Flocking to Mastodon: Tracking the Great Twitter Migration

Jiahui He*
Hong Kong University of Science and Technology (GZ)
Guangzhou, China
jhe976@connect.hkust-gz.edu.cn

Haris Bin Zia*
Queen Mary University of London
London, United Kingdom
h.b.zia@qmul.ac.uk

Ignacio Castro
Queen Mary University of London
London, United Kingdom
i.castro@qmul.ac.uk

Aravindh Raman
Telefonica Research
Barcelona, Spain
aravindh.raman@telefonica.com

Nishanth Sastry
University of Surrey
Guildford, United Kingdom
n.sastry@surrey.ac.uk

Gareth Tyson
Hong Kong University of Science and Technology (GZ)
Guangzhou, China
gtyson@ust.hk

ABSTRACT

The acquisition of Twitter by Elon Musk has spurred controversy and uncertainty among Twitter users. The move raised both praise and concerns, particularly regarding Musk's views on free speech. As a result, a large number of Twitter users have looked for alternatives to Twitter. Mastodon, a decentralized micro-blogging social network, has attracted the attention of many users and the general media. In this paper, we analyze the migration of 136,009 users from Twitter to Mastodon. We inspect the impact that this has on the wider Mastodon ecosystem, particularly in terms of user-driven pressure towards centralization. We further explore factors that influence users to migrate, highlighting the effect of users' social networks. Finally, we inspect the behavior of individual users, showing how they utilize both Twitter and Mastodon in parallel. We find a clear difference in the topics discussed on the two platforms. This leads us to build classifiers to explore if migration is predictable. Through feature analysis, we find that the content of tweets as well as the number of URLs, the number of likes, and the length of tweets are effective metrics for the prediction of user migration.

CCS CONCEPTS

• Computing methodologies → Supervised learning by classification; **• Human-centered computing → Social networks**;

KEYWORDS

Twitter, Mastodon, User Migration, Topic Modeling, Machine Learning

*Joint first authors

IMC '23, October 24–26, 2023, Montreal, QC, Canada

ACM ISBN 979-8-4007-0382-9/23/10…$15.00
https://doi.org/10.1145/3618257.3624819

ACM Reference Format:
Jiahui He, Haris Bin Zia, Ignacio Castro, Aravindh Raman, Nishanth Sastry, and Gareth Tyson. 2023. Flocking to Mastodon: Tracking the Great Twitter Migration. In *Proceedings of the 2023 ACM Internet Measurement Conference (IMC '23), October 24–26, 2023, Montreal, QC, Canada.*

1 INTRODUCTION

In October 2022, Elon Musk, a self-declared "free speech absolutist" [3] acquired Twitter — the social network that he regarded as the "de facto town square" where public debate takes place. Musk's takeover has been controversial and highly publicized. Some users admire Musk and his takeover, regarding it as crucial for free speech [12]; others have expressed concerns over increased misinformation and toxicity.

Regardless of one's stance, it is undeniable that the acquisition has led to a series of noteworthy events. On November 04, Musk fired half of the 7,500 employees previously working at Twitter. Two weeks later (November 17), hundreds of employees resigned in response to an ultimatum to commit to "extremely hardcore" work or leave. These events and the associated public backlash, prompted many users to search for alternatives. For context, Figure 1a presents a time series of Google trend search interest for "Twitter alternatives". We observe a large spike on October 28, the day after Musk's takeover. Similarly, Figure 1b shows equivalent search interest for other popular alternatives to Twitter, *e.g.* Koo (an Indian micro-blogging service), and Hive (a micro-blogging service that permits "Not Safe For Work" mature content). One platform that stands out as being particularly prominent is *Mastodon*, a decentralized micro-blogging platform. Anecdotally, Mastodon has gathered significant attention since October 2022. Indeed, on November 12, Mastodon announced that over 1 million users had registered in the prior two weeks [20]. This is confirmed in Figure 2, which shows the weekly number of registrations, logins and statuses during the period.[1] As expected, we see a large increase in all three activity metrics after the Twitter acquisition.

Mastodon is part of the wider *fediverse*, in which any person can create and operate their own Mastodon server (aka "instance"). Each Mastodon instance operates as an independent micro-blogging service, where users can create local accounts and enjoy similar functions to Twitter (*e.g.* posting, following). Importantly, these

[1] This was gathered from 2,879 Mastodon server's Weekly Activity Endpoint.

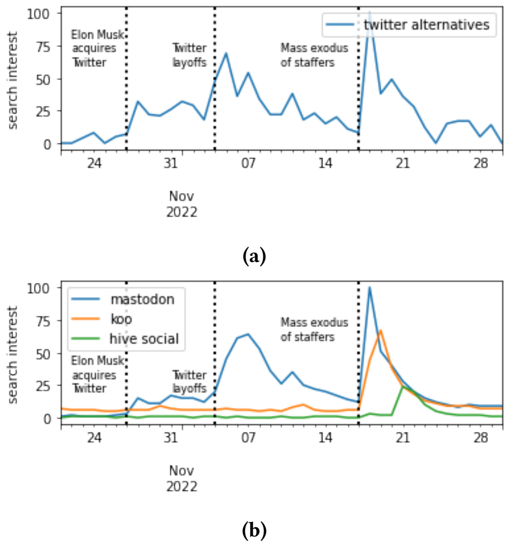


Figure 1: Interest over time for the search terms (a) Twitter alternatives and (b) Mastodon, Koo & Hive Social.

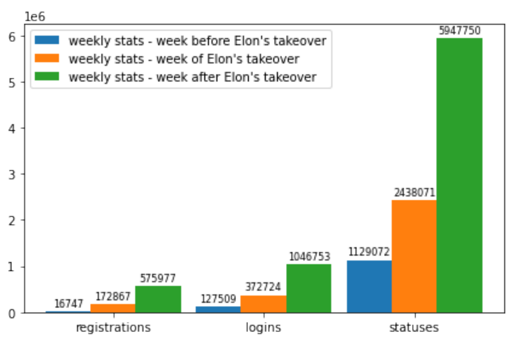


Figure 2: Weekly activity on Mastodon instances.

instances can also federate together, allowing users on one instance to follow users on another. This means that Mastodon operates in a decentralized fashion (with people joining independent instances), while retaining the ability to interact across the entire globe. This new paradigm has attracted significant attention and has made it an obvious candidate for users who are unhappy with Musk's acquisition (and the associated centralization of power in the hands of one individual).

We argue that this sudden interest in Mastodon offers a unique opportunity to study the migration of users between social networks. This is particularly the case due to the differing value propositions of the two platforms, with clear contrasts in the governance and ownership of Twitter vs. Mastodon. The unusual circumstances of the migration create further dimensions of analysis. Based on a dataset covering 136,009 Twitter users who move to Mastodon, we explore the migration from several angles. We start by inspecting how users are spread across instances, and to what extend this has a decentralizing (or centralizing) effect on Mastodon. We then explore the factors that impact the choice of migration, focusing on followee network influence. Following this, we investigate the behavioural patterns of users across the two platforms. Finally, we explore if it possible to predict users who might migrate, and exploit these results to quantify the key factors that can predict a user's choice to migrate. Our main findings include:

- The migration has created a user-driven pressure towards centralization on Mastodon. The top 25% most populous instances contain 96% of the users. Interestingly, this pressure is counterbalanced by the greater activity of users on smaller server instances: On average, users on single-user instances post 121% more statuses than users on larger instances.
- The social network of users on Twitter influences their choice of server instance on Mastodon. For example, for a migrated user, on average, 45.76% of user's Twitter followees have migrated to Mastodon before the user. Another unique feature of Mastodon is the ability for users to switch their account to a different instance. 4.09% of users later change their instance, often moving to another instance chosen by a large fraction (46.98% on average) of their Twitter followees.
- Users tend to post different topics across the two platforms. Only 1.53% of a user's Mastodon posts are identical to their Twitter posts. 33.6% of users do not have any overlap of hashtags on both platforms and 30.9% of users do not have any overlap of topics. Users have more diverse discussions on Twitter, while discussions about Fediverse and Migration are dominant on Mastodon during our measurement period.
- User migration is predictable (we attain an F1 score of 0.892). Unsurprisingly, the most predictive features are the content of user tweets, with common terminology, phrases and hashtags highly predictive of a user's likelihood to migrate. Further, the average number of URLs, the average number of likes received, and the average length of tweets correlates with the probability of migration.

2 MASTODON PRIMER

Mastodon is an open-source [19] federated server platform released in 2016. It offers micro-blogging functionality, allowing administrators to create their own independent Mastodon servers, aka **instances**. Each unique Mastodon instance works much like Twitter, allowing users to register new accounts and share statuses with their followers – equivalent to tweeting on Twitter. Users can also **boost** others' statuses – equivalent to retweeting on Twitter.

Instances can work in isolation, only allowing locally registered users to follow each other. However, Mastodon instances can also **federate**, whereby users registered on one instance can follow users registered on another instance. This results in the instance **subscribing** to posts performed on the remote instance, such that they can be pushed across and presented to local users. For simplicity, we refer to users registered on the same instance as **local**, and users registered on different instances as **remote**. Note that a user registered on their local instance does *not* need to register with the remote instance to follow the remote user. Instead, a user just creates a single account with their local instance; when the user wants to follow a user on a remote instance, the user's local instance performs the subscription on the user's behalf. This process is implemented using an underlying subscription protocol,

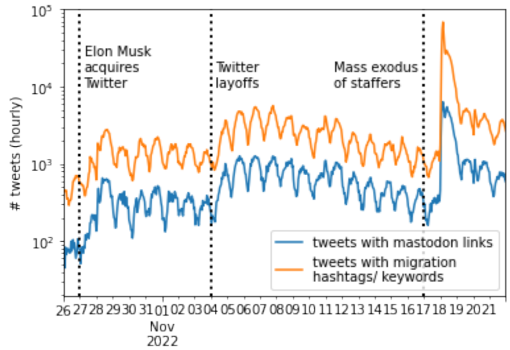


Figure 3: Temporal distribution of tweets containing (i) links to Mastodon instances and (ii) migration related keywords/ hashtags.

ActivityPub [1]. This makes Mastodon compatible with other decentralized micro-blogging implementations (notably, Pleroma). The **Fediverse**, refers to the growing group of ActivityPub compatible, and therefore interconnected, applications.

When a user logs in to their local instance, they are presented with three timelines: (*i*) a *home* timeline, with statuses shared by the accounts whom the user follows; (*ii*) a *local* timeline, listing the statuses generated within the same instance; and (*iii*) a *federated* timeline, with *all* statuses that have been retrieved from remote instances. The latter is not limited to remote statuses that the user follows; rather, it is the union of remote statuses retrieved by all users on the instance.

3 DATA COLLECTION

3.1 Mastodon Accounts from Twitter

We use an indexing website, `instances.social`, to identify a comprehensive list of Mastodon instances. We compile a set of 15,886 unique instances from this site. We then collect all tweets containing a link to any of these Mastodon instances using Twitter's Search API.[2] Additionally, we collect all tweets containing the following keywords related to the migration: 'mastodon', 'bye bye twitter', 'good bye twitter'; and hashtags #Mastodon, #MastodonMigration, #ByeByeTwitter, #GoodByeTwitter, #TwitterMigration, #MastodonSocial, #RIPTwitter. In total, we collect 2,090,940 tweets posted by 1,024,577 users between October 26, 2022 (*i.e.* a day before Musk's takeover) and November 21, 2022. Figure 3 shows the temporal distribution of these tweets.

We next retrieve the account metadata of any user that posted a tweet (*i.e.* display name, location, description, URLs, pinned tweet text). To map a Mastodon handle to a Twitter account, we perform a two-stage process: (*i*) We first search for Mastodon usernames in this user metadata (*e.g.* bio) and create a mapping between the Twitter account & Mastodon account if one is found. (*ii*) If the first step is unsuccessful, we then look for Mastodon usernames in each user's tweet text. If we find a tweet containing a Mastodon handle, we check if the handle String matches the Twitter handle exactly. If so, we label the user as migrated. We also collect information of what Mastodon instance a user has migrated to and whether the user has changed the instance they are registered with, as per the metadata in the user's Mastodon profile.

Using this methodology, we identify the Mastodon accounts of 136,009 Twitter users, which are registered across 2,879 unique Mastodon instances. We find that 72% of Twitter users that migrated created a Mastodon account with the same username that they use on Twitter. 4% of the Twitter users who create a Mastodon account, have a (legacy) verified status (*i.e.* authentic, notable, and active) on Twitter, suggesting that even well-established users have been migrating. To further verify the accuracy of the user mapping, we randomly select 100 users from the migrated users and manually check the user's information on the two platforms, as well as their following lists, to determine if they are the same user. The results show that among the 100 random users, 4 users are logged out or are inaccessible, and the remaining 96 are confidently the same user.

Of course, it is important to note that there are some limitations to our approach. Naturally, trying to collect definitions for all migrated users is an impossible task. We are unable to collect users who use different usernames on the two platforms. We also cannot detect users who do not mention their Mastodon account on Twitter. Although we miss some of these migrated users, our goal is to analyze the behavioral patterns of users before and after migration and their impact on the platform. Thus, we have focused on building a robust ground truth, rather than relaxing the matching criteria and introducing false positives.

3.2 Twitter and Mastodon Timelines

We next crawl both the Twitter and Mastodon timelines of the migrated users identified in the previous section. We use Twitter's Search API and Mastodon's Account's Statuses Endpoint.[3] For each user, we crawl all their tweets/statuses from October 01, 2022 to November 30, 2022. In total, we gather Twitter timelines for 94.88% of the users. The rest were suspended (0.08%), deleted/deactivated (2.26%), or the tweets were protected (2.78%). We also crawl the timelines of 79.22% of Mastodon users: the rest have either not posted a single status (9.20%) or their instances were down at the time of crawl (11.58%). In total, we gather 16,163,600 tweets, and 5,746,052 Mastodon statuses.

3.3 Followees

We crawl the user's followees for both Twitter and Mastodon accounts. We use the Twitter Follows API[4] and the Mastodon Account's Following Endpoint[5] respectively. Due to the rate limitations of the Twitter API, we crawl a sub-sample of 10% of the migrated users. For representativity, our sample relies on the followees distribution: we first calculate the median value of followees for migrated users, then we randomly select 5% from above the median value and 5% from below. In total, we gather followee data for 13,068 users. This covers 11,453,484 followee relationships.

[2]https://api.twitter.com/2/tweets/search/all

[3]https://docs.joinmastodon.org/methods/accounts/#statuses

[4]https://api.twitter.com/2/users/:id/following

[5]https://docs.joinmastodon.org/methods/accounts/#following

3.4 Toxicity

We also label all tweets and statuses using Google's Perspective API.[6] For a given post, Perspective returns a score between 0 and 1 for its toxicity (0 = non-toxic). We use the API's TOXICITY attribute that defines toxicity as "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion". Following the suggestion of the Perspective API official website,[7] we label any post with a score above 0.8 as toxic.

3.5 Non-Migrated User Baseline

Finally, as a baseline, we collect non-migrated user timeline data. We download tweets posted between October 26, 2022 and November 21, 2022 from the Internet Archive. We then filter out all users who mentioned Mastodon related keywords and Mastodon handle, and randomly select 136,009 users (the same number of users as migrated users) from the remaining tweets as the non-migrated user set. This non-migrated user set contains a total of 12,479,975 tweets. Note that as of January 2023, Twitter has 556 million users and Mastodon has less than 6 million users. This means that migration is the behavior of a small number of users, so we consider it feasible that no mention of Mastodon related content in that time period equals no migration.

3.6 Ethical Considerations

The datasets include both user and post information, and therefore might have privacy implications. To overcome any data mishandling, we exclusively collect publicly available data following well-established ethical procedures for social data [26, 27]. We have obtained a waiver from the ethics committee at the author's institution.

4 MIGRATION EFFECTS

4.1 The Centralization Paradox

Mastodon is a decentralized platform, in which users are spread across thousands of independent instances. We first test if the migration has reinforced decentralization or if, paradoxically the majority of users have migrated to a small number of servers. Overall, the Twitter users in our dataset migrate to 2,879 unique Mastodon instances. Figure 4 presents a histogram of the number of users in the top 30 Mastodon instances divided by whether they join before (blue) or after (orange) the Twitter acquisition. We find that, many Mastodon accounts (21%) advertised on Twitter in response to Elon Musk's acquisition predate the acquisition. Despite Mastodon's decentralization efforts, we observe a clear centralization trend with a large number of users migrating to a narrow subset of instances. For instance, mastodon.social, a flagship Mastodon instance operated by Mastodon gGmbH, receives the largest fraction (20.2%) of migrated Twitter users. 96% of users join the top 25% of largest instances (w.r.t. number of users). We show this in Figure 5, where we depict the distribution of users across the 25% largest instances.

We conjecture that users join large and well-known Mastodon instances because it helps them build larger social networks (*i.e.* number of followers & followees). We therefore examine the relationship between instance size and the social network of users on Mastodon. We analyze all the users who join Mastodon after the Twitter acquisition and have at least a 30 day old Mastodon account (to ensure a fair comparison). This covers 50.59% of all migrated users. We also divide the instances into quantiles according to their number of users.

Figure 6 presents the distribution of instances by size and CDFs of the number of followers, followees, and statuses of users on different-sized instances. Contrary to our hypothesis, users in the bigger instances tend to have smaller social networks. We find that 13.16% of instances have just one user. Paradoxically, the users of these single-user instances tends to have more followers, followees, and statuses than the users of larger instances. Users in smaller instances (lower quartile) have, on average, 19.19% more followers, 74.98% more followings, and 103.43% more statuses compared to those in bigger instances. This difference is even more significant for single-user instances, where users have 184.81% more followers, 99.04% more followings, and 121.14% more statuses than those in

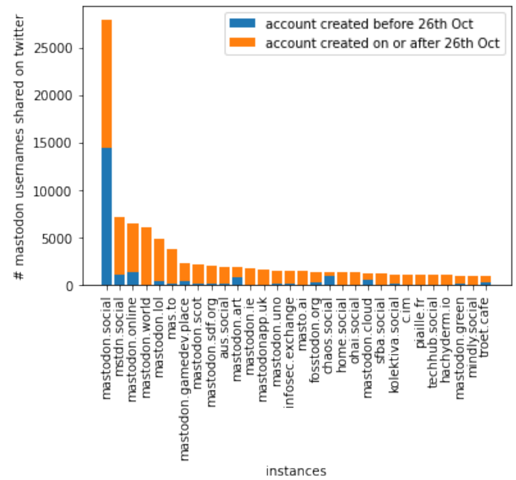


Figure 4: Top 30 Mastodon instances Twitter users migrated to.

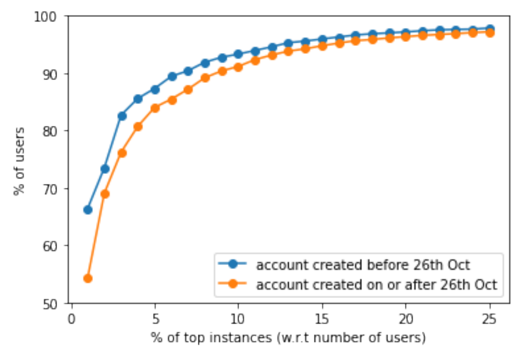


Figure 5: Percentage of users on largest 25% instances (w.r.t number of users).

[6]https://www.perspectiveapi.com
[7]https://developers.perspectiveapi.com/s/about-the-api-score?language=en_US

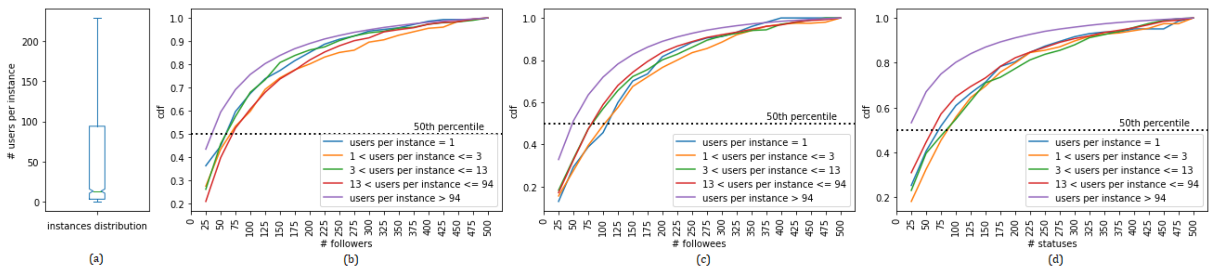


Figure 6: (a) Distribution of instances w.r.t. to the number of users. (b) CDF of number of followers of users on different-sized instances. (c) CDF of number of followees of users on different-sized instances. (d) CDF of number of statuses of users on different-sized instances.

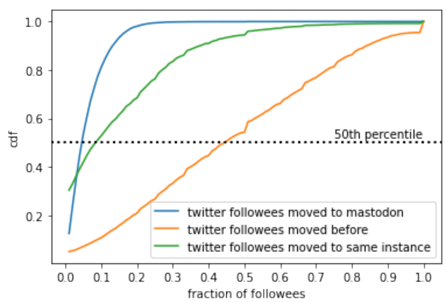


Figure 7: CDFs of the fraction of Twitter followees of each migrated user that (i) moved to Mastodon (blue) (ii) moved to Mastodon before the user (orange) and (iii) moved to the same instances on Mastodon as the user (green).

larger instances. This is because any user who sets up their own single-user instance is likely very dedicated. Indeed, some users create their own instance to give themselves greater control over the environment. Naturally, such users are more active.

4.2 Social Network Influence

There are at least two possible (non-exclusive) reasons for platform migration from Twitter to Mastodon, particularly after the Musk takeover: (*i*) A user might have decided to move for ideological reasons, *e.g.* when they disagree with Musk's actions after he gained control of Twitter; and (*ii*) A user might have decided to move because a large fraction of the accounts they follow move (and therefore Mastodon has become a relevant platform for them). Following on from this, we explore the impact that a user's social network has on their probability of migration.

To inspect this, we analyze the followees data from both Twitter and Mastodon for 10% of the migrated users (see §3.3). Figure 7 shows CDFs of the fraction of Twitter followees of each migrated user that (*i*) moved to Mastodon (blue); (*ii*) moved to Mastodon before the user (orange); and (*iii*) moved to the same Mastodon instances as the user (green). We notice that just 5.99% of each user's followees also migrate (on average). In fact, for 3.94% of the migrated users, none of their Twitter followees move to Mastodon. Thus, the majority of the social network of the migrated users seems reluctant to migrate, and sometimes they are the first in taking this step.

To better understand this, we compare the date on which each migrated user joined Mastodon with that of their Twitter followees who migrated as well. We find that, out of their followees, 4.98% of our migrated users were the first, and 4.58% were the last to migrate from Twitter to Mastodon. On average, 45.76% of the followees of a user migrated to Mastodon before the user actually did.

We also want to understand if users select the same Mastodon instance as their migrated social network. We therefore compare the instance of each migrated user with that of its migrated Twitter followees. On average, 14.72% of each migrated user's followees (that move to Mastodon) join the same instance. With 15K+ Mastodon instances, this is a considerable proportion, suggesting a clear network effect.

We also notice that these patterns are heavily impacted by one flagship instance: mastodon.social. This is the largest instance available, and is probably the best known. Of all the migrated users whose Twitter followees move to the same instance, 30.68% are on mastodon.social. That said, we also find small instances that attract significant proportions of a given user's Twitter followers. For example, 4.5% of the migrated users whose Twitter followees join them on the same instance are on mastodon.gamedev.place (a Mastodon instance focused on game development and related topics)

4.3 Instance Switching

A unique feature of Mastodon is that users can easily 'switch' instance. This involves migrating their data from one instance to another. We are curious to see if this is also driven by network effects. Overall, 4.09% of the users have switched from the Mastodon instance they initially created an account on (hereinafter first instance) to a new instance (hereinafter second instance). Curiously, 97.22% of these switches happened after Musk's Twitter takeover. This suggests that some users have backtracked on their instance choices, perhaps after finding a more suitable one.

Figure 8 shows the chord plot of switches from each user's first Mastodon instance to their second. A common pattern across these switches is that users move from general purpose/ flagship instances (*e.g.* `mastodon.social`, `mastodon.online`) to more topic specific instances, *e.g.* `sigmoid.social` (a Mastodon instance for people researching in Artificial Intelligence),

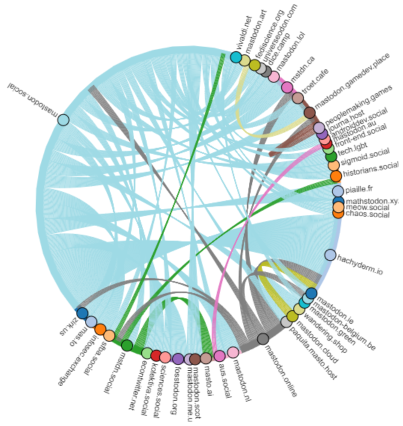


Figure 8: Chord plot of switching within Mastodon instances.

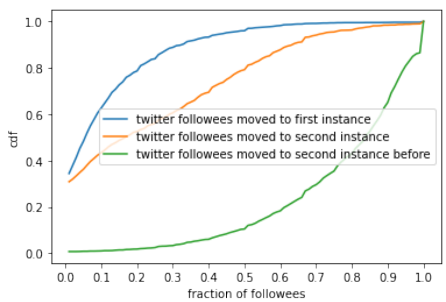


Figure 9: CDFs of the fraction of Twitter followees of each switched user that (i) moved to first instance (blue) (ii) moved to second instance (orange) and (iii) moved to second instance before the user (green).

historians.social (a Mastodon server for people interested in history) and infosec.exchange (a Mastodon instance for info/cyber security-minded people).

We notice a strong social network influence behind this switching. Figure 9 shows the CDFs of the fraction of Twitter followees of each switched user that (*i*) moved to the first instance (blue); (*ii*) moved to the second instance (orange); and (*iii*) moved to second instance before the user (green). On average, 46.98% of each user's followees (who moved to Mastodon) at some point also join the second instance. This is in contrast to just 11.4% who join the first instance. Interestingly, 77.42% of each switching user's followees (on average) join the second instance before the user. This suggests that the users switched from the first instance because a large fraction of their Twitter followees moved to the second one. This also suggest that switching is a popular function, and something that may be attractive for other social network implementations.

5 TIMELINES ANALYSIS

We are next curious to understand how people use their (two) accounts after migration.

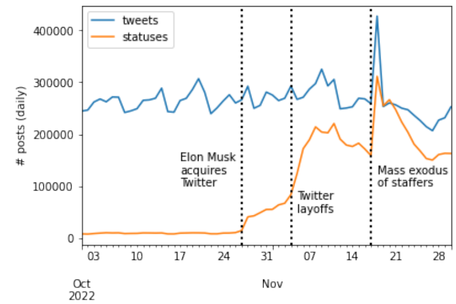


Figure 10: Temporal distribution of tweets and statuses posted by migrated users on Twitter and Mastodon respectively.

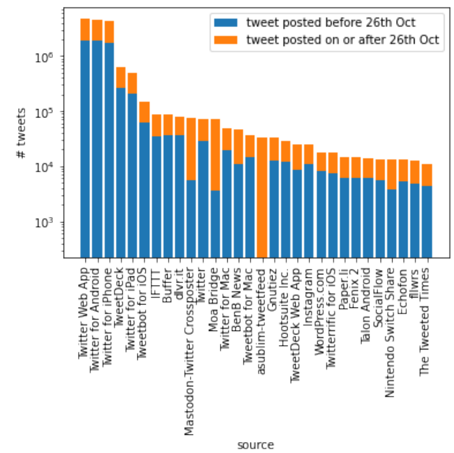


Figure 11: Top 30 sources of tweets. Note the log scale on the y-axis.

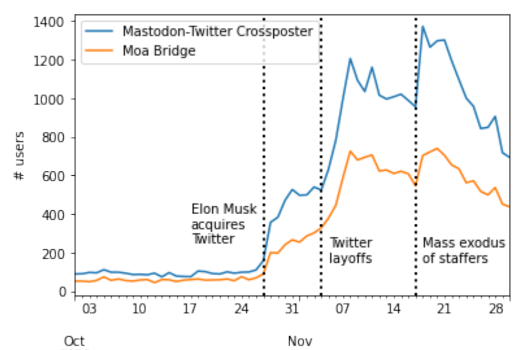


Figure 12: Number of users that use cross-posting tools daily.

5.1 Twitter vs. Mastodon Activity

We first analyze the timelines of migrated users from both Twitter and Mastodon. Figure 10 shows the number of tweets on Twitter and the number of statuses on Mastodon posted by migrated users

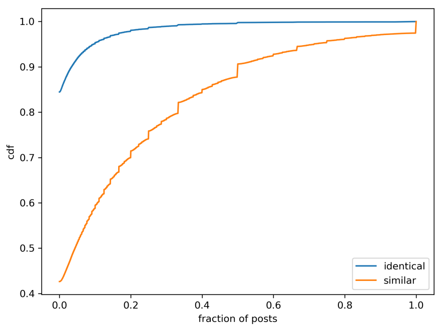


Figure 13: CDFs of fraction of each migrated user's Mastodon statuses that are identical or similar to its tweets.

each day in our dataset, *i.e.* from October 01 to November 30, 2022. We observe a significant growth in user activity on Mastodon after the acquisition of Twitter. However, the activity of migrated users on Twitter does not decrease in parallel, *i.e.* our migrated users use both their Twitter and Mastodon accounts simultaneously.

We conjecture that such users may simply be replicating identical or similar content across both platforms. Figure 13 plots the CDFs of the fraction of each migrated user's Mastodon statuses that are identical or similar to its tweets. We consider the Mastodon status similar to a tweet if the cosine-similarity of their sentence embeddings [23] is greater than 0.7. The similarity threshold is usually subjectively determined and varies depending on the specific task, usually between 0.5 and 0.9, so we chose a middle value of 0.7 for this task. On average, just 16.57% of each user's Mastodon statues are similar to their tweets. Just 1.53% of each migrated user's Mastodon statuses are identical, and 84.45% of the migrated users post completely different content in the two platforms. This suggests a mix of users, some of whom create different personas on the two platforms, and a smaller subset who mirror all their content.

A potential explanation for the latter is the use of cross-posting tools. Such tools allow users to automatically mirror their Mastodon statues on Twitter, and vice versa. To examine this, we compare the number of tweets posted via different sources before and after Musk's takeover in Figure 11. Naturally, the majority are posted by official Twitter clients such as the Twitter Web App. The two sources that increase most dramatically, however, are two well-known cross-posters, Mastodon-Twitter Crossposter and Moa Bridge — by 1128.95% and 1732.26%, respectively. Of all migrated users, 5.73% use one of the two cross-posters at least once. It is worth noting that asublim-tweetfeed also shows significant growth, but it is not a cross-poster. This suggests that such users see both Twitter and Mastodon as viable platforms, and have limited intention of creating multiple 'personas'. For completeness, Figure 12 also plots the number of users using cross-posters over time. We see that their usage increases rapidly after Musk's takeover. The downward trend towards the end of November is likely a result of the posting issues that cross-posters faced after their posting rate limit was revoked by Twitter [25].

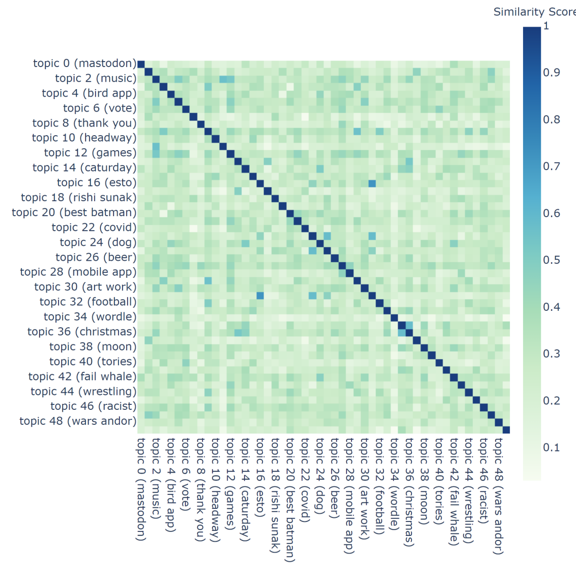


Figure 14: Similarity heat map of top 50 topics.

5.2 Content Analysis

To better understand the items of discussion, we next inspect the content of posts across both Twitter and Mastodon. Our goal is to understand how users utilize these platforms in parallel (if at all).

5.2.1 Methodology. We first extract all English language posts (60.0% of tweets, and 67.7% of Mastodon statuses). Note, the language code is part of the metadata. For the remaining languages, the top three are German, Spanish and French, accounting for 9.1%, 6.6% and 3.2% of all posts, respectively. We further filter all links and mentions from the text. Next, we use BERTopic [10] to train a topic model on the filtered text, which contain 4,558,369 Twitter tweets and 2,568,712 Mastodon statues. BERTopic employs sentence-transformers [24] and c-TF-IDF to generate compact clusters of information, facilitating the interpretation of topics while retaining significant words within the topic descriptions. The model also identifies some "outlier" topics, which are not considered in our analysis. A topic is considered an outlier if: (*i*) the topic has fewer than 1000 posts (this threshold is recommended by BERTopic); or (*ii*) the text does not have clear topical coverage (*e.g. "Quite a statement."*, *"Oh Lord. So apt."*).

We identify 492 topics, with 52.75% of all posts (*i.e.* tweets and statuses) classified as outlier topics as shown in Table §1. For the rest of our analysis we discard all the outlier topics. Of the remaining topics, 67 contain more than 10,000 posts, suggesting that a majority of the topics identified by BERTopic may represent more granular or niche areas of interest within the Twitter user base.

To check the semantic relationship between the topics, we briefly examine the cosine similarity calculated by BERTopic. We visualize this in Figure 14, where we show the cosine similarity between the top 50 topic embeddings. The average similarity between all themes is just 0.284, indicating that the correlation between different topics is low and there is little overlap in our topic modeling results. This confirms that the model has effectively segmented the textual

Topic	Overall Percentage		Representative words
	Twitter	Mastodon	
Outlier	34.58%	18.17%	twitter, much, me, don, your, do, you, about, really, will
0	2.22%	2.11%	mastodon, about mastodon, mastodon account, mastodon instance
1	0.82%	0.89%	twitter, tweet, riptwitter
2	1.07%	0.45%	music, songs, spotify, listening, soundtrack, playlist
3	1.01%	0.33%	ukraine war, ukraine, ukrainerussiawar, ukrainian
4	0.68%	0.53%	bird app, bird site, bird, birdsite, owl, app
5	0.66%	0.52%	book, novel, reading
6	0.88%	0.23%	vote, ballot, election day, polling, election, polls
7	0.68%	0.20%	elon musk, twitter musk, musk
8	0.20%	0.59%	thank you, thanks, you welcome, welcome thank
9	0.47%	0.16%	auspol, australia, sydney, melbourne, australian, korea, itaewon, seoul
10	0.39%	0.18%	headway slippage, buses, headway, cancellations announced, slippage
11	0.46%	0.10%	bbc6music, bbcradio2
12	0.29%	0.21%	games, boardgames, video game, playing
13	0.34%	0.16%	snow, snowing, first snow, snowfall, winter
14	0.23%	0.23%	caturday, catsofmastodon, happy caturday

Table 1: Top 15 topics distribution.

content. Finally, to accompany the topic analysis, we extract the 4,433,424 hashtags from all posts by migrated users from October 26, 2022 to November 21, 2022 (2,726,551 from Twitter, 1,706,873 from Mastodon). We next inspect the discussion content across platforms, individual users. and instances.

5.2.2 Cross Platform Analysis. We first inspect the hashtags and BERTopic topics discussed across both platforms. Figure 15 (a) and (b) shows the daily frequency of the top 30 most common hashtags on Twitter and Mastodon respectively while Figure 15 (c) shows the overall frequency of the top 30 hashtags on both platforms. Table §1 also lists the top 15 most frequent topics on Twitter and Mastodon with the percentages of tweets and statuses associated to each topic in the respective platform. We observe a range of hashtags and topical discussions, which vary across the measurement period. The migration seems to be the most common discussion in our dataset, partially due to our sampling approach. The most popular hashtag is `#mastodon`, appearing 59,367 times, and comprising 1.34% of all posts across both platforms. We observe a similar pattern in the topics where those related to Mastodon (`topic 0`) and Twitter (`topic 1`) are the most popular among migrated users, accounting for 4.33% and 1.71% of all posts across both platforms. The discussion of the migration across both platforms shows strong similarities. There is a similar percentages of posts associated to these topics on both platforms (*i.e.* `topic 0` and `1`), and similar daily trends in the use of `#mastodon`, `#twitter` and `#riptwitter` hashtags, as shown in Figure 15. We observe a substantial increase in these migration related topics and hashtags after three key time points: (*i*) Elon Musk acquires Twitter, (*ii*) Twitter layoffs, and (*iii*) Mass exodus of Twitter staffers.

To better understand the similarity of the topics discussed across both platforms, in Figure 16 we calculate the distance between all topics and plot them on a 2D map. The size of each circle represents the frequency of that topic and the color represents whether its frequency is higher on Twitter (dark blue) or Mastodon (bright orange). We observe a clear clustering of topics. In the large clusters, there is a blend of topics dominated by Twitter and Mastodon, respectively. In the small clusters, most of them are dominated by Twitter, while a small portion is dominated by Mastodon, *e.g.* `topic 28` (mobile apps), `topic 53` (toots), `topic 211` (cross-poster), etc. This indicates that the topics that dominate on Twitter tend to be general topics, whereas those that dominate on Mastodon tend to be related to Mastodon and how it works. This disparity in hashtag and topic distribution can potentially be attributed to the thematic nature of many Mastodon instances. For example, 40.63% of posts containing arts-related hashtags originated from art-themed instances such as `mastodon.art` and `photog.social`.

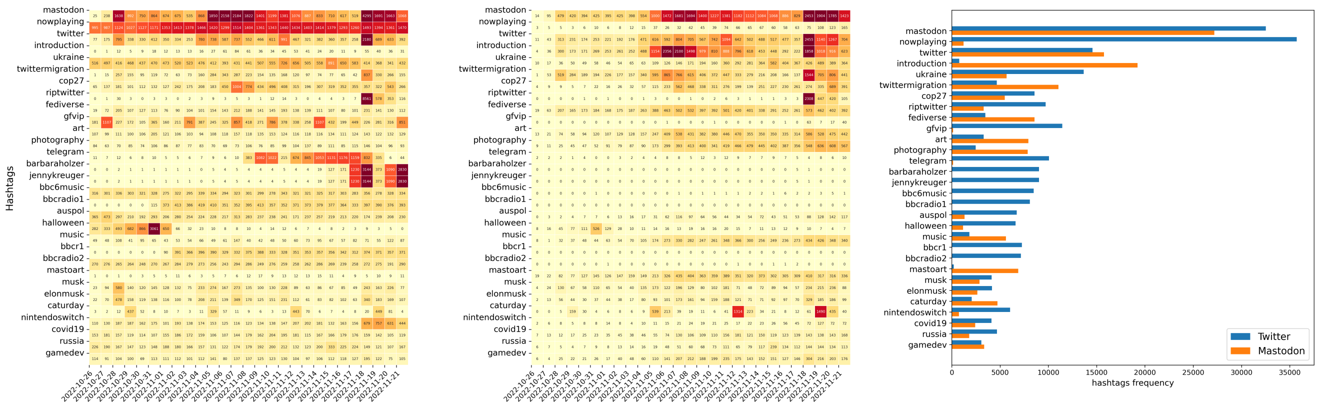
5.2.3 Per-User Analysis. We now focus on the user level. We hypothesize that migrated users may bring some of their topical preferences from Twitter to Mastodon. To confirm this, we calculate for each user how many hashtags and topics are overlapping on both platforms and how many are unique on one of the platforms. Figure 17 presents the CDF of the hashtags overlap and their uniqueness. We perform the same analysis for the topics: for each user, we calculate the percentage of their topics on Mastodon that they also discuss on Twitter. Figure 18 presents the distribution per user with outlier topic (solid line) and without outlier topic (dashed line). We focus on the latter, as including the outlier topics naturally overestimates the overlap.

We observe a diversity of user behaviors. 33.6% of users do not share *any* of the same hashtags on Mastodon as they do on Twitter, suggesting a distinct discourse across platforms. That said, 31.6% of users have hashtags that overlap by more than half on both platforms, and 10.3% of users use exactly the same hashtags across both Mastodon and Twitter. Similarly, we also see that 30.9% of users have no topic overlap across platforms. That said, 28.4% of users in Mastodon have, each, more than half of the topics common across both platforms.

The above reveals a small subset of users with extremely similar timelines on both their Mastodon and Twitter accounts: 11.4% of users have at least 90% overlapping topics on both platforms. For these users with similarity higher than 90%, we explore the content of their posts further. Mastodon related topics (`topic 0`) still accounts for the largest share (3.8%). Further, `topic 2` (music) and `topic 3` (Russia-Ukraine War) account for 0.70% and 0.35%, respectively. These findings suggest that, for these migrated users, they chose to bring popular Twitter topics to Mastodon for discussion.

5.2.4 Per-instance Analysis. We now examine whether users of different instances have different usage patterns. Recall that users must choose which specific Mastodon instance to register with, and instances are frequently theme-specific. For this purpose, we identify the 25 instances with the most users.[8] For users who migrated

[8]mastodon.social, mstdn.social, mastodon.online, mastodon.world, mastodon.lol, mas.to, mastodon,gamedev.place, mastodon.scot, mastodon.sdf.org, and mastodon.art, aus.social, mastodon.ie, mastodonapp.uk, mastodon.uno, infosec.exchange, masto.ai,



(a) (b) (c)

Figure 15: (a) Daily frequency of top 30 hashtags on Twitter. (b) Daily frequency of top 30 hashtags on Mastodon. (c) Top 30 hashtags distribution over two platforms.

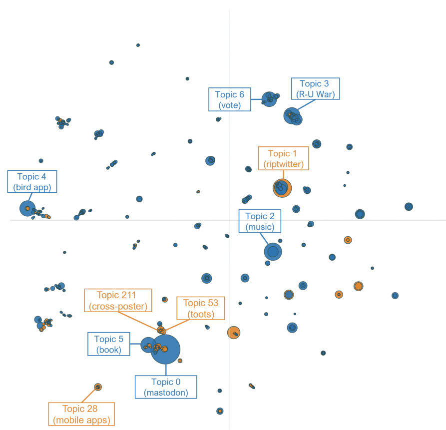


Figure 16: Inter-topic distance 2D map for all topics.

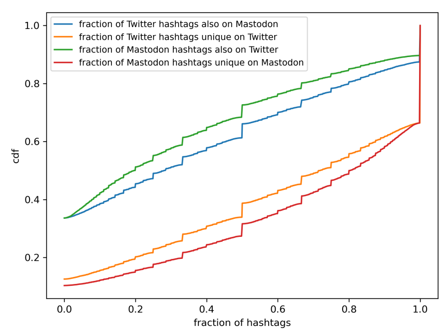


Figure 17: CDF of percentage of the hashtags of each migrated user that overlap across Twitter and Mastodon.

to these top 25 instances, we calculate the content similarity of their

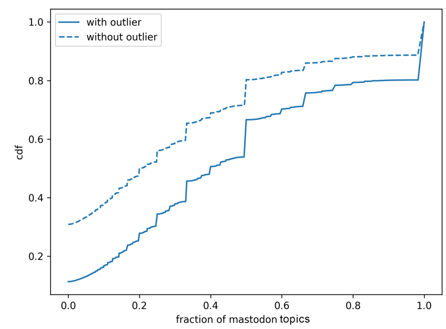


Figure 18: CDF of each migrated user's Mastodon topics that overlap to user's Twitter topics.

Mastodon timelines with their Twitter timelines. Similarly to Section 5.1,we consider posts to be similar when the cosine-similarity of their embeddings is greater than 0.7. Figure 19 presents the distribution of content similarity (upper) and topic overlap (lower) over two platforms for users who migrated to the top 25 largest instances. We observe similar distributions, but with clear outliers.

The average content similarity of all migrated users is 16.7%. But the average content similarity of all users migrating to the top 25 instances is 17.3%, suggesting that the users in larger instances are slightly more likely to have similar posts on Twitter and Mastodon. We observe the same phenomenon in the topic overlap: the average topic overlap rate is 30.55% and for users who migrated to the top 10 instances, the average topic overlap rate is 31.06%. This suggests that users who migrate to popular instances are more likely to discuss similar content on both platforms. In particular, users of `mastodon.social` have the highest mean values of content similarity and topic overlap, 22.9% and 36.8%, respectively. This is likely because `mastodon.social`, as the flagship instance, attracts many new users, who may have little experience with Mastodon. Therefore, it is intuitive that such experimental users may continue to post on Twitter with similar content.

---

fosstodon.org, chaos.social, home.social, ohai.social, mastodon.cloud, sfba.social, kolektiva.social, c.im, piaille.fr

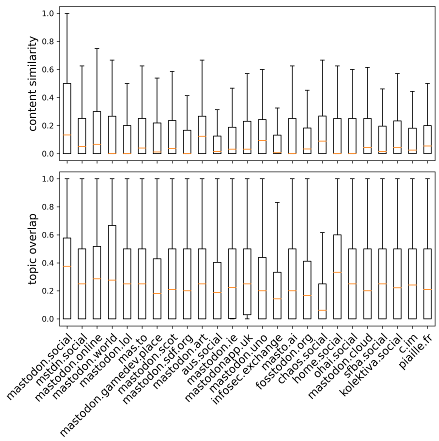


Figure 19: Box plots of content similarity (upper) and topic overlap (lower) over two platforms for users who migrated to the top 25 largest instances.

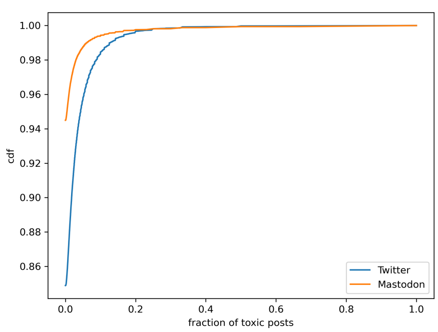


Figure 20: CDFs of each migrated user's toxic posts on Twitter and Mastodon.

5.2.5 Toxicity Analysis. We next briefly analyze content toxicity across both platforms (see Section 3.4). Note, this is motivated by anecdotal media coverage about moderation problems on Mastodon in the face of the new surge of users [4]. Figure 20 shows the CDFs of the percentage of each migrated user's toxic posts on Twitter and Mastodon. Overall, only 2.19% of Twitter tweets are toxic, Mastodon is substantially less toxic, with 0.19%. On average, each user posts 1.29% toxic tweets on Twitter vs. just 0.25% toxic statuses on Mastodon. Even though discussions on both platforms are non-toxic during our measurement period, we notice that 9.21% of migrated users post at least one toxic post on both platforms. While this may not be problematic for Twitter which has its own moderation team, it might present greater challenges for Mastodon, where volunteer administrators are responsible for content moderation.

6 MIGRATION PREDICTION

Finally, we quantify the factors that may predict if a Twitter user migrates to Mastodon. To identify these factors, we build a set of classifiers that strive to discriminate between users who migrate and users who do not (see Section 3.1 and Section 3.5).

6.1 Methodology

We extract 24 features for each user, including user statistics (*e.g.* number of Twitter followers and tweets) and text information (*e.g.* average length of text, number of URLs, number of hate words, sentence embedding). All features and their descriptions are in the Appendix. For the sentence embedding, we first filter all user tweets to remove stopwords, special symbols, emojis, etc. Then we use Word2Vec in *gensim* [22] to compute word embedding for each word in the filtered text, and aggregate the resulting vectors into a single vector representation of the entire text using max pooling. This technique retains the most important features and is immune to changes in the input sequence length. We use this method to construct embedding features for each user.

Next, we train several machine learning models with the sklearn library and use GridSearchCV to perform 5-fold cross-validation to find the optimal hyper-parameter settings. The task is to discriminate between migrated vs. non-migrated users using a users' Twitter timeline. Note, this only includes the users' Twitter timeline, *prior* to migration. The models we train are: (*i*) Logistic Regression (LR), (*ii*) Decision tree (DT), (*iii*) Random Forest (RF), (*iv*) K-nearest Neighbors (KNN) (*v*) Multilayer Perceptron (MLP), (*vi*) Naive Bayes (NB), and (*vii*) CatBoost (CB). Figure 21 (a) shows the F1 scores of the seven models on the task of predicting whether users migrate or not. We can confidently predict the likelihood of user migration with F1 scores of up to 0.892 for Random Forest, our best performing model.

6.2 Results

To quantify the most important factors, we next use our classifiers to understand the features that correlate with a users' decisions to migrate. To do this we extract feature importance for the three best performing models: Random Forest, Catboost and Decision Tree. The results are shown in Figure 21 (b) (c) and (d).

We find that the most important feature is the *sentence_embedding*, *i.e.* the content of the user tweets is the most predictive of their migration. This is perhaps intuitive, and we observe that the related topics such as discussing Mastodon are unsurprisingly key determinants in the classification. Among the migrated users, 75.31% mention keywords like "Mastodon", "Twitter migration" and "Bye Bye Twitter" with obvious migration intentions. Among the non-migrated users, only 0.72% of them mentioned these keywords. This is expected, as our experiences show that non-migrating users rarely talk about migration specifically. That said, we find other less clear-cut examples. Most notably, 65.44% of migrated users mention "RIP Twitter" and "#riptwitter"; yet, confusingly, 31.09% of the non-migrated users also mention them. One potential explanation is that "RIP Twitter" only indicates the user's opinion of Twitter and does not necessarily contain their intention to migrate. Another possible explanation is that the model is overfitted to identify keywords used in the original data collection. To explore the generality of our model, we also retrain the model with all migrated user users that *do not* mention any related keywords and hashtags from our data collection (33,580 in total) vs. the same number of randomly selected non-migrated users. The feature importance is shown in Figure 21 (e). Our best performing model is still Random Forest, with an F1 score of 0.778. Thus, although there

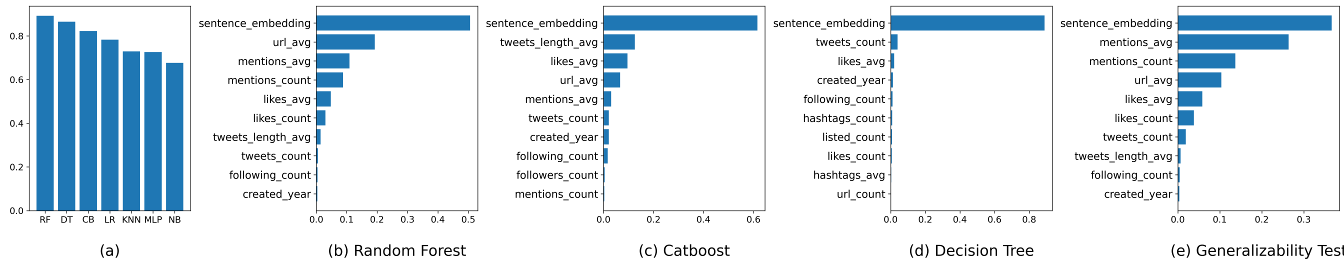


Figure 21: (a) F1 scores of the seven models on migration prediction. (b) Random Forest top 10 normalized feature importance. (c) Catboost top 10 normalized feature importance. (d) Decision Tree top 10 normalized feature importance. (e) Random Forest top 10 normalized feature importance without keyword/hashtag features.

is a decrease in performance, our high F1-score confirms that our classifiers are capable of discerning between these populations.

Briefly, we also observe several other features that possess predictive power. Features such as *url_avg*, *likes_avg*, and *tweets_length_avg* all play an important role. These features represent the average number of URLs of user posts (*url_avg*), the average number of likes received (*likes_avg*), and the average length of all tweets (*tweets_length_avg*). Therefore, we further examine these three features for both user groups. The average number of URLs for migrated users is 0.80 vs. 0.14 for non-migrated users. The average number of likes received by migrated users is 26.1, while for non-migrated users is 416.1, which is 16x higher. The average length of tweets is 129.2 for migrated users compared to 98.5 for non-migrated users. These results indicate that migrated users prefer to include URLs in their tweets and that the tweets are longer in length. In addition, non-migrated users receive far more likes than migrated users, which is perhaps to be expected. One potential explanation is that the high number of likes for non-migrated users may indicate that they have a higher influence and social network on Twitter. This may lead them to not want to leave Twitter for other platforms. We further argue that this tooling and insight is valuable for social media companies wishing to understand the departure of users.

7 RELATED WORK

Many previous efforts have been made to build decentralized online social platforms. In the earliest days, there were many peer-to-peer online social networks, such as Safebook [6], PeerSoN [5], LotusNet [2], and LifeSocial.KOM [9]. New federated social networks, such as Mastodon, Pleroma, and PeerTube, have since emerged. These platforms are collectively referred to as the *Fediverse* [18]. These social network applications use ActivityPub [1], a W3C protocol, to implement server federation.

Recent work has looked into these new federated social networks. Most related are two interesting studies that have explored the migration of users from Twitter to Mastodon since Musk's acquisition [12, 13]. An early study further explored centralization trends in Mastodon [21]. In contrast to us, they focused on infrastructural and resilience aspects of Mastodon. We found similar trends, but focus on the increased level of centralization driven by the migration. Another study [14] investigated the growth of Mastodon in its earlier years. We confirm that this growth has continued. Paradoxically, we also found that while centralization occurs in terms of how many users are attracted to an instance, smaller instances attract more active users. Other works focus on user behavior across instances [15, 16]. Our work also touches upon the need for decentralized moderation, but we show that Mastodon is less "toxic" than Twitter still. This has also been investigated in prior work on Pleroma (another Fediverse microblogging service). Hassan et al identify novel challenges [11] and propose a strawman solution. Zia et al. [4] also propose a model sharing solution to help automate moderation. Our work confirms the presence of toxic content in Mastodon, though the numbers identified do not show a trend towards greater toxicity than Twitter.

There have also been a number of studies of migration on other social networks. For example, [7] measured the migration activity of Fandom, tracking migrating users and the reasons behind their migration. The authors find that policy and value-based aspects are determinant in the migration. Gerhart et al. [8] analyze user migration from traditional social networks to anonymous social networks perspective. They identify that social norms drive migration. Zhong et al. [28] inspected the transfer of social network links across platforms. They show that transferring links from Facebook to Pinterest and last.fm improve their resilience of their social networks. Otala et al. [17] study the migration of Twitter users to Parler. The results show that, although Parler is not widely used, it has a significant impact on political polarization. Our work also studies the migration of Twitter users. However, to the best of our knowledge, it is the first to systematically measure and analyze the migration of users from a centralized platform like Twitter, to a decentralized equivalent.

8 CONCLUSION

In this paper, we have explored the migration of users from Twitter to Mastodon, prompted by Elon Musk's acquisition of Twitter. We have made several key observations. We found that 2.26% of users have completely left Twitter, deleting their accounts. Despite Mastodon's decentralized architecture, the 25% largest instance on Mastodon attract 96% of the users. Paradoxically, while larger instances attract more users, smaller ones attract more active users, reinforcing Mastodon's decentralization. We further observed the impact of the social network in migration, with an average of 14.72% of a user's Twitter followees migrating to the exact same Mastodon instance as the user. This led us to explore the topics that people discuss across both Mastodon and Twitter. We found that users tend to post *different* topics across the platforms. Only 1.53% of a user's Mastodon posts are identical to their Twitter posts. 33.6% of users do not have any overlap of hashtags on both platforms, and 30.9% of users do not have any overlap of topics. Hence, during our observation period, users have more diverse discussions on Twitter, while discussions about Fediverse and Migration are dominated on Mastodon. We note that this may differ across more extensive time periods. Finally, our findings inspired us to explore the prediction of user migration. Thus, we employed several classifiers, attaining an F1 score of 0.892. Using the feature importance of these models, we quantified the factors that predict if a user will migrate. We found that user tweets can effectively predict if a user will migrate.

There are a number of lines of future works. We would like to investigate more longitudinal trends to explore whether users retain their Mastodon accounts or return to Twitter. Unfortunately, this has proven difficult in the face of Elon Musk's shutdown of the free Twitter API. It would also be interesting to explore the longevity of these user-driven centralized Mastodon instances, and how their administrators operate them over longer time periods. We emphasize that this study provides the first step in studying the migration of Twitter users to alternative decentralized platforms. We hope that it will inspire further exploration of follow-up work.

ACKNOWLEDGEMENTS

This work was supported by EPSRC SODESTREAM (EP/S033564/1), AP4L (EP/W032473/1), REPHRAIN's "Moderation in Decentralised Social Networks" (EP/V011189/1), and EU Horizon Framework grant agreement 101093006 (TaRDIS).

REFERENCES

[1] ActivityPub. 2018. https://www.w3.org/TR/activitypub/. (2018).
[2] Luca Maria Aiello and Giancarlo Ruffo. 2012. LotusNet: Tunable privacy for distributed online social network services. *Computer Communications* 35, 1 (2012), 75–88.
[3] Bond Benton, Jin-A Choi, Yi Luo, and Keith Green. 2022. Hate speech spikes on twitter after elon musk acquires the platform. *School of Communication and Media, Montclair State University* (2022).
[4] Haris Bin Zia, Aravindh Raman, Ignacio Castro, Ishaku Hassan Anaobi, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2022. Toxicity in the Decentralized Web and the Potential for Model Sharing. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 6, 2 (2022), 1–25.
[5] Sonja Buchegger, Doris Schiöberg, Le-Hung Vu, and Anwitaman Datta. 2009. PeerSoN: P2P social networking: early experiences and insights. In *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*. 46–52.
[6] Leucio Antonio Cutillo, Refik Molva, and Thorsten Strufe. 2009. Safebook: A privacy-preserving online social network leveraging on real-life trust. *IEEE Communications Magazine* 47, 12 (2009), 94–101.
[7] Casey Fiesler and Brianna Dym. 2020. Moving across lands: Online platform migration in fandom communities. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–25.
[8] Natalie Gerhart and Mehrdad Koohikamali. 2019. Social network migration and anonymity expectations: What anonymous social network apps offer. *Computers in Human Behavior* 95 (2019), 101–113.
[9] Kalman Graffi, Christian Gross, Dominik Stingl, Daniel Hartung, Aleksandra Kovacevic, and Ralf Steinmetz. 2011. LifeSocial. KOM: A secure and P2P-based solution for online social networks. In *2011 IEEE Consumer Communications and Networking Conference (CCNC)*. IEEE, 554–558.
[10] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).
[11] Anaobi Ishaku Hassan, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2021. Exploring content moderation in the decentralised web: The pleroma case. In *Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies*. 328–335.
[12] Ujun Jeong, Paras Sheth, Anique Tahir, Faisal Alatawi, H Russell Bernard, and Huan Liu. 2023. Exploring platform migration patterns between twitter and mastodon: A user behavior study. *arXiv preprint arXiv:2305.09196* (2023).
[13] Lucio La Cava, Luca Maria Aiello, and Andrea Tagarelli. 2023. Get Out of the Nest! Drivers of Social Influence in the# TwitterMigration to Mastodon. *arXiv preprint arXiv:2305.19056* (2023).
[14] Lucio La Cava, Sergio Greco, and Andrea Tagarelli. 2021. Understanding the growth of the Fediverse through the lens of Mastodon. *Applied Network Science* 6 (2021), 1–35.
[15] Lucio La Cava, Sergio Greco, and Andrea Tagarelli. 2022. Information consumption and boundary spanning in decentralized online social networks: the case of mastodon users. *Online Social Networks and Media* 30 (2022), 100220.
[16] Lucio La Cava, Sergio Greco, and Andrea Tagarelli. 2022. Network analysis of the information consumption-production dichotomy in mastodon user behaviors. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1378–1382.
[17] Jacqueline M. Otala, Gillian Kurtic, Isabella Grasso, Yu Liu, Jeanna Matthews, and Golshan Madraki. 2021. Political polarization and platform migration: a study of Parler and Twitter usage by United States of America Congress Members. In *Companion Proceedings of the Web Conference 2021*. 224–231.
[18] Aymeric Mansoux and Roel Roscam Abbing. 2020. Seven theses on the fediverse and the becoming of FLOSS. (2020).
[19] Mastodon. 2022. Mastodon. hhttps://github.com/mastodon/mastodon. (2022).
[20] Mastodon. 2022. Over 1 million people have joined Mastodon since October 27. https://twitter.com/joinmastodon/status/1591519312338210816. (2022). Accessed: 2023-01-04.
[21] Aravindh Raman, Sagar Joglekar, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2019. Challenges in the decentralised web: The mastodon case. In *Proceedings of the Internet Measurement Conference*. 217–229.
[22] Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3, 2 (2011).
[23] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084
[24] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).
[25] Renato Lond Cerqueira. 2022. Timeline for the shutdown of the mastodon-twitter crossposter instance at crossposter.masto.donte.com.br. https://write.as/renatolond/timeline-for-the-shutdown-of-the-mastodon-twitter-crossposter-instance-at. (2022). Accessed: 2023-01-04.
[26] Leanne Townsend and Claire Wallace. 2016. Social media research: A guide to ethics. *University of Aberdeen* 1, 16 (2016).
[27] Leanne Townsend and Claire Wallace. 2017. The ethics of using social media data in research: A new framework. In *The ethics of online research*. Emerald Publishing Limited, 189–207.
[28] Changtao Zhong, Mostafa Salehi, Sunil Shah, Marius Cobzarenco, Nishanth Sastry, and Meeyoung Cha. 2014. Social bootstrapping: how pinterest and last. fm social communities benefit by borrowing links from facebook. In *Proceedings of the 23rd international conference on World wide web*. 305–314.

APPENDIX A

Feature	Description	Mean	Std Dev	Median
followers_count	Number of followers of user	7715.02	141053.62	611
following_count	Number of users that user follow	1743.52	6158.66	719
listed_count	The number of public lists that this user is a member of	59.29	544.45	4
tweets_count	Number of posts of user	61606.21	173498.44	18802
tweets_length_avg	Average posts length of user	115.78	43.59	110
hatewords_count	Number of hate words from hatebase.org	1.83	5.82	0
hatewords_avg	Average number of hate words from hatebase.org	0.04	0.12	0
url_count	Number of URLs in posts of user	31.69	189.35	7
url_avg	Average number of URLs in posts of user	0.51	0.48	0.48
hashtags_count	Number of hashtags in posts of user	35.73	198.50	6
hashtags_avg	Average number of hashtags in posts of user	0.49	0.85	0.17
mentions_count	Number of mentions in posts of user	45.08	93.61	8
mentions_avg	Average number of mentions in posts of user	0.62	0.78	0.31
likes_count	Number of likes received by users	30888.95	100533.25	344
likes_avg	Average number of likes received by users	632.07	3490.73	8.03
reply_count	Number of posts replied by user	18.68	24.87	5
reply_avg	Average number of posts replied by user	3.83	13.78	4.28
retweet_count	Number of reblogs by user	13.05	138.52	16
retweet_avg	Average number of reblogs by user	6.70	84.01	2.93
sensitive_count	Numebr of sensitive posts of user	2.19	63.91	0
sensitive_avg	Average numebr of sensitive posts of user	0.02	0.11	0
created_year	The year the user created the Twitter account	7.71	5.25	8
language_index	The most frequently used languages by users	-	-	-
sentence_embedding	Aggregation of all posts embedding by the user	-	-	-

Table 2: Summary of all extracted features used for model training