A System for Dynamically Tracking Content Moderation on Reddit

George Baker[1, 2] and **Bharadwaj Kadiyala**[1]
[1]Department of Operations and Information Systems, University of Utah
[2]Department of Pediatrics, University of Chicago
george.baker@bsd.uchicago.edu | bkadiyala@eccles.utah.edu

Abstract

Recent work in natural language processing, human-computer interaction, and computational social science takes interest in the study of decentralized content moderation, in which individual communities largely determine their own norms, rules, and enforcement thereof. A key challenge to this body of work is that, once moderated, content and related variables become difficult or impossible to recover; previous work often relied on 3rd-party historical data sources, but recent world events, legal disputes, and policy shifts have significantly disrupted these services, practically disabling their research use-cases. As a result, in order to conduct new research and reproduce previous results, researchers must record content as it's created, and monitor variables of interest over time. In this paper we present and publicly release a software system[1] for the dynamic monitoring of Reddit posts, communities, and moderation actions, to enable scalable and reproducible research on decentralized platform governance and content moderation. To the authors' knowledge, at the time of publication this system is the only available solution for general-purpose, real-time, policy-compliant longitudinal data collection on Reddit. Furthermore, the system's integration with the official Reddit API enables the collection of authentication-gated data such as community engagement metrics and moderation team information, which was unavailable in previous historical data sources.

1 Introduction

Recent research in Computational Social Science (CSS), Human-Computer Interaction (HCI), and Natural Language Processing (NLP) takes particular interest in decentralized content moderation, in which social media platforms largely delegate content moderation decisions to users and communities. Recent events including transitions from centralized moderation by X (formerly Twitter) (Lyons, 2021) and Meta (Kaplan, 2025) spur on this body of work, provoking questions about how human moderators behave (Huang et al., 2024), how rules and norms differ across communities (Chandrasekharan et al., 2018; Reddy and Chandrasekharan, 2023), and how Large Language Models (LLMs) can be effectively integrated into the content moderation process (Cao et al., 2024; Cho et al., 2024; Franco et al., 2024; Kumar et al., 2024; Ye et al., 2023).

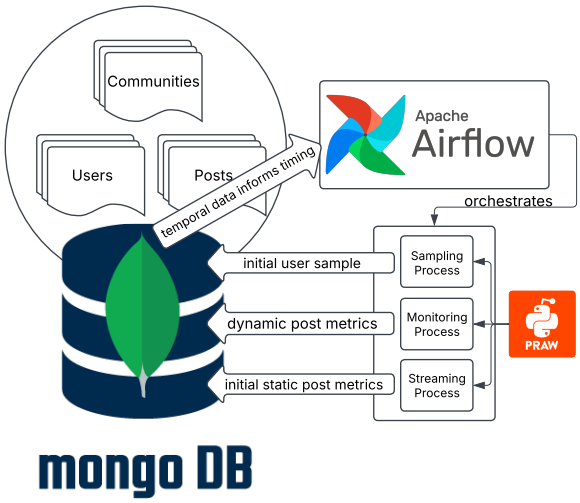

Figure 1: Overview of content moderation monitoring system architecture.

Reddit[2], a community-based, discussion-oriented social media platform, has been central to much of this research due to the variety and volume of publicly accessible data, as well as its relatively developer- and researcher-friendly policy compared to that of other social media platforms. However, recent upsets to this policy caused by concerns over user-data ownership and usage (Counts, 2024; Reddit, 2024) have resulted in the prohibition of the distribution of deleted content

[1]https://github.com/kadiyala-lab/moderation-monitoring

[2]https://www.reddit.com/

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 428–435

and interdicted third-party services doing so (namely the Pushshift project; Baumgartner et al. 2020), effectively destroying the reproducibility of past work using these sources of Reddit data, as well as disrupting future work.

In this work, to enable researchers to collect their own temporal data from Reddit for the purpose of studying content moderation, we introduce and publicly release (under a CC BY-NC-SA license) our data collection system. Our system consists of a highly configurable Apache Airflow directed acyclic graph (DAG) written in Python, with the purpose of recording static and dynamic variables with respect to posts, users, communities, and their moderators, in a manner that is compliant with Reddit's Public Content Policy[3]; unlike previous approaches, our system does not automatically distribute the content it collects, and our license prohibits commercial use. We hope that this system will not only re-enable reproduction of previous work that relied on now defunct data sources, but also facilitate ongoing and future research.

2 Related Work

2.1 Retroactive Reddit Data Sources

In contrast to real-time data collection approaches such as our own, some past work relied on retroactively available data. Retroactive data sources such as the Internet Archive have the advantage of instantly available historical data, but researchers are limited to the inconsistent and arbitrary data collection intervals of the platform, and the long-term availability of such services is also threatened by cyber-attacks (Kahle, 2024), legal disputes (Koebler, 2023), and the platforms' reliance on grants and donations to continue operation. These archival services also did not preserve certain authentication-gated data, which may be critical to the study of these communities; for example, information about a community's moderation team or engagement metrics such as weekly active users[4].

Pushshift The Pushshift platform was a "social media data collection, analysis, and archiving platform" (Baumgartner et al., 2020) that made historical data from Reddit publicly and freely available, enabling a wide range of social media and natural language processing research. As a result of disputes with Reddit over Data API Terms[5] violations, the Pushshift website became unavailable to the public in 2023 (Koebler, 2023), later returning with severely restricted data access such that only moderators of a given community had access to its corresponding data[6], motivating the present work to enable researchers to collect their own data.

The Internet Archive[7] is a public-access digital archive of online content including web pages (Kahle, 2002), which are preserved as snapshots at irregular intervals by web crawlers. Reddy and Chandrasekharan (2023) use historical data from the Internet Archive to study the evolution of rules in Reddit communities over time (described further in Section 2.2), observing that the availability of historical web scrapes depends on the popularity of the studied communities, and even for popular communities there are often gaps between scrapes of months to over a year, making analysis at small time intervals or of niche communities impossible. Thelwall and Vaughan (2004) study international coverage bias, finding that the Internet Archive disproportionately favors U.S.-based web pages for preservation, further limiting its application to research on international communities.

2.2 Content Moderation on Reddit

To highlight the wide range of content moderation research conducted using data from Reddit, here we summarize relevant recent works, many of which relied on the now defunct Pushshift project (Baumgartner et al., 2020) for reproduction, and now require a novel means thereto.

Huang et al. (2024) study the effects of politically biased moderation in Reddit communities, finding community moderators to make biased moderation decisions against users expressing political views opposite to their own, which acts as a driving factor in the formation of echo chambers, communities in which users only encounter beliefs that match their own.

Reddy and Chandrasekharan (2023) study how rules change in Reddit communities over time, finding that the majority of rule additions and modifications occur early in the community's lifespan, and that the type of rules added also changes over time.

Chandrasekharan et al. (2018) study norms (implicit rules) across Reddit communities, using con-

[3] https://support.reddithelp.com/hc/en-us/articles/26410290525844-Public-Content-Policy

[4] https://redditinc.com/news/new-ways-to-see-community-activity-on-reddit

[5] https://redditinc.com/policies/data-api-terms

[6] https://www.reddit.com/r/pushshift/comments/14ei799/

[7] https://web.archive.org/

tent moderation as a heuristic for norm violation. They find that norms on Reddit exist at Micro (specific to individual communities), Meso (shared across groups of communities), and Macro (applying to most communities) scales.

Gurkan and Suchow (2022) study how Reddit users learn the norms of the communities they participate in over time with the Cultural Consensus Theory framework (Romney et al., 1986), using historical data from Pushshift. Experimental results demonstrate that as time elapses and users gain more experience in a community, their calibration to the community-specific norms improves; additionally, the users' initial calibration to the community predicts the length of their tenure (users with low initial alignment with norms are unlikely to continue to participate in the community).

Ye et al. (2023) study the application of language models to automatic content moderation, using a multilingual dataset of 1.8 million comments from Reddit. They find that previous approaches based on toxic content identification struggle with comments in non-English languages. Qualitative analysis additionally finds that human moderation labels are highly noisy; many highly offensive comments are never moderated, and many benign comments are moderated.

Deolankar et al. (2024) monitor how soft peer feedback in the form of Reddit's peer-assigned ratings (*upvotes* and *downvotes*) affects the future posting behavior of recipients, finding that negative peer feedback can act to regulate future posting behavior without any explicit moderation action.

3 System Overview

Data Collection For interactions with Reddit's API, we use the PRAW (Python Reddit API Wrapper)[8] library, which ensures compliance with Reddit's API Rules[9]. For future extensibility, our system is designed with modular metrics for posts, users, and communities, such that custom metrics can be added by defining a Python method that takes the respective PRAW class as an input, and adding the name of the method to the configuration file. Our system categorizes metrics as either static or dynamic; static metrics (e.g. post author name, user join date) are collected only once when a data point is first observed, while dynamic metrics (e.g. a post's score, the number of moderators in a community) are collected at configurable intervals. To reduce unnecessary API calls, dynamic metrics may define a default initial value, for example, it is safe to assume that when a post is first made that it has no replies and has not been moderated.

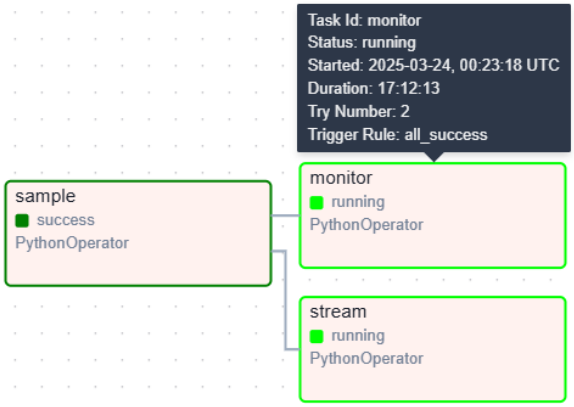

Figure 2: A screenshot of the Graph View of the content moderation monitoring system's DAG as it runs in Airflow.

Task Scheduling & Automation Apache Airflow DAG files define a number of tasks and their inter-dependencies to manage their scheduling and execution. In our main DAG we consider three primary tasks (shown in Figure 2):

1. An initial sampling task[10] which populates the database with the users or communities to be followed throughout the study. Our default sampling implementation follows the specifications of Deolankar et al. (2024) by first randomly sampling N unique public communities, and then recording up to the M most recent posters in that community (in cases where there are fewer than M recent posters, all recent posters are included).

2. A streaming task, which constantly streams new posts and related variables from the previously selected users or communities into the database. The logic of the streaming task is shown in Algorithm 1.

3. A monitoring task, which queries the database for dynamic variables that are ready to have new time-series values recorded, and does so. The streaming and monitoring tasks run simultaneously as parallel processes after the initial sampling task.

[8] https://github.com/praw-dev/praw
[9] https://github.com/reddit-archive/reddit/wiki/API

[10] The initial sampling step can be skipped if a fixed list of users or communities of interest are specified.

Algorithm 1 Stream Task

Require: streams $\subseteq$ {Subreddit, Redditor}
  while now − start < duration **do**
    for all stream ∈ streams **do**
      if # $\text{posts}_{\text{stream, day}}$ < limit **then**
        new_posts = getPosts(stream)
        insert new_posts into DB
        for all post ∈ new_posts **do**
          if post.author ∉ DB **then**
            insert post.author into DB
          end if
          if post.sr ∉ DB **then**
            insert post.sr into DB
            insert post.sr.mods into DB
          end if
        end for
      end if
    end for
  end while

Data Storage Several data storage options were considered in the design of our system, including file-based storage, relational databases, and document-oriented databases. Raw JSON file storage would be functional and easy to implement, however, due to the volume of data required (e.g. 2.8M comments in Chandrasekharan et al. (2018)), file sizes would grow excessively over time, threatening to fill memory, and increasing read and write times over the course of data collection. Relational databases such as MySQL, Postgres, and SQLite address the issue of having to read and write to large files from memory, and have improved data integrity, but operate on strictly defined schemas which inhibit researchers' abilities to rapidly iterate on experimental setups and variables of interest. On the other hand, document-oriented databases like MongoDB allow for flexible schemas, and are often faster to read from and write to than relational databases (Jose and Abraham, 2020). An additional benefit of MongoDB is its support of time-series data, which our use case of monitoring variables over time relies on. Figure 3 demonstrates an example document and corresponding time-series measurements.

Observability Although not strictly necessary for functional data collection, we additionally implement a monitoring stack to track the progress and health of the data collection system over the course of the run. The monitoring stack (pictured

Figure 3: An anonymized MongoDB document from the static posts collection (top) with a corresponding time-series entry (bottom), as collected by our system.

in Figure 5) consists of an open-source query exporter [11] which regularly collects metrics from the MongoDB database and stores them in an auxiliary Prometheus[12] database; the auxiliary database in turn serves as a data source for real-time interactive visualizations presented in a Grafana[13] dashboard (shown in Figure 4). An auxiliary database is used over direct connection to the MongoDB database because:

1. Interactions with the dashboard trigger additional queries which may take precedent over read and write operations related to data collection; the consistently timed metric exports ensures that the additional computational burden is uniform and doesn't interfere with data collection, and that the dashboard remains interactive and responsive to real-time inputs.

2. Direct connection to MongoDB using the official plugin requires a paid Grafana Enterprise subscription; we intend for our system to be able to run without any paid software or subscription services.

Failure Recovery Apache Airflow's "retry" mechanism automatically reattempts tasks that fail due to transient issues including VM restarts, network downtime, resource contention, and unhandled exceptions. The number of times to retry a given task is configurable within Airflow, and it is additionally possible to configure automated email alerts with respect to task status. Failed API calls, for example, due to exceeding Reddit's rate limits, are handled by retrying with exponential backoff, up until a configured limit; in cases in which this

[11] https://github.com/raffis/mongodb-query-exporter
[12] https://prometheus.io/
[13] https://grafana.com/

Figure 4: A screenshot of the Grafana dashboard during data collection.

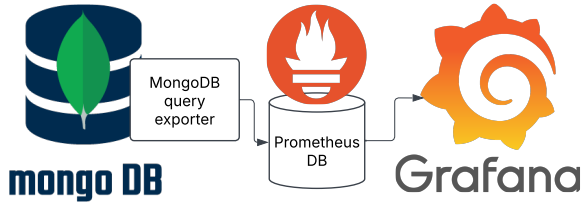

Figure 5: Monitoring stack overview.

limit is reached, a placeholder string is returned with the name of the exception to enable debugging and data completeness checks.

4 Evaluation

4.1 Qualitative Comparison to Other Systems

In Table 1, qualitative differences between the proposed system and the most widely used alternatives (The Internet Archive and Pushshift) are presented.

4.2 Timing Consistency

In temporal data analysis, an important assumption is often made that the time interval between observations remains relatively constant. This assumption can practically never be perfectly met, particularly when experiments rely on network communications, due to API rate limits, network latency, and hardware constraints.

To evaluate the proposed system's efficiency in terms of maintaining appropriate time interval widths, we run an experiment in which we sample (following Deolankar et al. (2024)) 1,807 users, and follow their posts and the communities in which they are made for three days, recording dynamic variables (details in Appendix A) with an intended interval of 24 hours.

We deploy our system to a virtual machine provisioned through the University of Utah's Center for High Performance Computing[14] with 16 gigabytes of RAM, 4 AMD EPYC 9334 CPU cores, and 100 gigabytes of storage.

We report a mean monitor timing error of 34.3 seconds with a standard deviation of 28.1 seconds. The mean timing error represents 0.03% of the targeted 24 hour monitoring interval.

5 Conclusion

In this work, we present and publicly release a data collection system for dynamically tracking variables on Reddit, specifically aimed at the study of content moderation and decentralized platform governance; however, despite this focus, the system is designed with extensibility in mind such that future work can easily adapt it to their use-case by defining custom metrics within our system in a plug-and-play manner. Our system is designed to re-enable reproduction of previous works that relied on the now defunct Pushshift project (Baumgartner et al., 2020), as well as to facilitate ongoing and future work, all while respecting Reddit's API and content usage policies.

In Section 4.2, we conduct a small-scale evaluation to demonstrate that our system is functionally capable of reproducing past data collection efforts, and that it is flexible in terms of the metrics it can capture, and performant in terms of maintaining

[14] https://www.chpc.utah.edu/

	Ours	Internet Archive	PushShift
Data Availability	High	Low*	None (Defunct)
Consistent Time Intervals	✓	✗	✗
Reddit Policy Compliant	✓	✗	✗
Authentication-Gated Data	✓	✗	✗
Access Mode	Real-time Collection	Historical	Historical

Table 1: A comparative evaluation of data collection systems for Reddit. *"Low" data availability denotes that there are often large, inconsistent temporal gaps between data points, and that not all variables of interest are necessarily captured.

consistently-timed monitoring intervals, a statistic which is often omitted in previous papers.

In future work, we intend to leverage our system to explore research questions about community dynamics and moderator selection, and how these impact content moderation and user engagement within Reddit communities. Substantial ongoing efforts will also be devoted to maintaining and improving the data collection system and its documentation.

Limitations

Here we acknowledge limitations to our system which bound its application.

As our system is built on top of the PRAW library, it is applicable only to Reddit, and no other social media platforms. Most recent work using Reddit data doesn't attempt cross-platform analysis, and so it is the opinion of the authors that this limitation isn't severe. Future work may take interest in generalizing findings across social media platforms; for such use cases our system is not directly applicable without substantial modification.

Our system makes the hard assumption that monitoring intervals should be consistently timed. We aren't presently aware of any use cases that would require non-constantly timed monitoring intervals, but concede the possibility.

Finally, a core limitation to the performance and scalability of our system are Reddit's API limits, which indirectly dictate a trade-off between sample size, number and complexity of metrics, and monitoring-interval-width; larger sample sizes and increased number and complexity of metrics inherently use more API calls, which limits the rate at which data can be collected and processed. In the event that this rate exceeds the monitoring interval width, in order to maintain data integrity it will be necessary to decrease the sample size, number of metrics, and/or complexity thereof. At the time of publication, Reddit imposes a per API key limit of 1,000 requests per 10 minutes, or 144,000 per day.

Ethical Considerations

For continued academic access to Reddit's data, we believe that it is necessary to respect Reddit's policy on data ownership and content usage[15], including their requests to not publicly share content that has been removed or made private on the platform, and not to train machine learning models on their data without appropriate permission. Our system was deliberately designed to not directly violate this policy, although it does facilitate the collection and analysis of such data, presenting the potential for misuse. As such, we distribute our system under a CC BY-NC-SA license to discourage commercial use, but still enable academic endeavors. To the best of our knowledge, our system does not violate any of Reddit's policies on its own, which we believe is an integral step towards preserving good relations between Reddit and academia.

6 Acknowledgments

The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged.

References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Yang Trista Cao, Lovely-Frances Domingo, Sarah Gilbert, Michelle L. Mazurek, Katie Shilton, and Hal Daumé Iii. 2024. Toxicity detection is NOT all you need: Measuring the gaps to supporting volunteer content moderators through a user-centric method.

[15] https://support.reddithelp.com/hc/en-us/articles/26410290525844-Public-Content-Policy

In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3567–3587, Miami, Florida, USA. Association for Computational Linguistics.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25.

Hyundong Cho, Shuai Liu, Taiwei Shi, Darpan Jain, Basem Rizk, Yuyang Huang, Zixun Lu, Nuan Wen, Jonathan Gratch, Emilio Ferrara, and Jonathan May. 2024. Can language model moderators improve the health of online discourse? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7478–7496, Mexico City, Mexico. Association for Computational Linguistics.

Aisha Counts. 2024. Reddit releases content policy while seeking more licensing deals (rddt).

Varad Deolankar, Jessica Fong, and S Sriram. 2024. Content generation on social media: The role of negative peer feedback. *Available at SSRN 4522092*.

Mirko Franco, Ombretta Gaggi, and Claudio E Palazzi. 2024. Integrating content moderation systems with large language models. *ACM Transactions on the Web*.

Necdet Gurkan and Jordan W Suchow. 2022. Learning and enforcing a cultural consensus in online communities. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Justin T Huang, Jangwon Choi, and Yuqin Wan. 2024. Politically biased moderation drives echo chamber formation: An analysis of user-driven content removals on reddit. *Available at SSRN*.

Benymol Jose and Sajimon Abraham. 2020. Performance analysis of nosql and relational databases with mongodb and mysql. *Materials Today: Proceedings*, 24:2036–2043. International Multi-conference on Computing, Communication, Electrical Nanotechnology, I2CN-2K19, 25th 26th April 2019.

Brewster Kahle. 2002. The internet archive. *RLG diginews*, 6(3).

Brewster Kahle. 2024. Learning from cyberattacks.

Joel Kaplan. 2025. More speech and fewer mistakes.

Jason Koebler. 2023. The reddit protest is a battle for the soul of the human internet.

Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 865–878.

Kim Lyons. 2021. Twitter launches birdwatch, a fact-checking program intended to fight misinformation.

Reddit. 2024. Publishing our public content policy and introducing a new subreddit for researchers.

Harita Reddy and Eshwar Chandrasekharan. 2023. Evolution of rules in reddit communities. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 278–282.

A Kimball Romney, Susan C Weller, and William H Batchelder. 1986. Culture as consensus: A theory of culture and informant accuracy. *American anthropologist*, 88(2):313–338.

Mike Thelwall and Liwen Vaughan. 2004. A fair history of the web? examining country balance in the internet archive. *Library & information science research*, 26(2):162–176.

Meng Ye, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. 2023. Multilingual content moderation: A case study on Reddit. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3828–3844, Dubrovnik, Croatia. Association for Computational Linguistics.

Appendix A Experimental Details

For the purposes of our evaluation we track a wide range of static and dynamic variables, all of which are available in the code repository.

At the user level we record username, account creation date, and comment and link "karma" (peer-assigned scores on Reddit) as static variables.

For each user, we record posts during the observation period. As static variables, we record the post type (comment or submission), the community (Subreddit) in which it was posted, the time of creation, and the body and title (if applicable). As dynamic variables, we record the number of top level replies, total replies, deletion (whether the user deleted their own post), removal (whether a moderator removed the post), and post score.

For each recorded community, we record as static variables whether the community is self-reported as an 18+ (NSFW) community, the name of the community, and the creation date of the community. As dynamic variables, we record the number of rules, number of subscribers, number of active users, and information about the moderation team including the number, names, and permissions of all moderators.

For each recorded moderator, we record the moderator's username and account creation date as static variables; number of communities moderated, and vector embeddings of the moderator's recent posting behavior as dynamic variables.

In total, we observe 12,212 posts across 3,453 communities with 19,889 unique moderators.