Truthiness and Activity Counts: The Challenges and Opportunities of Enumerating Behavior in Decentralized Social Media

Matthew N. Nicholson, Casey Fiesler, Brian C. Keegan

University of Colorado Boulder
Department of Information Science

Abstract

Basic measures of activity and use are central to computational social science research and policy, and in a decentralized domain it is more difficult to track these fundamental measures. In this work, we illustrate this difficulty by comparing several secondary sources of this data on the federated social network Mastodon, and demonstrate that there is widespread disagreement. We briefly discuss the how ground truth becomes contextual and political in this less centralized setting, and argue that new modes of infrastructure and infrastructuring are needed in this less centralized space.

Background

Elon Musk's acquisition of Twitter in October 2022 marked a significant turning point in the history of social media generally. Musk's disruptive changes to Twitter drove waves of migrations to alternative microblogging platforms like Mastodon, Bluesky, and Threads. Though the movement of users between online communities and social platforms is not new, but the scale and speed of users' departures from Twitter was notable (Huang 2022).

In response, researchers have sought to document these changes in activity (*e.g.*, number of users) and use (*e.g.*, number of posts), in both Twitter and Mastodon (He et al. 2023; Jeong et al. 2024). These studies often study the most visible subset of accounts (*i.e.* through hashtags and profile information), and rely on the (now paywalled) Twitter/X platform API for data collection.

While these are important steps towards a more complete understanding of the phenomenon, the ability for researchers to track the complete dynamics of the "Twitter migration" is further complicated by the architecture of these alternatives. Rather than a centralized source of truth publishing digital trace data over a public API, these decentralized social platforms require complex retrieval, filtering, and aggregation across decentralized networks resulting in divergent counts of users, instances, and posts over time.

Methods for estimating activity start by selecting a few starting seed instances and performing a breadth-first search, aggregating instance-level statistics along the way (La Cava, Greco, and Tagarelli 2021). The resulting counts therefore depend on the choices of seeds, implementation of the crawler, and respecting the preferences of instances and users to be included in enumerations. Even the "official" statistics reported by Mastodon's parent non-profit organization[1] rely on these same crawling techniques and come with similar caveats.

We ask two research questions: (1) how many accounts are on Mastodon? and (2) how many instances are on Mastodon? We describe our methods of post hoc and trace analysis, finding that several prominent data sources differ substantially from one another. We then discuss the implications of the difficulty in answering this fundamental question of counting in a federated context and highlight its importance to the principles of a sound computational social science. Finally, we conclude with a call for more investment and in longitudinal data collection and research infrastructure in the distributed web.

Methods

We collected a dataset from several publicly available secondary sources that monitor both the number of accounts and instances on Mastodon. A primary source is one produced (*e.g.*, those reported on an instance's "About" page, or returned by calling the `GET api/v2/instance` endpoint for a given instance). A "secondary" source aggregates and reports these primary sources in some way. This data was collected by a combination of scripts conducting regular retrievals using GitHub Actions since January 2023 and others collected using the Internet Archive's Wayback Machine.

Our focus on secondary sources in this work is intentional. We elected to instead draw from publicly available monitoring data—specifically the tools instances.social (Rousseau 2024), the Fediverse Observer (noa 2024b), and FediDB (noa 2024a)—to illustrate the challenges of answering fundamental research questions in computational social science and social computing.

Findings

Figure 1 visualizes the counts of users and instances from the three data sources between January 2023 and March 2024. We observe substantial differences across all three

[1] https://joinmastodon.org/servers

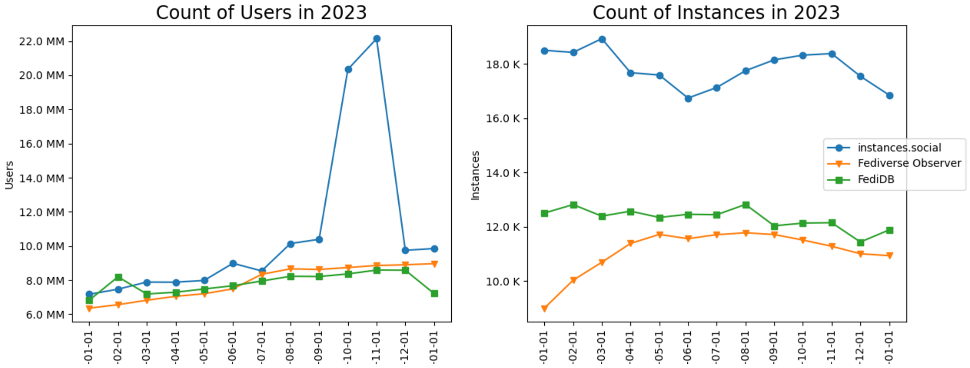

Figure 1: The user and instance counts in Mastodon, from three prominent sources.

counts, reflecting differences methods for sampling, retrieving, filtering, and aggregating data. Each of these sources of data are counting differently, despite all presenting themselves as a census.

User Counts. Although there are general positive trends over time, the actual counts of user vary substantially across the data sources. The `instances.social` count was impacted by a suspected duplication of records, leading to a reported doubling of users in a month and a subsequent precipitous drop after these duplicates were removed. Compared to the Fediverse Observatory, the source with the lowest initial reported user counts in Jan 2023 (6.4 million), differences in reports from instances.social range between 2.4% and 150% more users in the same month, while FediDB reports a range between 19.5% fewer and 24.8% more.

Instance Counts. While there is greater agreement in the estimates of the total number of instances, the trends between all sources disagree. Instance.social and FediDB report decreases in instance counts year over year (of 8.9% and 5.0%, respectively), while Fediverse Observatory reports an increase of 21.5% year over year.

Discussion and Conclusion

The disagreements on fundamental activity metrics between several reputable sources of information is indicative of the ongoing challenges of sampling, reproducibility, and the ethical use of data in computational social science that have only become more salient with decentralized architectures (Lazer et al. 2020).

Validity and Reproducibility

Social computing and internet studies researchers have largely constructed our disciplines on an assumption of centralized architectures providing a single and reliable source of truth. Emerging decentralized social media platforms (protocols) like Mastodon (ActivityPub) and Bluesky (AT Proto) destabilize this assumption. If fundamental constructs like counts of users, instances, and posts cannot be reliably defined or reproducibly observed, more complex constructs face risks of being contested as invalid.

Sampling across the federated social media is heavily skewed by the choice of seed instances and assumes consistent server uptime. These temporal differences presents a challenge to the reproducibility of this research beyond traditional concerns around "data, code and method" (Hutton and Henderson 2015). Performing a full census of instances at a regular interval for the sake of full data coverage is both a costly computational endeavor and may risk violating norms of privacy held in Mastodon if approaches like "polite data crawling" (La Cava, Greco, and Tagarelli 2021) are not followed. We could count on Mastodon's official API statistics, but platform-level statistics are often not suitable for the kind of analyses useful to researchers.

The decisions about research design, retrieval, and processing of digital trace data from decentralized social media demand stronger scrutiny than data from centralized platforms. A crawler's initial choice of seed can influence which connected components potentially network are examined. The possibilities of a "fediverse" of interoperable instances with different affordances emulating microblogging like Twitter, image sharing like Instagram, social filtering like Reddit, and video hosting like YouTube illustrate the challenges of defining boundaries. Whether to include instances that reject basic moderation expectations or using derivatives of Mastodon's open source code (*e.g.*, former president Donald Trump's Truth Social) are other edge case examples that researchers must consider. These are illustrative rather than exhaustive examples of how basic methods of counting activity on decentralized social media are not objective and neutral but deeply contextual and political.

Measurement in the Post-API Age

Interest in decentralized social media has been influenced by the deterioration of other social media services, including interoperability and accessibility via APIs. Facebook, Instagram, YouTube, Twitter, and Reddit have all dramatically curtailed researchers' independent access to platform data since 2016. This "post-API age" is characterized by technological and financial limits on access to data, heightened concerns among users about non-consensual uses of their data, and greater threats to researchers for violating privacy policies or terms of service (Freelon 2018; Bruns 2019). The challenges faced by researchers of decentralized social architectures and the practices and solutions they generate are likely to be of interest to other researchers grappling with an increasingly fragmented and difficult-to-analyze social media ecosystem. Efforts should be made by the former to ensure their methods, tools, and perspectives can inform similar and growing challenges confronting the latter.

Governance and Ethics

Decentralized social media as a popular alternative to centralized platform architectures also arrives at a distinctive historical moment in the history of computing. Large language models have emerged as disruptive technologies, but their value is contingent on their access to enormous volumes of training data typically scraped from corpora like publicly-available social media posts (Bender et al. 2021). Many Mastodon users and instances are deeply invested in new models of privacy, moderation, and governance as reactions to centralized social platforms' practices. They can be very sensitive to the kinds of non-consensual data collection about their content (Gehl and Zulli 2023).

Though formal user protections exist like the GDPR and CCPA, these specific statutes apply only to certain entities of a particular size or hosted in a specific jurisdiction. Questions of agency over data use remain—including "who can access this data," "for what purpose," and "at what level should these decisions be made?" remain open and challenging. New infrastructure and practices need to be developed to ensure accountable governance and consensual ethical use of the data in decentralized social media. This will likely involve a combination of social (practices, licenses, policies, *etc.*) and technical (interface design, API parameters, databases, *etc.*)

Infrastructuring Quantitative Description

The institutions and mechanisms for funding, supporting, and sustaining research on decentralized social media still need to be established. Unlike major industry players, there are neither data science teams nor funding opportunities at Mastodon or Bluesky to support student interns and researchers. We are also in a "post-LLM age" where high-resolution social data is increasingly used to train valuable AI models. Polices and laws for governing how users' data can be utilized remain under-developed. Users and governors of decentralized social media should be proactive in setting the terms of how their data will be used.

For scientists, it is more important than ever that research communities can agree on how to answer to fundamental quantitative questions like "how many users are on a decentralized platform?" Answering questions like this may require metrics incorporating uncertainty and acknowledging computational and ethical limits on data collection. Within a decentralized architecture, this will require iteratively developing and openly documenting methods, sources, and metrics to ensure validity and reproducibility and developing social and technical systems to ensure accountable governance and ethical data collection.

Acknowledgments

Supported by NSF awards #2309485 and #2240196.

References

2024a. FediDB - Fediverse Network Statistics.

2024b. Fediverse Observer checks all servers in the fediverse and gives you an easy way to find a home using a map or list.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. Virtual Event Canada: ACM. ISBN 978-1-4503-8309-7.

Bruns, A. 2019. After the 'APIcalypse': Social Media Platforms and Their Fight against Critical Scholarly Research. *Information, Communication & Society*, 22(11): 1544–1566.

Freelon, D. 2018. Computational Research in the Post-API Age. *Political Communication*, 35(4): 665–668.

Gehl, R. W.; and Zulli, D. 2023. The digital covenant: non-centralized platform governance on the mastodon social network. *Information, Communication & Society*, 26(16): 3275–3291.

He, J.; Zia, H. B.; Castro, I.; Raman, A.; Sastry, N.; and Tyson, G. 2023. Flocking to Mastodon: Tracking the Great Twitter Migration. In *Proceedings of the 2023 ACM on Internet Measurement Conference*, IMC '23, 111–123. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703829.

Huang, K. 2022. What Is Mastodon and Why Are People Leaving Twitter for It? *The New York Times*.

Hutton, L.; and Henderson, T. 2015. Making Social Media Research Reproducible. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(4): 2–7. Number: 4.

Jeong, U.; Sheth, P.; Tahir, A.; Alatawi, F.; Bernard, H. R.; and Liu, H. 2024. Exploring Platform Migration Patterns between Twitter and Mastodon: A User Behavior Study. ArXiv:2305.09196 [cs].

La Cava, L.; Greco, S.; and Tagarelli, A. 2021. Understanding the growth of the Fediverse through the lens of Mastodon. *Applied Network Science*, 6(1): 1–35. Number: 1 Publisher: SpringerOpen.

Lazer, D. M. J.; Pentland, A.; Watts, D. J.; Aral, S.; Athey, S.; Contractor, N.; Freelon, D.; Gonzalez-Bailon, S.; King, G.; Margetts, H.; Nelson, A.; Salganik, M. J.; Strohmaier, M.; Vespignani, A.; and Wagner, C. 2020. Computational social science: Obstacles and opportunities. *Science*, 369(6507): 1060–1062.

Rousseau, A. 2024. TheKinrar/instances. Original-date: 2017-04-06T09:10:34Z.