



This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101135437



AI services for Continuous trust
in emerging Digital Environments

D2.1 – Report on technological and socio-behavioral characterization of next-generation social media

Grant Agreement no.:	101135437
Project Title:	AI-CODE - AI services for Continuous trust in emerging Digital Environments
Contractual Due date:	30/11/2024
Actual submission date:	29/11/2024
Lead beneficiary:	FBK

Deliverable number	2.1
Deliverable Title:	Report on technological and socio-behavioral characterization of next-generation social media
Description:	This deliverable will summarise the research performed in T2.1 and outline the emerging trust and disinformation risks.
Type/Nature	☒ R - Document, report ☐ DEM - <i>Demonstrator, pilot, prototype</i> ☐ DEC - <i>Websites, patent filings, videos, etc.</i> ☐ OTHER ☐ ETHICS – <i>Ethics requirement</i> ☐ ORDP – <i>Open Research Data Pilot</i>
Dissemination Level	☒ Public – fully open ☐ Sensitive — limited under the conditions of the Grant Agreement ☐ EU classified — RESTREINT-UE/EU-RESTRICTED, CONFIDENTIEL-UE/EU-CONFIDENTIAL, SECRET-UE/EU-SECRET under Decision 2015/444. For items classified under other rules (e.g. national or international organization), please select the equivalent EU classification level.
Work Package number	WP2
Work Package leader	KInIT
Author(s) <i>in alphabetical order</i>	Beatriz Arregui-García (FBK) Anees Baquir (FBK) Santiago Barrio (UPM) Viktoras Daukšas (Debunk) Riccardo Gallotti (FBK) Andreas Giefer (DW) Paula Gürtler (CEPS) Jakub Kubs (Debunk) Lucio La Cava (UNICAL) Francisco Moreno (UPM) Pablo Perdomo (UPM) Ivan Srba (KInIT) Eleni Tsalapati (ATC)
Internal Reviewers <i>in alphabetical order</i>	M.A. Barbero-Álvarez (UPM) Manos Schinas (CERTH)

Authors (in alphabetical order)

Name	Partner	e-mail
Beatriz Arregui Garcia	FBK	barreguigarcia@fbk.eu
Anees Baqir	FBK	abaqir@fbk.eu
Santiago Barrio	UPM	s.barrio@upm.es
Viktor Daukšas	Debunk	Viktoras@Debunk.org
Riccardo Gallotti	FBK	rgallotti@fbk.eu
Andreas Giefer	DW	andreas.giefer@dw.com
Paula Gürtler	CEPS	paula.gurtler@ceps.eu
Jakub Kubs	Debunk	jakub.kubs@debunk.org
Lucio La Cava	UNICAL	lucio.lacava@dimes.unical.it
Francisco Moreno	UPM	fran.moreno@upm.es
Pablo Perdomo	UPM	p.perdomo@upm.es
Ivan Srba	KInIT	ivan.srba@kinit.sk
Eleni Tsalapati	ATC	e.tsalapati@atc.gr

History of changes

Version	Date	Changes
0.1	30/09/2024	Table of Content
0.2	08/11/2024	Individual contributions
0.3	12/11/2024	Draft ready for reviewers
0.4	15/11/2024	Draft reviewed by UPM
0.5	25/11/2024	Review by the coordinator
1.0	29/11/2024	Final version prepared

Table of contents

Executive Summary.....	6
1 Introduction and Background	7
1.1 New risks and opportunities associated with federated social media	7
1.1.1 The Fediverse and Mastodon.....	7
1.1.1 Bluesky	9
1.2 Early characterization of immersive social media	11
1.2.1 Main features and taxonomy.....	11
1.2.2 More diffused platforms and userbase.....	12
2 Data Collection from Decentralized Social Media	14
2.1 Mastodon API	14
2.2 Bluesky API & Firehose	15
2.2.1 The Bluesky Firehose	15
2.2.1.1 API	15
2.2.1.2 Capturing Data	15
2.2.1.3 Evolution of the number of posts per minute.....	16
2.2.1.4 Languages	16
2.2.1.5 Hashtags.....	16
2.2.1.6 Labels.....	16
2.2.1.7 Accounts.....	16
2.2.1.8 The database structure.....	17
2.2.1.9 The tables and their relationships.....	17
2.3 77.lt	18
3 Social Network Reconstruction	20
3.1 Mastodon instances network	20
3.2 Copresences in Decentraland	21
3.2.1 Data collection	21
3.2.2 Copresence network construction.....	21
3.2.3 User activity and spatial dynamics	23
3.2.4 Blockchain transaction analysis.....	24
4 Methodologies for the Analysis of Disinformation Spreading and Hate Speech on Federated Social Media.....	26
4.1 Signed ban networks	26
4.2 Information transmission model	28
4.2.1 Results.....	29
4.2.2 Discussion.....	31
4.3 Patterns of moderation in Mastodon's network	33
4.3.1 Data collection	33
4.3.2 Methodology.....	33

4.3.3	Results.....	34
4.4	Social bots	36
4.5	Graph-based methodologies.....	38
4.5.1	Clustering analysis.....	38
4.5.2	Propagation graphs	39
4.6	Hate speech analysis on Mastodon.....	41
4.6.1	Data collection from Mastodon	41
4.6.2	Hate Speech Detection Tool	43
4.6.3	Results of hate speech detection in Mastodon	43
4.7	Bluesky superspreaders.....	44
4.7.1	Superspreaders	44
4.7.2	Detecting hate speech using an Interactive Data Exploration.....	44
4.7.2.1	The Dashboard	44
4.7.2.2	Top hate speech accounts.....	45
5	Privacy and Data Governance in Decentralized Networks.....	46
6	Conclusions	49

Executive Summary

This report summarizes the research carried out in task *T2.1 Decentralised and immersive social media*. More specifically, it examines emerging trends in social media platforms, focusing on federated systems and immersive virtual environments. Our analysis reveals both promising developments and significant challenges that will guide the development of the AICODE services.

Federated platforms like Mastodon and Bluesky represent a significant shift toward decentralised social media infrastructure, offering users enhanced control over personal data and platform governance. While this decentralisation empowers users and administrators, it also increases system complexity, creating new challenges for consistent content moderation.

In parallel, immersive environments are reshaping social interaction online. Gaming platforms such as Fortnite¹ and VRChat² are evolving into social spaces, while virtual worlds like Decentraland³ demonstrate novel patterns of user engagement and economic activity. However, the predominantly young user demographic and game-centric nature of these experiences suggest that, while representing an interesting and growing social space, it is premature for AI-CODE services to focus on these platforms for solutions to reduce disinformation spread.

Cross-cutting challenges emerge in content moderation and data governance. Community-driven moderation approaches show promise but lack consistency across federated instances, potentially facilitating the formation of echo chambers and new pathways for misinformation spread. While decentralized platforms offer enhanced user control, they introduce increased complexity in data governance, with server-specific decisions affecting data accessibility.

The emergence of these platforms represents a significant shift in social media's evolution, with implications for user protection, content moderation strategies, and platform governance models. These findings suggest the need for balanced policy development as these platforms evolve. Monitoring tools suitable for decentralized and immersive contexts will be crucial for understanding how these spaces develop and mature into new forms of digital social interaction.

¹ <https://www.fortnite.com/>

² <https://hello.vrchat.com/>

³ <https://decentraland.org/>

1 Introduction and Background

1.1 New risks and opportunities associated with federated social media

Decentralized Online Social Networks (DOSNs) are emerging as an alternative to the popular centralized platforms such as Facebook and Twitter (now X), given the growing concerns about owner scalability and side effects of centralization (e.g., unexpected policy changes). The key idea behind DOSNs is to bring back the user at the center of the stage (user-centric perspective), aiming at fostering more spontaneous interactions, no longer driven by algorithms.

1.1.1 The Fediverse and Mastodon

Decentralization leads to the formation of independent, yet cooperating servers scattered throughout the Web. This is possible thanks to the availability of open-source software, that allows anyone to create a new server, dubbed in this contest an “instance”, and specific communication protocols that allow different servers—even pertaining to different platforms—to seamlessly communicate, similar to the functionality provided by emails.

Among the different protocols, ActivityPub⁴ (a W3C recommended standard) represents one of the most widely used in the Decentralized Web. It allows server-to-server interactions, to enable decentralized platforms sharing information with each other, and a client-to-server interaction, allowing users, bots and automated processes communicate with ActivityPub through servers. ActivityPub is based on boxes (i.e., in-box and out-box), through which users and servers can communicate with the rest of the world, as illustrated in Figure 1.

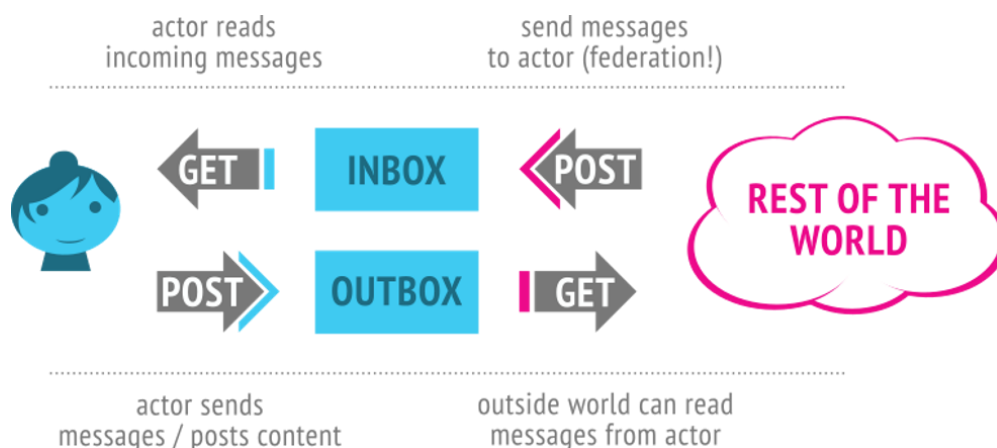


Figure 1: Schematic representation of the ActivityPub protocol. Source: w3c.org.

During time, several services started embracing the ActivityPub protocol, developing the so-called Fediverse, i.e., the federated universe of decentralized services. An illustration of the main services being deployed on the Fediverse is reported below in Figure 2.

⁴ <https://www.w3.org/TR/activitypub/>

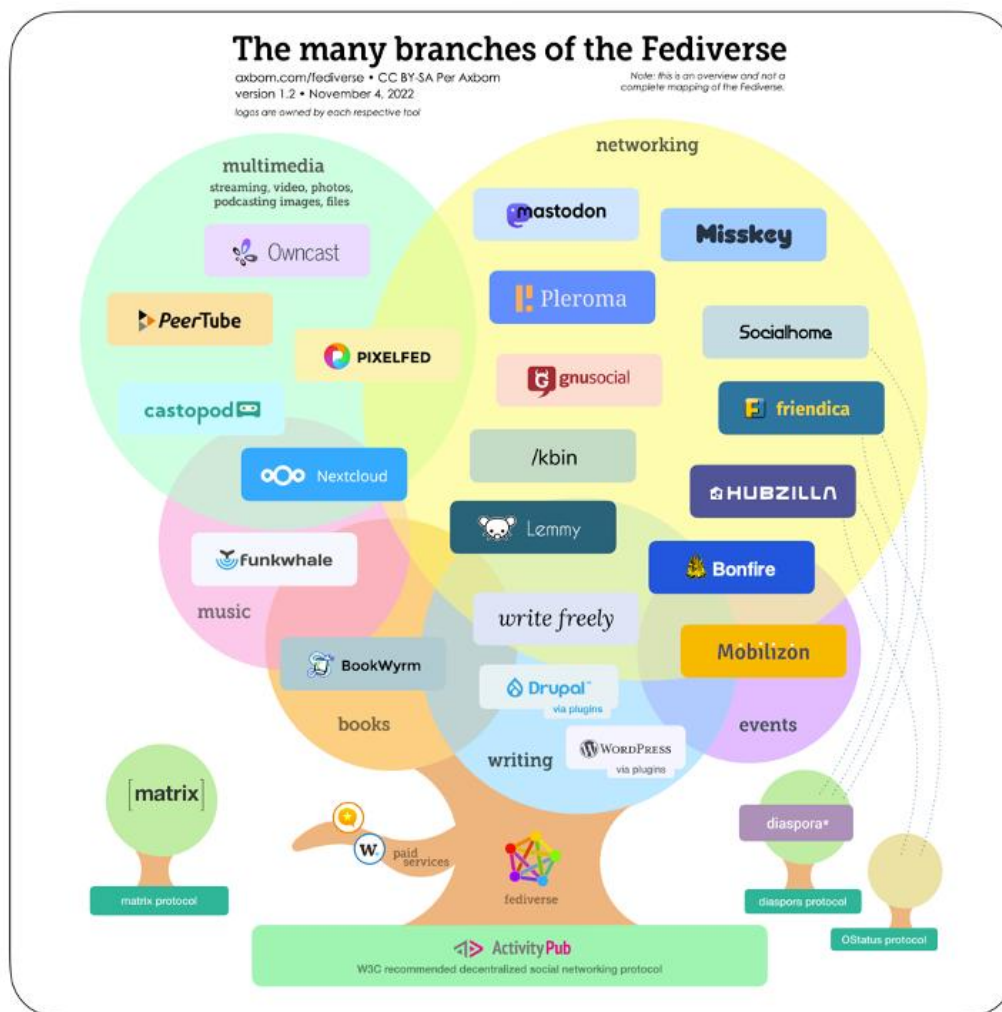


Figure 2: Main services in the Fediverse. Source: jaxbom.org.

In such a federated scheme coexist multiple platforms. In turn, each platform can create multiple instances, communicating with each other. Moreover, they can also communicate with instances from other services embracing the same protocol, as illustrated in the following Figure 3.

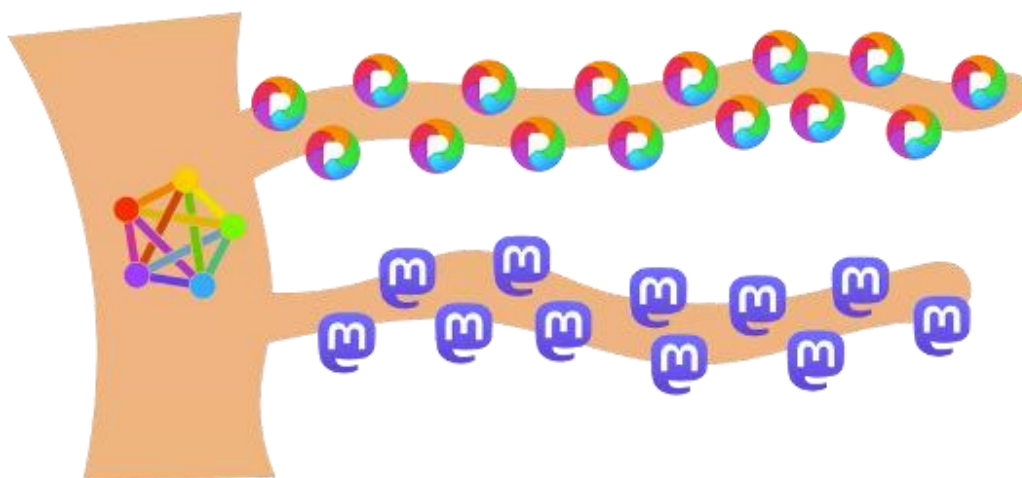


Figure 3: Different Fediverse services may communicate with each other given the same protocol. Source: jaxbom.org.

Among the various services in the Fediverse, Mastodon emerged as the most prominent one,⁵ standing out as

⁵ <https://fedidb.org/>

the decentralized alternative to Twitter. Mastodon is a microblogging platform that offers a Twitter-like user experience. Developed in Europe upon the ActivityPub protocol, it features niche community formation and moderation (akin to Reddit). Users can choose one among thousands of instances and register there, while, as said above, being able to interact with other instances as well as other services in the whole Fediverse. Interestingly, Mastodon offers different timelines to users including the one containing content by the people followed (even from different instances), and a “local” one that shows all content being created on the instance where a user is registered. As a result, content created on a given instance can easily flow throughout the Mastodon servers, and even cross the Mastodon boundaries, spanning the entire Fediverse. This interesting feature represents a key point for new risks in the social media landscape. Indeed, while on the one hand it is valuable given the possibility to create content that can easily traverse services, it adds a layer of complexity in tracking the content flow. This is further exacerbated by the fact that, while anyone can create a new server and join the Fediverse, the detection of newly created servers is not an easy task, and most of them can remain undetected for a long time.

One potential risk deriving from this ease in creating new servers is the possibility of developing “isolated niches of disinformation”, whereas registered users mainly consume local (misinformative) content, letting their instance becoming a giant information bubble.

All the points discussed above bring new challenges in the (decentralized) moderation. First, moderation remains mostly local to the instance, while the proliferation of new servers throughout the Fediverse increases the likelihood of receiving exogenous content, created outside the local instance. This determines an increased effort in content moderation, as local moderators (typically volunteers as well as instances administrators) must face both local and “external” content. Moreover, moderation is no longer managed by a cohesive board as in the case of centralized social media, rather it is demanded to local decisions in each instance, which means that different servers can have different policies, and let diverse content flow.

Nonetheless, Mastodon grants administrators and moderators a new set of decentralized moderation tools aimed at overcoming or mitigating such issues. In particular, each instance can declare certain policies determining the scope of the posts that are permitted on that server, and registered users must adhere to these policies. More interestingly, Mastodon allows administrators to moderate network interactions (i.e., via ActivityPub) with instances. Thanks to this new feature, Mastodon instances can decide to silence or ban specific instances (e.g., those spreading misinformation), preventing their users from interacting with them, and so reducing the likelihood of receiving their content.

1.1.1 Bluesky

Bluesky is another emerging decentralized alternative to Twitter that has been gaining momentum in recent times. Differently from Mastodon, it uses the ATProtocol and fosters quasi-decentralization through custom domains. It offers a feed-based timeline structure, where people can use the feed containing content created by followed users, or subscribe to custom feeds curated by other users on specific topics. An example of custom feeds is reported below in Figure 4.

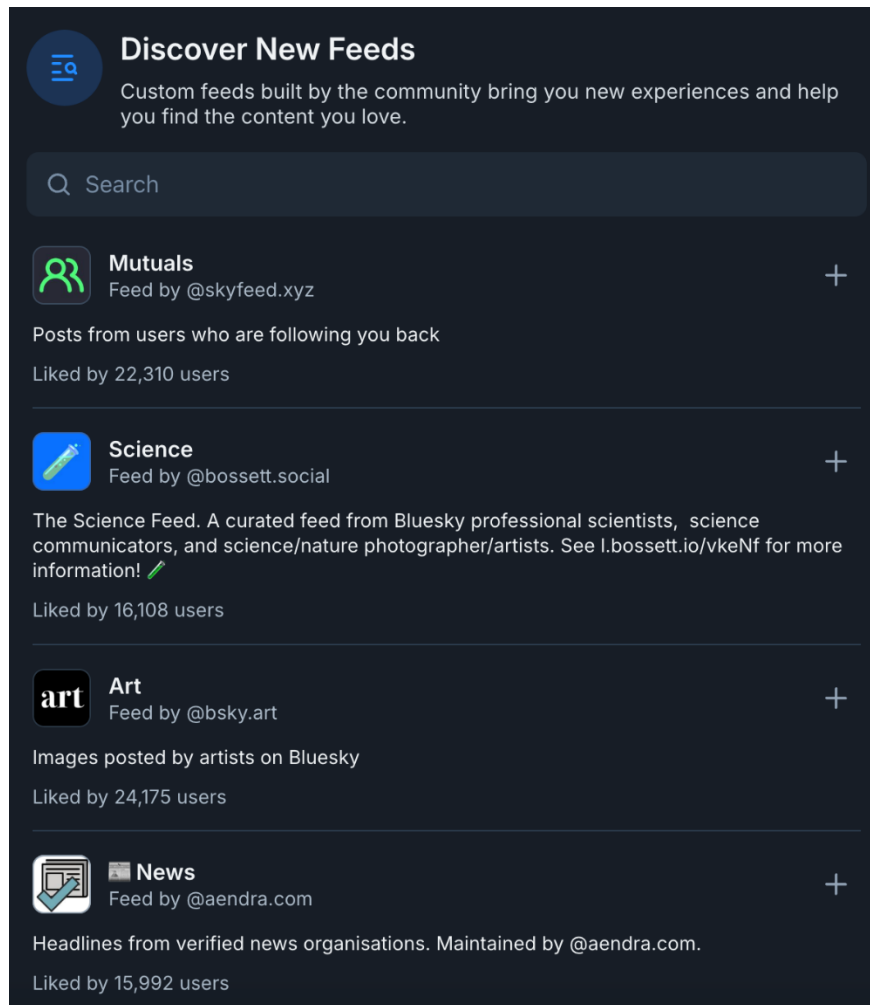


Figure 4: Example of BlueSky feed. Source: bsky.app.

Bluesky puts particular emphasis on user-driven content moderation, letting individual users decide whether and how to moderate content to be shown on their feeds. Indeed, as reported in Figure 5, users can use their account settings to filter out specific contents, or to put warnings on certain categories of content. The responsibility for filtering is shared. On one side, the Bluesky network labels content automatically, probably using a proprietary model. On the other side, each user can decide how labels content is displayed: is it shown? Is it hidden? Does the client application display a warning?

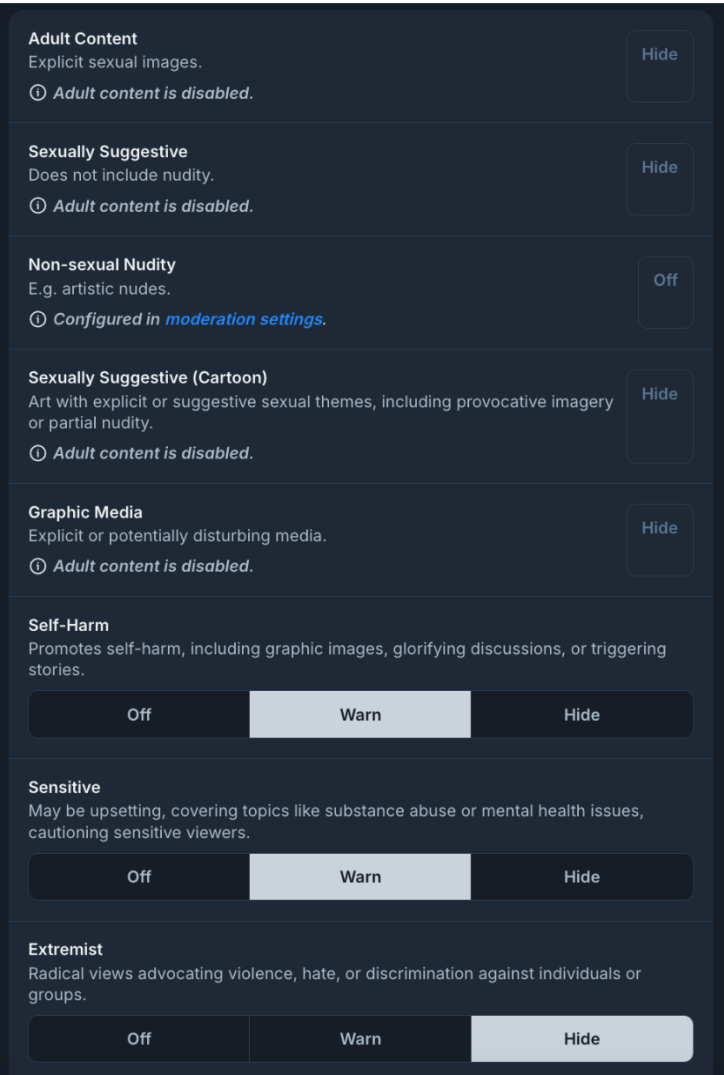


Figure 5: Examples of how to moderate the own feed in Bluesky. Source: bsky.app.

1.2 Early characterization of immersive social media

1.2.1 Main features and taxonomy

An immersive social media platform, often referred to by the general public as a Metaverse, represents a type of online space that has progressively transitioned from the realm of science fiction (as the term Metaverse was first used in the novel *Snow Crash* by Neal Stephenson) to commercial endeavours by major players in the media and technology sectors.

While the vision of this type of online space is that of a persistent virtual world accessed via Virtual Reality (VR) headsets, this is only one of the possible designs available and not necessarily the one that will become the most prominent.

From a technological perspective, the commercial availability of VR headsets has expanded significantly over the past few years, driven by advancements in technology, increased consumer interest, and a competitive market that now includes major tech players such as Microsoft, Sony, and Apple. Affordable and user-friendly VR devices from companies like Meta (Oculus), Pico, and Sony have made immersive experiences more accessible to the public. These companies have primarily catered to gaming enthusiasts while also attempting to capture the attention of professionals in education, training, and enterprise. Conversely, Apple's Vision Pro headset stands out as a premium offering, targeting a professional audience and positioning itself as an Augmented Reality (AR) tool for business applications, design, and advanced immersive experiences.

Nevertheless, at the present time most users of immersive online spaces are not accessing via Virtual or Extended reality, but using various devices including mobile phones, tablets and PC. In the following table, we present an overview of some of the most popular virtual spaces to highlight the wide diversity present in the market.

1.2.2 *More diffused platforms and userbase*

The landscape of virtual worlds is dominated by popular gaming platforms like Fortnite, Minecraft, and Roblox, each attracting massive user bases. Fortnite has pioneered large-scale virtual events and concerts⁶, and following Epic Games' recent agreement with Disney⁷, is poised to become a hub for immersive experiences featuring Disney's extensive intellectual properties. Minecraft, in its thirteenth year, maintains its position as the most played video game in history, offering users unprecedented creative agency to build and modify their environment. Notable examples like the long running "anarchy" server 2b2t⁸, can be considered as natural social experiments demonstrating the possible positive and negative consequences of unregulated virtual spaces. Roblox, while primarily attracting younger users, offers cross-platform accessibility from mobile to VR devices. While these spaces are starting to mature and social spaces and have witnessed emerging social phenomena including organized political expressions such as pro-Palestinian protests⁹, we have not found any report of misinformation or disinformation systematically being spread in virtual worlds. The early academic analysis in this warns however on the multimodal nature of immersive virtual worlds as AI systems are becoming more and more able to create realistic but fake audiovisual digital content¹⁰.

Virtual reality-focused platforms like VRChat and Meta's Horizon Worlds operate within the constraints of the growing but still limited VR market. Horizon Worlds has expanded both geographically and demographically, recently lowering its age restriction from 18+ to 13+, and subsequently to 10+¹¹. Similarly, VRChat maintains a nominal 13+ age requirement. Our preliminary research on VRChat revealed that most spaces were predominantly occupied by younger users, with limited adult interactions rarely happening at the margins of gaming-oriented environments.

A distinct category comprises platforms built around blockchain-based economies. Decentraland (using MANA tokens), The Sandbox (SAND cryptocurrency), and Somnium Space (CUBE tokens) enable users to purchase, develop, and trade virtual real estate and assets. While these spaces primarily attract cryptocurrency enthusiasts and promote the concept of "play-to-earn," they offer unique research value through their transparent, blockchain-recorded transactions - currently the only open and readily available data source we identified for analysing behavioural patterns in virtual worlds.

Beyond these blockchain-based examples, studying virtual worlds presents significant challenges that suggest comprehensive analysis or service development may be premature. The prevalence of minor users, particularly in popular platforms like Roblox and VRChat, raises complex ethical considerations for data collection and analysis. Despite industry efforts to position these spaces as professional environments for remote work, user interaction remains predominantly gaming focused. The emergence of room-specific community governance, where individual space creators implement and enforce distinct rules, presents an interesting self-regulatory model. However, the fragmented nature of these communities and their youthful demographic make it premature to implement robust anti-disinformation services. Nonetheless, ongoing research remains crucial as these platforms evolve and potentially mature into more diverse digital environments.

⁶ <https://www.theverge.com/2024/11/4/24288014/fortnite-chapter-2-remix-snoop-dogg-times-square-music>

⁷ <https://thewaltdisneycompany.com/disney-and-epic-games-fortnite/>

⁸ <https://www.independent.co.uk/tech/there-s-alternative-minecraft-server-without-any-rules-gaming-tech-a7323201.html>

⁹ <https://www.cnn.com/2023/10/26/kids-are-attending-pro-palestinian-protests-on-roblox.html>

¹⁰ Wang, J., Makowski, S., Cieřlik, A., Lv, H., & Lv, Z. (2023). Fake news in virtual community, virtual society, and metaverse: A survey. IEEE Transactions on Computational Social Systems.

¹¹ <https://www.theverge.com/2024/8/28/24230531/meta-pretens-horizon-worlds>

Table 1: List and characteristics of popular virtual spaces.

	Tech	Focus	Ownership	Community	Economy	Users	Target Audience	Unique Features	Monetization	Content Creation Tools	Avatar Interaction	Community Governance	Regional Popularity
Fortnite	PC, Console, Mobile	Battle Royale, social events	Centralized	Gaming	In-game currency (V-Bucks), Battle Pass, cosmetic items	Over 400 million registered players globally	Teens & young adults	Live events and concerts	In-game purchases for skins and passes	Limited custom content creation	Limited emotes and dances during gameplay	Centralized, with developer control	Global, very popular in the U.S. and Europe
Minecraft	PC, Console, Mobile	Sandbox game, creativity	Centralized	Gaming & Creative	In-game currency (Minecoins), marketplace for user-generated content	Over 300 million copies sold, estimated 170M+ monthly active users	All ages	Extensive modding community	In-game purchases, game sales	Built-in editor and mods	Basic avatar visibility and movement	Centralized, with some community moderation	Global, especially popular in the U.S., Europe, and Asia
Roblox	Browser-based, PC, Console, VR	Gaming, user-generated content	Centralized	Largely younger	In-game currency (Robux), user-generated marketplace	Over 71.5 million daily active users as of Q4 2023	Kids & teens	Vast user-generated worlds	In-game purchases for items and game passes	Roblox Studio for game creation	Extensive customization and interaction in games	Centralized, with player moderation	Global, very popular in North America
Zepeto	PC, Mobile	Avatar-based social platform, virtual fashion	Centralized	Fashion-focused	In-game purchases, virtual goods marketplace	Over 300 million registered users	Teens & young adults	3D avatar customization	In-game purchases for outfits and items	Limited customization tools	Avatar interactions in social settings	Centralized, with some moderation	Popular in Asia, especially South Korea
VRChat	VR, PC	Social interaction, virtual worlds	Centralized	Tech-savvy, Casual	VRChat Plus subscriptions, virtual goods marketplace	Over 10 million registered users	Teens & adults	User-created worlds and avatars	VRChat Plus subscriptions and in-game purchases	Custom avatar creation tools	High interactivity with gestures and emotes	Centralized, with community reporting	Global, particularly popular in the U.S. and Europe
IMVU	PC, Mobile	Virtual fashion, social interaction	Centralized	Fashion-focused	Virtual currency (Credits), virtual goods marketplace	Over 20 million registered users	Adults	Customizable avatars and rooms	In-game purchases for items and features	Limited room design tools	Avatar interactions through chat and activities	Centralized, with community guidelines	Popular in the U.S. and Europe
Second Life	PC	Virtual world, social interaction	Centralized	Long-standing	Virtual currency (Linden Dollars), virtual goods marketplace	Estimated 200,000-500,000 active monthly users	Adults	Extensive user-created content	Land sales and virtual goods marketplace	Building tools for user content	Rich social interactions with customizable avatars	Centralized, with community moderation	Popular in North America and Europe
Horizon Worlds	VR, PC	Social interaction, virtual worlds	Centralized	Tech-savvy	Virtual goods marketplace, future monetization planned	Over 500,000 monthly active users	Adults	Integrated with Meta's ecosystem	Virtual goods marketplace	World building tools	Social interactions with avatars	Centralized, with developer oversight	Growing in popularity in North America
Decentraland	PC, Web	Virtual real estate, blockchain-based economy	Decentralized	Crypto-enthusiasts	Decentralized, blockchain-based economy (MANA)	Approximately 10,000 daily active users	Adults	Ownership of virtual land	Sales of virtual land and NFTs	User-built content creation	Basic interactions within a decentralized environment	Decentralized governance by users	Popular in crypto-enthusiast circles
Rec Room	VR, PC, Mobile	Social games, virtual worlds	Centralized	Casual	In-game currency (Tokens), virtual goods marketplace	Over 3 million monthly active users	Teens & adults	Game creation and sharing	In-game purchases for cosmetics	Game creation tools	Engaging avatar interactions in games	Centralized, with player reporting	Gaining popularity in North America
The Sandbox	PC, Web	User-generated content, blockchain-based economy	Decentralized	Crypto-enthusiasts, creators	Decentralized, blockchain-based economy (SAND)	Approximately 30,000 monthly active users	Adults	Virtual land ownership and monetization	Sales of land and NFTs	Game creation tools	Basic avatar interactions tied to user-generated content	Decentralized governance by users	Popular in crypto-enthusiast circles
Somnium Space	VR, PC	Virtual real estate, blockchain-based economy	Decentralized	Crypto-enthusiasts	Multi-cryptocurrency support, primary token (CUBE)	Under 10,000 monthly active users	Adults	Immersive VR experiences	Sales of virtual land and NFTs	Limited building tools	Avatar interactions in immersive environments	Decentralized governance by users	Niche popularity, growing in crypto circles

2 Data Collection from Decentralized Social Media

In this section, we report the efforts done in AI-CODE for the automated data collection of social media data coming from decentralized sources. The development of automated data collection procedures is in particular fundamental for the activities of the T4.1 AI-CODE service ("Disinformation detection in next-generation social media"). We explored three possible sources: i) the highly decentralized Mastodon, that act as the main template for federated social media; ii) Bluesky, that is intensively growing in active population, is technically based on a federated protocol but at the time of the writing is mostly cantered around its main instance, and the smaller Lithuanian news website 77.lt, which is part of the Cross.social platform (built on ActivityPub) that however represent a strong example of a likely source of disinformative content.

2.1 Mastodon API

Mastodon offers a rich set of free API endpoints for collecting public data about instances and users via simple HTTP requests. These are usually performed on specific target instances of interest, e.g., if the desired information is available at the `/api/v1/example_info` endpoint from the Mastodon APIs, it is possible to retrieve it for a given instance (e.g., <https://mastodon.example/>) by performing a GET HTTP request to the https://mastodon.example/api/v1/example_info endpoint. Similar considerations apply to users, whereby requests are typically to be performed toward the specific instances hosting them. Table 2 summarizes the main Mastodon endpoints, along with their description, taken parameters, and response details.

Table 2: Main Mastodon API endpoints.

Endpoint	Description	Parameters	Response Details
GET /api/v2/instance	Retrieves all key metadata about a given server	None	- Software version - Server description - Aggregated usage statistics (e.g., monthly active users) - Contact details - List of rules - Technical metadata
GET /api/v1/instance/rules	Retrieves the list of rules enforced within the server	None	List of rules in JSON format
GET /api/v1/instance/extended_description	Retrieves the “extended description” containing more detailed information about the instance	None	Text containing the extended description
GET /api/v1/instance/activity	Provides weekly aggregated statistics about the instance’s activity over the last three months	None	- Weekly number of created statuses (posts) - Weekly user logins - Weekly user registrations
GET /api/v1/instance/domain_blocks	Retrieves information on moderated instances (blocked or silenced) and the reasons for these actions	None	DomainBlocks objects in JSON format, with details on moderated instances and reasons
GET /api/v1/accounts/:id/followers	Retrieves the list of followers of a given user ID	User ID Optional: limit (default 40,	List of followers in JSON format, with metadata about each user

		max 80)	
GET /api/v1/accounts/:id/following	Retrieves the list of accounts followed by a given user ID	User ID Optional: limit (default 40, max 80)	List of following accounts in JSON format, with metadata about each user
GET /api/v1/accounts/lookup	Retrieves user information by their Mastodon handle	acct (Mastodon handle)	User information in JSON format, including the ID (needed for most endpoints)

Finally, it should be noted that the distributed nature and the continuous emergence of new instances poses challenges in tracking all instances, which reflects into both (server or user) scenarios. To address this issue, there exist some platforms that continuously discover and index Mastodon instances, such as *instances.social*, that currently tracks approximately 16,000 servers. These platforms exhibit specific API endpoints that allow the straightforward retrieval of the instances list.¹²

2.2 Bluesky API & Firehose

Bluesky data was retrieved via two different methods – via authenticated and targeted API requests, as well as through the Bluesky Firehose.

2.2.1 The Bluesky Firehose

2.2.1.1 API

Bluesky’s Firehose data is easily accessible through an access point described in the public documentation¹³. Most notably, no credentials are required to get access to the following real-time messages *posts*, *reposts*, *likes* and *follows*. It should be noted, however, that the firehose does not provide data to an arbitrary level of detail. As an example, user authentication needs to be used for image retrieval: while an opaque image reference is provided through the firehose, the actual image needs to be retrieved via a separate API call, requiring authentication. The advantage of using the Firehose API is that no rate limits apply. Instead, the network’s activity can be monitored in real-time.

In the following subsection, we characterize a sample of data collected from the Firehose. The goal of investigating this data was to assess whether trending misinformation can be identified based on this data, possibly augmented with a low number of authenticated API calls that do not run against the rate limits. The results of this investigation can later be used to setup a more target collection of data, concentrating on content that potentially carries disinformation.

2.2.1.2 Capturing Data

The data was captured using a set of Python scripts based on the *atproto* Python library¹⁴. The scripts capture the incoming data stream through a WebSocket, then extract relevant data and finally store this data in a SQLite3 database. The statistics of the collected data are provided in Table 3.

Table 3: Statistical description of the Bluesky data collected.

The size of the uncompressed sqlite3 file	255 MB
Total number of posts	260,775
Start of data collection	2024-05-17 13:18:30 UTC
End of data collection	2024-05-17 19:52:55 UTC
Time range	6h 34m 25s

¹² <https://instances.social/api/>

¹³ <https://docs.bsky.app/docs/advanced-guides/firehose>

¹⁴ <https://github.com/MarshalX/atproto/tree/main>

Average number of post messages observed	11 messages per second
--	------------------------

2.2.1.3 Evolution of the number of posts per minute

During the observation period, an average of 658 posts were captured every minute. The graph in Figure 6 shows a slight decrease in activity over time. We also note a series of the activity spikes falling at exact hours, which represent a clear marker of automated posting activity, possibly associated with the diffusion of social bots.

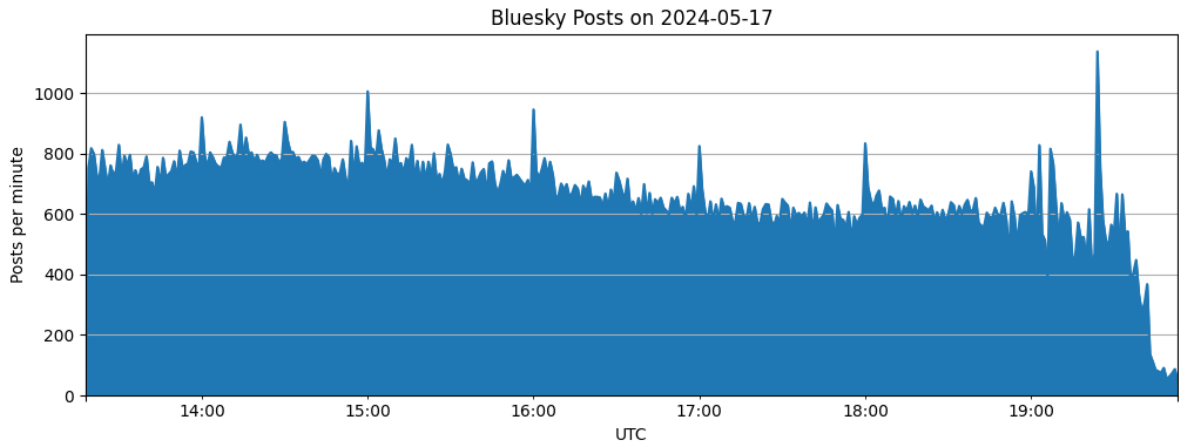


Figure 6: Activity graph of an example of scraping of BlueSky posts over time.

2.2.1.4 Languages

A total of 88 languages were identified by the Bluesky system, covering 95.2% of the posts. In other words, about 5% of posts were not mapped to a specific language by the Bluesky system. The most diffused languages where: English (48%), Japanese (23%), German (10%), Portuguese (4%), French (2%), Spanish (1%), Dutch (1%) and Turkish (1%).

2.2.1.5 Hashtags

Out of the total number of 260,775 posts,

- 17,434 posts or 6.7% contained at least one hashtag;
- a total of 45,847 hashtags were detected;
- after conversion to lower-case, this could be further reduced to 19,821 distinct tags.

The most frequent hashtags were: *totp*, *art*, *nowplaying*, *furry*, *fursuitfriday*, *aiart*, *furryart*, *photography*, *fcf*, *vtuber*, *ai*, *nsfw*, *digitalart*, *bandcamp*, *sinkan*, *ff14*

2.2.1.6 Labels

The Bluesky moderation system features labels placed on content by moderation services¹⁵. Out of a total number of 260,775 posts, 1439 or 0.55% were labeled by the Bluesky system. Four distinct labels were detected: *porn*, *nudity*, *sexual*, and *spoiler*.

2.2.1.7 Accounts

A preliminary analysis reveals that 260,775 posts recorded were created by 71,768 distinct accounts. This means that on average, each account created 3.6 on average posts during the observation period of less than 7 hours.

Unsurprisingly, the frequency of sending posts is not equally distributed between authors. The following Figure 7 shows the average time that passed between posts by the same author (here referenced by author number on the x-axis).

¹⁵ <https://docs.bsky.app/docs/advanced-guides/moderation>

It shows that the most prolific authors posted at a rate of slightly over 1 post per minute (bottom left of the graph), while the least active users only posted once (top right of the graph). For the latter, the number of minutes between posts is the same as the entire recording duration, which was 394 minutes.

The graph also shows that about 600 users, or 0.8%, posted at average time intervals of 10 minutes or less.

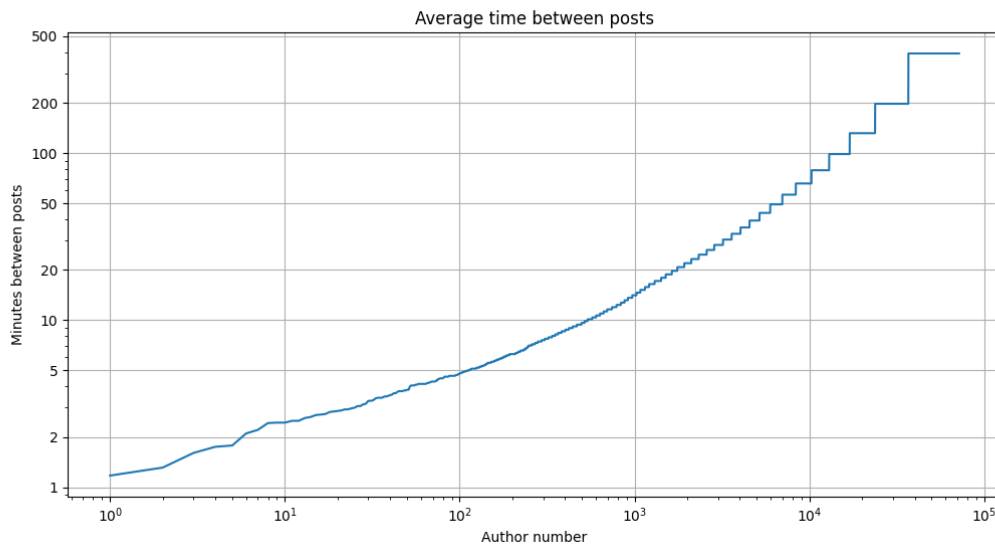


Figure 7: Average time passing between posts by the same author (author number on the x-axis).

2.2.1.8 The database structure

The output of the Firehose has been archived in a database following the schema shown in Figure 8.

2.2.1.9 The tables and their relationships

The schema's central table is the **posts** table. It represents a Bluesky *skeet* (equivalent of a *tweet* in Twitter/X terms). A post can contain several links and images, each of them are extracted into their own **links** and **images** tables. Each link or image uses the *postsRowId* foreign key to link to its related post's *rowId*. Note that the images table in fact only contain image identifiers – The actual images still need to be retrieved using authenticated calls from the Bluesky servers.

This data is stored in the **interactions** table. The *postUri* identifies the message that was interacted with. Note that this message might have been captured in the **posts** table, but not necessarily so. For each post contained in the **interactions** table, the number of observed likes and reposts are counted via *likesCounter* and *repostsCounter*, respectively.

Note that the counters only track the number of observed *likes* and *reposts* messages while running the scripts. This means that any previous interaction related to the same post is not taken into account here.

The **evolution** table is an extended version of the **interactions** table. While **interactions** simply sum up the number of observed likes and reposts, the **evolution** table tracks the timestamp (*addedToDB*) of each individual interaction. This allows an assessment of how the popularity of messages evolve over the observed time period.

Note that an evolution is tracked only for posts for which at least two likes or two reposts messages are observed.

Follows messages are added to the **follows** table. Note that the user identifiers stored in *followingUser* and *followedUser* can also appear in the **posts** table's *author* field, but not necessarily so. The reason is that a given post's author might not receive any followers during the observation period, nor might the author follow another user.

The **users** table provides a perspective on the popularity of users that are followed by others (i.e. the

followedUsers). Each time a user is followed, the user's entry in the **users** table is either created or updated (using the *followersCount* field).

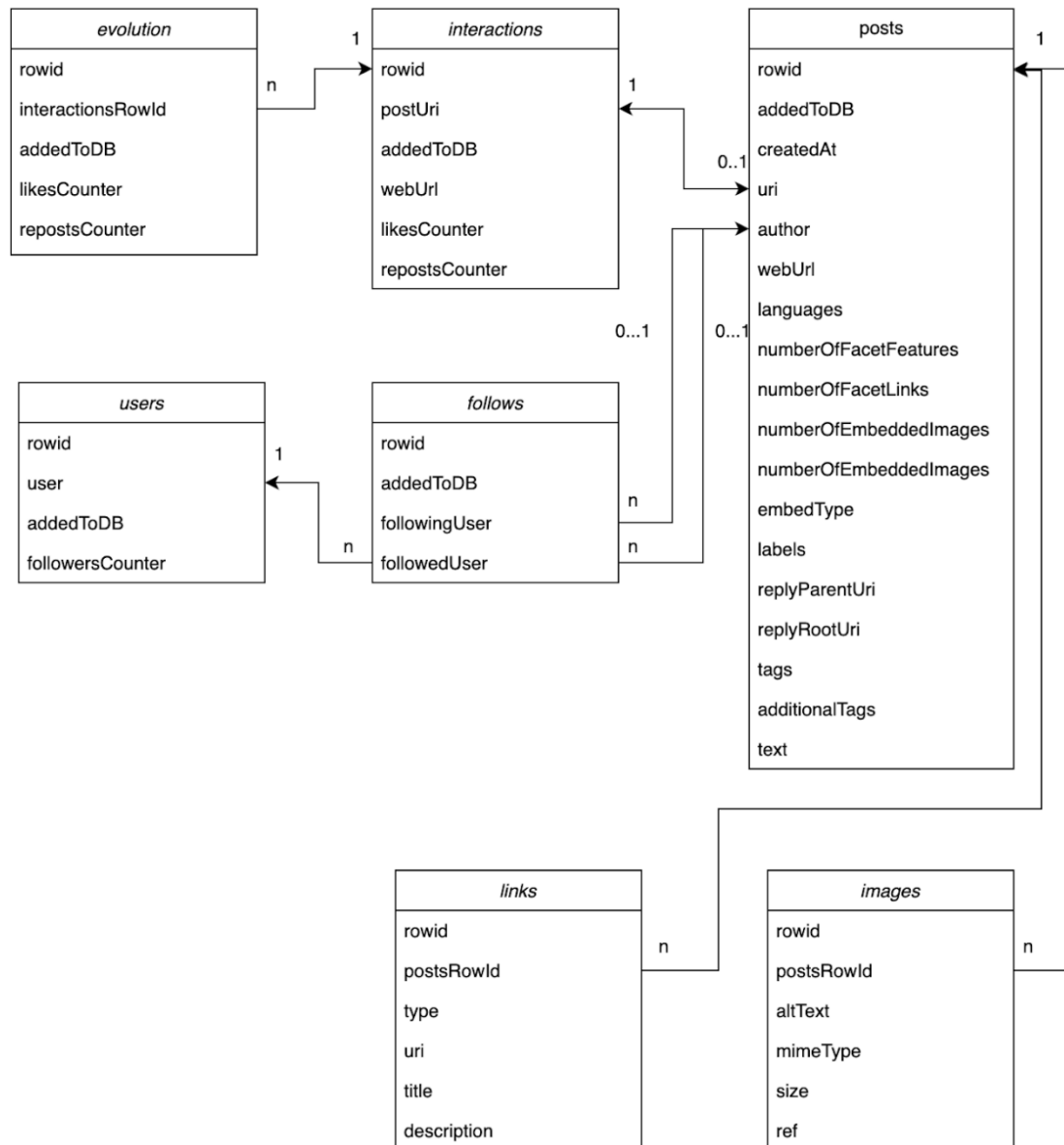


Figure 8: Functioning of the database for the outputs of Firehose.

2.3 77.lt

77.lt is a Lithuanian social media platform that operates on a decentralised social networking model, but is also known for spreading disinformation content, such as anti-vaccination or pro-Russian content. An exploratory analysis by Debunk.org suggests that most users on this platform are in fact fake accounts.

For this analysis, Debunk.org gathered information about 22,533 via webscraping. Of these, 13,684 accounts (60.7%) were created on the same day, and all had the same name: 'Varis Atgalis' (Figure 9). The forthcoming Debunk.org report will provide a full description of the methodology that showed that the 77.lt platform is a *de facto* bot farm.

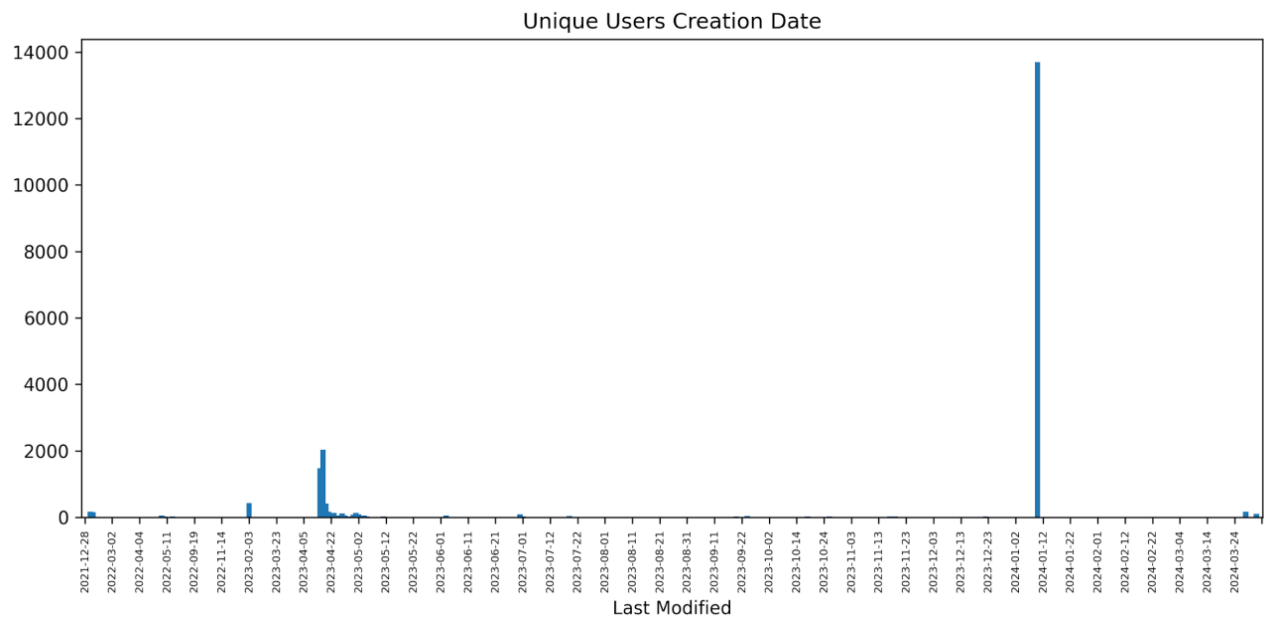


Figure 9: 77.lt's amount of users created by date.

Debunk.org retrieved 28,742 comments, which came from 966 unique users. Users were then flagged based on an analysis of their activity. Factors considered in determining whether an account exhibited the characteristics of a bot included primarily the metadata of the account, rather than the content of the comments. Considerations included overactivity (an above-average number of comments in a short period of time), the generic nature of the comment and the use of profile pictures generated using the Generative Adversarial Networks (GAN) technique (Figure 10). Classification of the images in terms of the representation of the face generated by the GAN was carried out using an open-sourced image classification tool¹⁶.

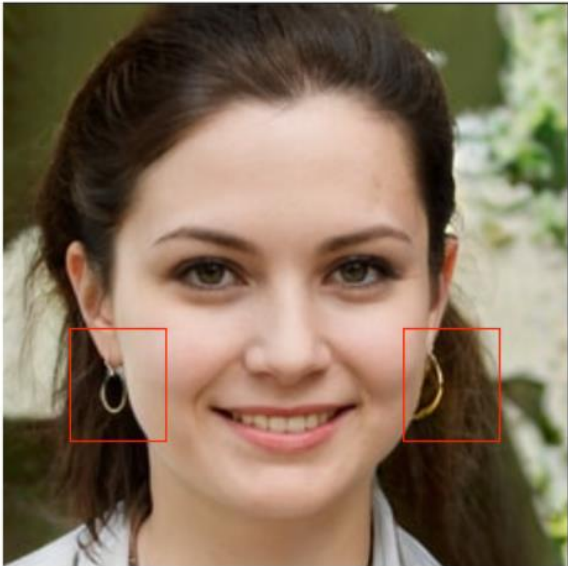


Figure 10: Example of a profile with a GAN-created profile picture, found in 77.lt. Source: 77.lt.

¹⁶ Yang, K. C., Singh, D., & Menczer, F. (2024). Characteristics and prevalence of fake social media profiles with AI-generated faces. *arXiv preprint arXiv:2401.02627*.

3 Social Network Reconstruction

In this section, we describe the delicate process between data collected and the reconstruction of the social structure. In this case, besides the two federated example of Mastodon and Bluesky, we add the results of an exploratory analysis on the crypto-based virtual world Decentraland.

3.1 Mastodon instances network

In section 3.1 we described how Mastodon data can be collected. Being among the goals of AI-CODE to be able to track the diffusion of disinformation, in Federated social media this can be mapped not only as it spreads between different users, but also as it propagates from instance to instance. This is relevant as all users of an instance can access in a particular page the content from other instances on which one of the users engaged through their personal network: if someone in an instance engages with unreliable content, all other users might also read the same unreliable news

One very useful piece of information available is the network of instances, representing how the information is able to flow (or not) between different instances. To build this, we can define a directed network modeling the relations between a set of users U from a set of instances I as $G = \langle V, E \rangle$, where the node set V contains pairs (u, i) with $u \in U$ and $i \in I$, and the edge set $E \subseteq V \times V$ contains the following relations among users, such that any $(x, y) \in E$, with $x = (u, i)$ and $y = (v, j)$ indicates that user u registered on the instance i follows the user v registered on the instance j . It should be noted that users from different instances might have the same username, i.e., $u = v$ provided that $i \neq j$. Similarly, users can be registered in the same instance, if they have different usernames, i.e., $u \neq v$ if $i = j$.

From this initial network modeling, we can infer a network model aimed at shaping relations among the different Mastodon (and non-Mastodon) instances I as follows. We define a directed-weighted network $G_I = \langle V_I, E_I, s \rangle$ where $V_I = I$ represents the set of instances, and E_I is the set of edges such that $(i, j) \in E$ means that there is at least one user registered on instance i following one user from instance j . Finally, $s: E \mapsto R$ is an edge weighing function that assigns each edge with the corresponding number of users from the source instance following users on the target instance.

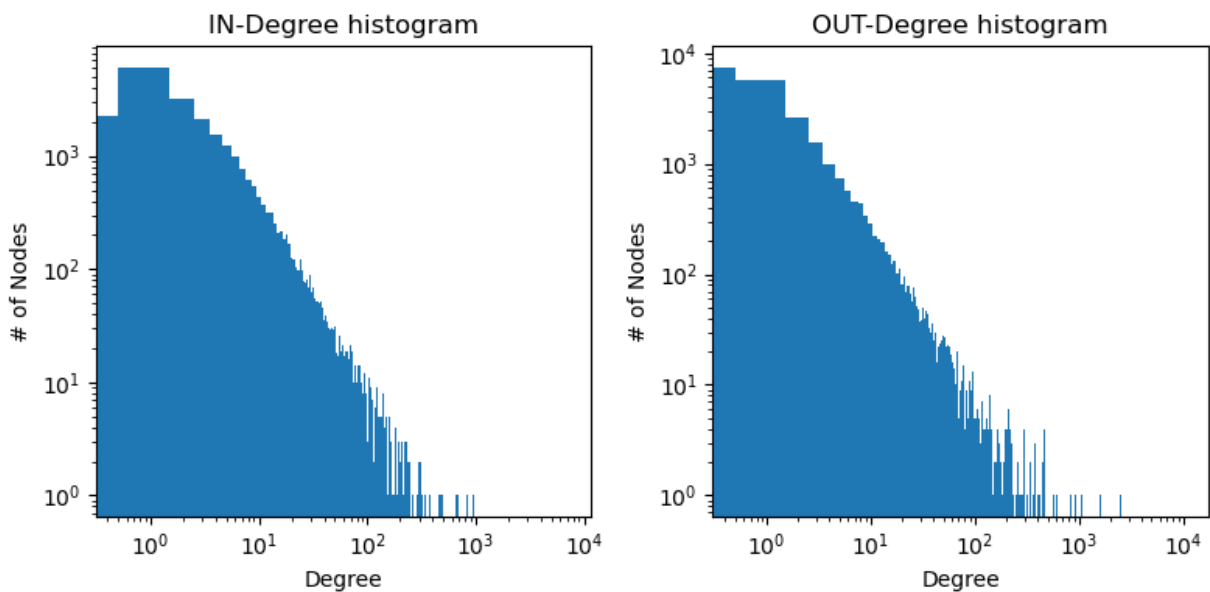


Figure 11: In-degree and out-degree histograms.

Figure 11 shows the in-degree and out-degree distribution of the resulting network.

3.2 Copresences in Decentraland

At this early stage of their diffusion, little is known of how information might propagate in Virtual Worlds and Metaverses. As to the physical world, one can have epidemics spreading by contact¹⁷ and, similarly, people disseminating information with the word-of-mouth. Knowing the position of users in virtual worlds allows us in principle to reconstruct the social network of their contacts and to study how information can spread across this virtual community through word-of-mouth.

For doing this, we focus on Decentraland, a blockchain-based virtual world where users can buy, sell, and create digital assets, including land, artwork, and immersive experiences. Built on the Ethereum blockchain, it allows users to explore themed districts, attend virtual events, and interact in real-time. Each user navigates the space using a unique digital identity, engaging with other participants across a dynamic and ever evolving metaverse. As user buy and sell assets, this recorded in blockchain together with where the transaction happened, allowing us to reconstruct the position of users and thus their contact network.

3.2.1 Data collection

Data was gathered using the methodology recently introduced by Vasan and colleagues¹⁸, focusing on two primary sources: Decentraland's public API and blockchain APIs, specifically Etherscan¹⁹ and Polygonscan²⁰. The Decentraland API provided essential details about active islands and user interactions. Information collected included island IDs, center coordinates and user-specific attributes such as user IDs, blockchain addresses, last activity timestamps, and 3D position coordinates. These data was collected continuously, with each entry timestamped to track user activity over time. The collection process saved data every 30 minutes to manage memory efficiently and included robust error handling to ensure the process remained uninterrupted despite occasional network issues.

Complementing this, blockchain transaction data was retrieved using Etherscan and Polygonscan, focusing on NFT transfers involving Decentraland users. Transaction data, such as block number, timestamp, transaction hash, wallet address, contract address, token ID, and gas values, were fetched in manageable batches. The collected information was then organized into separate DataFrames for Ethereum and Polygon, enabling a clear view of user interactions across both networks.

3.2.2 Copresence network construction

Using raw data from these interactions, we have constructed a copresence network to analyze user engagement and connections within Decentraland. In this network, each node represents a unique user ID, and an edge between two nodes signifies that those users were present in the same land parcel at the same day and hour. The network consists of 612 nodes, with an average of 125 interactions per day across all nodes. Most interactions occur between 3 a.m. and 10 a.m., with a pronounced peak at 9 a.m (Figure 12).

When aggregating the temporal dimension of the network, we observe that 79% of the nodes make up the largest component, while the remaining nodes interact in smaller groups or primarily in pairs. The figure below (Figure 13) shows a snapshot of the aggregated network, where each node's color represents its degree: darker colors indicate nodes with lower connectivity.

Future analysis of this copresence network will potentially provide valuable insights into user behavior and social dynamics within Decentraland, helping to map patterns of interaction and spatial engagement across the platform. Similarly to what is done in epidemics in the physical world, these copresence networks can be at the base of the study of socially transmitted "virtual viruses"²¹. At the same time, models describing on spreading behavior on this copresence network can also describe the diffusion of reliable or unreliable information in this virtual social environment as information in virtual worlds can be transmitted, similarly to

¹⁷ Lofgren, E. T., & Fefferman, N. H. (2007). The untapped potential of virtual game worlds to shed light on real world epidemics. *The Lancet infectious diseases*, 7(9), 625-629.

¹⁸ Vasan, K., Karsai, M., & Barabasi, A. L. (2024). Human Mobility in the Metaverse. *arXiv preprint arXiv:2404.03071*.

¹⁹ <https://etherscan.io/>

²⁰ <https://polygonscan.com/>

²¹ Lofgren, E. T., & Fefferman, N. H. (2007). The untapped potential of virtual game worlds to shed light on real world epidemics. *The Lancet infectious diseases*, 7(9), 625-629.

real words, with the word of mouth.

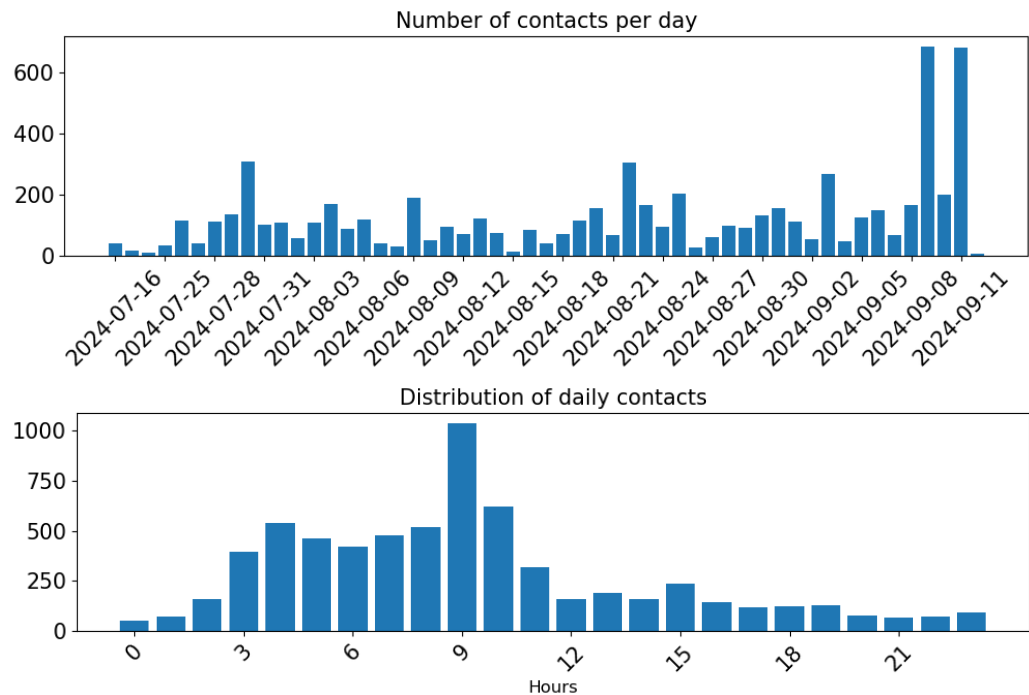


Figure 12: Distribution of interactions.

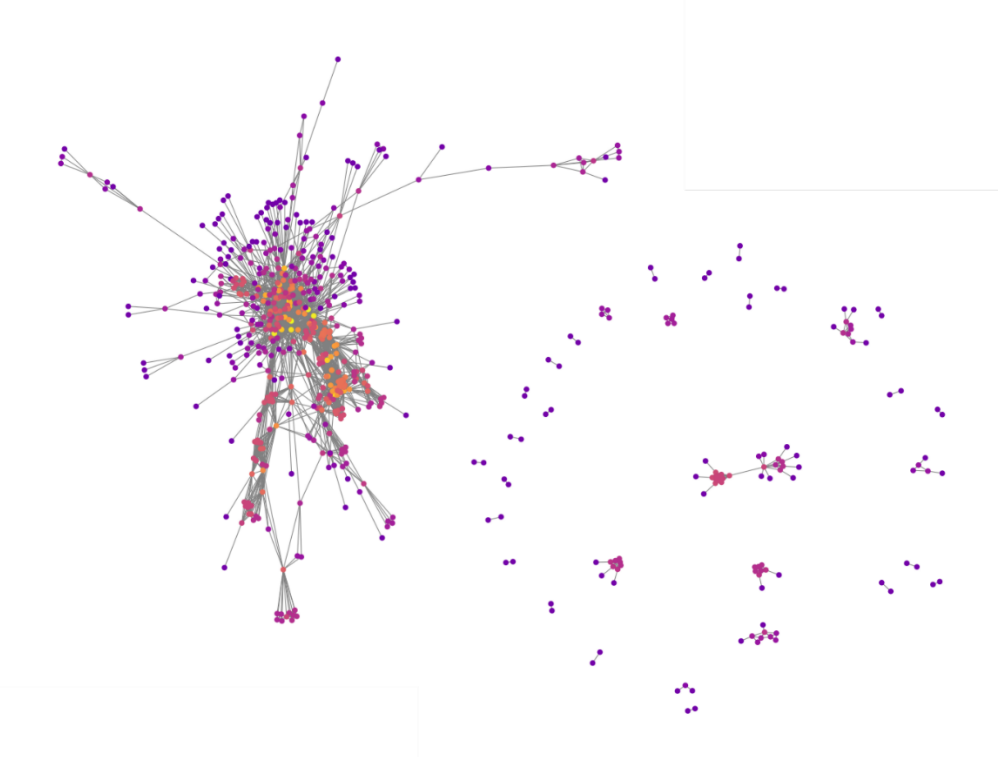


Figure 13: Graph representing the aggregated network.

Future analysis of this copresence network will potentially provide valuable insights into user behavior and social dynamics within Decentraland, helping to map patterns of interaction and spatial engagement across the platform. Similarly to what is done in epidemics in the physical world, these copresence networks can be at the base of the study of socially transmitted "virtual viruses". At the same time, models describing on spreading behavior on this copresence network can also describe the diffusion of reliable or unreliable information in this virtual social environment as information in virtual worlds can be transmitted, similarly to real words, with the word of mouth.

3.2.3 User activity and spatial dynamics

The collected data also highlight the user behavior trends within Decentraland. Figure 14 showcases the trend of unique daily users, highlighting the evolution of active user engagement over time. This trend reveals periods of growth and possible stagnation, linked to key updates or events within the metaverse.

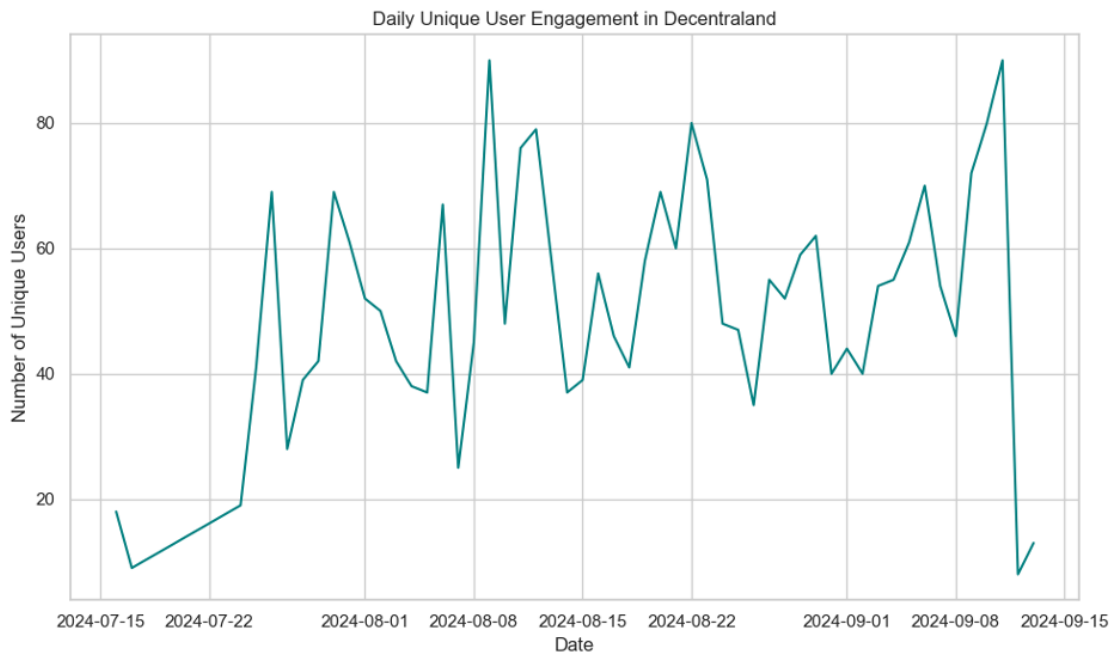


Figure 14: Distribution of unique users interacting within Decentraland.

Furthermore, hourly activity patterns shown in Figure 15 emphasize peak hours from 08:00-10:00, with the highest interactions occurring around 09:00 with comparatively less activity from 06:00-07:00 and 11:00-14:00.

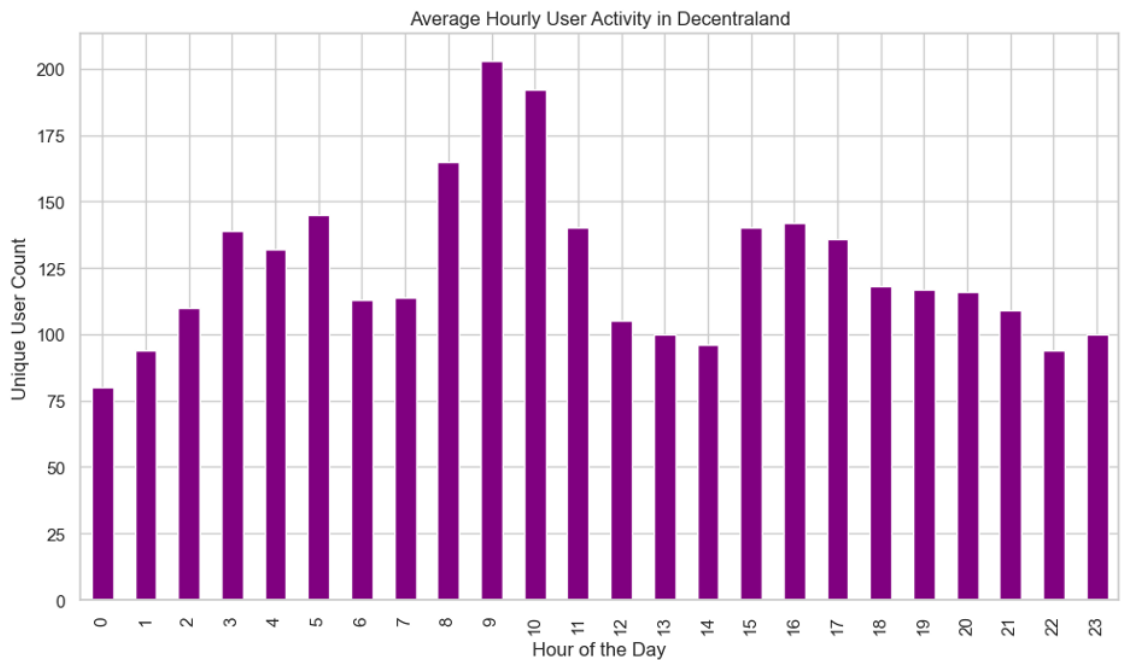


Figure 15: Distribution of users active by hour of the day.

The spatial analysis further identifies the top 10 islands (i.e. the spatial unit in Decentraland) by unique user count, as shown in Figure 16. These islands represent critical hubs of interaction, likely serving as centers for social and economic activities.

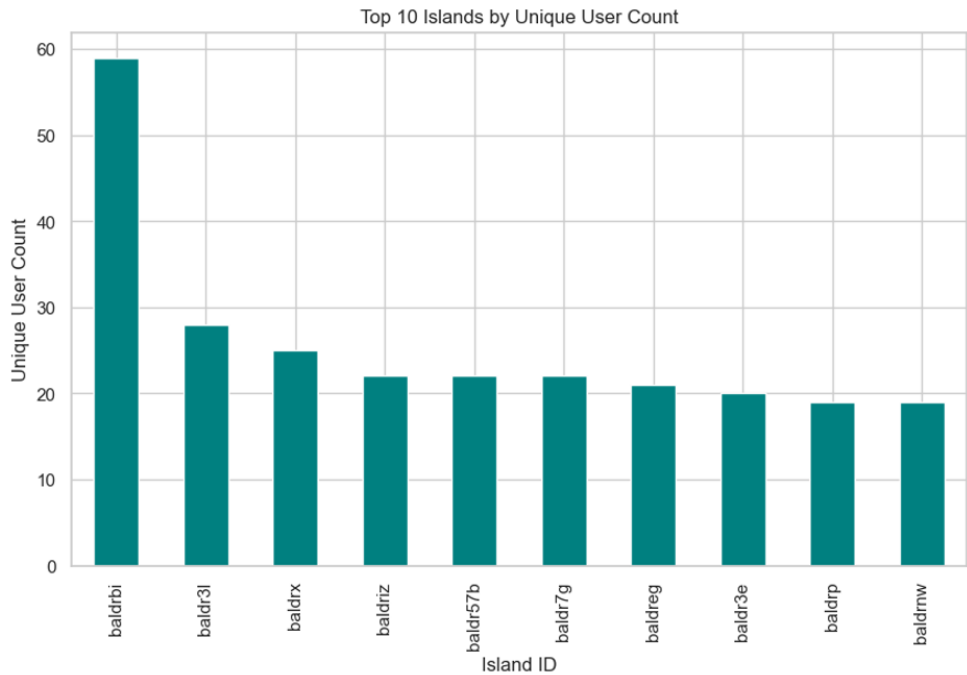


Figure 16: Distribution of top 10 Islands based on the unique user count.

Building on to the distribution of unique users on Islands, Figure 17 provides a spatial map of user density across island centers, highlighting where engagement is most concentrated, that also represent the areas where it is more likely that people meet, communicate and here malicious actors could more efficiently disseminate disinformation.

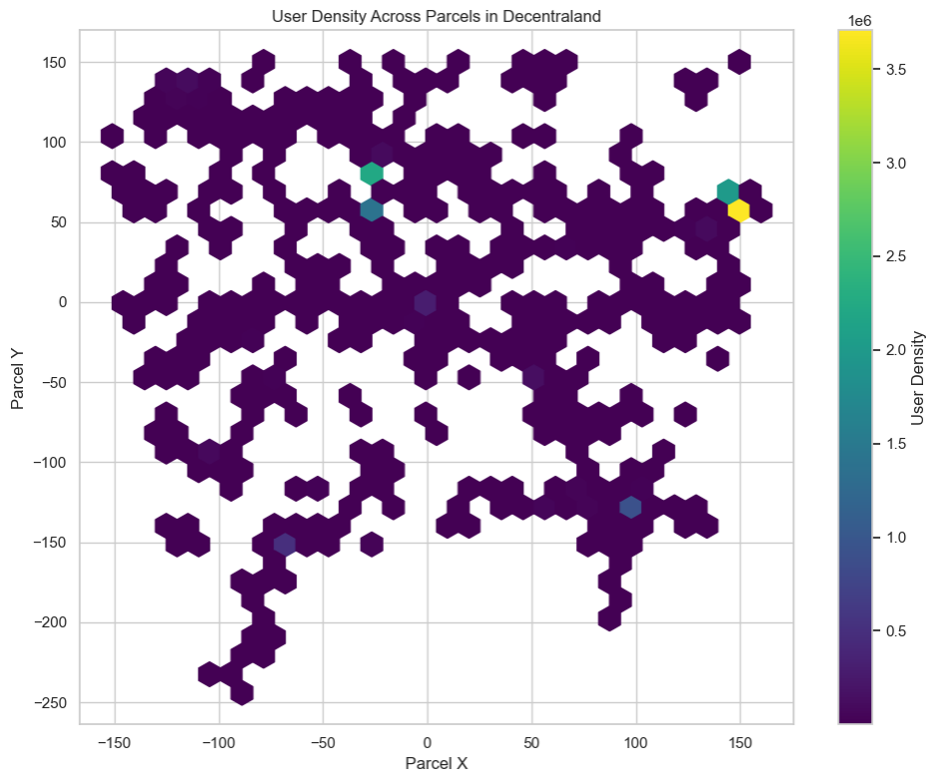


Figure 17: Density distribution of users across the centers of all Islands highlighting regions of high activity.

3.2.4 Blockchain transaction analysis

Our blockchain analysis reveals diverse user engagement across Etehereum and Polygon. Among 1,157 users,

240 engaged in 88,628 Ethereum transactions, while 514 conducted 279,032 transactions on Polygonscan. This variation highlights different usage patterns across the two networks. Figure 18 contrasts the daily unique user counts for Decentraland, Ethereum and Polygon, illustrating engagement fluctuations and identifying peak activity periods.

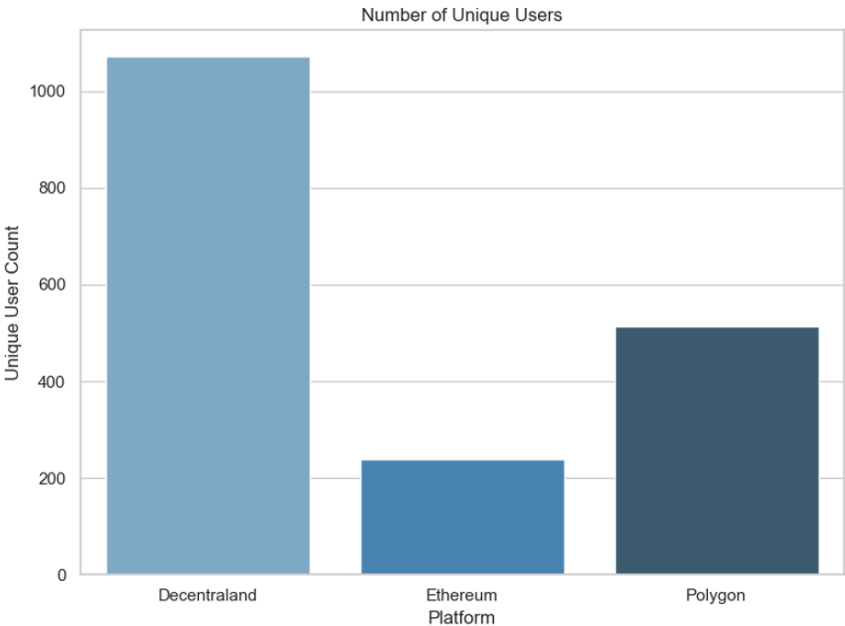


Figure 18: Distribution of unique users across Decentraland, Ethereum and Polygon.

Furthermore, Figure 19 shows the time series distribution of transitions performed on Ethereum and Polygon where Polygon has dominated the transaction landscape over time with significant peaks, reaching more than 800 transactions in the first week of August 2024.

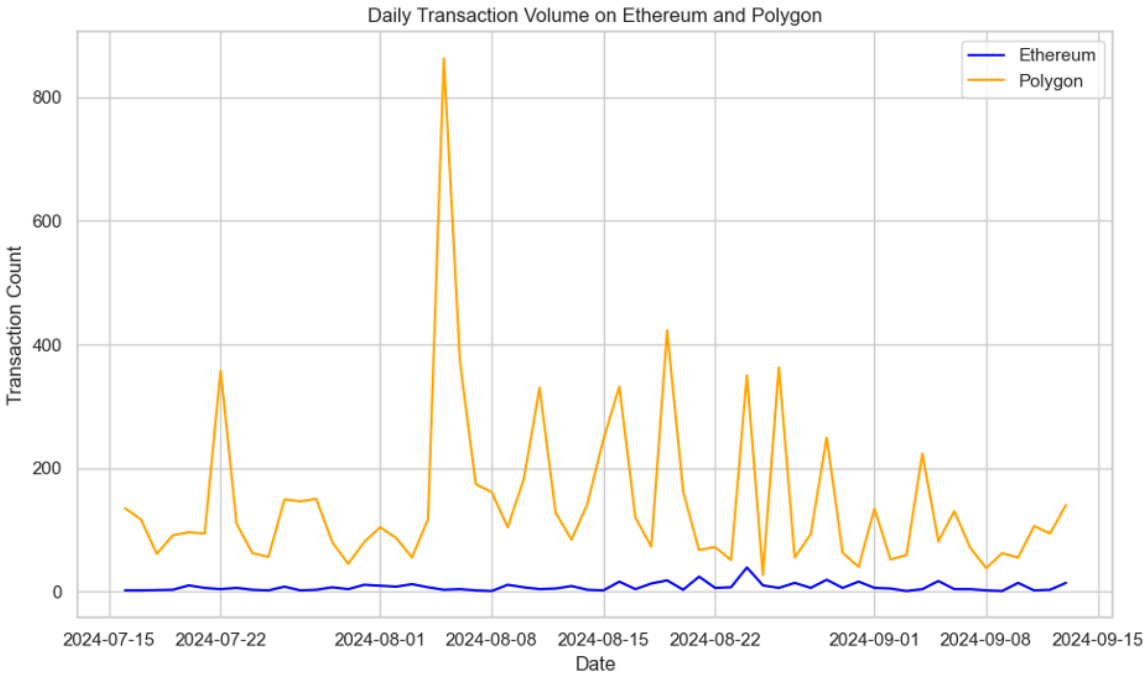


Figure 19: Distribution of Transaction Volume on Ethereum and Polygon.

Whereas Ethereum had comparatively almost non-existent transactions with only a few of them occurring after the second week of August in 2024.

Moving on to the interacted contracts on both networks, Figure 20 presents the top interacted contracts, offering

insight into user preferences and investment trends.

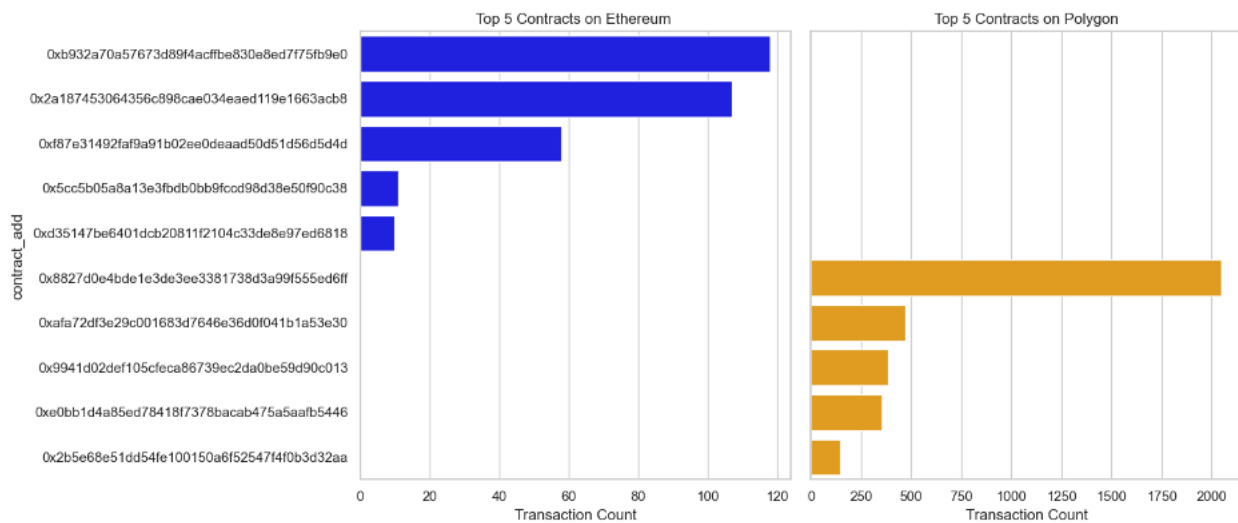


Figure 20: Bar plots distribution for top 5 contracts in both networks.

To conclude, the data collection and copresence network analysis provides a detailed understanding of social interactions, spatial engagement and blockchain transactions within Decentraland. These findings offer insights into the dynamics of blockchain-based Metaverses and the opportunity for studying this virtual social system and to track how (dis)information could propagate in it via word of mouth.

4 Methodologies for the Analysis of Disinformation Spreading and Hate Speech on Federated Social Media

In this section, we describe a set of novel methodological solutions that we designed specifically or adapted from previous studies for the analysis of the disinformation risks of federated social media. These methodologies will be integrated in the services currently being developed in WP4 and WP6.

4.1 Signed ban networks

Decentralized social media and their moderation capabilities introduce novel interaction types involving instances, namely positive and negative ones. The former stems from the following relationships between users, that determine positive interactions among instances, whereas the latter derives from enforced moderation strategies (e.g., bans) among instances. Positive links among instances can be reconstructed starting from the API endpoints reported in Section 3.1 and modeled according to the process described in Section 4.1, whereas the negative ones can straightforwardly collect using the DomainBlock objects as described in Section 3.1. While decentralization makes detecting instances tough, there exists services like instances.social²² that serve as indexes for newly released instances and can be hence leveraged to initialize a set of Mastodon instances to be modeled.

Formally, given a set of Mastodon instances I , their signed ban network can be shaped as follows. We define a directed-weighted network modeling the positive links among instances as $G^+ = \langle V^+, E^+, w \rangle$, where $V^+ \subseteq I$, and $E^+ \subseteq V^+ \times V^+$ is the set of positive edges, such that $(i, j) \in E^+$ indicates that there is at least one user from instance i following a user from instance j . Besides, w is a weighing function such that, for each edge (i, j) , $w(i, j)$ indicates the number of users from instance i following users on instance j .

Likewise, we define a directed network modeling the negative links among instances as $G^- = \langle V^-, E^- \rangle$,

²² <https://instances.social/>

where $V^- \subseteq I$ is the set of instances, and E^- contains the set of negative edges, such that $(i, j) \in E^-$ means that instance i enforced moderation toward instance j .

Finally, the signed ban network can be created as $G_s = \langle V_s, E_s, s \rangle$, where $V_s \subseteq I$, $E_s = E^+ \cup E^-$, and $s: E \mapsto \{+1, -1\}$ is a function that assigns each edge in E with the corresponding sign, such that $s(i, j) = +1$ if $(i, j) \in E^+$, -1 otherwise.

In the negative links network G^- , the directionality of edges captures the roles within moderation interactions. The source node represents the instance enforcing moderation (the *banner*), while the target node represents the instance subject to moderation (the *banned*). Figure 21 illustrates the evolution of moderation patterns over the course of one year. It provides statistics on the daily number of negative interactions (*bans*), the number of banned instances, and the number of banners, highlighting trends and changes in moderation dynamics over time.

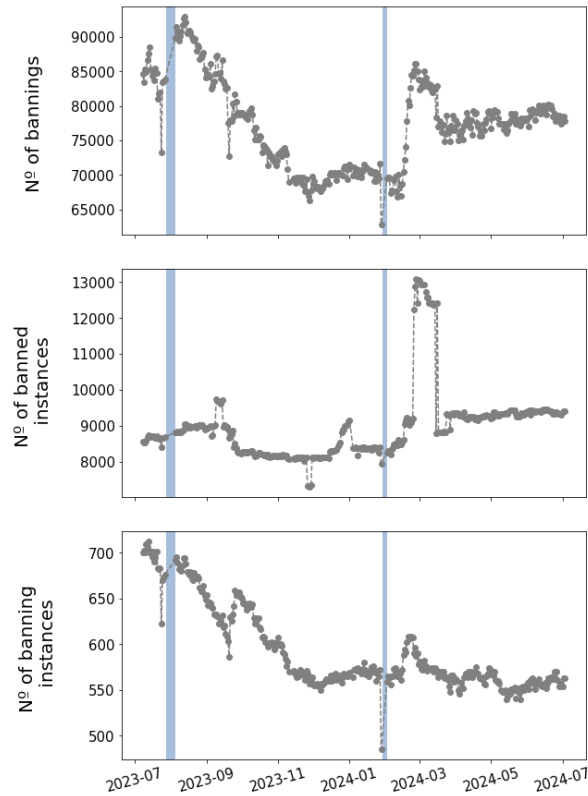


Figure 21: Trends of the moderation patterns. Points in the blue area are excluded due to missing data.

The fact that a small group of instances is responsible for banning near 12000 is also reflected in the in- and out-degree distribution of the network (Figure 22).

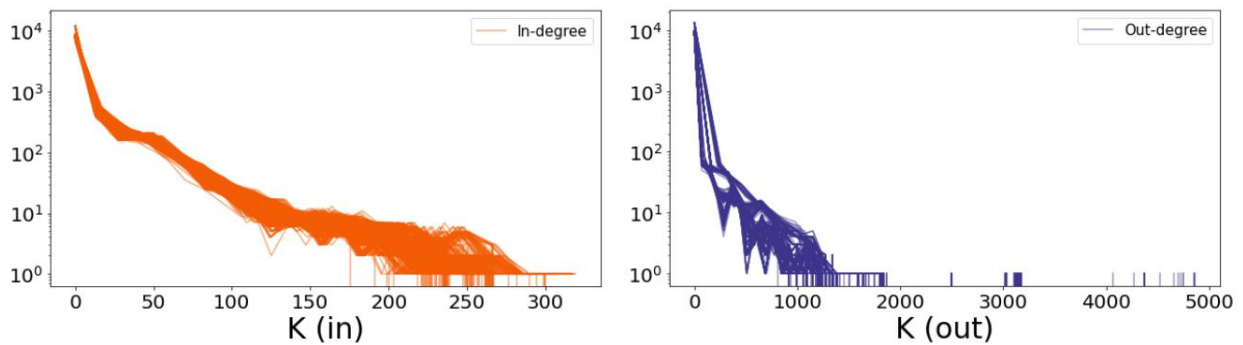


Figure 22: in- and out-degree distribution of the network.

4.2 Information transmission model

As discussed in section 3.1, for federated social media such as Mastodon, it is relevant to understand how information flows not only among users but also among instances, as the social network's feed is directly informed by content on which other members of the same instance engaged.

In this section, we model how information could potentially spread across the Mastodon instances network incorporating both positive and negative interactions among nodes. The network is composed of two layers: a positive network, representing interactions such as following relationships, and a negative network, representing bans. The primary objective is to analyze how information propagates between two groups of nodes, with group membership determined by the structure of the negative network. As we have seen in the previous section, there is little overlap between the domains that enforce bans and those that are banned. This distinction allows us to categorize the roles of domains in the network in two factions: the minority group (m) which comprises the instances that execute bans and the majority group (M) which consists of those receiving the bans. Domains that both execute and receive bans are still considered part of the minority group (see Figure 23). Once we assign each domain the role of 'banner' or 'banned', several questions arise. How does information travel between these two groups? Assuming that the banned group is the source of misinformation, disinformation, or other harmful content, is the moderation process sufficient to contain the spread of their messages? To address these concerns, we propose a model to quantify how information flows within and between these two groups.

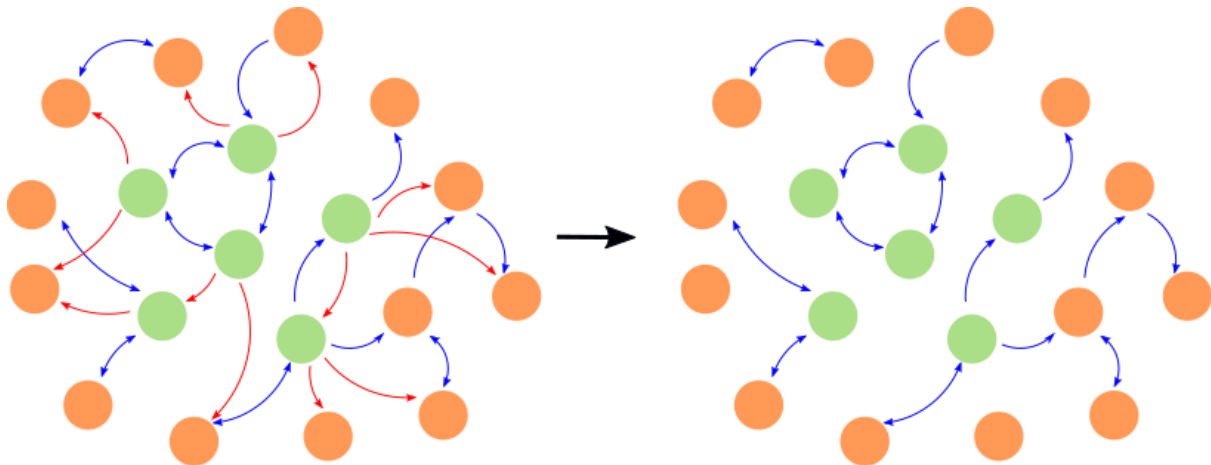


Figure 23: Construction of the network: Nodes are divided based on their role in moderation. Nodes that receive bans form the majority (orange), while those initiating bans form the minority (green). Information spreads exclusively through positive interactions ('follow').

Negative interactions in the network only serve to define the groups to which the nodes belong, as we assume that information can only spread through positive interactions, such as a 'follow'. We study the transmission of information within the network by modeling two types of contagion processes—simple contagion and complex contagion—occurring simultaneously. This hybrid approach, as proposed by Diaz-Diaz and collaborators²³, accounts for both the possibility that an individual may become informed after a single interaction with an informed neighbor (simple contagion) and the fact that adopting a new idea may require multiple exposures from different sources (complex contagion) before the information is transmitted.

We select one of the groups, either the majority (M) or the minority (m), to serve as the source. The source group follows a simple contagion dynamic modeled via a SIR compartmental model. This model considers that the nodes can be either in one of these three states: susceptible (S), informed (I) or recovered (R). A susceptible node which interacts with an informed node will get informed with probability β . Besides, we assume that an informed node will try to spread the idea to their neighbors just once, after which it gets recovered (Figure 24). The nodes who belong to the group which is not the source can get infected through a complex contagion mechanism, implemented via a threshold model, where a susceptible node may get infected

²³ Diaz-Diaz, F., San Miguel, M., & Meloni, S. (2022). Echo chambers and information transmission biases in homophilic and heterophilic networks. *Scientific Reports*, 12(1), 9350.

if the fraction of its informed neighbors is higher than a certain threshold T , ($T \in [0,1]$).

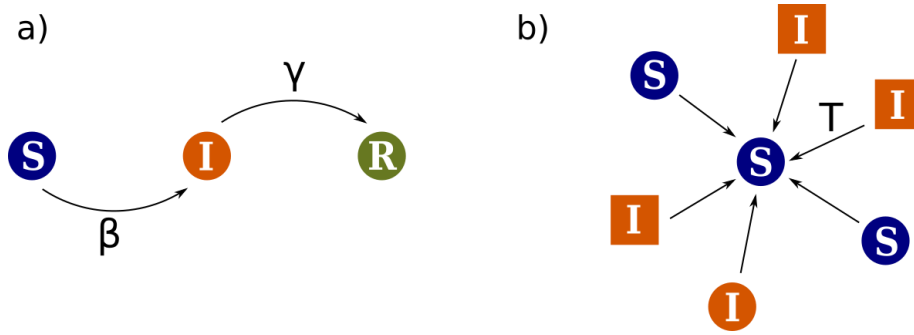


Figure 24: Spreading model: (a) Simple contagion: SIR compartmental model. (b) Complex Contagion, with threshold T . The central node will become informed if it has a fraction of informed neighbors, no matter which group they belong, higher than T .

Since we are interested in understanding how the information is shared within and between groups, after selecting the source group we also designate a target group. The target group may be the same as the source group if we wish to study how information generated by the minority (or majority) spreads within the minority (or majority). Alternatively, the target group may be different from the source group, allowing us to observe how information from the minority spreads within the majority, or vice versa. Summing up, we can analyze four different cases of the information transmission, IT_{st} , where s and t represent the source and target group. These four cases are: IT_{MM} , IT_{Mm} , IT_{mM} , IT_{mm} . We measure the information transmission using the final attack rate which is the fraction of informed nodes in the target group over the total number of nodes in that group. This metric is equivalent to computing the probability of transmitting the information from the source group to the target one, IT_{st} .

4.2.1 Results

We start by assuming that simple contagion occurs with probability $\beta = 1.0$. In the following figure we observe the final attack rate for the four different cases IT_{st} . When the source of the information comes from an instance in the minority group, it reaches almost all (97%) the nodes in that group. In contrast, if the source is in the majority group, less than the 85% of the nodes in that group become informed. Examining the crossed cases (IT_{mM} and IT_{Mm}) when $T = 0$, i.e. when the nodes getting informed under the complex contagion mechanism need only one of their neighbors being informed to stop being susceptible, we observe that 98% of the nodes in the majority would get informed of a message generated within the minority. However, in the IT_{Mm} case, only 80 % of the target group's nodes are reached. For $T > 0.2$, IT_{Mm} decays rapidly, making it easier to transmit information from the minority to the majority (Figure 25).

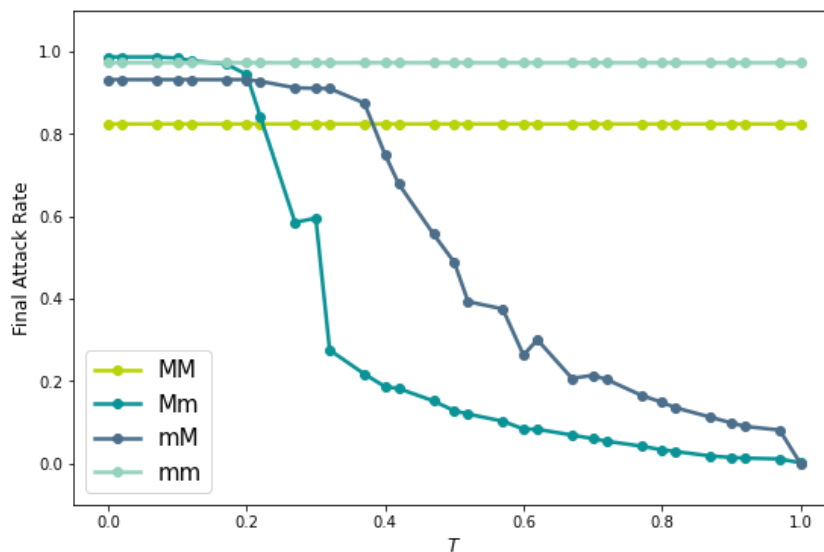


Figure 25: Final attack rate: fraction of informed nodes in the target group for the four different IT cases depending on the value of the threshold T .

We further analyze the final attack rate by examining its behavior under varying values of T and also β . These results are shown in the following figure, where we can observe the effect of varying the probability β . As we had previously seen, the spreading is more effective within the minority group where even for $\beta \sim 0.2$, over 75% of the nodes become informed. The case IT_{Mm} is very efficient for high values of β and low values of T . However, as T increases, there is an abrupt transition, making it difficult for the minority group nodes to become informed by the majority. In the opposite case, IT_{mM} the observed transition is smoother (Figure 26).

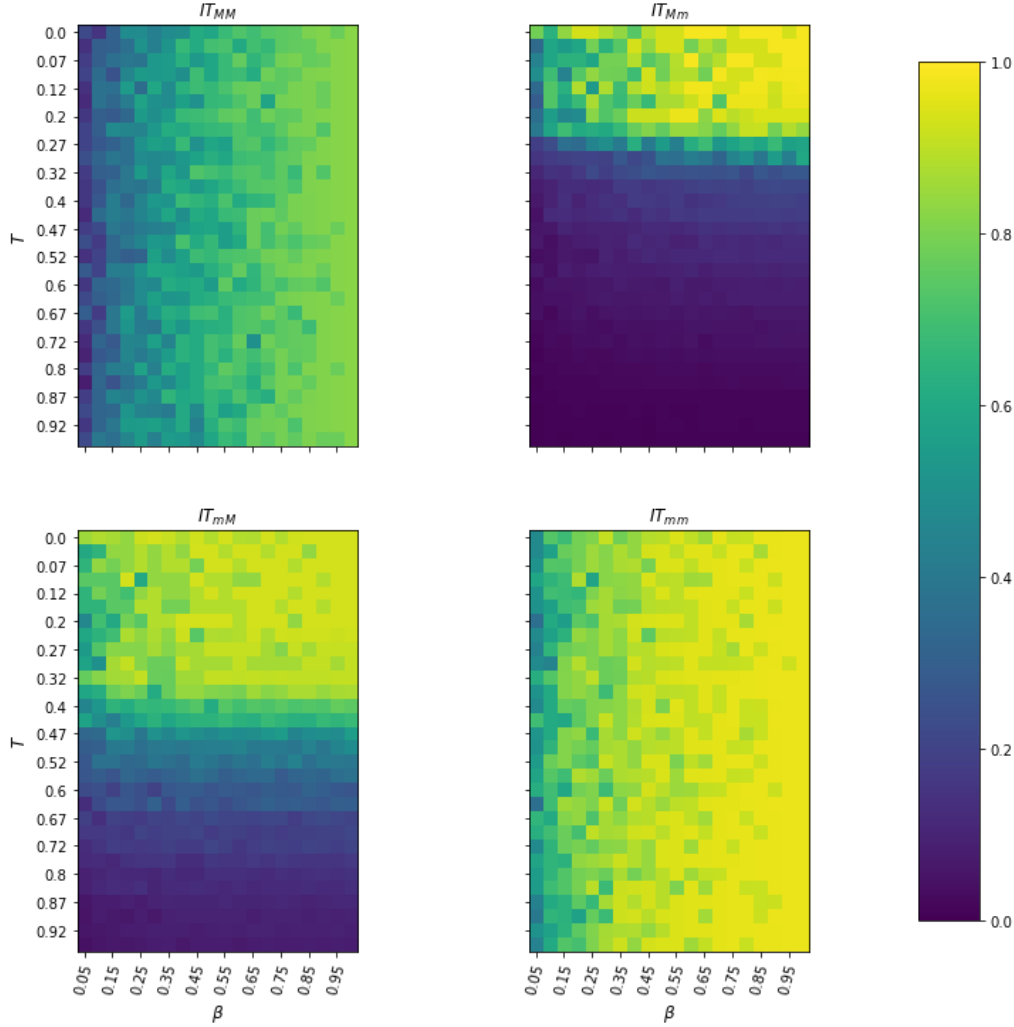


Figure 26: Final attack rate: exploration of the parameters space (T, β) for the fraction of informed nodes in the target group for the four different IT cases.

We can define some further observables in order to assess the information transmission in the Mastodon Network:

$$\bar{IT} = \frac{1}{4}(IT_{mm} + IT_{mM} + IT_{Mm} + IT_{MM}) \quad (1.1)$$

$$B_E = \frac{1}{2}(-IT_{mm} - IT_{mM} + IT_{Mm} + IT_{MM}) \quad (1.2)$$

$$B_R = \frac{1}{2}(-IT_{mm} + IT_{mM} - IT_{Mm} + IT_{MM}) \quad (1.3)$$

$$B_{EC} = \frac{1}{2}(IT_{mm} - IT_{mM} - IT_{Mm} + IT_{MM}) \quad (1.4)$$

The mean information transmission, $\langle IT \rangle$, is an average over the IT_{st} observables and assesses how effective

is the network in transmitting the information, regardless of the role of the nodes. We have that $0 < \langle IT \rangle < 1$. The observables in Eq. 1.2-1.4 have a domain ranging from -1 to 1. B_E and B_R are the emissivity and the receptivity biases, respectively. When their values are equal to zero it means that neither group is more efficient at passing or receiving information. A positive B_E indicates that information originating in the majority propagates more effectively, while a positive B_R suggests that the majority has a greater probability of receiving information than the minority. Analogous observations can be made for the minority group when $B_E, B_R < 0$. Finally, we introduce the echo chamber bias B_{EC} which assesses whether the information spreads exclusively within nodes of the same group or across different groups. A value close to one suggests that information is constrained within each group, whereas a negative value signals the presence of an "anti-echo chamber," where information flows predominantly between nodes of opposing groups.

The measures of average information transmission display a logical distribution as illustrated in Figure 27. As we increase T or decrease β , the spread of news in the network becomes more difficult. The emissivity and receptivity biases show no trivial results, both showing a non-monotonous pattern. In particular, B_E reaches its minimum at $T = 0.3$ and $\beta = 0.15$, increasing with higher values of β and as T approaches 0. An inverted pattern is found for B_R where we can observe the maximum values for $T = 0.3$ and high values of β . The fact that the overall exploration of B_E results in negative values highlights the fact that the nodes within the minority are more efficient in emitting the information. Regarding the echo chamber effect, we observe both echo chamber and anti-echo chamber regions, depending on the tuning of the network parameters. When information has a high probability of being transmitted from the source to its neighbors, and as we increase T , an echo chamber is formed.

4.2.2 Discussion

The current rise and adoption of Decentralized Online Social Networks (DOSNs) raises several questions, particularly regarding how these new platforms will respond under the spread of fake news. While numerous studies have tackled the understanding of how information is transmitted in centralized social networks, it is still not clear how the spreading of misinformation and its moderation works in decentralized networks.

In this study we use an instance-based approach to analyze the spreading of information in the Mastodon's federated network. In particular, we focus on a subset of the network where the instances play two distinct roles. On one hand, we have the instances who act as moderators and block certain domains preventing the spread of content on their part. On the other hand, we have the instances who are being banned. We model a hybrid contagion where the information propagates following a simple or complex contagion mechanism depending on whether the node adopting the new information belongs to the same group as the source or not.

From Figure 27 and Figure 28 we observe that the fraction of nodes reached by information originating from their own group is consistently higher within the minority. This suggests that the banned instances, which form the majority, are less well-connected than the smaller faction which is actively moderating the network. Regarding the cross information spreading, we find that when information spreads quickly (high β), a larger fraction of nodes from the minority will get informed by the majority than in the opposite case. If we consider the majority group as a source of harmful content, the scenario is not optimal. However, this particular situation happens only for low values of T and high values of β . A minimal change of these parameters shows that in fact, the nodes in the majority get rapidly isolated while the spreading from the minority to the majority is more resilient under changes of the network. This suggests that majority group nodes, in general, rely on passing information through minority nodes to stay informed.

Observing Figure 29 we examine potential biases in the spreading process. Remarkably, the emissivity bias is always negative indicating that the minority group is more efficient at spreading information. However, there is no clear bias when it comes to receiving information, with B_R remaining between $[-0.2, 0.2]$ after a detailed exploration of the parameter space. It is also interesting that both echo-chamber and anti-echo-chamber effects can be observed. Even in scenarios of high network permeability (low T), the formation of an echo-chamber can be appreciated. This evidences the fact that the moderation process is effective and even if it is enforced in a decentralized way it is effective in isolating the Mastodon users spreading harmful content.

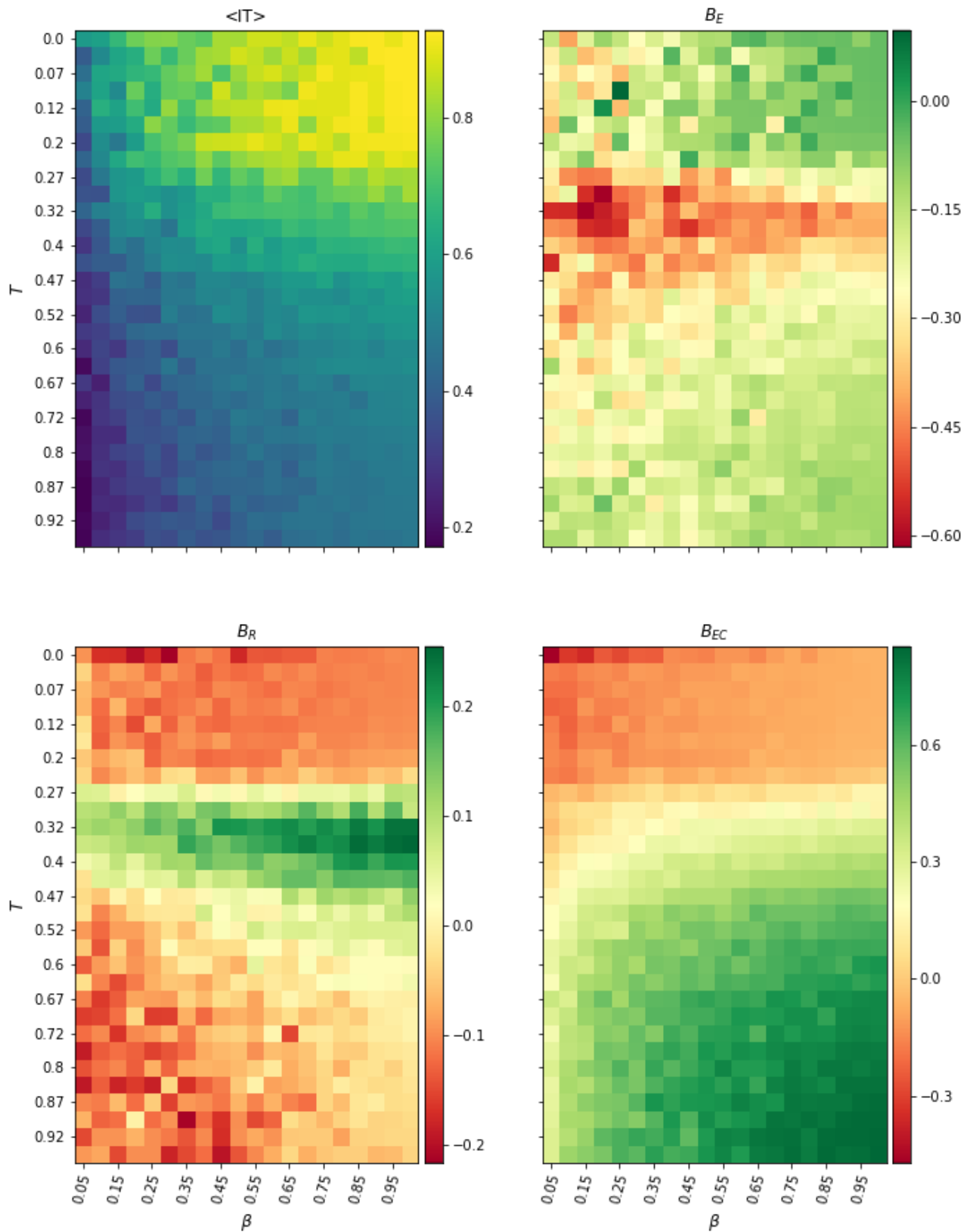


Figure 27: Information biases: Average Information Transmission, Emissivity, Receptivity and Echo-chamber bias.

4.3 Patterns of moderation in Mastodon's network

4.3.1 Data collection

As outlined in Section 3.1, we utilized Mastodon's API endpoints to collect extensive public data about instances. This comprehensive data gathering was crucial to understanding the moderation dynamics across Mastodon instances. Notably, during this process, we encountered challenges related to data obfuscation. For example, some instances deliberately obscured the names of blocked instances to prevent inadvertently promoting them. To address this, we matched these obfuscated names with known domains from our dataset whenever a clear match could be established, such as converting "ma***don.example" to "mastodon.example".

This data collection process was conducted daily over a specified duration, resulting in a total of 80,649,891 moderation activities. After filtering out entries that lacked reasons for moderation, we were left with 27,397,146 entries. These entries were then categorized into four distinct groups: Harassment (H), Hate Speech (Hs), Misinformation (M), and Spam (S), based on the keywords outlined in the table below.

Category	Keywords	Count
Harassment	harass*	1,268,238
Hate Speech	hate*	3,180,915
Misinformation	misinfo* disinfo* fake*	302,928
Spam	spam*	2,881,362

The highest number of moderation entries fell under the Hate Speech category, with 3,180,915 entries, followed by Spam with 2,881,362 entries. Harassment and Misinformation had significantly lower counts, with 1,268,238 and 302,928 entries, respectively.

4.3.2 Methodology

To explore patterns of moderation within the Mastodon federated network, we constructed a signed, temporal, and directed network denoted as $G^- = \langle V, E^- \rangle$, where the set V represents all instances identified during our data collection, and E^- denotes the moderation relations. This network was treated as a multiplex network, with each layer corresponding to a specific reason for moderation: Harassment (H), Hate Speech (Hs), Misinformation (M), and Spam (S).

Our analysis focused on examining the extent and frequency of moderation activities for each category within the network. We evaluated the total number of connections of each instance across the network (referred to as the aggregated topological degree k_i) as well as their specific connections within each layer (referred to as the layer-specific degree $k_i^{[\alpha]}$, where $\alpha \in \{H, Hs, M, S\}$).

To quantify how evenly each instance's connections were distributed across the different categories, we computed the participation index (P) using the formula:

$$P_i = \frac{M}{M-1} \left[1 - \sum_{\alpha=1}^M \left(\frac{k_i^{[\alpha]}}{o_i} \right)^2 \right]$$

where M represents the total number of layers and o_i is the overlapping degree of node i .

Based on the values of P_i , instances were classified into three categories:

1. Focused: $0 \leq P_i \leq \frac{1}{3}$
2. Mixed: $\frac{1}{3} < P_i \leq \frac{2}{3}$

3. Multiplex: $P_i > \frac{2}{3}$

Furthermore, we utilized the Z-score of the overlapping degree to distinguish nodes as either hubs or regular instances within the network based on the following equation:

$$z(O_i) = \frac{O_i - \langle O \rangle}{\sigma_O}$$

Here, $\langle O \rangle$ is the average overlapping degree of the nodes in the system and σ_O is the standard deviation. Based on the Z score of their overlapping degree, we classify nodes with $z(O_i) \geq 2$ as *hubs*, and those with $z(O_i) < 2$ as regular nodes. By considering both the multiplex participation coefficient P_i and the total overlapping degree O_i , we define six categories of nodes as shown in Figure 5.2.3.1, where each node is plotted as a point in the $(P_i, z(O_i))$ plane.

4.3.3 Results

In the following Figure 28, we plot the distribution of the Participation coefficient P_i for the moderation activities within the Mastodon network. The distribution of P_i ranges from 0 to 1 and is quite broad. This suggests a wide variation in how instances engage across the four moderation categories. The P_i reflects the extent to which an instance's moderation activities are spread across these categories, while the Z-score of overlapping degree indicates the overall number of connections (moderation activities) for each instance. Instances with $Z(O_i) \geq 2$ are classified as hubs, meaning they are highly active, while those with $Z(O_i) < 2$ are classified as regular nodes. The distribution of $Z(O_i)$ is asymmetric, indicating that most instances have fewer connections than the network average, but a small number of hubs are significantly more connected.

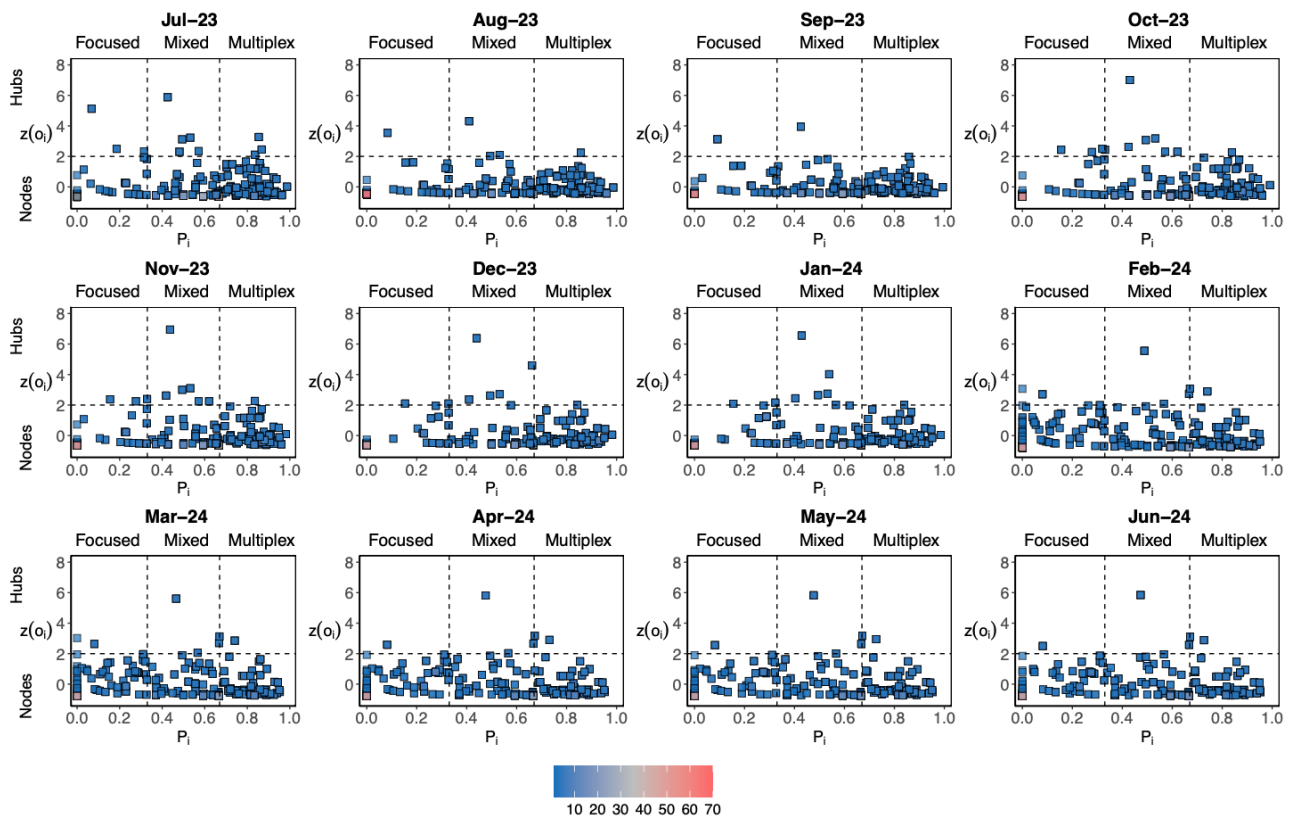


Figure 28: The participation indices and their z-score.

The analysis highlights distinct patterns in moderation activities across different content layers and overtime, revealing key insights into how moderation efforts are distributed and sustained. On average, the P_i is 0.513, indicating that act of moderation is moderately spread across layers. However, Misinformation and Harassment

layers show the most balanced participation i.e., $P_i = 0.632$ and 0.601 respectively, suggesting that these types of content are being handled by a broader range of moderating instances. In contrast, Spam has a lower P_i of 0.425 , indicating that it is managed by a more concentrated group of instances. This is further reflected in the large number of blocked instances in the Spam category (2,353 moderations), which significantly exceed the numbers for other layers.

A total of 495 unique moderating instances are involved in the network, with considerable overlap between layers. Instances in focused category, who concentrate on a single moderation type, are the most in numbers (287), while instances who fall in Mixed category, i.e., they are performing moderation for more than one and less than all reasons, are second in number with 170 and instances in Multiplex category are at third with 121. Importantly, Multiplex moderators demonstrate significant long-term engagement, with 74 remaining active for a full 12 months, compared to 41 Focused and 19 Mixed moderators. These patterns indicate that Multiplex moderators, though smaller in number, play a critical role in ensuring cross-layer moderation over time.

The result also shows that moderation activity varies by duration. Focused moderators consistently account for more than half of the activity, i.e., 53-57% across the year, while Mixed and Multiplex moderators maintain stable yet smaller contributions. This suggests that Focused moderators are essential for addressing specific types of content, while Multiplex moderators address the issues of multiple reasons.

These findings reveal that moderation efforts are distributed unevenly across different content layers, with Misinformation and harassment benefiting from more diverse and widespread moderator participation. Spam, on the other hand, is handled by a more concentrated group of specialists, as reflected in its lower participation diversity and higher number of blocked instances. While most moderators focus on a single content type, a smaller but crucial group of Multiplex moderators plays a key role in managing multiple layers and maintaining a long-term engagement. These patterns suggest that different types of content require distinct moderation strategies, with some benefiting from broader participation and others from more focused, specialized attention.

In addition to understanding the distribution of moderation efforts across different content layers, it is important to examine how instances transition between moderation categories over time. While many moderators maintain a consistent focus on a single category, others exhibit more dynamic behavior, shifting between Focused, Mixed, and Multiplex roles. Analyzing these transitions reveals patterns of engagement that further enlighten the overall structure of moderation within the network.

We observe these transitions in the following Figure 29 which shows how instances shift between categories over time.

The figure highlights the dynamic nature of moderation roles, showing that while some instances remain within a single category, others transition between them. Specifically, from July-August to May-June, several key observations can be made which are as follows:

- **Focused to Focused:** A significant number of instances remain in the Focused category throughout the period. Transition percentages range from 25.71% to 33.40%, suggesting that many instances prefer to concentrate on a single moderation type consistently. This stable participation highlights the importance of specialized moderating instances in managing specific content.
- **Focused to Mixed or Multiplex:** Transitions from Focused to Mixed or Multiplex categories are relatively rare, reflecting minimal cross-category shifts for these instances. For instance, less than 2% of Focused moderating instances transition to Mixed in any given period, and shifts to Multiplex are even rarer. This indicates that those who focus on one type of moderation activity tend to remain specialized.
- **Mixed to Focused or Multiplex:** Moderators in the Mixed category show more flexibility, with occasional transitions to either Focused or Multiplex. However, the majority remain Mixed, with steady participation ranging from 10.53% to 16.19%. This suggests that while instances in Mixed categories handle more than one type of content, they often maintain this broader focus over time.

- **Multiplex to Focused or Mixed:** Instances classified as Multiplex demonstrate a similar consistency, with most staying in this category. However, small transitions to Focused or Mixed categories occur abundantly (up to 1.82%), indicating that some Multiplex moderators may shift to more specialized roles over time.
- **New and Inactive Instances:** A notable number of instances become inactive or leave the moderation network over time, particularly after brief periods of engagement. For example, the NA category in the figure indicates inactivity or exit from the network, grows from 19.03% in July-August to 36.64% in May-June, reflecting an increasing dropout rate. This trend contrasts with the relatively stable engagement of long-term moderators, especially in Multiplex category.

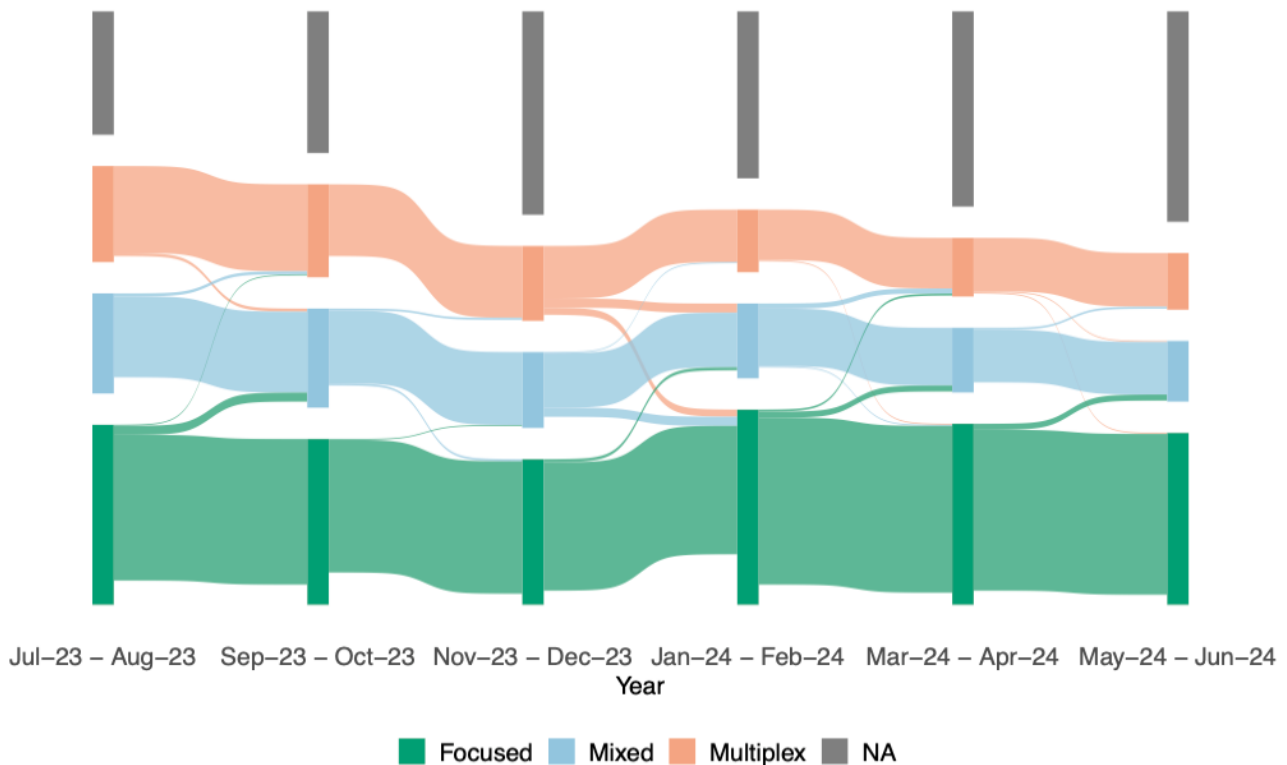


Figure 29: Instance transitions across different categories overtime.

Overall, this analysis of transitions highlights the stability of Focused and Multiplex moderators, while the Mixed category shows moderate flexibility. The transitions underscore that while some moderators shift between categories, the majority exhibit consistent behavior, particularly in specialized or broad moderation roles. These trends suggest that long-term engagement, particularly from instances in Multiplex category, plays a crucial role in maintaining moderation across multiple content layers, while Focused moderators provide essential coverage for specific issues.

4.4 Social bots

Social bots are automated accounts on social media platforms designed to imitate human behaviour to a believable degree and influence online content, discussions and overall information sources coming from end users. These bots are programmed to carry out tasks like spreading misinformation, amplifying specific views, or coordinating efforts to manipulate public opinion. They pose significant threats to the integrity of online conversations by artificially boosting the visibility and influence of certain narratives while suppressing others. Social botnets represent a darker aspect of social media, where automated content amplification distorts public discourse, undermines trust, and weakens democratic processes.

The problems with social bots go beyond their influence on public opinion. They can also be used for malicious purposes such as spreading propaganda, manipulating stock markets, or facilitating cyberattacks. Detecting

and addressing social botnets is a delicate task due to their growing sophistication, employing AI techniques to mimic human behaviour more realistically. Differentiating between real user interactions and bot activity requires advanced detection technologies.

Whilst predominantly abundant in traditional social media platforms such as Twitter/X or Instagram, due to their maturity in today's digital environment and the majoritarian userbase, bots in platforms located in the Fediverse are even more difficult to track.

While malicious bots are prevalent on centralized platforms, they are much harder to find on the Fediverse for several reasons. Firstly, instances that host harmful content, such as spam, NSFW material, or those promoting problematic ideologies like unrestricted free speech, are quickly banned, as suggested by Bono et al.²⁴. Additionally, the Fediverse is less polarized than centralized platforms, with 97% of instances being neutral servers, as stated by La Cava and collaborators²⁵. This decentralized nature creates smaller and more cohesive communities that are easier to moderate and more protective of their spaces, allowing them to address harmful behaviours more quickly.

Unlike traditional centralized networks, where suspicious activities and manipulation come often from economic and political interests, most Fediverse organizations are non-profit. This lack of commercial motivation implies that these platforms are less open to the manipulative tactics and fake engagement, commonly seen on larger networks. In fact, researching into the top ten companies (as ranked by Google), selling likes on social networks revealed that none offer services for Mastodon or Bluesky. This suggests that large-scale bot farms have not yet infiltrated the Fediverse.

Most of the Fediverse platforms are open-source and have transparency policies. Since their code is publicly available, developers and community members can identify potential vulnerabilities and implement security patches to mitigate the risk of bot infiltration or detect suspicious behaviour. Their transparency and communication of policies and actions against malicious behaviour provide a proactive environment for protecting their communities, as administrators and moderators are known and reachable.

Bots on the Fediverse, particularly on Mastodon, also differ significantly from those on centralized platforms. On Mastodon, bots are typically self-identified and perform non-malicious, community-focused tasks. For example, some bots provide news aggregation, delivering daily summaries, while others act as educational bots, sharing scientific facts, art, or historical events. There are also weather bots, community alert bots, and bots that replicate content from official centralized accounts, tasks that serve the community rather than exploit it.

To further understand bot behaviour, a study using a Twitter bot detector was conducted by Gutierrez and collaborators²⁶, who used the profile attributes of Twitter accounts, with an F1 score of 0.77, applying it to a Mastodon dataset of 2300 accounts (56% self-identified bots, 44% not self-identified). By mapping similar profile attributes from Mastodon and Twitter/X, a precision of 0.8, recall of 0.65, and an F1 score of 0.72 were achieved. After these results, we retrained the model with new Mastodon data of 16092 accounts, (50% self-identified bots, 50% not self-identified) and a validation set of 20% we achieve a precision of 0.96, recall of 0.93, and an F1 score of 0.94. These results suggest that while the detector works reasonably well on Mastodon, it only identifies whether an account is a bot, without determining if the bot is malicious.

Given the differences highlighted in the previous paragraphs, detecting malicious bots is clearly more complex on the Fediverse. Future work should focus on identifying bots that are not self-identified and analysing their behaviour and posts to determine whether they are malicious, as detecting bots alone does not reveal their

²⁴ Carlo Alberto Bono, Lucio La Cava, Luca Luceri, and Francesco Pierri. 2024. An Exploration of Decentralized Moderation on Mastodon. In Proceedings of the 16th ACM Web Science Conference (WEBSCI '24). Association for Computing Machinery, New York, NY, USA, 53–58. <https://doi.org/10.1145/3614419.3644016>

²⁵ Lucio La Cava, Domenico Mandaglio, and Andrea Tagarelli. 2024. Polarization in Decentralized Online Social Networks. In Proceedings of the 16th ACM Web Science Conference (WEBSCI '24). Association for Computing Machinery, New York, NY, USA, 48–52. <https://doi.org/10.1145/3614419.3644013>

²⁶ D. Martín-Gutiérrez, G. Hernández-Peñaloza, A. B. Hernández, A. Lozano-Diez and F. Álvarez, "A Deep Learning Approach for Robust Detection of Bots in Twitter Using Transformers," in *IEEE Access*, vol. 9, pp. 54591-54601, 2021, doi: 10.1109/ACCESS.2021.3068659.

intent on these social networks.

4.5 Graph-based methodologies

4.5.1 Clustering analysis

The main goal of this section is to contextualize the clustering methods in the environment of decentralized social media and explore its possibilities of application. The application of clustering algorithms can vary from classic techniques (without machine learning) to more specialized services, where Graph Neural Networks (GNNs) or deep learning approaches are applied to preprocess data or to analyse the structure of the social media in question.

Clustering is a technique that can be used to group users, content, or interactions based on their trustworthiness and credibility. This approach is essential in social networks, where large volumes of content flow quickly and the credibility of information is a critical point to maintaining an informed public. By applying clustering algorithms, platforms can analyse patterns in user behaviour, interaction history, and content engagement to identify trustworthy clusters, such as credible sources of news, while isolating suspicious activities linked to misinformation or harmful content. Thus, clustering can be regarded as another asset to identify trustability. It can help in identifying and reducing the spread of misinformation, worthy of testing in environments such as the Fediverse or next-generation social media, especially when concerning user analysis.

In centralized social networks, clustering not only helps in detecting misinformation but also helps flagging potential trust issues within communities. For instance, users with erratic engagement patterns or those partaking in coordinated efforts to manipulate discussions can be grouped together, indicating the need for closer monitoring. Furthermore, clustering methods such as k-means or DBSCAN can help finding influencers or trusted nodes within a network. For instance, the DeepTrust model combines clustering with machine learning to evaluate user trustworthiness by ²⁷[OBJ:OBJ]. This is especially relevant for combating disinformation, as it allows social platforms to identify both credible nodes (e.g., trusted news sources) and suspicious clusters linked to coordinated manipulation. Currently, there are not many relevant studies applying such techniques in a decentralized social network. Nevertheless, and given the advantages detected in centralized versions, it is worthy of noting that such techniques could help assess credibility especially in large communities (large servers/instances).

While centralized social networks like Twitter/X and Facebook have used clustering for trust assessment, the decentralized nature of the Fediverse presents additional challenges and opportunities. With smaller, restricted communities, clusters are easier to manage and moderate but detecting malicious behaviour can be harder due to the lack of uniform algorithms across instances. However, clustering might still be an effective tool in these networks²⁸. An example is the use of Social Network Analysis (SNA) and network science metrics to detect misinformation by analysing user behaviour and content dissemination patterns. A software architecture including modules for community detection, centrality calculation, and misinformation analysis is proposed, demonstrating how clustering can reveal coordinated misinformation efforts²⁹.

Focusing on the decentralized architecture of the next-generation social media, their architecture must be taken into account as the main factor that influences the clustering analysis. The architecture may vary in between a centralized version with a fully connected network where all the users interact (much alike traditional social media), compared with the setup for decentralized versions where the users are spread in different servers (or instances). There is even a possibility for users to be in different servers at the same time and the users can also interact with users from other servers, as seen before. Technically speaking, the input data should consist in a set of users which are interconnected in a certain way, which can vary in between topics, groups, or specific contents/users (e.g. a fan group of an artist in Facebook). With the connection of different users and contents,

²⁷ Alrubaian, M., Al-Qurishi, M., Omar, S., & Mostafa, M. A. (2021). Deeptrust: a deep learning approach for measuring social media users trustworthiness. *arXiv preprint arXiv:2101.07725*.

²⁸ Tao Huang, Decentralized social networks and the future of free speech online, *Computer Law & Security Review*, Volume 55, 2024, 106059, ISSN 0267-3649, <https://doi.org/10.1016/j.clsr.2024.106059>.

²⁹ Duzen, Z., Riveni, M., & Aktas, M. S. (2023). Analyzing the Spread of Misinformation on Social Networks: A Process and Software Architecture for Detection and Analysis. *Computers*, 12(11), 232.

a "sub-network" or graph can be built including the information of the relations in between the users, and this is used by clustering algorithms to find relevant groups. In the context of decentralized versions, the situation offers a new set of possibilities as users are already spread into communities, so the focus of the analysis can be a specific server, or a larger group of different servers which interact in between each other. The definition of the graph and their relations is based on the goal the model wants to achieve and usually includes features about the user and their interactions. Additionally, it can also use features about the posts and their spreading patterns.

In this direction, some classic techniques (non-AI) could offer promising results, e.g. average clustering, a method based on the mean of the clustering coefficients of each node in the graph, understanding the clustering coefficient as a fraction of the neighbouring node links and the actual node links. This technique is also interesting in combination with other graph analysis techniques such as "Pagerank"³⁰ and "Betweenness Centrality"³¹ to provide information on how the users build communities and behave, alongside measuring the amount of influence in their communities

With regards to the AI-based methods, GNNs are the most relevant models that can be used to perform clustering. The key difference in using GNNs to perform the clustering analysis is the usage of embeddings for the nodes in the graph, where information about the user, interactions or posts is encoded. More information will be provided in the following section.

With all this into consideration, the potential of graph clustering analysis can be useful to further understand the structure of social networks and to identify key communities or users and their behaviour, in addition to possible disinformation threats or malignant users. Experiments are planned combining the techniques described in this same section in order to shed light on trustability in decentralized social networks.

4.5.2 Propagation graphs

Content propagation analysis in social media is a vital area of study that focuses on understanding how information, ideas, or content spread across networks of users. Evidently, social media are a prime example of a rapid and widespread propagation of content. In relation with the previous point, low-credibility content propagation is crucial to be analysed to ensure a trustable social experience. It is especially important to track the virality of low credibility content by suspicious sources and social botnets.

In this context, propagation of content can be modelled with graph theory. A piece of information (e.g. a post, image, or video) moves along the edges of the graph, reaching various nodes. The structure of the graph, alongside the weight properties of individual nodes, influences how far and fast the content spreads. For example, influential users (with high connectivity or many followers) could be regarded as "hubs" through which pieces of content can go viral.

GNNs are a class of neural network models specifically designed to handle data in structures compatible with the description given in the previous paragraph. Unlike traditional neural networks, that deal with flat, grid-like data (such as images or sequences), GNNs reach to capture the relationships and dependencies in graph-structured data. They can learn to aggregate information from neighbouring nodes and edges, making them suitable for social media analysis.

In this project a GNN model known as GraphSAGE³², trained in Twitter, was adapted with the aim to detect the propagation of low credibility content such as fake news in BlueSky and Mastodon. Given the similitude of characteristics between Twitter and Mastodon/BlueSky, the aim is to extrapolate already known graph structures to understand better the nature of virality in Fediverse platforms. This way, the problem was framed as a binary classification task, with the model trained using binary cross-entropy as the loss function to optimize

³⁰ Agryzkov, Taras, Francisco Pedroche, Leandro Tortosa, and José F. Vicent. 2018. "Combining the Two-Layers PageRank Approach with the APA Centrality in Networks with Data" ISPRS International Journal of Geo-Information 7, no. 12: 480. <https://doi.org/10.3390/ijgi7120480>

³¹ Zhang, Jinhua, Qishan Zhang, Ling Wu, and Jinxin Zhang. 2022. "Identifying Influential Nodes in Complex Networks Based on Multiple Local Attributes and Information Entropy" Entropy 24, no. 2: 293. <https://doi.org/10.3390/e24020293>

³² William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 1025–1035.

its predictions.

The GNN GraphSAGE model was trained on the User Preference-aware Fake News Detection (UPFD) dataset³³, which contains propagation graphs derived from Twitter. Each graph represents a post sharing a piece of news (classified as either fake or true). To adapt the model for BlueSky and Mastodon, the training graphs were replicated with some modifications. Specifically, since reposting timestamps—used to construct the original graphs—are unavailable on these platforms, certain assumptions were made to compensate for this limitation.

Propagation graphs were created with the following criterion. The root node corresponds to the user that has shared the content, and all the users that have shared/reposted the original node are the leave nodes. If one of these users follows the original poster, they are linked together. Similarly, if a user does not follow the original poster but follows one of the users that are already link to the original poster, they are linked together. If none of these conditions are satisfied, the user is connected to the user with the highest number of followers. An example of a graph created using this method can be seen in Figure 30.

For each user in the graph several node features were extracted. For the root node, the original post was served as input to BERT³⁴ and a feature vector was obtained. For each other user, this vector was created using the last 200 posts as input to BERT. Both the creation of the graph and the extraction of node features try to replicate the pipeline followed by³⁵, which was used to train the GNN on Twitter data.

This method aims to significantly reduce the spread of misinformation on decentralized social networks like Mastodon and BlueSky by providing an automated and reliable way to identify and flag low credibility or false content. By improving the accuracy and speed of misinformation detection, it contributes to create a more trustworthy information environment, enabling users to make more informed decisions. Future steps involve obtaining a labelled dataset of posts from Mastodon and BlueSky and testing the pipeline in it. The workflow will then be adjusted to maximize the performance of the models. Should there be availability of enough samples, a specific model that specialises in Mastodon and BlueSky will be trained.

³³ Yingdong Dou, Kai Shu, Congying Xia, Philip S. Yu, and Lichao Sun. 2021. User Preference-aware Fake News Detection. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 2051–2055. <https://doi.org/10.1145/3404835.3462990>

³⁴ Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1, 4171–4186. <https://arxiv.org/abs/1810.04805v2>

³⁵ Dou, Y., Shu, K., Xia, C., Yu, P. S., & Sun, L. (2021). User Preference-aware Fake News Detection. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.

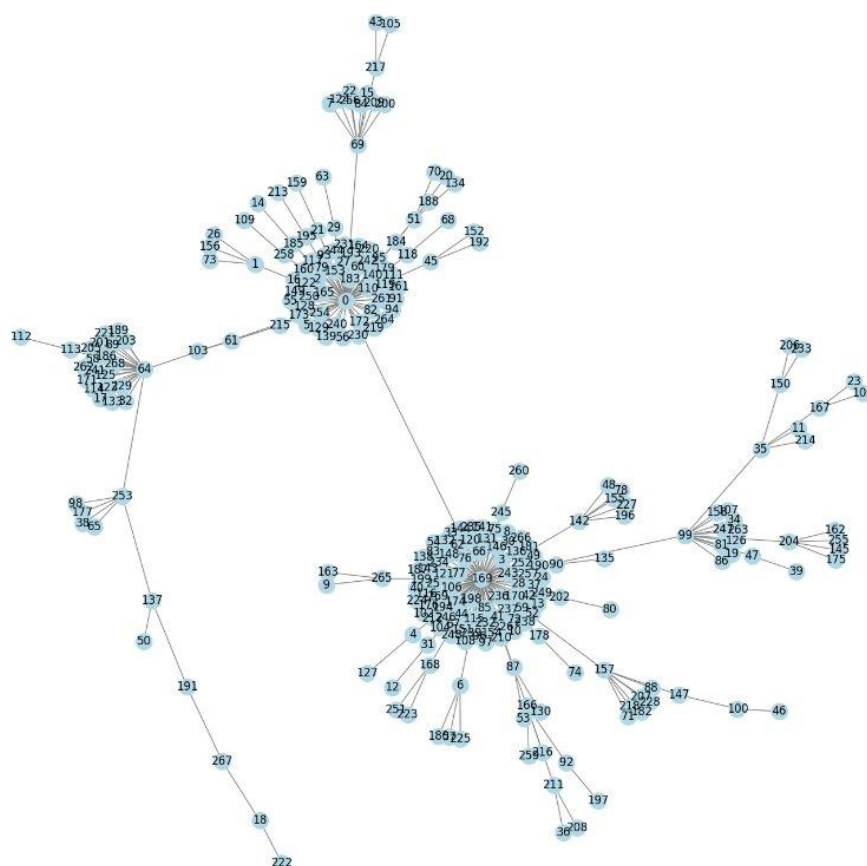


Figure 30: Example of how a propagation graph looks like. Node 0 corresponds to the original post. All other nodes represent nodes that have reposted the original post.

4.6 Hate speech analysis on Mastodon

4.6.1 Data collection from Mastodon

For Mastodon posts, called “toots”, the Python package “Mastodon.py” was used. Mastodon.py is a Python wrapper for the official Mastodon API and enables the user to perform actions programmatically. One such action is to search a Mastodon channel for specific keywords. In our case, the keywords that performed the search are “climate change”, “European union”, “elections”, “immigration”, “Ukraine”, “Israel”, “LGBTQ+”. We focused on these keywords as according to fact-checks available in Truly Media, these were the most frequent disinformation topics for the first five months of 2024. We collected the 2,000 most recent posts for each keyword, resulting in 13,847 total number of collected toots.

Mastodon offers many instances, each one dedicated to specific topics of interest like politics, music, sports, etc. Many instances are private and cannot be accessed unless the user is a member. For our case, the search was conducted on the instance “mastodon.social” which is an instance of general interest, thus the chance of finding a toot that contains one of the keywords mentioned above is higher and is open for all users. The collection of data was performed using the pagination technique, which is used in API design and development to retrieve large data sets in a structured and manageable manner, in smaller chunks.

The method that performs a search for Toots that contain a keyword, using the pagination method is:

```
mastodon.search(keyword,
    resolve=True,
    result_type="statuses",
    account_id=None,
    offset=Pagination_number,
    min_id=None,
    max_id=None,
    exclude_unreviewed=False)
```

The result of this method is a dictionary of the following form:

```
{
  'accounts': # List of user dicts resulting from the query
  'hashtags': # List of hashtag dicts resulting from the query
  'statuses': # List of toot dicts resulting from the query
}
```

Because we have instructed the API to only include statuses, the lists of the “accounts” and “hashtags” keys were zero, and only the “statuses” list was populated. A “status” has the following metadata:

```
{
  'id': # Numerical id of this toot
  'uri': # Descriptor for the toot
    # EG 'tag:mastodon.social,2016-11-25:objectId=<id>;objectType=Status'
  'url': # URL of the toot
  'account': # User dict for the account which posted the status
  'in_reply_to_id': # Numerical id of the toot this toot is in response to
  'in_reply_to_account_id': # Numerical id of the account this toot is in response to
  'reblog': # Denotes whether the toot is a reblog. If so, set to the original toot dict.
  'content': # Content of the toot, as HTML: '<p>Hello from Python</p>'
  'created_at': # Creation time
  'reblogs_count': # Number of reblogs
  'favourites_count': # Number of favourites
  'reblogged': # Denotes whether the logged in user has boosted this toot
  'favourited': # Denotes whether the logged in user has favourited this toot
  'sensitive': # Denotes whether media attachments to the toot are marked sensitive
  'spoiler_text': # Warning text that should be displayed before the toot content
  'visibility': # Toot visibility ('public', 'unlisted', 'private', or 'direct')
  'mentions': # A list of users dicts mentioned in the toot, as Mention dicts
  'media_attachments': # A list of media dicts of attached files
  'emojis': # A list of custom emojis used in the toot, as Emoji dicts
  'tags': # A list of hashtag used in the toot, as Hashtag dicts
  'bookmarked': # True if the status is bookmarked by the logged in user, False if not.
  'application': # Application dict for the client used to post the toot (Does not federate
    # and is therefore always None for remote toots, can also be None for
    # local toots for some legacy applications).
  'language': # The language of the toot, if specified by the server,
    # as ISO 639-1 (two-letter) language code.
  'muted': # Boolean denoting whether the user has muted this status by
    # way of conversation muting
  'pinned': # Boolean denoting whether or not the status is currently pinned for the
    # associated account.
  'replies_count': # The number of replies to this status.
  'card': # A preview card for links from the status, if present at time of delivery,
    # as card dict.
  'poll': # A poll dict if a poll is attached to this status.
}
```

4.6.2 Hate Speech Detection Tool

For completeness of this deliverable, in this section we briefly describe the hate speech detection tool which was developed by ATC as part of the ATLANTIS project (Horizon-IA, GA 101073909).

The hate speech detection tool is a classifier that takes as input a text and classifies it into the five classes presented next, along with a prediction score. The five classes are the following:

- Racism, i.e. related to the race of an individual or a group
- Sexism, related to expressions of misogyny/misandry
- Sexual orientation, connected to sexual orientation and gender identity
- Religious chauvinism, referring to the religion of an individual or a group
- None

For its development the lightweight Transformer model DistilBERT³⁶ has been fine-tuned with the publicly available datasets CONAN³⁷ (COunter NArratives through Nichesourcing) and the HKUST-MLMA³⁸ (Multilingual and Multi-Aspect Hate Speech Analysis). The CONAN is a multilingual and expert-based dataset of hate speech/counter-narrative pairs for English, French and Italian. It contains various types of metadata like expert demographics, counter-narrative types and hate speech topics. The HKUST-MLMA dataset is composed of a pilot corpus of 100 tweets per language, and comparable corpora of 5,647 English tweets, 4,014 French tweets, and 3,353 Arabic tweets. It contains information about the hostility type, the hostility directness (direct vs indirect), the target attribute (i.e. sexual orientation, disability), the target group (i.e. gay, immigrants, woman) and the annotator's sentiment (i.e. shock_disgust). It should be noted that only the English subset from both datasets has been used in the analysis.

To build the dataset, we use the aforementioned publicly available datasets annotated with labels of Hate Speech and consisting of texts predominantly from Twitter. Additionally, we obtain data with crawling techniques, e.g. via the Twitter python API 1, using lists of relevant post IDs. We subsequently merge partial results from all the different sources via a conversion and curation procedure that applied data cleaning, label mapping and relevant information extraction. After this process, we arrive at the 5 labels of interest and a dataset of 15680 instances.

The accuracy of the Hate speech detection tool is 87%.

4.6.3 Results of hate speech detection in Mastodon

The distribution of toots per type of hate speech is presented in the following figure: almost 3% of the collected toots hindered some sort of hate speech (with prediction score more than 90%), i.e., 387 toots out of the 13,847 collected ones, with Racism (1.86%, 257 toots) being the most often type of hate speech content (Figure 31).

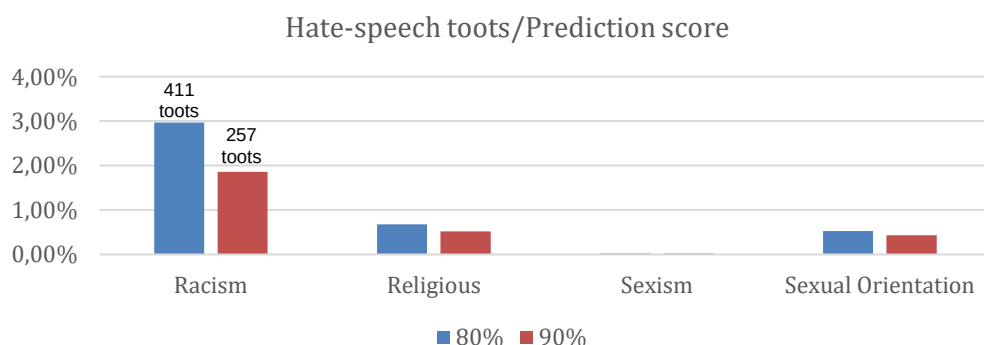


Figure 31: Distribution of toots per hate speech type in Mastodon (for prediction scores 80% and 90%).

³⁶ https://huggingface.co/docs/transformers/model_doc/distilbert

³⁷ <https://github.com/marcoguerrini/CONAN#Multi-hate-target-knowledge-grounded-hate-countering-dataset>

³⁸ https://github.com/HKUST-KnowComp/MLMA_hate_speech/blob/master/hate_speech_mlma.zip

4.7 Bluesky superspreaders

4.7.1 Superspreaders

Superspreaders are defined as highly active, visible, and productive users who spread information on social media. They often have large following and produce significant content but may not always be accurate or trustworthy.

In the initial phase of this project, we concentrated on identifying superspreaders of hate speech. These are accounts that a) produce hate speech and b) manage to distribute it to many followers who in turn share the content with their followers. We expect these superspreaders to be instrumental in pushing disinformation narratives which makes this analysis particularly relevant. At the time of writing this report, a) was addressed using the interactive visualization described below. The analysis corresponding to b) is intended to be an extension of the dashboard, to be developed as part of Task 2.4.

4.7.2 Detecting hate speech using an Interactive Data Exploration

4.7.2.1 The Dashboard

The 260.774 Bluesky posts were run through ATC's hate speech detector, further described in section 4.6.2. The latter scores each post on a scale from 0 to 1 with respect to four potential hate speech categories: *racism*, *religious*, *sexism* and *sexual orientation*. In addition, the detector also scores to what degree a given post does not contain any hate speech. This is the fifth *non-hateful* category.

We developed an interactive dashboard with the goal of identifying those accounts that post a considerable amount of hate speech. A screenshot of this dashboard is shown below in Figure 32. The aim of this dashboard is two-fold: first, it allows to identify the top hateful accounts in an interactive fashion; second, it serves as playground to develop suitable visualizations that can be used later in some of AI-CODE's services.

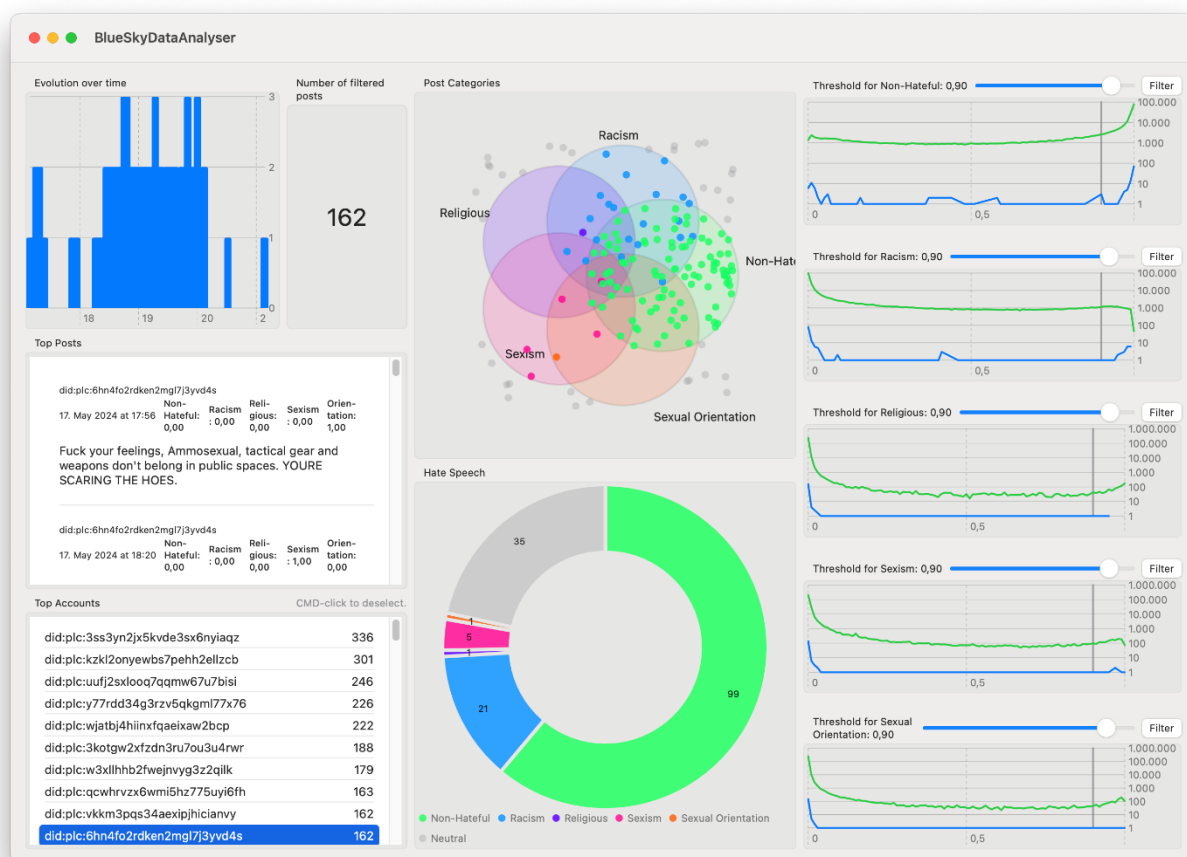


Figure 32: Interactive dashboard generated for this task.

In this example, account `did:plc:vkkm3pqs34aexipjhicianvy` is selected at the bottom-left account table. Right clicking the row opens up a browser which reveals the account's name and public handler.

The Hate Speech pie chart reveals that while 99 (or more than half) of this account's captured posts are categorized as non-hateful, there are still 21 hateful posts, as well as 5 sexists posts.

The five-dimensional Venn-diagram (top middle) provides another perspective on this account's post Categories: each post corresponds to a dot whose color and position identifies it to belong to either of the five categories – or overlaps thereof. Note that gray dots represent posts that could not be mapped into any of the five hate speech categories.

Other graphs are the evolution of the number of filtered posts over time (top left), the number of selected posts (next to it), as well as a list of the top posts created by the selected account, listed in the order of the degree of hateful content. The latter is calculated based on the geometric average of the post's scores.

Finally, the histograms on the right show the distribution of posts matching the respective category – calculated based on *all* collected posts (green), as well as the *filtered* posts (blue) associated with the user. The slider determines the score threshold above which a given post is mapped into the respective category. In this example, a post with a racism score of at least 0.9 is considered to be racist.

4.7.2.2 Top hate speech accounts

The following information has been extracted with a threshold of 0.9 for each hat speech category. As shown in Figure 33, all in all, a total of

- 151.723 (58%) posts were categorized as non-hateful
- 95.365 (37%) posts could not be categorized, since none of the thresholds were exceeded
- 13.689 (5%) posts were categorized as hateful

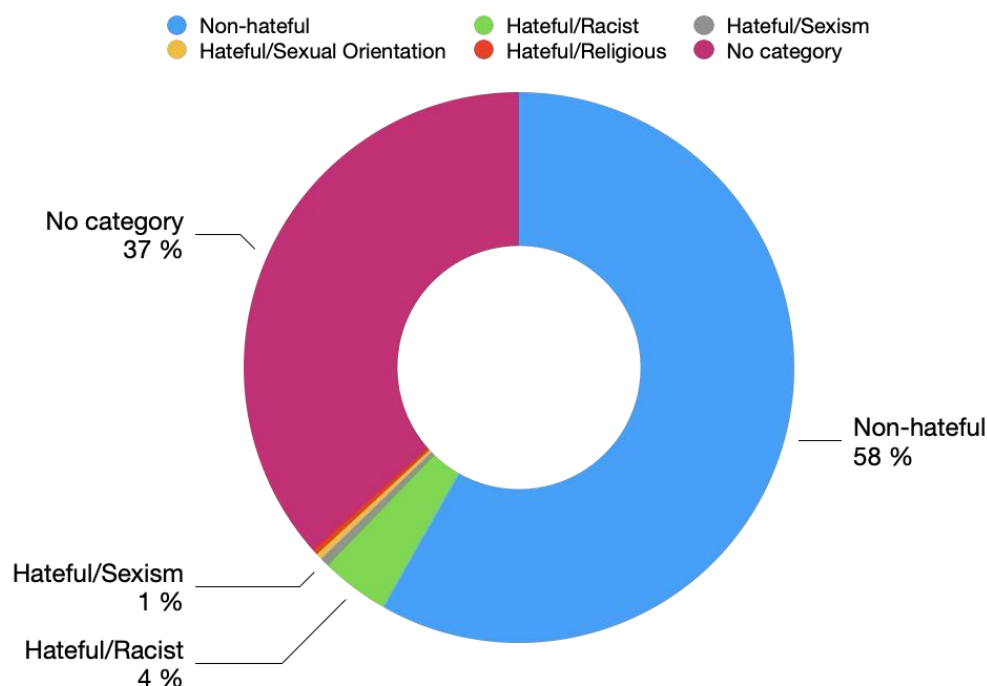


Figure 33: Pie chart of the results of the performed hate speech analysis.

A total of 10.529 posts (4%) were categorized as **racist**. The top accounts with the highest number of racist posts were:

#	Name	User ID	Posts
1	BBC Asian Network Now Playing Bot	@bbcasiannetworkbot.bsky.social	27
2	şifosu ☺☆☺	@eilliebilish.bsky.social	27
3	nashoba holba	@urlocalgooddog.bsky.social	21

Note that it is unlikely that the BBC Asian Network Bot really posts hateful content, this is an example where the hate speech categorisation model yielded false positives.

A total of 780 posts (0.3%) were categorized as **religious**. The top accounts with the highest number of religious posts were:

#	Name	User ID	Posts
1	BBC Radio 1Xtra Now Playing Bot	@bbcradio1xtrabot.bsky.social	7
2	Daniel Goldman	@bbcradio1xtrabot.bsky.social	6
3	Dr. Exmaahal	@exmaahal.bsky.social	5

A total of 1437 posts (0.6%) were categorized as **sexism**. The top accounts with the highest number of sexist posts were:

#	Name	User ID	Posts
1	Tilly	@tillyx.bsky.social	7
2	Gapozathens	@gapozathens.bsky.social	6
5	nashoba holba	@urlocalgooddog.bsky.social	5

Note that account #3 was deleted in the meantime, whereas #4 does not contain any posts anymore.

A total of 943 posts (0.03%) were categorized as being explicit about **sexual orientation**. The top accounts with the highest number of posts were:

#	Name	User ID	Posts
1	Chaturbate Gay - Free chat & Gay Sex on webcam	@chaturbate-gay.bsky.social	22
2	ismua Bluesky Elder 🏳️🌈🚢💜	@ismua.bsky.social	11
3	lena wife	@nebulena.net	5

5 Privacy and Data Governance in Decentralized Networks

Federated social media platforms are motivated by a number of grievances with the commercial social media platforms. Among them are issues related to privacy and data governance that empowers the platform users. Privacy is enshrined as a fundamental right in Art. 7 of the EU Charter on Fundamental Rights³⁹. The same Charter also contains a right on personal data protection (Art. 8), which can be understood as specifying how to understand privacy in the context of online data⁴⁰. This right emphasises: strengthened rights for the data subject, imposes obligations on those who process personal data, and establishes monitoring of compliance with the respect of rights. This is further developed in the European Data Protection Regulation (GDPR)⁴¹. If ‘data privacy’ is mentioned in this section, it is meant here in the sense of Art. 8 of the EU Charter, as personal data protection that empowers the individual data subject.

³⁹ European Union. 2010. Charter of Fundamental Rights of the European Union. Official Journal of the European Union C83. Vol. 53. Brussels: European Union.

⁴⁰ Gloria Gonzalez Fuster and Hielke Hijmans, ‘The EU Rights to Privacy and Personal Data Protection: 20 Years in 10 Questions’, Vrije Universiteit Brussel, 2019, 1–2, https://cris.vub.be/ws/portalfiles/portal/45839230/20190513.Working_Paper_Gonza_lez_Fuster_Hijmans_3_.pdf.

⁴¹ Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data

With data governance, this section refers to those mechanisms that allow the platform users to exercise the type of control that empowers users, i.e. determining, access, use, storage, and sharing of one's data⁴².

5.1 State of play in centralised networks

The argument for decentralised platforms, such as Mastodon, relies on the shortcomings of centralised platforms in a number of areas, among them privacy and data governance that favours platform profits over user empowerment. The business model of centralised social media platforms – to sell targeted advertising space to third parties – incentivises them to collect as much (behavioural) data on their users as possible to improve the advertising services they offer to third parties⁴³. Their business model is built on collecting and analysing – mining – data and profiling users so that they can be targeted with relevant advertisement in decisive moments⁴⁴. The user does not control their data in this model and rarely understands the extent of the collection and mining processes going on in the background.

GDPR, which entered into force in 2016, was aimed at addressing some of these issues. However, as researchers suggest, the information asymmetry⁴⁵ related to the data-driven business models of social media platforms and search engines, was not sufficiently addressed by the regulation.

GDPR contains provisions on the legal purposes for the collection of specific categories of data and establishes legal grounds for subjects to demand transparency from the data controller (art. 15). The report by the 'Multistakeholder Expert Group on the GDPR evaluation' from June 2024⁴⁶, overall emphasises a positive impact of the GDPR on data privacy practices and consumer protection. However, they also take note of some shortcomings, pointed out by the literature more broadly.

For example, Van de Wardt⁴⁷ argues that effectiveness of these provisions and rights is limited in practice, because "much more data is being amassed than the consumer can reasonably oversee, by an exorbitant number of interconnected parties, through a variety of means which are far from self-evident."

One solution that is offered by the GDPR is the idea of "informed consent" as a legal basis for data processing. Thus, users of social media platforms are now presented with the privacy policy of the platform, that they can accept – or not use the platform. However, on the one hand this does not always seem to meet the conditions for 'informed consent'⁴⁸, and research suggests that few internet users make the significant time investment to read privacy policies in full⁴⁹, or they are not always findable⁵⁰. Further, it remains uncertain in how far data collectors, such as Google, adhere to these privacy policies⁵¹.

Another issue that remains unaddressed on centralised platforms is a limited use of the right to data portability of data subjects (art 20). The GDPR Expert Group suggests that consumers are not aware of this right, and that controllers do not support this right technically⁵².

⁴² Mark Ryan, Paula Gürtler, and Artur Bogucki, 'Will the Real Data Sovereign Please Stand up? An EU Policy Response to Sovereignty in Data Spaces', International Journal of Law and Information Technology 32, no. 1 (1 June 2024): 15, <https://doi.org/10.1093/ijlit/eaee006>.

⁴³ ZHILIN Zhang et al., 'Trouble in Paradise? Understanding Mastodon Admin's Motivations, Experiences, and Challenges Running Decentralised Social Media', Association for Computing Machinery, 2024; Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (London: Profile books, 2019).

⁴⁴ Peter J. Van De Waerdt, 'Information Asymmetries: Recognizing the Limits of the GDPR on the Data-Driven Market', Computer Law & Security Review 38 (September 2020): 105436, <https://doi.org/10.1016/j.clsr.2020.105436>; Zuboff, The Age of Surveillance Capitalism.

⁴⁵ Van De Waerdt, 'Information Asymmetries' argues that there is a twofold information asymmetry: i) the information asymmetry when the companies collect data on the unassuming data subject, and ii) the extent and purposes to which the data is being analysed by the data controller (p. 3).

⁴⁶ Multistakeholder Expert Group to support the application of Regulation (EU) 2016/679, 'Contributions from the Multistakeholder Expert Group to the Commission 2024 Evaluation of the General Data Protection Regulation (GDPR)', Report (European Commission, 10 June 2024), <https://ec.europa.eu/transparency/expert-groups-register/screen/meetings/consult?lang=en&meetingId=54422&fromExpertGroups=3537>.

⁴⁷ 'Information Asymmetries', 5.

⁴⁸ Nathalie A. Smuha, 'Beyond the Individual: Governing AI's Societal Harm', Internet Policy Review 10, no. 3 (30 September 2021), <https://doi.org/10.14763/2021.3.1574>.

⁴⁹ Aleecia M McDonald and Lorrie Faith Cranor, 'The Cost of Reading Privacy Policies', Isjlp 4 (2008): 543; Van De Waerdt, 'Information Asymmetries'.

⁵⁰ Mukund Srinath et al., 'Privacy Lost and Found: An Investigation at Scale of Web Privacy Policy Availability', in Proceedings of the ACM Symposium on Document Engineering 2023 (DocEng '23: ACM Symposium on Document Engineering 2023, Limerick Ireland: ACM, 2023), 1–10, <https://doi.org/10.1145/3573128.3604902>.

⁵¹ Van De Waerdt, 'Information Asymmetries', 13.

⁵² Multistakeholder Expert Group to support the application of Regulation (EU) 2016/679, 'GDPR Expert Group Report', 8.

This short discussion on the limitations of current privacy and data governance practices of centralised social media platforms, despite GDPR, illustrate that there is room for improvement. This is where Fediverse platforms aim to offer a socio-technical solution.

5.2 The ideal of decentralised networks

Recent research finds that Fediverse community administrators are motivated among other things by “enhancing users’ privacy”⁵³. In the conceptualisation of Zhang et al.⁵⁴, the main obstacles in centralised platforms are the data-driven business model mentioned above, as well as a lack of interoperability, which creates platform lock-in effects for users. Fediverse platforms, such as Mastodon, have several features that are supposed to overcome these challenges.

This section focusses on describing some of these features of Fediverse platforms in their ideal form and how they enhance privacy and optimise data governance for users. The final section analyses how well they deliver in practice and which privacy needs persist for Fediverse platforms, identifying four main points:

1. **There is no immediate profit-incentive for platform providers for data mining**⁵⁵ : Mastodon is a not-for-profit company. Instead, it is open source and community supported. The user’s feeds are not algorithmically mediated, which already reduces the privacy-infringement through algorithmic processing, and there are instances that do not allow for any advertising at all. This overall infrastructure results in a lack of incentives for data mining and the trading of insights to third parties.
2. **Data is distributed across different nodes**⁵⁶ : Whereas all user data is hosted by one company in the case of commercial social media platforms, in the case of Mastodon, it is hosted by different instance providers on a number of servers. This means, “privacy-intrusive data mining” is possibly made more difficult⁵⁷. If you don’t trust one party with your data, you can simply opt for a server run by another. In theory, you could even just open your own instance.
3. **Instances can determine their own data privacy policies**: As an open-source platform, Mastodon enables instance administrators to tailor privacy policies so that they are in alignment with the principles and values of the specific community⁵⁸.
4. **Users have meaningful data portability**⁵⁹: In theory, users are free to switch between instances without losing their following and other data. Thus, providing substantive data portability and significantly more control over one’s data.

5.3 Realistic view on persisting privacy needs on decentralised networks

Unfortunately, however, the reality of privacy practices and data portability in decentralized networks has not quite achieved the ideal outlined above, yet. This section focusses on two privacy needs that persist, and that are well documented in the literature.

1. **Instance availability might vary, limiting access and control over one’s data**: Raman et al.⁶⁰ conducted a broad investigation into the availability of instances, as many instances regularly experiences outages/failure. Instance going offline means that the user content (such as posts on

⁵³ Zhang et al., ‘Trouble in Paradise? Understanding Mastodon Admin’s Motivations, Experiences, and Challenges Running Decentralised Social Media’, 523.

⁵⁴ *ibid.*

⁵⁵ *ibid.*

⁵⁶ Aravindh Raman et al., ‘Challenges in the Decentralised Web: The Mastodon Case’, in Proceedings of the Internet Measurement Conference (IMC ’19: ACM Internet Measurement Conference, Amsterdam Netherlands: ACM, 2019), 217–29, <https://doi.org/10.1145/3355369.3355572>; Zhang et al., ‘Trouble in Paradise? Understanding Mastodon Admin’s Motivations, Experiences, and Challenges Running Decentralised Social Media’.

⁵⁷ *ibid.*

⁵⁸ Sohyeon Hwang, Priyanka Nanayakkara, and Yan Shvartzshnaider, ‘Whose Policy? Privacy Challenges of Decentralized Platforms’, SSRN Electronic Journal, 2023, <https://doi.org/10.2139/ssrn.4416746>; Zhang et al., ‘Trouble in Paradise? Understanding Mastodon Admin’s Motivations, Experiences, and Challenges Running Decentralised Social Media’.

⁵⁹ Zhang et al., ‘Trouble in Paradise? Understanding Mastodon Admin’s Motivations, Experiences, and Challenges Running Decentralised Social Media’, 520:9.

⁶⁰ Raman et al., ‘Challenges in the Decentralised Web’.

Mastodon) and user accounts hosted by that server become unavailable for that period of time. That implies, of course, that users with an account on the affected instance has no control over their account during the outage. Raman et al. also emphasise that open questions remain on how users can be protected from data loss during instance outages/failures.⁶¹

However, the outage of one instance does not only affect the users who are registered there, but it has ripple effects: “e.g., if three users on different instances follow each other, $U1 \rightarrow U2 \rightarrow U3$, then the failure of the instance hosting $U2$ would also disconnect $U1$ and $U3$ (assuming that no other paths exist).”⁶² A similar argument also applies if this was e.g. a cyber-attack. One way to enhance system resilience, which Raman et al. propose, is to replicate posts on instances where people are subscribed or even on random instances, so that they remain available even if the original instance goes offline. It should be noted that the Mastodon technical infrastructure could support replication on subscribed instances⁶³.

On the one hand, this type of replication would enhance users control over the content they interact with, as they can revisit it any time. On the other hand, this implies a serious trade off in data privacy for the original poster. The original poster might have chosen their hosting server for its particularly stringent privacy policy. The instance where it is replicated to enhance system resilience might have a different privacy policy and/or lower cybersecurity standard. Raman et al. find that there is a lot of uncertainty on how well your privacy and ability for controlling access to your data is protected in this system of content replication⁶⁴.

2. **Instances do not actually customize privacy policies:** Another challenge that persists with regards to privacy in the Fediverse is, that while privacy policies might be customisable, it remains uncertain in how far server owners and users pay increased attention and care to these policies. For example, Hwang et al.⁶⁵ present initial results that show that very few servers actually customise the privacy policies they offer. In total, “81% of servers used the exact same text, with an additional 3% using a minor variation; another 12% of servers used highly similar variations of another text.” These numbers do not yet reveal anything about the actual privacy practices of these different communities, or expectations different instance users might have.

Hwang et al.’s research further reveals that there is considerable ambiguity in the privacy policies that are adapted, including how copies of posts are stored on other servers and who has access to the posts on those other servers⁶⁶. This finding was also supported by the interviews carried out by Zhang et al.⁶⁷.

In conclusion, one has to recognise that a need for addressing privacy and data governance for user empowerment remains topics that deserve more attention in future research and technology development.

6 Conclusions

In this deliverable, we have summarized the research performed in task *T2.1 Decentralised and immersive social media*.

As demonstrated, the current social media landscape is evolving along two distinct trajectories: *federated platforms* and *immersive virtual environments*. While immersive virtual worlds show promise for future social interaction, their current gaming-centric nature and predominantly young user base suggest it is premature to

⁶¹ *ibid.*

⁶² *ibid.*

⁶³ *ibid.*

⁶⁴ *ibid.*

⁶⁵ Hwang, Nanayakkara, and Shvartzshnaider, ‘Whose Policy?’, 1.

⁶⁶ *ibid.*

⁶⁷ Zhang et al., ‘Trouble in Paradise? Understanding Mastodon Admin’s Motivations, Experiences, and Challenges Running Decentralised Social Media’. “I know that there’s no one central repository of data about a user within the fediverse [...] Content is disseminated via the federated connections that your server has and then cached or downloaded and stored on many different servers [...] You don’t know where your information ends up, and you can’t get rid of your data because it’s everywhere. (P10, General)”(520:12)

deploy comprehensive disinformation countermeasures in these spaces.

In contrast, federated platforms present both unique opportunities and challenges for addressing disinformation. The decentralized nature of platforms like Mastodon and Bluesky creates smaller, more cohesive communities that facilitate faster response to harmful behaviours. The non-profit nature of most Fediverse organizations significantly reduces the economic incentives that often drive manipulation on traditional platforms. Indeed, our research indicates an absence of commercial bot services targeting these platforms, suggesting they have so far avoided large-scale automated manipulation.

The federated structure supports effective moderation through instance-level decisions about bans and labels, with administrators being typically known and reachable within their communities. The open-source nature of these platforms enables community oversight of code and security measures. However, this same decentralization can hinder coordinated responses to disinformation, as detection algorithms and policies vary across instances. While one might leverage the homophilic nature of connected instances to predict similar behaviour patterns, technical challenges persist, including frequent server outages and limited customization of privacy policies.

Notably, the bot ecosystem in federated platforms differs fundamentally from centralized social media. Mastodon bots typically self-identify and serve community-focused purposes, such as news aggregation or educational content sharing, rather than engaging in manipulative behaviour. This transparency aligns with the broader governance model of these platforms, where community-driven moderation and open communication of policies create a more protective environment.

Our remaining work in WP2 will be carried out as a part of (already running) task *T2.4 AI-based credibility and trust assessment and disinformation detection tools*. The goal of this task is to critically evaluate and coalesce the findings and outcomes from T2.1 presented in this deliverable with the findings from tasks *T2.2 Generative AI methods for text generation* and *T2.3 Trustworthy AI methods for synthetic media detection* presented in the deliverable D2.2 together with the current capabilities of the state-of-the-art solutions on both, credibility and trust assessment as well as disinformation detection.

More specifically, as a part of our remaining work, **we will build on the outcomes and findings of T2.1 and investigate how the existing tools can be adapted to the next-generation social media** (especially federated platforms). We will also identify any necessary adaptation changes (given the researched characteristics of these platforms) that will serve as a valuable input for development of services envisioned in the AI-CODE project. The findings presented in this deliverable will specifically feed into the *Disinformation detection in the next-generation social media* and *Trustability and credibility assessment for content and sources* services.

Overall, based on the outcomes of our already achieved results and findings, we can conclude that the AI-CODE services are well-positioned to address the specific challenges that media experts are going to experience more and more on federated platforms. Our project is building tools that respect and enhance their community-driven nature of decentralized social media while helping to coordinate responses to disinformation across instances. As these platforms continue to evolve, maintaining this balance between local governance and system-wide resilience will be crucial for their sustainable development as alternatives to traditional social media.