotted that the amount of damage tolerable from a creature (bottom-left) and its health score (bottom-right) may negatively impact on the price prediction. As concerns the textual components, we report high importance for the terms "Roar", "Ward", "Protected" and "Flank", i.e., those describing cards' effects.

Figure A.11 (b) shows an NFT of the *Somnium-Space* collection (ID = 2026), i.e., a Metaverse project, correctly classified by both visual and textual modules in MERLIN. Here we spotted that the PVM actually benefits from the meaningful writings in the image (e.g., the square footage of the area), along with its shape and positioning in the mini-map. Conversely, the description seems to be less informative than the previous case (as it focuses mainly on the name of the Metaverse project), with scores produced by *LIME* for the PLM that are lower than an order of magnitude w.r.t. the previous example.

Looking at the example in Figure A.11 (c), we observe an NFT whereby only PVM correctly predicted the target class. The *LIME* interpretation indeed provides some clues on that: the PVM seems to concentrate on the eyes, the silhouette and other characteristics of the profile of the drawing typical of the *Axie* collection; conversely, the description is poorly informative

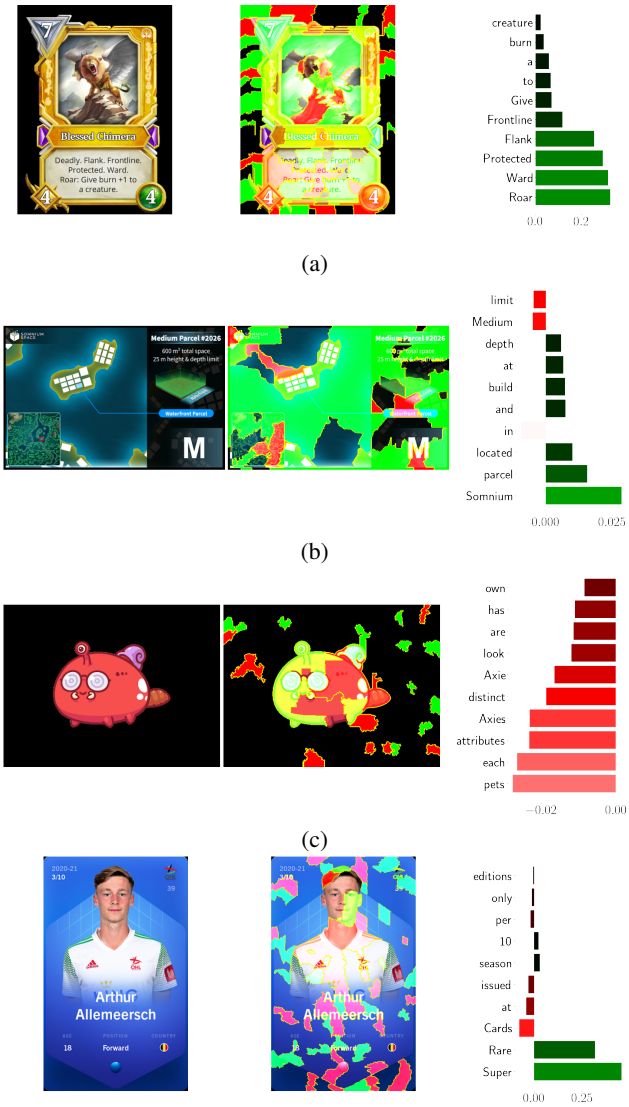


Fig. A.11. Example NFT images (left) and LIME interpretation of PVM features (center) and PLM features (right). Red, resp. green, denote a negative, resp. positive, impact of features on the model prediction

since it is the same across all NFTs in that collection, thus having a negative impact on the PLM performance.

Finally, in Figure A.11 (d) we show an NFT of the *Sorare* collection, a football game in which players use NFTs to trade and manage virtual teams. Both PVM and PLM correctly predicted the target class. The *LIME* interpretation suggests that the PVM mostly focuses on the age and face of the player, as well as on his position on the pitch and team, as reported in the card. Besides, the "Super" and "Rare" terms are successfully leveraged by the PLM to discern that such an NFT is particularly rare within its collection.

A.5 Exploring the Role of Inspiration in Non-Fungible Tokens

SHAP is used to explain the outcomes of a model designed for the following classification task: Given two PVMs with shared weights, and an input pair of tokenized images $(\mathcal{T}_i, \mathcal{T}_j)$, the goal is to predict the probability that the two images are similar or not similar. The class probabilities are defined as $\langle sim(\mathbf{h}_i, \mathbf{h}_j), 1 - sim(\mathbf{h}_i, \mathbf{h}_j)\rangle$, where $sim(\mathbf{h}_i, \mathbf{h}_j)$ is the cosine similarity between $\mathbf{h}_i$ and $\mathbf{h}_j$, i.e., the embedding for image i and j, respectively (cf. Sect. *Data Extraction and Network Modeling*).

Shapely regression values are represented as a linear model over importance scores assigned to the features [151]. In our setting, features correspond to groups of pixels extracted from each input pair of images. These raw features are then corrupted through blurring functions which are aimed at creating two coalitions of images for each given feature: a coalition where a particular feature is present and one where it is corrupted. More precisely, for each feature $f \in F$ (with F indicating the space of features), the Shapley value Φ_f for f can in principle be computed by estimating the difference between the prediction of a model $\mathcal{M}$ (i.e., pair of PVMs) when the feature is used and the prediction of the model without that feature:

$$\Phi_f = \sum_{S \subseteq F \setminus \{f\}} \frac{|S|(|F| - |S| - 1)}{|F|} [\mathcal{M}_{S \cup \{f\}}(x_{S \cup \{f\}}) - \mathcal{M}_S(x_S)]$$

where x denotes a feature representation of the pair of images which depends on each of the *coalitions* S selected from the feature space F. To efficiently computing the explanations, Shapely values are estimated by sampling approximations, thus allowing to avoid computing $2^{|F|}$ possible coalitions and retraining the model $\mathcal{M}$.

Before computing explanations, we resized each image to 512×512 pixels. Upon this, we configured SHAP to compute K=10 000 samples for each explanation, where a Gaussian kernel of size 64×64 is used to mask features by generating a Gaussian blurring.