Online Social Networks and Media 51 (2026) 100342

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

A longitudinal analysis of misinformation, polarization and toxicity on Bluesky after its public launch ☆

Gianluca Nogara [a,*], Erfan Samieyan Sahneh [a], Matthew R. DeVerna [b], Nick Liu [b], Luca Luceri [c], Filippo Menczer [b], Francesco Pierri [d], Silvia Giordano [a]

[a] ISIN - DTI, SUPSI, Lugano, Switzerland
[b] Observatory on Social Media, Indiana University, Bloomington, USA
[c] USC Information Sciences Institute, Los Angeles, CA, USA
[d] Dip. Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy

ARTICLE INFO

Keywords:
Bluesky
Decentralization
Online social media
Misinformation

ABSTRACT

Bluesky is a decentralized, Twitter-like social media platform that has rapidly gained popularity. Following an invite-only phase, it officially opened to the public on February 6th, 2024, leading to a significant expansion of its user base. In this paper, we present a longitudinal analysis of user activity in the two months surrounding its public launch, examining how the platform evolved due to this rapid growth. Our analysis reveals that Bluesky exhibits an activity distribution comparable to more established social platforms, yet it features a higher volume of original content relative to reshared posts and maintains low toxicity levels. We further investigate the political leanings of its user base, misinformation dynamics, and engagement in harmful conversations. Our findings indicate that Bluesky users predominantly lean left politically and tend to share high-credibility sources. After the platform's public launch, an influx of new users — particularly those posting in English and Japanese — contributed to a surge in activity. Among them, several accounts displayed suspicious behaviors, such as mass-following users and sharing content from low-credibility news sources. Some of these accounts have already been flagged as spam or suspended, suggesting that Bluesky's moderation efforts have been effective.

1. Introduction

Decentralized social media platforms have emerged as alternatives to traditional, centralized networks, offering users greater control over their data, content moderation, and online interactions. Unlike mainstream platforms, such as Twitter, Facebook, and Reddit, which rely on centralized governance and proprietary algorithms, decentralized platforms distribute authority across independent servers or protocols, reducing single points of failure and promoting transparency [1–4].

Bluesky Social[1] is a novel decentralized social media platform for microblogging based in the United States. Founded by Jack Dorsey as a Twitter-like alternative, Bluesky aims to provide a more open and user-controlled social networking experience. Initially available through an invite-only phase, the platform required a free invitation code to join Bluesky officially opened to the public on February 6th, 2024 [5]. With only a few thousand users active in January 2024, the platform registered more than one million new users on the first day of its opening [6]. Recent events, such as the banning of X/Twitter in Brazil and the 2024 U.S. presidential election, have contributed to a surge in new users on Bluesky, particularly among those who have grown distrustful of X and are seeking an alternative platform for political discussions and online discourse [7–9]. As a result, Bluesky has attracted millions of new users, solidifying its position as a viable alternative to mainstream platforms [10].

As decentralized platforms like Bluesky gain traction, they present new opportunities and challenges in online social dynamics, user behavior, and moderation strategies. On the one hand, they empower users with greater control over their online interactions, reduce reliance on centralized authorities, and promote transparency in content moderation [4]. On the other hand, the decentralized nature of these platforms raises concerns about the effectiveness of moderation, the spread of misinformation, and the potential for fragmented online communities with varying moderation standards [11].

☆ This article is part of a Special issue entitled: 'disinformation-toxicity-harms' published in Online Social Networks and Media.

* Corresponding author.
E-mail address: gianluca.nogara@supsi.ch (G. Nogara).

[1] https://bsky.social/about

https://doi.org/10.1016/j.osnem.2025.100342
Received 16 April 2025; Received in revised form 4 November 2025; Accepted 9 November 2025
Available online 27 November 2025
2468-6964/© 2025 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Here, building upon the preliminary findings of our previous work [12], we conduct a comprehensive analysis of user activity on Bluesky following its public launch on February 6th, 2024. Several characteristics we analyze align with findings from previous studies on related platforms [13,14]. Particularly, we confirm a broad distribution of user activity, minimal sharing of low-credibility sources, and a predominantly center-left-leaning user base. Our findings confirm these patterns in the context of Bluesky and extend them by examining temporal trends, community structures, and multilingual toxicity in a newly emerging decentralized environment. Specifically, we examine key characteristics of the platform, including temporal patterns of user engagement, the prevalence of different languages, and the structural properties of the follower network. We then investigate the estimated political leanings of Bluesky users and their relationship with sharing information from sources of varying reliability and participation in harmful conversations. Lastly, we analyze the largest communities on the platform and assess the extent and effectiveness of content moderation measures implemented during Bluesky's initial months of global availability. This work expands on our preliminary report [12] by adding an analysis of source credibility, examining the trend of high and low credibility sources along the timeline, and extending the study of toxicity from English and Japanese to other languages in the data. We also build a resharing network by linking users who shared posts, which allows us to analyze communities based on the characteristics of the users. Finally, we leverage user status information to analyze how many users have been moderated by the platform or are no longer present.

2. Related work

Distributed social networks aim for decentralization, allowing users to have more control and privacy. Early efforts like LifeSocial.KOM [15] and PeerSON [16] were based on the peer-to-peer model but faced challenges in performance and reliability. This led to a shift toward server-based federated models like Mastodon [2,3]. This approach balances flexibility and ease of use while maintaining some decentralization [1].

Mastodon, created in 2016, is a free and open-source social media platform that allows users to create their own servers ("instances") and connect with others across the globe. It is a decentralized social network, meaning that it is not owned by a single entity, but rather a network of independent servers that are connected together. A number of studies have identified striking features that make up Mastodon's distinct "fingerprint", distinguishing it from better-known online social networks [1,17]. Mastodon has, however, suffered from some natural pressures towards centralization, which can lead to potential points of failure [2].

To overcome this weakness, Bluesky developed its own decentralized AT protocol. Scalability, security, and ease of use make it an attractive option for building open and decentralized social media applications that prioritize user privacy and data security [4]. Using standard web technologies and re-using existing data models from the Web 3.0 protocol family also contributes to its efficiency and reliability [18]. Additionally, its federated networking model bolsters security by dispersing data across numerous servers, mitigating the risk of a single point of failure [2].

Existing literature on Bluesky primarily focuses on general platform analysis, highlighting its features and overall structure. Several studies have delved into the political dynamics and polarization within social networks, focusing on user activity patterns and interactions [13]. These analyses reveal how user interactions often lead to the formation of ideological groupings, which can evolve into echo chambers. Additionally, the impact of these dynamics on political division within decentralized social networks has been thoroughly examined. Other research has explored the platform's federated model [14], which allows users to curate their own experiences through customizable user feeds while maintaining interoperability between instances. Additional studies have focused on content analysis and user activities [19]. They have identified trends in user engagement and content virality, such as the types of posts that gain traction and the role of influencers in shaping discussions, highlighting differences from traditional centralized platforms. Finally, interactions were collected at the millisecond level by creating a multi-network temporal dataset [20] in order to perform an analysis of complex temporal dynamics such as community formation and social sanction patterns, i.e., user blocking.

Differently from previous work, our work focus on the evolution of the platform during its opening to the public to see if this has led to substantial changes in users or activities. Before the opening, the platform's controlled access reduced the risk of unauthorized or harmful usage. To study how this changed after the opening, we conducted an extensive longitudinal study of users and their activities before and after the opening.

3. Methods

3.1. Data collection

We collected data using Bluesky's free and publicly available *Firehose* endpoint, which grants developers real-time access to platform activities such as user posts, follows, and likes [21,22]. This endpoint ensures uninterrupted data collection by allowing reconnection and retrieval of up to 72 h of past data in case of disruptions, ensuring comprehensive coverage during the observation period.

We employed the `dart` library from the AT protocol [23], leveraging the Firehose endpoint[2] [24]. This data collection process was performed through the use of an existing open-source project [25]. The process is both straightforward and flexible, as the AT protocol does not require user authentication for access.

Bluesky enables the tracking of various user activities on the platform. Our analysis focuses on key actions: posts, replies, reposts, follows, and blocks, which constitute the primary forms of user interaction and communication. In this study, we considered quotes as posts since they contain original content from users and do not imply an endorsement of the quoted content. Users follow others to stay updated, create posts to share content, repost content from others, reply to engage in discussions, and block users to prevent other users from interacting with them. Active users are defined as individuals who have engaged in at least one activity, while passive users refer to individuals who have been subjected to an activity but have not independently participated in any activity.

While Bluesky's terms of service do not impose privacy restrictions on data collection, we strictly collect only publicly available user information, posts, and associated metadata in accordance with its Privacy Policy.[3] We do not publicly release the collected data and provide only anonymized information in this paper, except for select prominent accounts discussed in Section 4.7.

Fig. 1 presents key statistics of our dataset, collected over a 56-day period from January 9 to March 4, 2024, and encompassing 114 million user activities, including posting activities as well as follow and block actions.

To analyze the language distribution in shared content, we first removed irrelevant content (e.g., URLs, emojis, etc.) and then applied a language classifier using the `langdetect`[4] NLP library. Users were assigned to a language based on the language used in most of their posts.

[2] https://atprotodart.com/docs/lexicons/com/atproto/sync/subscriberepos/

[3] https://bsky.social/about/support/privacy-policy

[4] https://pypi.org/project/langdetect/

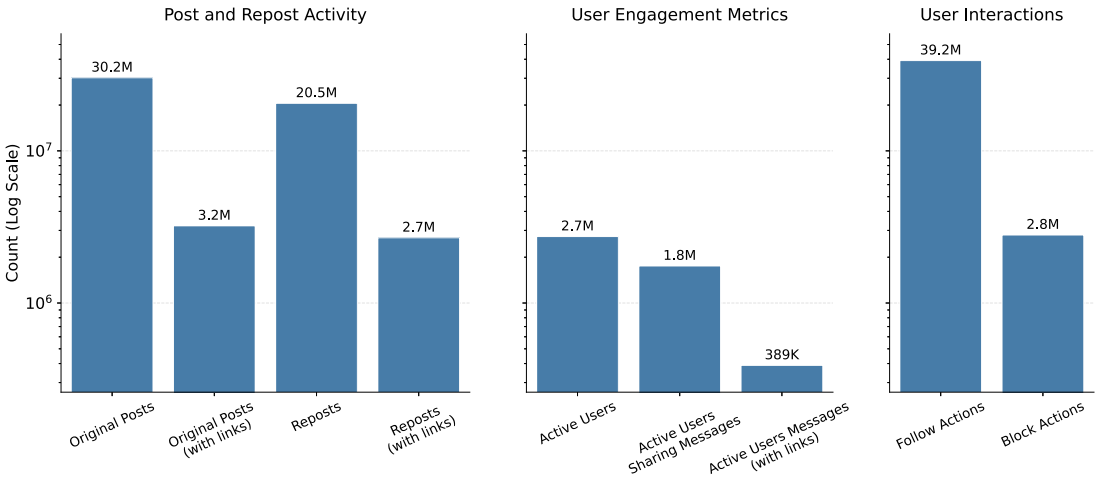

Fig. 1. Key dataset statistics over the whole observation period: total posts, posts containing links, active users, shared messages (with links), follow actions, and block actions. Values are presented on a logarithmic scale due to different orders of magnitude.

3.2. News source labeling

To evaluate the reliability of news outlets shared on Bluesky, we label web domains using NewsGuard[5] ratings, following a well-established approach in the literature [26–28]. NewsGuard is a reputable, independent organization that employs experts to assess news sources based on criteria such as transparency, accountability, adherence to journalistic standards, and error correction. Its ratings range from 0 (highly unreliable) to 100 (highly reliable), providing a standardized measure of news source credibility. Following prior work [29–31], we classify news outlets with a NewsGuard rating of 30 or lower as *low-credibility*, and those with a rating higher than 60 as *high-credibility* news sources.

We also leveraged political bias ratings from Media Bias/Fact Check (MBFC),[6] an independent organization that evaluates news media sources, to classify the political leaning of news websites shared on Bluesky. Sources are categorized along a seven-point political spectrum: Extreme Left (−3), Left (−2), Left-Center (−1), Least Biased (0), Right-Center (1), Right (2), and Extreme Right (3). We compute the political leaning score of a user by averaging the leaning of the sources shared by that user across their posts, following previous work [32].

The subset of posts containing URLs provides a valuable window into the types of external content shared within the broader dataset. Approximately 6.2 million posts, representing 8.7% of all posts, included at least one URL. Within this subset, we were able to apply source-level evaluations using two widely used external rating systems. Specifically, about one million posts (16% of posts containing URLs) were rated using NewsGuard reliability scores, and approximately 850,000 posts (13.6%) were rated using political bias classifications from Media Bias/Fact Check (MBFC). These ratings allow us to characterize the credibility and political orientation of a significant portion of the shared content and form the basis for several key analyses presented in the paper.

3.3. Community detection

To study user communities and interactions, we build a directed, weighted network based on user reshare activities. We chose the reshare network since it reflects the dynamics of content sharing and community endorsement, making it the most informative for analyzing active, functional communities, different networks such as follow and block do not give information on actual engagement or content propagation. Users are represented as nodes and an edge $(i \to j, w)$ represents user j resharing content by user i w times. We apply the Louvain community detection algorithm [33] to the undirected, unweighted version of this graph, obtaining 14,381 different communities. We studied the five largest communities, which represent 87% of users in our data. To gain a deeper understanding of the discussions within each community, we manually examined both users and the content they shared. We performed this manual analysis in order to also consider elements not visible through the data, such as images, profile pictures, and profile descriptions, so as to improve the accuracy of the analysis. For each community, we selected a sample of 30 users that included both randomly chosen members and those with the highest sum of in- and out-degree. We analyzed the posts shared by these users and, when possible, inspected their active profiles on the Bluesky web interface to observe their most recent activity and interactions.

4. Results

4.1. Temporal patterns of online activity

As shown in Fig. 1, original posts are the most common user activity on the platform, suggesting that users prefer creating new content over resharing or interacting with existing posts. This trend may be driven by the platform's rapid user-base expansion and it contrasts with existing centralized social media platforms, such as Twitter/X, where resharing via retweets is more prevalent [34,35].

Fig. 2A presents the daily number of unique active users along with follow and block actions. The platform's public opening on February 6th (dashed line) triggered sharp spikes in activity, peaking at one million active users and over 7 million follow actions the next day. These surges represent a nearly six-fold increase in active users and a 35-fold increase in follow actions compared to the previous day. Both trends declined rapidly in the following days, eventually stabilizing at slightly higher levels than those observed before the opening, likely reflecting the waning initial excitement around the platform.

The rise in users and follow actions was accompanied by a 3.6-fold surge in blocking activity, increasing from 33,269 to 120,054 instances. Blocking activity stabilized after a few days but remained slightly elevated compared to pre-February 6 levels. During the observation

[5] https://www.newsguardtech.com/
[6] https://mediabiasfactcheck.com/

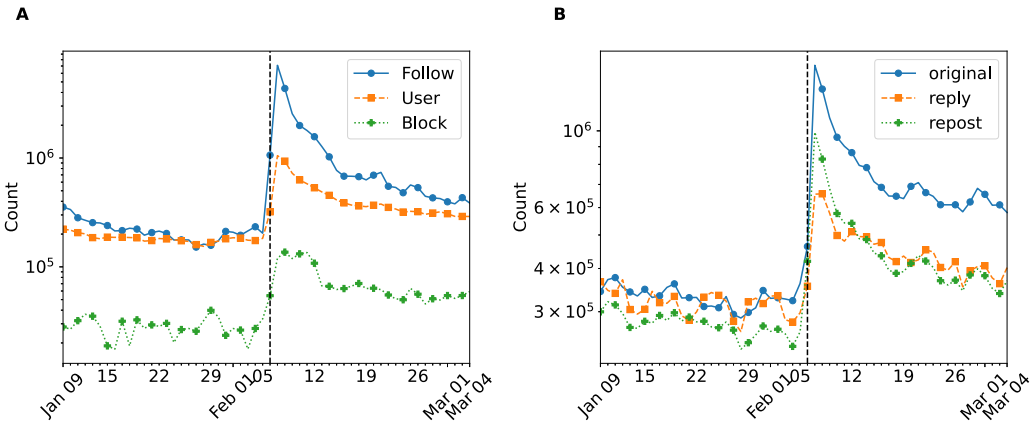

Fig. 2. Online activity on Bluesky before and after the public opening (Feb. 6th), indicated by the dashed line.

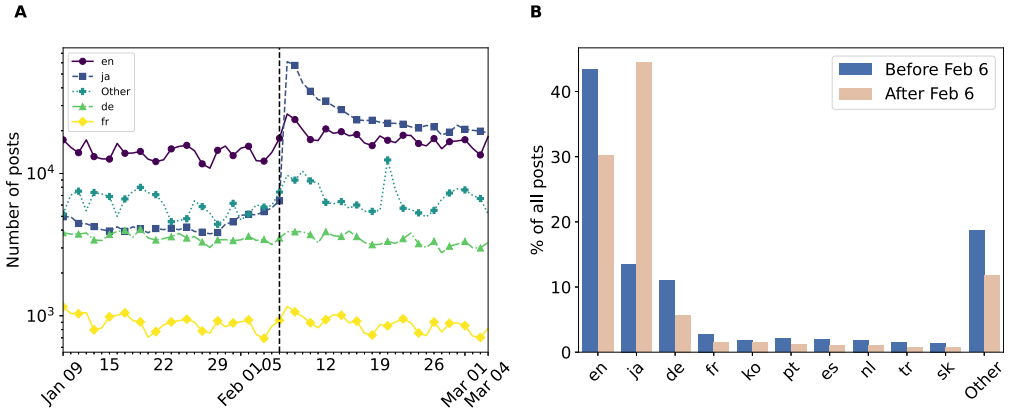

Fig. 3. **(A)** Trend of 5 top languages on Bluesky during the observation period. **(B)** Top 10 languages in Bluesky. Bars of the same color sum to 100%.

period, 287,539 users blocked a total of 758,681 accounts, accounting for 2.7 block actions received per user on average. However, due to the heterogeneous nature of blocking behavior, with some users blocking a disproportionately large number of accounts, the mean number of accounts blocked per user was higher (9.5).

Fig. 2B depicts the volume of shared posts over time, categorizing user activities into original posts, replies, and reposts. As previously noted in Fig. 1, original posts are the most prevalent form of shared activity, a trend that becomes even more pronounced following the platform's public launch.

As expected, the platform's opening on February 6th triggered a sharp increase in shared content across all types of posts, driven by the influx of new users. The volume of original posts surged 4.3-fold, from 362k on February 5th to 1.5M the next day. Reposts and replies also saw substantial growth, rising from 262,201 and 297,717 to 990,835 and 656,063, respectively. Despite these initial spikes, posting activity rapidly declined in the following days.

4.2. Prevalence of different languages

Fig. 3A shows the trends of the top five languages used throughout the observation period. Japanese posts surged immediately after the platform's opening, becoming the most used language. In contrast, content in other languages, including English, remained relatively stable, exhibiting only minor fluctuations following the opening.

Fig. 3B shows the prevalence of the top 10 languages in user posts, highlighting the dominance of English and Japanese, which together account for more than two-thirds of all content. The share of English content declined from 43% to 30% following the platform's opening, while Japanese content increased from 14% to 44% over the same period.

4.3. Changes in the follower network

Similarly to what we did for the reshare network, we build another network using the follow actions in the dataset. In this network, a directed edge $(i \rightarrow j)$ represents user i following user j. The sharp increase in *follow* activities after the public launch, shown in Fig. 2, is reflected in various follower network statistics measured before and after the platform's opening, as detailed in Table 1.

While the density of the follower network slightly decreased post-opening, the size of the strongly connected component more than tripled, and the average degree more than doubled. This suggests that Bluesky users tend to follow more accounts after the opening. The out-degree distributions in Fig. 4A further confirm this trend.

The average number of followers per user increased from 4.8 before February 6th to 10.5 after. Fig. 4 plots the distributions of out-degree (number of follow actions) and in-degree (number of followers). The ten accounts that gained the most followers and the ten accounts that followed the most users before and after the opening of the platform are presented in Table 2.

Notably, in each period, a single outlier user performed an exceptionally high number of follow actions — 50,570 before and 167,220 after, approximately 2–3 times the number of actions of other users

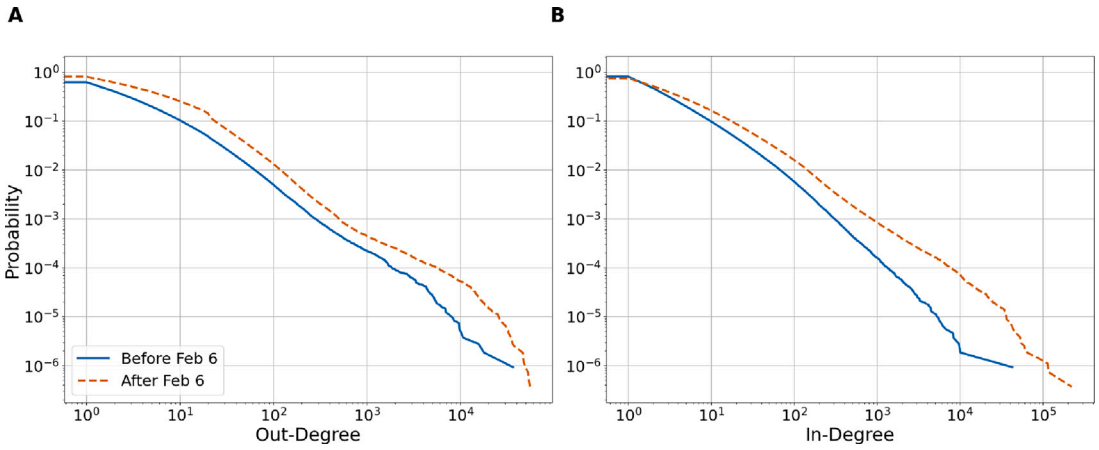

Fig. 4. Complementary cumulative distributions of node **(A)** out-degree and **(B)** in-degree in the follower network.

Table 1
Follower network statistics. LSCC stands for the largest strongly connected component.

	Before Feb. 6	After Feb. 6	Difference
Number of nodes	1,088,539	2,751,272	1,662,733
Number of edges	5,230,054	28,838,739	23,608,685
Density	4.4×10^{-6}	3.8×10^{-6}	-0.6×10^{-6}
Avg. in/out degree	4.8	10.5	5.7
LSCC size	~200k	~650k	~450k

listed in Table 2. Since the number of users performing follow actions before the opening is about one-third of those after (see Table 1), the probability in the CCDF tail is higher for the earlier group. When these two outlier users are removed, the crossover between the CCDF curves disappears.

Among the users who gained the most followers (Table 2), several remained consistent across both periods, with the majority being news outlets such as *The Washington Post*, *The New York Times*, and *Bloomberg*. This prevalence of news organizations among the most-followed accounts suggests that Bluesky may be evolving into a platform for news dissemination, potentially positioning itself as a replacement for Twitter. Despite this trend, three new Japanese accounts entered the top ten following the platform's public launch, with one rising to second place — surpassed only by the official Bluesky account.

Before the platform's opening, the users who followed the most new accounts were primarily English-language accounts and appeared to engage in normal activity, with the exception of one user who was suspended by Bluesky. After the opening, however, the composition of these users changed significantly: half were Japanese, two accounts were deleted, one was suspended, and two were flagged as spammers by Bluesky. Unlike the overall network behavior, these users were more engaged in reposting activities than in creating original posts.

4.4. Political leaning of Bluesky users

Fig. 5 presents the distribution of estimated political alignment among active Bluesky users, defined as those who shared at least five posts linking to rated websites. Each user's political alignment is determined by averaging the political bias scores of the websites they shared during the observation period (see Section 3.2). We can observe that the distribution is skewed towards liberal leaning, consistent with previous findings reported for Twitter before Musk's acquisition.[7] We present the distribution of political leanings computed over the entire observation period, as no significant differences were detected before and after the platform's opening (Mann–Whitney U test, $p = 0.05$).

We determine each user's political leaning based on the average bias score of their posts. In this analysis, we make the simplifying assumption that a negative bias score indicates a left-leaning, a score of zero indicates a centrist, and a positive score indicates a right-leaning, acknowledging that this thresholding is inherently arbitrary. The analysis revealed that the majority of active users on the platform are classified as left-leaning (bias score < 0), accounting for 74.61% of the sample. In contrast, 18.17% of users are categorized as centrist (bias score $= 0$), and only 7.21% are identified as right-leaning (bias score > 0). Accordingly, we report the distribution over the entire observation period, as the comparison between the pre- and post-opening phases of the platform did not yield a significant difference.

4.5. Credibility of information shared on Bluesky

Overall, the prevalence of low-credibility content shared on Bluesky during the period of analysis is negligible, comprising only 0.08% of all posts. Similarly, the proportion of users who shared at least one link to low-credibility sources is very small, accounting for just 0.13% of all users.

Fig. 6 presents the daily proportion of posts sharing links to high- and low-credibility websites. We observe a noticeable decline in the share of high-credibility domains after the platform's public opening, whereas the proportion of low-credibility domains remains relatively stable throughout the observation period. Despite these trends, high-credibility domains are shared 125 times more frequently than low-credibility ones on an average day. The median daily share rate is approximately 10% for high-credibility domains, compared to just 0.08% for low-credibility domains.

Fig. 7 presents the most frequently shared low- and high-credibility websites throughout the entire observation period. The composition of low-credibility domains remained largely consistent before and after Bluesky's public opening, with one notable exception: , a far-left socialist site known for publishing disputed claims, which saw its share decrease by half compared to pre-opening levels. However, this decline is not due to a reduction in the actual sharing of this domain but

[7] https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

Table 2
Annotated list of the ten users with the most new followers and the ten users who performed most follow activities before and after the platform's opening. Only the real names of prominent accounts are included for privacy reasons. 'JA' indicates Japanese-speaking users.

Accounts with most new followers				Accounts who followed most accounts			
Before Feb. 6		After Feb. 6		Before Feb. 6		After Feb. 6	
Account	Num.	Account	Num.	Account	Num.	Account	Num.
Bluesky	43,801	Bluesky	120,709	user	50,570	user	167,220
user	10,256	user JA	88,438	user	24,588	user JA	85,075
user	9886	user	66,535	user	23,482	user JA	55,073
Wash. Post	8,508	NY Times	62,087	user	21,907	user JA	45,848
NY Times	8,393	Wash. Post	60,904	user	13,414	user	44,588
user	6,809	user	55,081	user	12,310	user	44,460
Bluesky CEO	6,314	user JA	54,944	user	10,689	user	43,786
user	5,744	Bloomberg	54,800	user	9,950	user	40,926
user	5,621	user	52,455	user	8,211	user JA	37,209
Bloomberg	5,332	user JA	49,697	user	7,588	user JA	36,189

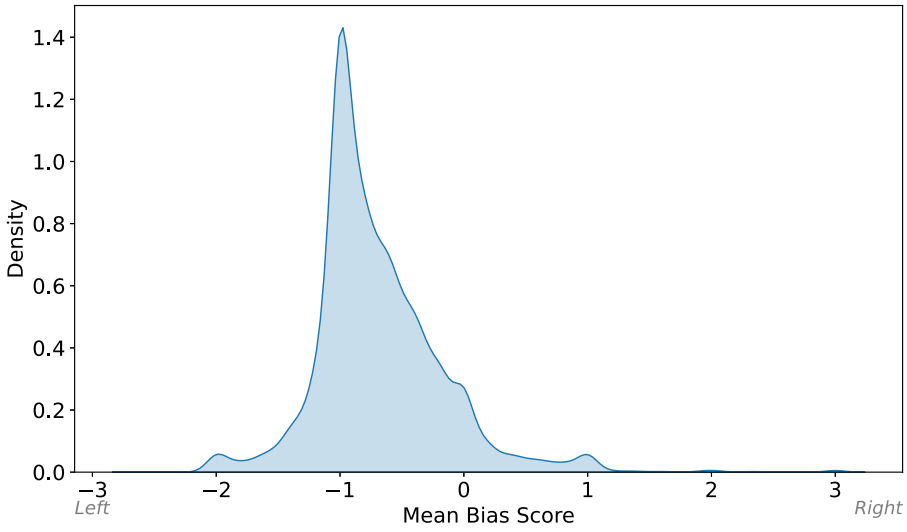

Fig. 5. Distribution of the average political leaning score of users that shared at least 5 links to rated domains.

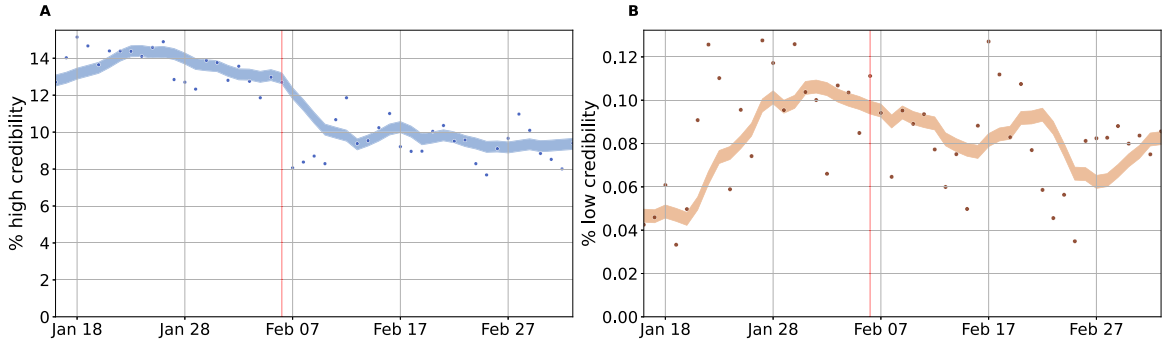

Fig. 6. Moving average (7-day) of the percentage of high credibility and low-credibility domains shared by users. **(A)** Percentage of posts sharing links to high-credibility domains, i.e., NewsGuard ratings ≥ 60. **(B)** Percentage of posts sharing links to low-credibility domains, i.e., NewsGuard ratings ≤ 30. The choice of different colors for high and low credibility content, light blue and orange, respectively, was made to help recognize the type of credibility of the figures.

rather to the overall increase in the sharing of all domains. As expected, many of these low-credibility sources are news agencies that have been previously accused of spreading disinformation [36]. Regarding high-credibility domains, a significant observation is the presence of German news agencies — e.g., and — alongside English-language sources, such as , and . This suggests that German users may be more active on Bluesky compared to other social networks, potentially reflecting regional differences in platform adoption.

Consistent with findings reported for other social media platforms [31,37–39], we observe evidence of so-called "superspreaders", i.e., a small subset of users responsible for the majority of unreliable content shared on Bluesky. Specifically, we find that ten accounts (1.8% of all users who shared low-credibility content) were responsible for disseminating 62% of links to low-credibility sources.

4.6. Toxicity of Bluesky conversations

We analyzed the toxicity of online conversations on Bluesky using Detoxify [40], a model used to detect toxic comments, focusing on the languages supported by the model (English, Italian, French, Russian, Portuguese, Spanish, and Turkish). Toxicity levels remained stable

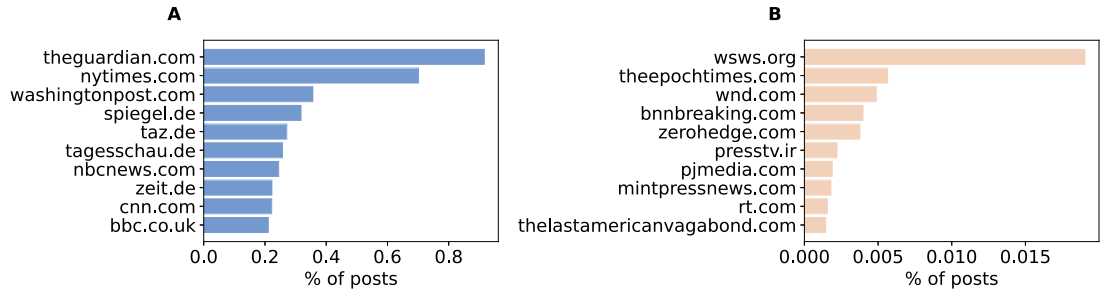

Fig. 7. Most shared **(A)** high-credibility and **(B)** low-credibility websites during the period of analysis.

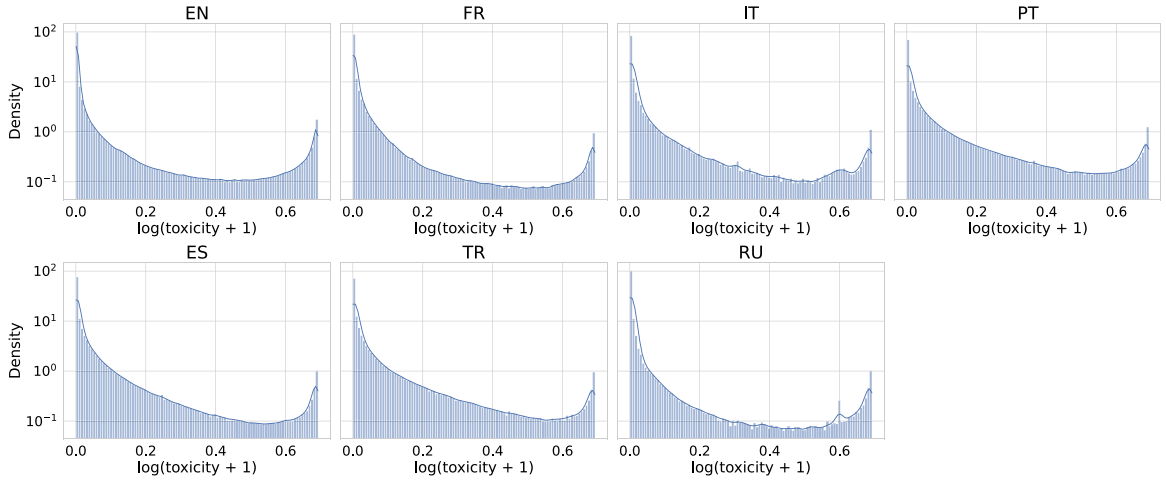

Fig. 8. Toxicity of Bluesky posts. Density distribution of toxicity for the different languages. The y-axis is scaled logarithmically to enhance visibility of low-density regions.

throughout the observation period, with no significant temporal trends emerging after the platform's public launch.

While the distributions of toxicity scores differ across languages ($p < 0.05$ according to a Kruskal–Wallis test), Fig. 8 shows that highly toxic posts are rare overall, and that only a small fraction of posts exhibit concerning levels of toxicity in all cases. This suggests that, despite some variation in average scores, overtly toxic language is not a dominant feature of discourse on the platform. The same pattern holds for user toxicity, which is computed by averaging the toxicity scores of all posts that each user shared. In this analysis, we did not use a threshold, but results are similar when we eliminate users with only one post (less than 7%). Overall, our findings suggest that, while some toxicity does exist on the platform, it is relatively uncommon, and that the tone of conversations may be influenced by language, and possibly by broader cultural or community dynamics.

After a Bonferroni correction, we collected the pairs of languages with significant differences in toxicity per user ($p < 0.05$): English (EN) differs from Spanish (ES), Russian (RU), Italian (IT), and French (FR); and Portuguese (PT) differs from Spanish (ES), Russian (RU), Italian (IT), and French (FR).

4.7. Interplay between misinformation, political leaning, and toxicity

To investigate whether these factors vary according to users' activity levels and permanence in the dataset, we classified users according to their activities and age, calculated as the number of days elapsed between their first and last activity. Users who posted between one and nine times were considered *low activity*, those with between 10 and 99 posts were classified as *high activity*, and users with 100 or more posts were classified into the *hyper activity* group. Regarding age, we classified users with age smaller than 7 days as *low age*, those with age between 7 and 29 days as *medium age*, and *high age* those with age older than 30 days. Analyzing the activity of users, we did not find any correlation between these parameters and more or less active users; similarly, no correlations were found between the same parameters and the age of users. The results are in Sec. 6.

As shown in Fig. 9A, we observe that extreme users, on the left and the right side of the political spectrum, are associated with a lower quality of shared information (computed as the average rating of all the links they shared) compared to other users.

Fig. 9B illustrates the correlation between toxic behavior and political leaning. We observe that left-leaning users exhibit slightly higher levels of toxicity than center- or right-leaning users; however, these differences are small, indicating that all user groups are aligned at similar toxicity values.

Finally, as shown in Fig. 9C, we do not observe significant patterns in the relationship between toxic behavior and misinformation sharing on the platform.

4.8. Analysis of communities

We now explore the five largest reposting communities identified using Louvain's algorithm (see Section 3.3 for details). These communities are numbered by size, with community 1 being the largest and community 5 the smallest. They are relatively homogeneous, organized

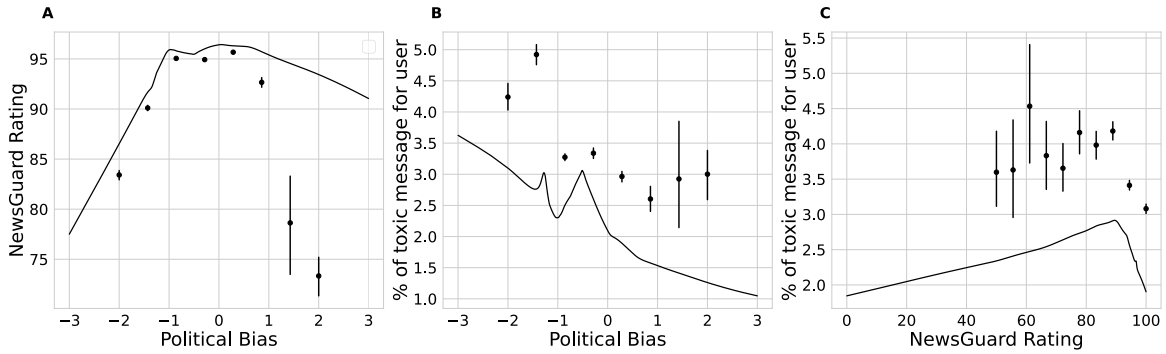

Fig. 9. Interplay between metrics. A locally weighted scatterplot smoothing (LOWESS) curve is overlaid to highlight non-linear trends. The error bars represent the standard deviation. **(A)** Relationship between political bias and reliability, **(B)** between political bias and toxicity, and **(C)** between reliability and toxicity.

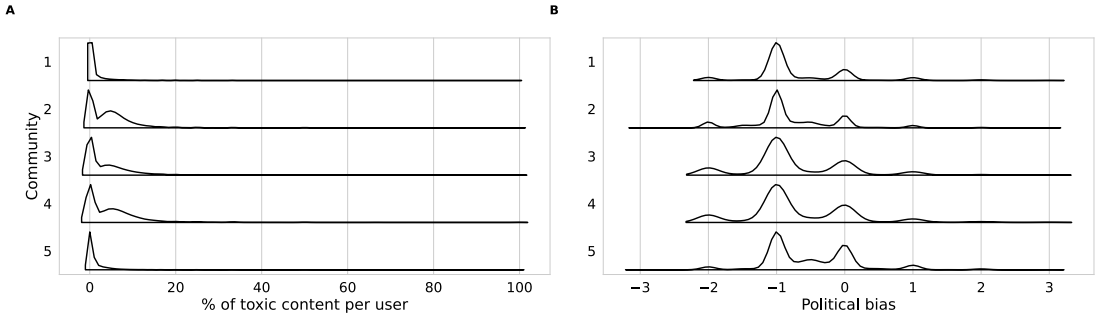

Fig. 10. User-level community analysis. Violins are split in half for a better understanding of the graph. **(A)** Distributions of percentages of toxic content. **(B)** Distributions of political bias, where -3 means extreme left, 0 center, and 3 extreme right.

Table 3
Annotated list of the five largest user-level communities with the respective percentages of users in the dataset, the dominant language of the community, and a brief description of the content shared by their users.

Community	% of users	Language	Description
1	35.5	JA	explicit art cartoons
2	19.7	EN	left-wing politics
3	13.7	EN	art cartoons
4	10.4	EN	explicit art cartoons
5	7.5	DE	left-wing politics

primarily by language groups and type of content shared. Table 3 shows the percentages of users in each community.

Through a manual review of hashtags, linked domains, and images (see Section 3.3 for details), we found that much of the content circulating within these communities is related to artistic and cartoon themes. A strong Japanese-language presence is evident in the dataset, particularly within community 1, which is characterized by apolitical, artistic, and explicit content. Similar themes appear in communities 3 and 4, which also share English-language artistic material, though community 4 is more explicit in tone and subject matter. In contrast, communities 2 and 5 are politically focused, with English and German as their dominant languages respectively. Reviewing shared domains within these communities reveals that both exhibit a left-leaning political orientation.

As shown in Fig. 10A, we computed the distributions of toxic messages shared on average by users across different communities, to measure the frequency with which users in different communities share toxic messages. To identify toxic texts, we used a toxicity score threshold of 0.5 [41,42]. Our analysis focused on the median proportion of toxic messages per user within each community. The results indicate that users in communities 2 and 4 tend to share more toxic content, with median toxicity rates of 3.7% and 3.5% per user, respectively. In comparison, community 3 shows significantly lower toxicity levels, with a median value of just 1.0%. Users within communities 1 and 5 display minimal to negligible toxic behavior relative to other communities, both presenting a median toxicity level close to 0.0%. Kruskal–Wallis tests with Bonferroni correction reveal statistically significant differences ($p < 0.001$) among all community pairs except for between communities 2 and 4.

We also analyzed the distribution of political leanings among users in different communities. As shown in Fig. 10B, the overall political orientation skews left, with community medians clustering between −0.67 and -1, in line with the general Bluesky population.

4.9. Content moderation

We further examined moderation on the platform by analyzing the status of user accounts. To do this, we used the `getProfiles` endpoint,[8] which provides information about user profiles, including their status.

A user's account status varies depending on their standing on the platform. Active accounts return a status of *Online*, while deleted accounts return *InvalidRequest*, indicating that the profile was likely canceled by the user. Moderated accounts produce one of two responses — *AccountDeactivated* or *AccountTakedown* — both signaling a policy violation. The former typically indicates a temporary enforcement action, while the latter reflects a more permanent removal.

As of November 2024, most users in our dataset (95.9%) remained active. In contrast, 3.6% of accounts had been deleted, 0.4% were temporarily deactivated, and 0.1% had been permanently removed — indicating that 0.5% of users had been subject to platform moderation.

To understand the reasons for user moderation, we analyzed the extent to which users in each of these groups shared toxic and low-credibility content. Fig. 11A suggests that moderation was done against

[8] https://docs.bsky.app/docs/api/app-bsky-actor-get-profiles

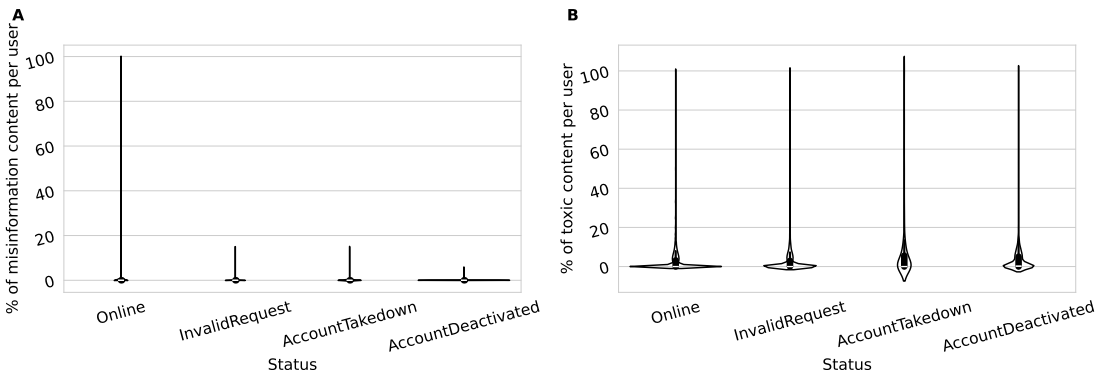

Fig. 11. Analysis of user status categories. **(A)** Percentage of misinformation shared. **(B)** Percentage of toxic content shared.

users who violated the terms of service, as users who shared links from low-credibility sources remain. Similarly, in the case of toxicity, the moderated categories do not exhibit different behavior from other users, as shown in Fig. 11B, likely because toxicity is not officially mentioned in the community guidelines [43]. Kruskal–Wallis tests with Bonferroni correction were performed on the category of account according to toxicity and misinformation per user. The tests reveal statistically significant differences ($p < 0.001$) between all categories except the pairs *AccountDeactivated-AccountTakedown* in both cases and *AccountDeactivated-InvalidRequest* in the case of misinformation.

5. Discussion

In this study, we provided the first large-scale analysis specifically examining how Bluesky's transition from an invite-only platform to a publicly accessible one impacted user activity and network structure. Our findings reveal that while Bluesky exhibits a broad distribution of activity similar to more established social media platforms, it stands out due to a higher volume of original content compared to reshared posts and low levels of toxicity. After the public opening, Bluesky experienced a surge in new activity. The influx of Japanese users is an intriguing trend that warrants further investigation, as it may indicate that Bluesky's conversational dynamics are particularly appealing to this demographic.

We uncovered evidence of suspicious user behavior, including accounts that followed large numbers of users and shared content from low-credibility news outlets. A small subset of users was responsible for the majority of unreliable content shared, a pattern previously observed on other platforms like Twitter/X. Following the platform's public opening, some of the most aggressive account followers were flagged as spam or even deleted or suspended, suggesting attempts to misuse the platform. The fact that some of these suspicious actors were swiftly banned or labeled as spam indicates that content moderation mechanisms are actively functioning on Bluesky. Our analysis of account statuses further reveals that only 0.5% of users have been moderated, while 3.6% are no longer on the platform, leaving 95.9% of users active [44].

We examined the reshare network, identifying that the five largest communities accounted for 87% of users. Our findings suggest that English-speaking communities exhibit higher levels of toxicity compared to other language-based communities, though overall, toxicity levels remained relatively stable across different languages. We also confirmed that Bluesky users lean predominantly left-wing politically, as indicated by the classification of shared domains and the political orientation of the largest communities.

Our findings have several important implications for understanding decentralized social media dynamics and content moderation. The relatively low levels of toxicity on Bluesky, combined with its high proportion of original content, suggest that alternative platforms may foster a different type of user engagement compared to mainstream counterparts. However, the emergence of suspicious activity, including users sharing low-credibility information and engaging in mass-following behaviors, indicates that Bluesky is not immune to manipulation attempts. The platform's ongoing moderation efforts, which have already flagged or removed some misbehaving accounts, highlight both the challenges and potential effectiveness of content governance in alternative social platforms.

The left-leaning political orientation of Bluesky's user base raises questions about ideological fragmentation and the extent to which decentralized platforms may cater to specific political communities. The significant influx of Japanese users also suggests that Bluesky may be attracting geographically and linguistically distinct user groups, which could influence the platform's future development and community structure.

Our results confirm the findings previously obtained by earlier studies [13,14], thus supporting the presence of wide distributions of activity, the limited number of low-quality links, and the fact that the platform is populated mainly by users with a center-left political orientation.

This study has several limitations. First, our analysis covers a relatively short period (56 days), during which the platform underwent a significant transformation as it opened to the public. As a result, our findings capture early trends that may continue to evolve over time. Second, approximately 8% of reposts could not be traced back to their original posts due to the constraints of our data collection period. Third, the credibility and political bias scores we used were available only for a subset of websites that we could assign a rating to only 16.0% of posts containing a URL using NewsGuard and 13.6% using MBFC, limiting the scope of our misinformation analysis. Finally, our toxicity analysis was restricted to languages supported by Detoxify — namely English, Spanish, French, Italian, Portuguese, Russian, and Turkish — leaving out other languages that may have distinct toxicity patterns.

Future research should extend the analysis over a longer period to assess the long-term evolution of Bluesky's user behavior, content sharing patterns, and moderation effectiveness. A deeper investigation into the rise of Japanese users on the platform could provide insights into regional adoption patterns of decentralized social media. In addition, the slopes of the two degree distributions should be compared so as to check for any decreases in slope, thus indicating the emergence of larger hubs. Further studies should also explore content moderation approaches in decentralized networks, comparing their effectiveness against those of centralized platforms. Additionally, expanding toxicity and misinformation analysis to a broader range of languages could offer a more comprehensive understanding of harmful content dynamics in multilingual decentralized spaces. Finally, the network analysis could be extended to study block actions among users.

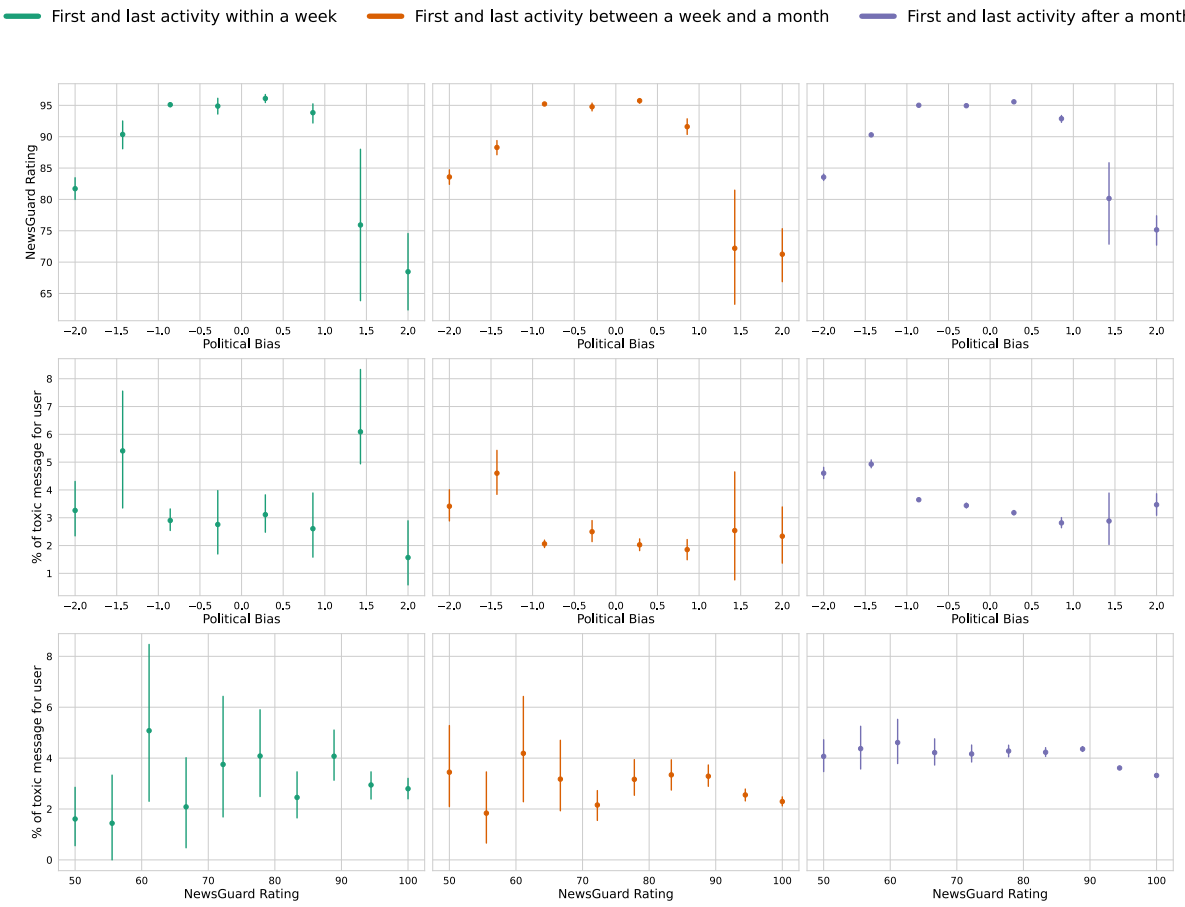

Fig. 12. Interplay between metrics based on the user's age. The error bars represent the standard deviation. The first horizontal row shows the relationship between political orientation and reliability, the second row shows the relationship between political orientation and toxicity, and the third row shows the relationship between reliability and toxicity.

CRediT authorship contribution statement

Gianluca Nogara: Writing – original draft, Visualization, Formal analysis, Methodology, Data curation. **Erfan Samieyan Sahneh:** Writing – original draft, Visualization, Formal analysis. **Matthew R. DeVerna:** Writing – review & editing, Supervision. **Nick Liu:** Data curation. **Luca Luceri:** Writing – review & editing, Funding acquisition. **Filippo Menczer:** Writing – review & editing, Supervision, Funding acquisition. **Francesco Pierri:** Writing – review & editing, Supervision, Methodology, Funding acquisition. **Silvia Giordano:** Writing – review & editing, Supervision, Methodology, Funding acquisition.

Ethical considerations

Our study leverages publicly available data collected from the Bluesky social network, in full accordance with its Terms of Service. We do not attempt to deanonymize users or infer private information, and we do not release any raw or potentially identifiable data as part of this work. Our analysis is limited to aggregate patterns and publicly shared content.

As our research exclusively involves the analysis of publicly available data and does not engage directly with human subjects, it was determined to be exempt from formal IRB review. Nonetheless, we conducted the study in line with best practices for ethical social media research, including data minimization and user privacy protection.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the Swiss National Science Foundation (grant number CRSII5_209250) and the Italian Ministry of Education (PRIN PNRR grant CODE prot. P2022AKRZ9 and PRIN grant DEMON prot. 2022BAXSPY).

Appendix

Here we provide supplementary figures and results referenced and commented on in the main text.

Fig. 12 shows the interplay between the political bias, toxicity, and NewsGuard rating with users classified using the age. Users have been classified according to their age within the dataset, obtained by performing a difference between the first and the last activity. Fig. 13 shows the interplay between the same metrics with users classified according to their activities.

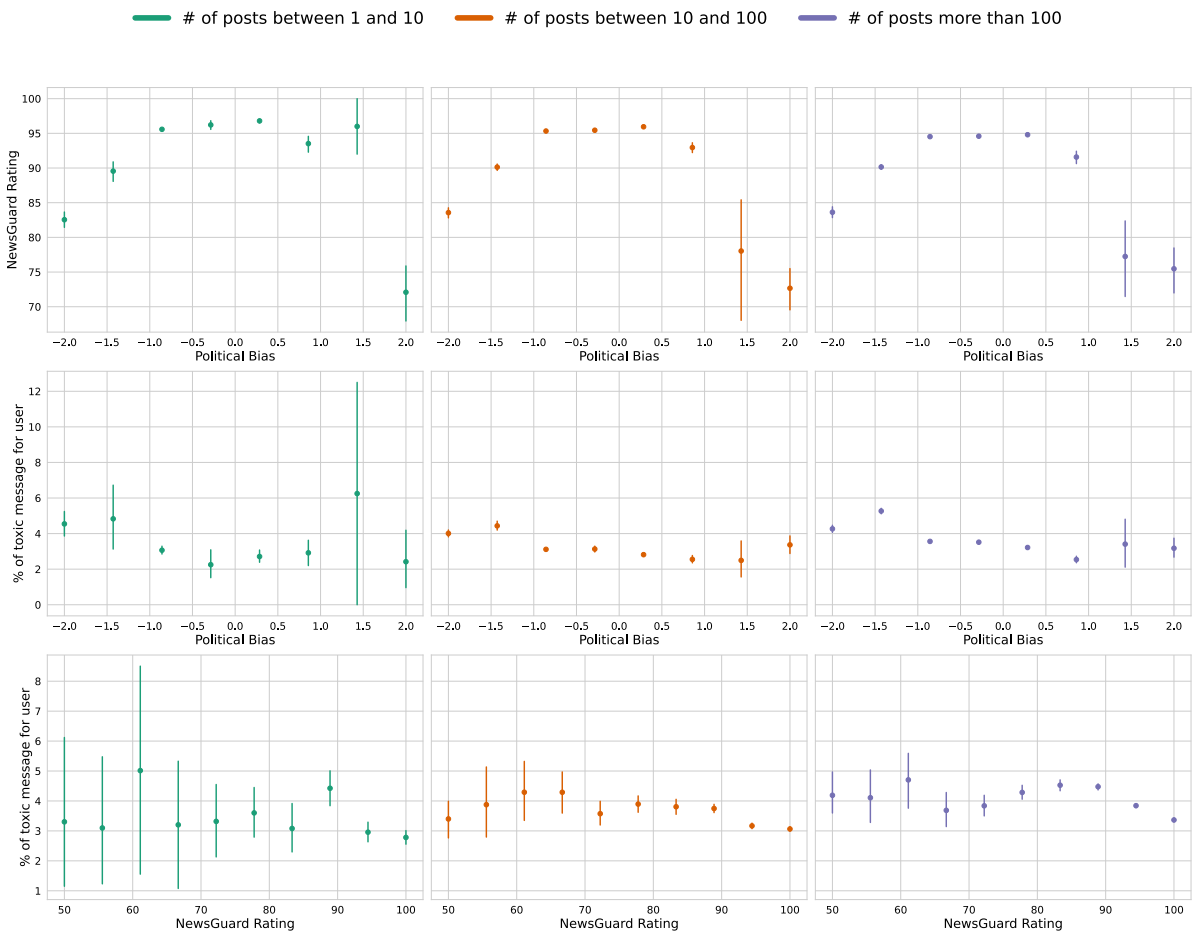

Fig. 13. Interplay between metrics based on the user's activity. The error bars represent the standard deviation. The first horizontal row shows the relationship between political orientation and reliability, the second row shows the relationship between political orientation and toxicity, and the third row shows the relationship between reliability and toxicity.

References

[1] C.A. Bono, L. La Cava, L. Luceri, F. Pierri, An exploration of decentralized moderation on Mastodon, in: WEBSCI '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 53–58, http://dx.doi.org/10.1145/3614419.3644016.

[2] A. Raman, S. Joglekar, E.D. Cristofaro, N.R. Sastry, G. Tyson, Challenges in the decentralised web: The Mastodon case, in: Proc. Internet Measurement Conference, 2019, URL https://arxiv.org/abs/1909.05801.

[3] M. Zignani, S. Gaito, G.P. Rossi, Follow the "Mastodon": Structure and evolution of a decentralized online social network, in: International Conference on Web and Social Media, 2018, pp. 541–550, http://dx.doi.org/10.1609/icwsm.v12i1.14988.

[4] M. Kleppmann, P. Frazee, J. Gold, J. Graber, D. Holmgren, D. Ivy, J. Johnson, B. Newbold, J. Volpert, Bluesky and the AT protocol: Usable decentralized social media, 2024, arXiv preprint URL https://arxiv.org/pdf/2402.03239.pdf.

[5] A. Silberling, Bluesky is now open for anyone to join, 2024, URL https://techcrunch.com/2024/02/06/bluesky-is-now-open-for-anyone-to-join/ (Last Accessed 16 April 2024).

[6] Bluesky, One million new users since we opened Bluesky yesterday!, 2024, Post on Bluesky. URL https://bsky.app/profile/bsky.app/post/3kkv3clm3su2h (Last Accessed 16 April 2024).

[7] L. La Cava, L.M. Aiello, A. Tagarelli, Drivers of social influence in the Twitter migration to Mastodon, Sci. Rep. 13 (2023) 21626, http://dx.doi.org/10.1038/s41598-023-48200-7, URL https://www.nature.com/articles/s41598-023-48200-7.

[8] D.D. Placido, The X (Twitter) Exodus to Bluesky, explained, 2024, Forbes, URL https://www.forbes.com/sites/danidiplacido/2024/11/19/the-x-twitter-exodus-to-bluesky-explained/.

[9] R. Boyd, From X to Bluesky: why are people fleeing Elon Musk's 'digital town square'?, 2024, The Guardian, URL https://www.theguardian.com/media/2024/dec/11/from-x-to-bluesky-why-are-people-abandoning-twitter-digital-town-square.

[10] D. Butts, Social media platform Bluesky attracts millions in Brazil after judge bans Musk's X, 2024, CNBC, URL https://www.cnbc.com/2024/09/04/social-media-platform-bluesky-attracts-millions-in-brazil-after-judge-bans-musks-x-.html.

[11] C. Berg, E. Morton, M. Poblet, Social media has huge problems with free speech and moderation. Could decentralised platforms fix this? 2021, Published online in The Conversation, URL https://theconversation.com/social-media-has-huge-problems-with-free-speech-and-moderation-could-decentralised-platforms-fix-this-157053.

[12] E.S. Sahneh, G. Nogara, M.R. DeVerna, N. Liu, L. Luceri, F. Menczer, F. Pierri, S. Giordano, The dawn of decentralized social media: An exploration of Bluesky's public opening, Lecture Notes in Comput. Sci. (2025) 422–437, http://dx.doi.org/10.1007/978-3-031-78541-2_26.

[13] D. Quelle, A. Bovet, Bluesky: Network topology, polarisation, and algorithmic curation, 2024, arXiv preprint arXiv:2405.17571.

[14] A. Failla, G. Rossetti, "I'm in the Bluesky tonight": Insights from a year worth of social data, 2024, arXiv preprint arXiv:2404.18984.

[15] K. Graffi, C. Gross, D. Stingl, D. Hartung, A. Kovacevic, R. Steinmetz, LifeSocial.KOM: A secure and P2P-based solution for online social networks, in: 2011 IEEE Consumer Communications and Networking Conference, CCNC, 2011, pp. 554–558, http://dx.doi.org/10.1109/CCNC.2011.5766541.

[16] S. Buchegger, D. Schiöberg, L.-H. Vu, A. Datta, PeerSoN: P2P social networking: early experiences and insights, in: Workshop on Social Network Systems, 2009, pp. 46–52, http://dx.doi.org/10.1145/1578002.1578010.

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

[18] C. Edwards, Social media and the distributed self, Eng. Technol. 18 (4) (2023) 40–46.

[19] L. Balduf, S. Sokoto, O. Ascigil, G. Tyson, B. Scheuermann, M. Korczyński, I. Castro, M. Król, Looking at the blue skies of Bluesky, 2024, arXiv:2408.12449.

[20] U. Jeong, B. Jiang, Z. Tan, H.R. Bernard, H. Liu, Descriptor: A temporal multi-network dataset of social interactions in Bluesky social (BlueTempNet), 2024, http://dx.doi.org/10.21227/yrsy-ee91, Ieee.org 1URL https://ieeexplore.ieee.org/abstract/document/10706594.

[21] Bluesky, Firehose, 2024, https://www.docs.bsky.app/docs/advanced-guides/firehose (Accessed 29 April 2024).
[22] AT Protocol, Event Stream, 2024, https://atproto.com/specs/event-stream.
[23] atprotodart.com, Bluesky | Dart package, 2024, Software library, URL https://pub.dev/packages/bluesky (Accessed April 2024).
[24] S. Kato, Bluesky Firehose endpoint, 2024, https://atprotodart.com/docs/lexicons/com/atproto/sync/subscriberepos/ (Accessed April 2024).
[25] K. Burghardt, Data collection with Bluesky Firehose endpoint, 2024, https://github.com/KeithBurghardt/bluesky_firehose/tree/main (Accessed April 2024).
[26] K.-C. Yang, F. Pierri, P.-M. Hui, D. Axelrod, C. Torres-Lugo, J. Bryden, F. Menczer, The COVID-19 Infodemic: Twitter versus Facebook, Big Data & Soc. 8 (1) (2021) http://dx.doi.org/10.1177/20539517211013861.
[27] F. Pierri, M.R. DeVerna, K.-C. Yang, D. Axelrod, J. Bryden, F. Menczer, One year of COVID-19 vaccine misinformation on Twitter: Longitudinal study, J. Med. Internet Res. 25 (2023) http://dx.doi.org/10.2196/42227.
[28] J. Lühring, H. Metzler, R. Lazzaroni, A. Shetty, J. Lasser, Best practices for source-based research on misinformation and news trustworthiness using NewsGuard, J. Quant. Descr.: Digit. Media 5 (2025) http://dx.doi.org/10.51685/jqd.2025.003, URL https://journalqd.org/article/view/4500.
[29] G. Nogara, F. Pierri, S. Cresci, L. Luceri, S. Giordano, Misinformation and Polarization around COVID-19 vaccines in France, Germany, and Italy, in: Proc. 16th ACM Web Science Conference 2024, 2024, http://dx.doi.org/10.1145/3614419.3644020.
[30] F. Pierri, The diffusion of mainstream and disinformation news on Twitter: the case of Italy and France, in: Companion Proc. of the Web Conference, WWW, 2020, pp. 617–622, http://dx.doi.org/10.1145/3366424.3385776.
[31] F. Pierri, A. Tocchetti, L. Corti, M. Di Giovanni, S. Pavanetto, M. Brambilla, S. Ceri, VaccinItaly: monitoring Italian conversations around vaccines on Twitter and Facebook, 2021, arXiv preprint arXiv:2101.03757.
[32] M. Cinelli, G.D.F. Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, The echo chamber effect on social media, Proc. Natl. Acad. Sci. 118 (2021) 1–8, http://dx.doi.org/10.1073/pnas.2023301118, URL https://www.pnas.org/doi/10.1073/pnas.2023301118.
[33] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. 2008 (10) (2008) P10008, http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.
[34] T. Alshaabi, D.R. Dewhurst, J.R. Minot, M.V. Arnold, J.L. Adams, C.M. Danforth, P.S. Dodds, The growing amplification of social media: measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020, Epj Data Sci. 10 (2021) http://dx.doi.org/10.1140/epjds/s13688-021-00271-0.
[35] L. Luceri, F. Cardoso, S. Giordano, Down the bot hole: Actionable insights from a one-year analysis of bot activity on Twitter, First Monday (2021) http://dx.doi.org/10.5210/fm.v26i3.11441.
[36] M. Hellman, Everyday disinformation: RT and sputnik news coverage, in: Security, Disinformation and Harmful Narratives: RT and Sputnik News Coverage About Sweden, Springer Nature Switzerland, Cham, 2024, pp. 59–99, http://dx.doi.org/10.1007/978-3-031-58747-4_3.
[37] M.R. DeVerna, R. Aiyappa, D. Pacheco, J. Bryden, F. Menczer, Identifying and characterizing superspreaders of low-credibility content on Twitter, PLOS ONE 19 (5) (2024) e0302201, http://dx.doi.org/10.1371/journal.pone.0302201.
[38] G. Nogara, P.S. Vishnuprasad, F. Cardoso, O. Ayoub, S. Giordano, L. Luceri, The disinformation dozen: An exploratory analysis of Covid-19 disinformation proliferation on Twitter, in: Proc. 14th ACM Web Science Conference 2022, 2022, http://dx.doi.org/10.1145/3501247.3531573.
[39] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on Twitter during the 2016 U.S. presidential election, Science 363 (6425) (2019) 374–378, http://dx.doi.org/10.1126/science.aau2706.
[40] L. Hanu, Unitary team, Detoxify, 2020, Github. https://github.com/unitaryai/detoxify.
[41] Y. Hua, T. Ristenpart, M. Naaman, Towards measuring adversarial Twitter interactions against candidates in the US midterm elections, in: Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, 2020, pp. 272–282, http://dx.doi.org/10.1609/icwsm.v14i1.7298, URL https://ojs.aaai.org/index.php/ICWSM/article/view/7298.
[42] M. Saveski, B. Roy, D. Roy, The structure of toxic conversations on Twitter, in: Proceedings of the Web Conference 2021, vol. 1229, 2021, http://dx.doi.org/10.1145/3442381.3449861.
[43] Community Guidelines, 2024, Bluesky, URL https://bsky.social/about/support/community-guidelines.
[44] D.A. Broniatowski, W. Zhong, J.R. Simons, M. Dredze, L.C. Abroms, Explaining Twitter's Inability to Reduce Vaccine Misinformation during the COVID-19 Pandemic, Springer Science and Business Media LLC, 2025, http://dx.doi.org/10.21203/rs.3.rs-5691823/v1, URL https://www.researchsquare.com/article/rs-5691823/v1.