Self-Moderation in the Decentralized Era: Decoding Blocking Behavior on Bluesky

Carlo Bono[1], **Chang Liu**[2], **Giuseppe Russo**[3], **Francesco Pierri**[1]

[1]Politecnico di Milano, Milan, Italy
[2]Indiana University, Bloomington, IN, USA
[3]EPFL, Lausanne, Switzerland
carlo.bono@polimi.it

Abstract

Centralized social media platforms combine top-down moderation with user-level mechanisms like blocking, but access to behavioral and moderation-related data is typically restricted. In contrast, emerging decentralized networks like Bluesky provide researchers with open access to blocking actions, offering a unique opportunity to study self-moderation at scale. Understanding blocking behavior is critical in order to assess how community safety and user autonomy can be balanced in these environments. This study examines blocking on Bluesky, analyzing more than 100M actions by nearly 2M users over three months. We construct behavioral profiles from 86 features capturing user activity, content, and network interactions, and address two questions: (1) Can a user's propensity to be blocked by others be estimated from in-platform behavior? and (2) Which behavioral features are most informative in predicting this propensity? Using state-of-the-art machine learning models, we achieve high accuracy in binary classification (AUROC > 0.85) and reliable performance in regression ($R^2 \approx 0.5$). Our findings show that a small subset of features is sufficient for robust prediction, even when accounting for higher activity on the platform. Through explainability analyses, we identify the behavioral signals most strongly associated with blocking outcomes, enabling a transparent understanding of why certain users are more likely to be blocked. Our study provides a framework and empirical evidence that advances understanding of both the mechanisms and feature signals underlying user self-moderation, while also offering insights relevant to moderation design.

Introduction

The rise of decentralized social platforms has caused significant shifts in the way online communities are managed, evolve and interact (Zignani, Gaito, and Rossi 2018). Unlike traditional social networks, decentralized platforms prioritize the principles of user autonomy and control, allowing for a more personalized and attuned experience (Raman et al. 2019). Emerging decentralized platforms have gained attention for their innovative approach to social networking and their potential to reshape online interactions, with mechanisms similar to blockchain technology (Guidi 2020). Decentralized approaches enable independent servers to host content, reducing the risk of centralized control and empowering users with greater autonomy over their data and content (Raman et al. 2019; Zignani, Gaito, and Rossi 2018). However, together with the freedom offered by a decentralized paradigm, also comes the challenge of moderating harmful and abusive behavior (Ghenai et al. 2025; Zhang and Gao 2024; Bonifazi et al. 2022; Bono et al. 2024). Online abuse, harassment, and toxic interactions have already been longstanding issues on centralized platforms, prompting various moderation strategies over time, ranging from algorithmic content filtering to user reporting systems (Jhaver et al. 2021). In centralized social networks, moderation is usually enforced through top-down and platform-driven mechanisms (Chandrasekharan et al. 2017; Pierri 2024; Di Giovanni et al. 2022). Decentralized platforms, where content moderation is often less centralized, rely more on user-driven moderation, with blocking actions serving as a tool for managing unwanted interactions (Zignani, Gaito, and Rossi 2018).

On online social platforms, users may become aware that they have been blocked, possibly through indirect indicators. On microblogging platforms, blocking a user prevents any form of interaction and removes their presence from the experience of the blocking user. Blocked users are unable to like, reply to, mention, or follow the blocker, and their posts, replies, and profile are no longer visible to the blocking user. However, information about block actions is not always publicly accessible. The approach of Bluesky, one of the emerging decentralized social platforms (Balduf et al. 2024, 2025; Failla and Rossetti 2024; Sahneh et al. 2024), is distinctive in that user blocks are currently public and enumerable, since all servers across the network must be aware of blocks in order to honor them. This approach presents a novel opportunity for research into user behavior and moderation dynamics within decentralized social networks. This transparency introduces both positive and negative potential implications. On one hand, users could be aware of possibly problematic interactions, but on the other hand, social dynamics could be influenced as blocks may carry stigmas or lead to public shaming. As a result, understanding the behavioral features that predict the propensity to be blocked is critical from the perspective of both users and platform developers, and researchers.

While a few studies have analyzed behaviors such as

muting and unfollowing on social media through experiments (Martel et al. 2024; Rathje et al. 2024), these actions have generally been difficult to study in detail, since platforms do not make this kind of user activity accessible. This paper aims to fill this gap by examining the behavioral features that correlate with being blocked, and investigating the extent to which behavioral patterns can predict being blocked. Our analysis seeks to contribute to the broader discussion on online moderation and user behavior in decentralized social networks. In doing so, we place particular emphasis on model interpretability, which enables us to identify and characterize behavioral patterns most strongly associated with blocking.

We investigate the following research questions:

RQ1 Can we estimate Bluesky users' propensity to be blocked based on their observed in-platform activity?

RQ2 Which behavioral features are most indicative of a user's propensity to be blocked?

To this end, we collect a longitudinal dataset of all the Bluesky activity over a three-month period (June-August 2024) and extract over 80 features that describe online behaviour from different perspectives, such as activity, content, and interactions. We frame the task as both a binary classification and a regression task to assess the predictability and intensity of blocking behavior within the observation period, in a retrospective manner. We also apply feature importance techniques to quantify the informativeness of features and feature groups for detecting blocked users. Our findings show that extremely blocked users can be identified with high accuracy, offering insights that can inform the design of moderation tools and interventions.

Related Work

Centralized and Decentralized Moderation Policies

Research on platform moderation has investigated the behavioral and ecological effects of interventions, typically distinguishing between "hard" and "soft" moderation strategies (Trujillo and Cresci 2022; Schneider and Rizoiu 2023). Hard moderation, such as banning communities or users, directly removes content or accounts (Young 2022; Rogers 2020), while soft moderation, including visibility adjustments or contextual interventions, reshapes user behavior without removal (Shen and Rosé 2022; Zannettou 2021; Attanasio et al. 2026). Reddit's quarantining mechanism is an example of soft moderation, limiting community growth while retaining activity, though it has been criticized for reinforcing echo chambers and allowing platforms to profit from controversial content (Copland 2020; Chandrasekharan et al. 2022). Community-driven initiatives like X's Community Notes similarly aim to contextualize potentially misleading content, increasing trust in fact-checking but struggling to prevent initial exposure to misinformation (Chuai et al. 2023; Drolsbach, Solovev, and Pröllochs 2024; Solovev and Pröllochs 2025). Banning, a common hard moderation approach, removes harmful communities entirely (Innes and Innes 2023; Ali et al. 2021). While effective at reducing harmful content on mainstream platforms, deplatformed users often migrate to fringe platforms, potentially reinforcing toxic communities and facilitating spillover of harmful behavior (Horta Ribeiro et al. 2021; Zuckerman and Rajendra-Nicolucci 2021). Targeted moderation at the individual level—deplatforming influencers (Ribeiro et al. 2024) or mitigating "superspreaders" of misinformation (Baribi-Bartov, Swire-Thompson, and Grinberg 2024)—can limit harm effectively without the broader disruptions caused by community-wide interventions. Decentralized platforms have drawn attention for their transparency and user-driven moderation mechanisms (Zia et al. 2023; Raman et al. 2019). For example, Bluesky has open-sourced Ozone, its official moderation tool, enabling users and developers to create independent moderation services and customizable content filters, expanding the ways in which moderation and self-curation can be implemented beyond individual blocklists (Balduf et al. 2024; Kleppmann et al. 2024), while prior work on Mastodon highlights how decentralized moderation often relies on on server-level blocklisting practices (**?**). This multi-tiered approach highlights the potential of community-involved moderation in decentralized contexts.

Bluesky in Research

Recent studies of Bluesky have characterized its architecture, user activity, and unique features. Balduf et al. (2024) provide a large-scale analysis of Bluesky's modular architecture and third-party provider ecosystem, while Failla and Rossetti (2024) release a high-coverage dataset encompassing 4 million users and 235 million posts, including social interactions and content recommendation outputs. Quelle and Bovet (2025) examine user-generated algorithmic feeds, noting widespread creation but limited adoption, with polarization emerging on topics such as the Israel-Palestine conflict. Sahneh et al. (2024); Nogara et al. (2026) longitudinally analyze activity during Bluesky's public rollout, observing patterns similar to established platforms, higher volumes of original content compared to resharing, and low toxicity levels, suggesting effective moderation during rapid growth. These studies collectively illustrate Bluesky's value as a platform for examining decentralized social dynamics.

Self-Moderation in Online Social Platforms

Self-moderation mechanisms, such as blocking or muting, allow users to reduce exposure to unwanted content or harassment (Vogels 2021; Seering 2020). Research indicates that user blocking is often politically or ideologically motivated, with users disproportionately blocking counter-partisan accounts or misinformation spreaders, thereby increasing network polarization (Kaiser, Vaccari, and Chadwick 2022; Martel et al. 2024; Baysha 2020). Studies also highlight the persistence of exposure to misinformation despite blocking or unfollowing behaviors, pointing to the limits of reactive strategies (Ashkinaze, Gilbert, and Budak 2024). Psychological and social factors further influence blocking behaviors. Blocking can protect mental well-being (Hunt et al. 2018; Fox and Moreland 2015), reflect cultural norms (Zhang and Fu 2020), and be guided by peer influence.

Position of Our Work

Prior work has examined self-moderation through surveys, experiments, or small-scale analyses, but has not tested whether users' behavioral traces can systematically explain or predict blocking outcomes. Bluesky's public access to self-moderation data provides a rare opportunity to analyze these behaviors at scale, offering insights into dynamics that were previously difficult to study on centralized platforms. Our study addresses this gap by formulating the problem of blocking behaviour as both a binary classification and a regression task, applying state-of-the-art models to Bluesky's publicly accessible self-moderation data. By evaluating the predictive power of behavioral features in a retrospective setting, our work provides the first large-scale evidence that blocking behavior can be inferred from per-user activity patterns, offering a foundation for future efforts to anticipate blocking events and to investigate their causal effects on user behavior.

Dataset and Methods

The methodology employed in this study is outlined in Figure 1 and can be summarized as follows. We first collect all Bluesky public data within a specified timeframe. We then build user profiles for all the active users observed, incorporating several characteristics of their activity, interactions, and shared content. To evaluate whether these behavioral signals explain blocking outcomes, we formulate the task as both a binary classification problem and a regression problem, and we leverage state-of-the-art machine learning methods to assess the informational value of these features in relation to users' propensity to be blocked by others.

Data Collection

We collected data continuously over a period of roughly three months, from June 1st, 2024, to August 28th, 2024. Statistics for the resulting dataset are provided in Table 1.

Data has been obtained through Bluesky's publicly accessible Firehose endpoint, which provides unauthenticated, real-time access to granular platform activity. A key feature of the endpoint is the ability to resume data retrieval after connection interruptions, allowing data collectors to recover missing records from up to 72 hours prior. This ensures a continuous data stream throughout the observation period, mitigating potential gaps caused by temporary network failures. To perform data acquisition, we utilized the `bluesky` Dart library (Bluesky 2024) and an established open-source library (Burghardt 2024) to interact with the AT Protocol ecosystem and consume the `com.atproto.sync.subscribeRepos` endpoint — commonly known as the Firehose[1] stream (Kato 2024).

Bluesky enables the monitoring of public user activities, which offers valuable insights into social interactions within the platform. Firehose data include user-generated content such as posts, replies, follows, likes, and other interactions (Bluesky 2024; AT Protocol 2024). Our study considers five primary forms of interaction on the platform:

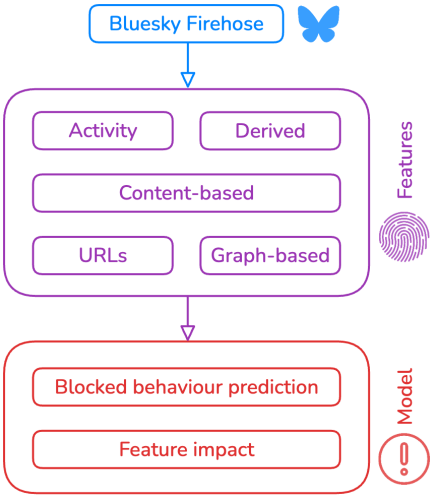

Figure 1: Diagram describing the workflow of our research.

- **Posts** – Original content shared by users.
- **Replies** – Direct responses to existing posts, fostering discussion.
- **Reposts** – Sharing of content originally published by other users.
- **Follows** – Directed social connections where one user subscribes to another user's updates.
- **Blocks** – Symmetric user actions restricting content visibility and interactions among specific accounts.

Specifically, blocking a user restricts all forms of interaction and removes their presence from the experience of the user performing the block. A blocked account cannot like, reply to, mention, or follow the blocking user. Additionally, their posts, replies, and profile will be hidden from search results. Unlike other platforms, blocks on Bluesky are public[2], enabling the study of the propensity of being blocked by other users. Mute operations, which are asymmetrical operations causing users' posts to be excluded from a feed, are instead private on Bluesky and are not included in this study.

Although Bluesky's terms of service do not impose restrictions on data collection from publicly accessible content, we adhere to strict ethical guidelines. Our dataset comprises only publicly available data collected in accordance with the platform's Privacy Policy.[3] To uphold user privacy, we do not make any raw data available. Instead, this paper presents only anonymized, aggregate observations.

Feature definition and preprocessing

In our analysis, we focus on the subset of users who published at least 10 posts during the observation period, irrespective of the account creation date and whose posts are most frequently written in English, resulting in a total of

[1] https://docs.bsky.app/docs/advanced-guides/firehose

[2] https://docs.bsky.app/blog/block-implementation

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

Item	Count
Blocks	3 278 406
Follow actions	33 942 018
Likes	292 388 501
Posts	79 737 148
Reposts	30 284 394
Unique users	1 979 713

Table 1. Summary of the collected dataset (June 1 – August 28, 2024).

427 118 users. For each of these users, we extract 86 numerical features covering multiple aspects of their behavior. These features are derived from both the collected data and external sources, providing a comprehensive view of user activity during the observation period. The complete list of features, with descriptions and logical groupings, is provided in the Appendix. Taken together, these features aim to capture meaningful behavioral dimensions that distinguish different types of users, such as high-frequency posters, socially central accounts, or users who share more toxic or politically polarized content. This breadth enables us to characterize user profiles in a way that is interpretable and behaviorally grounded, helping to contextualize why some users may be more likely to be blocked. Features related to network centrality and hyper-activity have, for instance, been linked to bot-like or abusive users on Bluesky (Nogara et al. 2026).

Action Features These 18 features are obtained by counting the number of relevant events observed in the dataset. They include counts for the following actions types: likes, posts, reposts, follows, and blocks. Each event can be a create or a delete action, which are counted separately. Moreover, since the source and target user handles are known, they are used to calculate **Derived** counts for actions directed at each user, that is, the number of events in which a user has been liked, reposted, followed, or blocked. The counts of replies, as an author or as a subject, are derived from the post contents.

Post-derived Features The following 13 textual features are extracted from the content of each post: total number of characters, number of lowercase and uppercase characters, digits, spaces, and emojis. We compute the mean and standard deviation of these features aggregating them at the user level. We also compute the variability of the declared post languages as the normalized Shannon entropy of the observed languages.

We associate several toxicity dimensions (Identity attack, Insult, Obscene, Toxicity, Severe toxicity, Threat, Sexually explicit) with each post using Detoxify (Hanu and Unitary team 2020), restricting the application to posts written in its supported languages[4]. Multilingual toxicity classification models can, in general, show variable performance across languages. This observation is partially mitigated by the fact that, in our experimental setup, user-level toxicity aggregates are dominated by English content, for which Detoxify has been extensively validated[5].

The mean and standard deviation of these toxicity indicators are calculated at the user level, resulting in 14 features. Each dimension has a value in the interval $(0, 1)$. Since our analysis only includes users with at least 10 posts, no missing values are present in the data.

Additionally, when posting original content, users may include URLs in their posts. We extract all the URLs contained in the posts and parse the domain names from these URLs. For each user, we calculate the average number of URLs per post and the overall variability of the domains, again as the normalized Shannon entropy of the observed domains.

The computation of all the post-derived features considers only original post objects, excluding reposts.

URL-derived Features We utilize the domains of the shared URLs to characterize users in terms of misinformation and credibility For each user, we count the number of domains shared, grouped according to the categories defined by Media Bias/Fact Check (MBFC[6]), considering dimensions of *bias* (Extreme-left, Left, Center-left, Center-right, Right, Extreme-right, Satire, Conspiracy, Pro-science), *credibility* (Low, Medium, High), and *factuality* (Very low, Low, Medium, Mostly, High, Very high, Mixed), resulting in 19 features. These counts are then normalized by the total number of posts published by the user. Additionally, we compute the average domain quality score based on the ratings provided by (Lin et al. 2023), substituting missing values with the average of the valid scores. The number of shared domains matching these two external sources is also used as a feature, again normalized by the number of posts.

Graph-based Features Centrality measures are widely used to capture users' influence and position within a network (Newman 2010). We compute graph-based centrality measures — Coreness, total Degree, and PageRank — from the networks of the following interactions: follows, likes, replies, and reposts, considered separately. Coreness measures the depth of a node within the network, Degree quantifies the number of direct connections a user has, and PageRank evaluates a user's importance based on the probability that a random walk over the network lands on them.

Target variables and predictive tasks

We formalize blocking prediction as both a binary classification task and a regression task, enabling us to examine whether users' behavioral features can explain not only the occurrence but also the intensity of blocking behavior. We remark that this analysis is conducted retrospectively on existing blocking events, and not to predict future blocks, which we leave for future work.

Since no standard threshold exists for classifying a user as *blocked*, we consider two target variables to capture blocking behavior: the absolute number of blocks ($raw_{blocked}$) and the activity-normalized number of blocks ($norm_{blocked} = raw_{blocked} \,/\, \#posts$). The normalization

[4] 'en', 'pt', 'ru', 'es', 'fr', 'tr', 'it'.

[5] We recall that users included in the analysis are required to have English as the most frequent language of their posts.

[6] https://mediabiasfactcheck.com/

is motivated by the fact that highly active users may appear to be blocked more often simply due to their overall activity level. We remark that we focus only on active users who shared at least 10 posts over the observation period (427 118).

Task 1: Binary Classification The first task is a binary classification problem: using percentiles of the two variables defined above ($raw_{blocked}$ and $norm_{blocked}$) as thresholds, we predict whether a user belongs to the positive *blocked* class.

We employ Random Forest classifiers as tree-based ensemble methods are well-suited for heterogeneous, nonlinear feature spaces. We implement them via `XGBoost` with 500 estimators, evaluated with the average of 10 independent runs of 10-fold cross-validation, with the label distribution balanced by randomly undersampling the majority class. This procedure aims at a greater robustness, particularly at higher thresholds where the label imbalance becomes more extreme. Feature importance is quantified with SHAP (Lundberg and Lee 2017), which provides model-agnostic, locally consistent attributions and allows us to interpret how individual behavioral features contribute to predictions. Both XGBoost and SHAP value analyses are state-of-the-art approaches for modeling and interpreting user behavior in social media settings (Shevtsov et al. 2022; Mathew et al. 2019; Yang et al. 2020). Performance is measured with Receiver Operating Characteristic Area Under the Curve (ROC AUC), where 0.5 indicates random predictions and 1 perfect classification.

Task 2: Regression The second task is a regression problem, where we obtain a model estimate $\hat{y}$ of the actual number of blocks received y, according to both raw and activity-normalized definitions. We use Random Forest Regression (XGBRFRegressor) and AutoGluon TabularPredictor (Erickson et al. 2020) with the "high quality" setting. AutoGluon's automated model selection and ensembling provide strong performance on heterogeneous tabular data with minimal tuning. Model performance is assessed using R^2 and Mean Absolute Error (MAE):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,$$

using an 80:20 random train-test split.

Results

Exploratory analysis

We first provide descriptive statistics of the dataset, offering an overview of Bluesky users' activity during the period of analysis.

The top panel of Figure 2 reports the temporal distribution of actions performed by Bluesky users during the observation period. Likes are the most frequent action, with a daily median of 3 240 712. Original posts occur nearly three times as often (daily median = 873 760) than reposts (daily median = 338 544). Users also frequently follow each other, with a daily median of 248 255 follow actions. In contrast, block actions are significantly less common, with a daily median of 31 934, making them several orders of magnitude less frequent than the other interactions. In terms of temporal

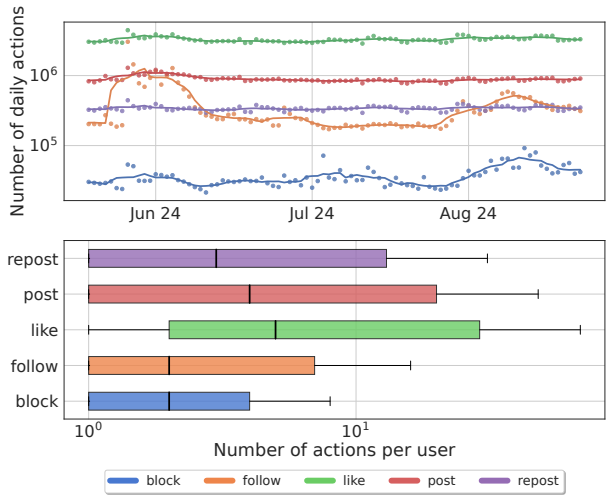

Figure 2: **(top)** Time series of the number of daily actions performed on Bluesky during the observation period. Daily observations and a 7-day moving average are reported. The scale of the y-axis is logarithmic. **(bottom)** Boxplot of the number of actions (log scale) performed by Bluesky users. Median values of each distribution in the bottom plot are: blocks = 2, follows = 2, likes = 8, posts = 6, reposts = 4.

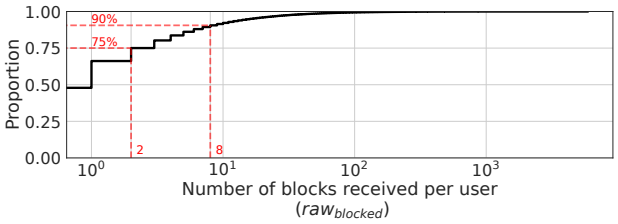

Figure 3: Empirical cumulative distribution of the number of blocks received by a user (log scale), considering users that shared at least 10 posts over the observation period.

trends, we observe a spike in follow actions in June, likely driven by a surge in the platform's popularity. Additionally, an increasing trend in follow and block actions can be observed toward the end of August, which can be attributed to growing political engagement ahead of the 2024 U.S. Presidential elections.

Bottom panel of Figure 2 illustrates the distribution of user actions on Bluesky during the observation period. As expected in online social networks, all distributions exhibit a power-law behavior, where the majority of users perform only a limited number of actions, while a small fraction of users exhibit significantly higher activity levels, with extreme outliers carrying out up to 74 000 blocks, 517 000 likes, 145 000 follow actions, 316 000 original posts, and 82 000 reposts. This suggests the presence of potentially misbehaving users, likely engaging in harassment, spam, or automated activity on the platform.

Our primary objective is to investigate and predict the

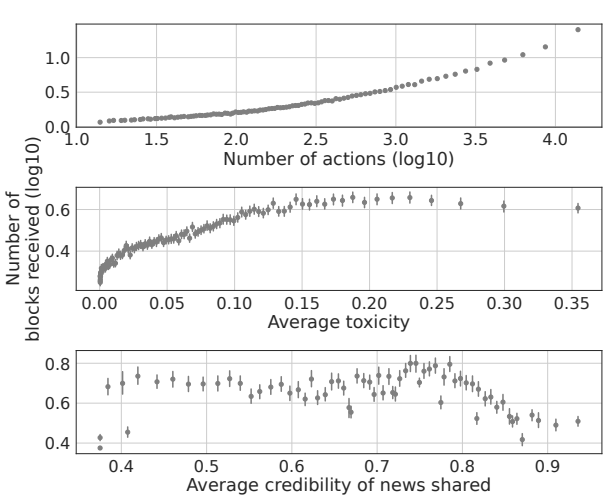

Figure 4: Correlation between different user features — number of actions, average toxicity and average credibility of news shared — and the number of blocks received ($\log_{10}$). We bin the x variable into 100 discrete bins and then estimate the mean value of the y variable with 95% confidence interval using bootstrapping. The analysis is performed on 427 118 users sharing at least 10 posts.

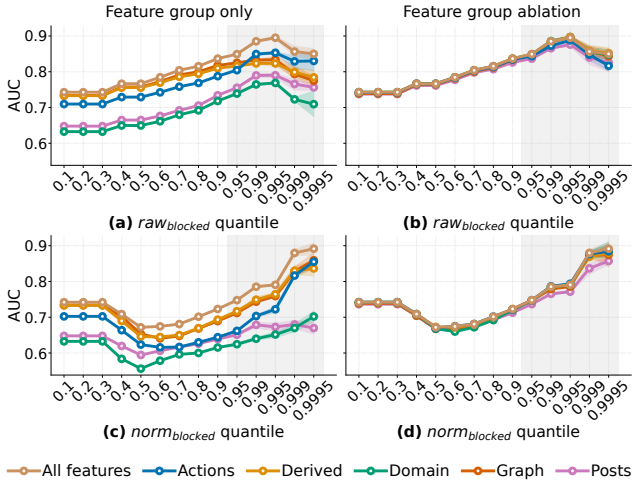

Figure 5: ROC AUC of the XGBoost classifier on the positive class (*blocked* users) for different quantiles of the **(a)** raw and **(c)** normalized block count, trained using *only* one subset of features at a time. Panels **(b, d)** show the performance when training models using all *except* one subset of features at a time. All panels also report the ROC AUC of a predictor trained with all the features, for comparison purposes. Labels in the legend indicate groups of features as defined in the Methods. The grey area highlights performance for the thresholds in the range $[0.9, 0.9995]$.

propensity of being *blocked* on Bluesky. As specified in the Methods, we focus on over 400k users who shared at least 10 posts over the observation period. Figure 3 shows the distribution of the number of blocks received by these users. We can observe that the large majority of users (90%) received less than 10 blocks, with a small but non-negligible proportion of users that received over 100 blocks ($\approx 0.6\%$). Figure 4 illustrates the relationship between the number of blocks ($\log_{10}$ scale) received by users and three features for user characterization: total actions performed, average toxicity of posted messages and average credibility of shared news URLs. A strong positive association, likely exponential given the logarithmic scale, between a user's total activity on the platform and the number of blocks received can be observed, suggesting that highly active users are more likely to be blocked (Pearson's $R = 0.55$, Spearman's $R = 0.53$), potentially due to misbehavior, spamming, or as a result of their increased visibility. Similarly, the average message toxicity of users is moderately correlated with the number of blocks received (Pearson's $R = 0.20$, Spearman's $R = 0.23$). In contrast, no clear relationship emerges between sharing less credible news sources and being blocked.

Classification task

As specified in the Methods, we formulate a binary classification task to assess how well *blocked* and *non-blocked* users can be distinguished. We perform the same experiments independently for the $raw_{blocked}$ and $norm_{blocked}$ measures. Users below a threshold are labeled as the negative class (*non-blocked*), while those above are labeled as the positive class (*blocked*). To evaluate the informativeness of features for this task, we train classifiers on different subsets of features, grouped by type.

Panels **a, c** of Figure 5 show the performance, in terms of ROC AUC, of the binary classifiers across different thresholds for $raw_{blocked}$ and $norm_{blocked}$, respectively. We evaluate a classifier for each feature group and compare its performance against using all features. Results are averaged over 10 different runs for each classifier, each run consisting of an independent random 10-fold cross-validation. We first observe that the classifier trained on all features consistently outperforms reduced models across thresholds in both settings, reaching max AUC values of 0.892 and 0.875, respectively. Performance improves with larger thresholds, though with distinct patterns. In the raw block count setting, AUC steadily increases up to the highest thresholds, after which it slightly declines (quantiles 0.995–0.9995). In contrast, in the normalized block count setting, performance dips at mid-range quantiles (0.3–0.7) before rising sharply, reaching its peak at the highest quantile (0.9995). These trends suggest that more extreme users are generally easier to distinguish, particularly when block counts are normalized by activity. In the same panels, we also observe that groups of behavioral features like Action, Derived, and Graph, are highly effective at predicting blocked users even when used in isolation, achieving AUC values up to 0.856 and 0.846 in the $raw_{blocked}$ and $norm_{blocked}$ settings respectively, and mimicking the overall trend of the classifier trained on all features. In contrast, features like Domain and Posts appear to be less informative, with the weakest performance observed in the $norm_{blocked}$ setting, where their AUC remains below 0.7.

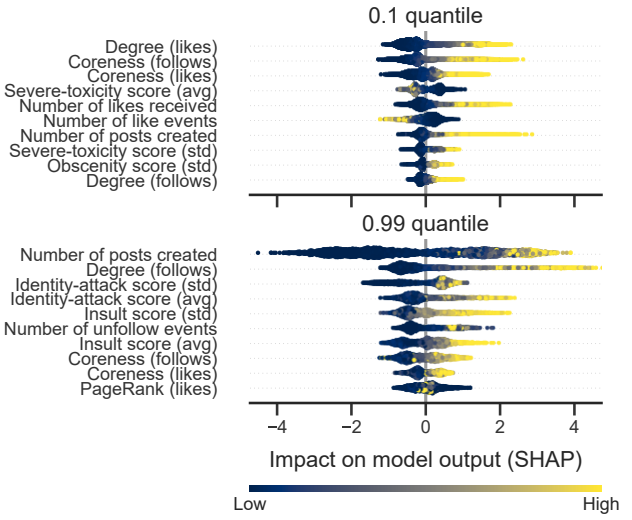

Figure 6: SHAP beeswarm plots comparing the 0.1 and 0.99 quantile setups, top 10 features reported.

To gain insight into the contribution of individual features to blocking, Figure 6 reports SHAP beeswarm plots for two representative thresholds (quantiles 0.1 and 0.99), showing the top 10 most influential features for each model. It is worth noting that, when focusing on more extreme users with the higher quantile, substantially larger SHAP magnitudes indicate stronger and more polarized feature effects, in line with the previous observations.

We further conducted a feature ablation study, training binary classifiers over the same thresholds while removing a single feature group at a time. As shown in panels **b, d** of Figure 5, most settings yield performance comparable to the full model, with the absence of Action or Post features causing the largest decline (≈ -5 p.p.). These results highlight a non-trivial interplay among features: for example, post-related features alone have limited predictive power, yet their removal from the full set produces a noticeable drop in performance. This effect is most evident at higher quantiles, while at lower quantiles, the absence of a feature group appears to be compensated by the remaining features. We examine the contribution of individual features in more detail in the next section.

Feature importance analysis

We investigated the importance of individual features in driving the classification performance using SHAP analysis as described in the methods.

Since feature importance depends on how *blocked* is defined, Figure 7 reports the union of the top-ranked features across thresholds for both classification settings, yielding nine features within the top eight per setting, indicating strong overlap among the most informative signals. Some features that are important at lower thresholds lose relevance for heavily blocked users, such as the degree in the *likes* graph and the number of *likes* received. Conversely, other features gain importance at higher thresholds. In the

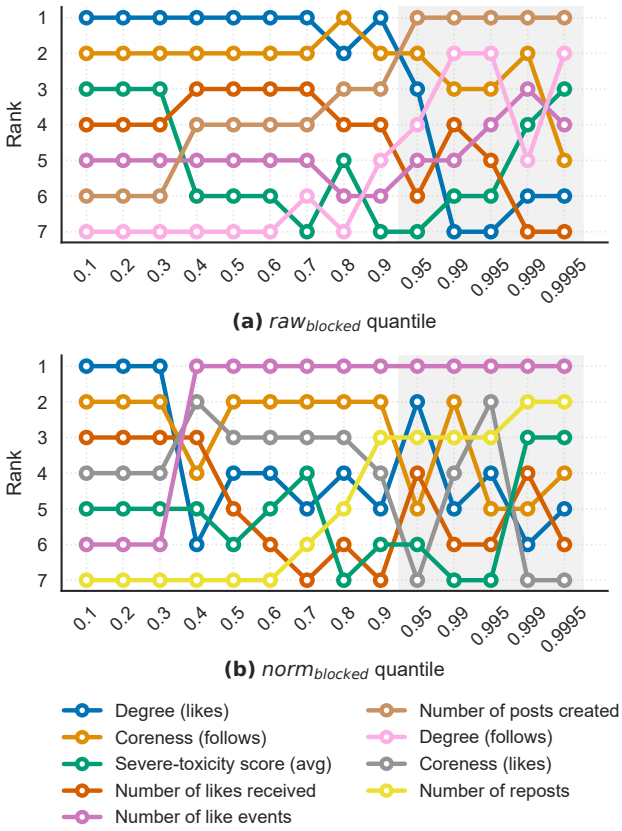

Figure 7: Ranking of the union of the top 8 features, resulting in 9 unique features, by SHAP importance in the raw block count setting **(top)** and the normalized block count setting **(bottom)**, thresholding on progressive quantiles.

raw block count setting (panel **a**), this includes degree in the *follows* graph, number of *likes*, and number of *posts*. In the normalized block count setting (panel **b**), the most discriminative features are the number of *likes*, number of *reposts*, and the *toxicity* score. These patterns suggest that frequently blocked users exhibit distinctive behavioral fingerprints, consistent with the classification results discussed in the previous section.

Figure 8 shows how feature importance at the group level varies with the threshold. Both settings exhibit similar patterns: Posts features are the most important groups on average (respectively $\approx 37\%$ and $\approx 28\%$). However, at higher thresholds, Graph features appear to be less discriminating. At the same time, as blocking behavior becomes more pronounced, Actions gain relative importance and Domain and Derived features become less relevant. This indicates, overall, that activity patterns and content are strong indicators of frequent block targets, especially for extreme users.

The classification results of the feature ablation experiment are reported in Figure 9, which contains four panels: the top row shows classifiers trained on the *best* n features (ranked by SHAP importance), and the bottom row shows classifiers trained on the *worst* n features; the left and right

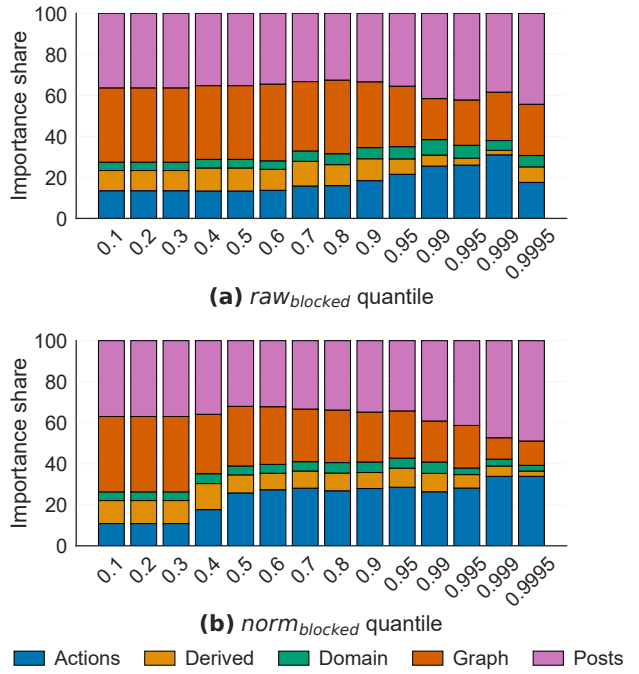

Figure 8: SHAP importance values aggregated by feature group across thresholds in the raw **(a)** and normalized block count setting **(b)**. Labels in the legend indicate groups of features as defined in the Methods.

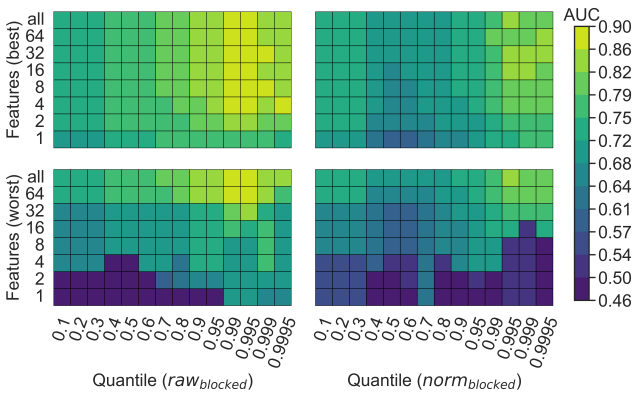

Figure 9: Classification performance with increasing numbers of features. The **top** row shows classifiers trained on the best n features (ranked by importance), and the **bottom** row on the worst n features; left and right columns correspond to the raw and normalized block count settings, respectively.

columns correspond to the raw and normalized block count settings, respectively. Ten classifiers are trained on independent random samples, and the average ROC AUC is reported. In the top row, results show that a limited set of features can achieve reasonable performance. At higher thresholds, as few as four features are sufficient to achieve near-maximum performance, with additional features contributing only marginally. In the bottom row, the opposite pattern

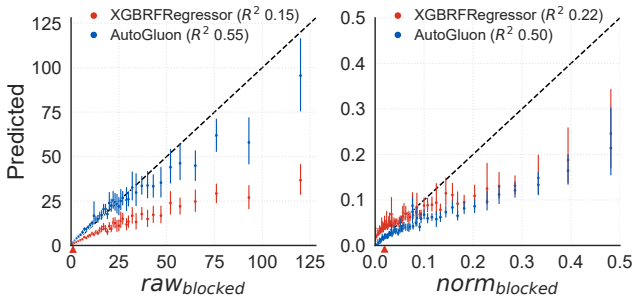

Figure 10: **(left)** Regression of raw block count per user, and **(right)** normalized block count, comparing XGBRFRegressor (red) and AutoGluon (blue). Markers correspond to percentile bins of the true values, the red triangle denotes the median.

can be observed: using only a few of the least informative features yields performance close to random classification. For both low and high thresholds, performance comparable to that of the top two features requires including at least 64 of the worst features. These results highlight the value of feature selection: a small subset of informative features is sufficient for robust predictions, while adding less informative features provides little benefit. At the same time, the degree of informativeness depends on the definition of *blocked* users.

Regression analysis

Given the strong classification performance, particularly for frequently blocked users, we next assess whether regression models can estimate the actual number of blocks received. Figure 10 reports results for the two targets — $raw_{blocked}$ (left) and $norm_{blocked}$ (right) block counts — using Random Forest Regression (red) and AutoGluon (blue).

Across both settings, AutoGluon consistently generalizes better than the baseline Random Forest model, though at a higher computational cost (about one hour of training versus under one minute for the baseline). Prediction accuracy is higher for users with fewer blocks, while both models underperform for the most heavily blocked users, a comparatively rare group, where estimates of $raw_{blocked}$ and $norm_{blocked}$ deviate more substantially from the true values. Nevertheless, the median values of both variables are very low, meaning that most users fall within the range where predictions are more reliable. To illustrate this practically, approximately 84% of users with zero $raw_{blocked}$ blocks are correctly predicted to receive fewer than one block.

Discussion

Contributions This work presents the first large-scale longitudinal study of blocking behavior on Bluesky, based on more than 3 million block actions carried out by nearly 2 million users over a three-month period. We develop an interpretable, ML-based framework for retrospective identification of blocking outcomes, providing a foundation for future early-warning models and for causal investigations into

the effects of blocking on user activity.

Our descriptive analysis shows that blocks, a less frequent interaction compared to likes, follows, and posts, are largely employed by Bluesky users and follow a heavy-tailed distribution. A small group of highly active or potentially misbehaving accounts stands out with extreme block-receiving activity. We also find that blocked users tend to be more active and moderately more toxic, though blocking does not appear to be systematically linked to sharing low-credibility content.

Building on these insights, we investigated the extent to which blocked and non-blocked users can be distinguished. We framed this as a binary classification task using two target definitions: raw block counts and activity-normalized block counts. We then used quantiles of their respective distributions as classification thresholds. A Random Forest classifier consistently achieved strong performance, with more extreme users being the easiest to identify, particularly in the normalized setting, suggesting that being blocked is associated with distinctive behavioral fingerprints. However, because interpretability is closely linked to model performance, insights for users in the lower block quantiles, where predictive accuracy is more limited, should be treated with caution. At the same time, the consistency of feature-importance patterns across thresholds provides additional evidence of robustness.

We further examined feature importance at group and individual levels using SHAP values, which provide locally consistent, model-agnostic explanations of complex ensemble predictions and are widely used in social computing research to interpret behavioral and content-based models (Shevtsov et al. 2022; Mathew et al. 2019; Yang et al. 2020). Results highlight how the relative importance of highly informative features shifts depending on the definition of *blocked* and the selected threshold. Moreover, we show that graph-based features are more predictive at lower thresholds, while posting activity and content become more important for heavily blocked users. We also discuss how high classification performance can be achieved using fewer features.

Finally, we modeled the number of blocks received as a regression task. AutoGluon outperformed Random Forest Regression in both raw and normalized settings, though at a greater computational cost. Both models perform reliably for the majority of users with few blocks but struggle to capture extreme cases. However, given that the median block counts are very low, this limitation is not a major concern in practice.

Our work draws on research in moderation, decentralized governance, and self-moderation dynamics. While prior studies (Vogels 2021; Seering 2020; Kaiser, Vaccari, and Chadwick 2022; Martel et al. 2024; Baysha 2020; Ashkinaze, Gilbert, and Budak 2024; Hunt et al. 2018; Fox and Moreland 2015; Zhang and Fu 2020) have examined user-driven moderation behaviors such as muting or unfollowing on platforms like Twitter, our work is the first to analyze behavioral patterns in relation to blocking on Bluesky at scale, enabled by the platform's unique data openness. By framing blocking as both a classification and a regression task, we show that behavioral traces can retrospectively predict blocking with meaningful accuracy, providing a methodological bridge toward future early-warning systems and causal analyses of how blocking shapes user behavior. In doing so, our study demonstrates how decentralized transparency can advance evidence-driven moderation research and practice.

Implications There are several implications of our findings. First, our study highlights that self-moderation in the form of user blocking is a widely used mechanism on Bluesky, suggesting that users actively shape their online environments by filtering interactions. We demonstrate that users who receive a high number of blocks exhibit distinctive behavioral traits that set them apart from the general user population, even when accounting for higher activity.

Second, these distinctive traits can be effectively encoded and leveraged by machine learning models, demonstrating not only that blocking outcomes are retrospectively predictable but also that such predictability provides a foundation for future early-warning or flagging systems that could help moderation teams surface potentially problematic users before issues escalate. Notably, comparable performance can be achieved with parsimonious models that rely on only a handful of features, opening avenues for lightweight, privacy-preserving implementations and for future work aimed at understanding the causal pathways linking user behavior and blocking outcomes.

An important implication of our study lies in the value of data transparency. Our analysis was made possible by the open nature of Bluesky's data ecosystem, which provides access to behavioral and moderation-related data. Other major social platforms also employ blocking and similar self-moderation mechanisms, yet such data are typically inaccessible to researchers, even on other decentralized platforms like Mastodon. Broader data sharing would significantly advance the scientific understanding of online social dynamics and enable collaborative, evidence-based approaches to moderation. Platforms could benefit from shared insights into behavioral markers of problematic interactions, while researchers and policymakers could better assess the effectiveness and consequences of moderation strategies. In this sense, Bluesky serves as a compelling case study for how openness can catalyze innovation in trust and safety research.

Limitations The analysis has been conducted over a three-month period, which may not capture long-term trends or account for seasonal variations in user behavior. A limited time frame could miss important shifts in blocking activity that might emerge over a longer observation period. Moreover, we only analyze users' activity from a static perspective, i.e., by aggregating their behaviour over the entire period of analysis.

Our study focused on a large but simple set of behavioral features to analyze blocking activity, potentially overlooking other influential factors. Incorporating more advanced features, such as text embeddings from social media posts or graph embeddings representing user interaction networks, could enhance the predictive power of the models. Alterna-

tive toxicity assessment strategies could also be explored, differentiating multiple information sources (e.g., replies and reposts) and considering additional languages.

While this study examines the patterns and characteristics of users who are blocked, it does not explore the underlying motivations behind these actions. Understanding why users block others or get blocked — whether due to harassment, misinformation, or other factors — could provide more context and actionability for the findings.

Lastly, due in part to technical constraints and in part to the current demographics of the social network, this study is limited to users who predominantly interact in the English language, which may not fully represent global user behavior. Language, cultural norms, and social dynamics can significantly influence how users interact and block others. Moreover, the Bluesky user base itself is not necessarily representative of other platforms or countries, limiting the generalizability of our findings beyond this specific context.

Future Work A key area for future research is related to the underlying motivations behind user blocking behavior. While our study focused on estimating the number of blocks received, it did not investigate the reasons users engage in blocking actions. Understanding the psychological, social, or contextual factors that drive users to block others — such as harassment, misinformation, or disagreements — could provide insights into the root causes of blocking behavior. A promising direction would be to incorporate data from Bluesky's Labelers, community- or service-run classifiers that assign safety, content, or policy-relevant labels to posts and accounts, including the one operated by the platform itself. These labels could provide valuable ground truth signals, complementary to toxicity and media-bias annotations, enriching our analysis. A qualitative analysis of the most frequently blocked users could reveal more about who these users are and the specific use cases — such as harassment, misinformation, or conflict patterns — that quantitative features alone cannot capture, informing more targeted moderation strategies. This would help develop more accurate models of user behavior and enable platform developers to create more sophisticated systems for managing toxic or disruptive interactions.

Our models demonstrate that it is possible to estimate users' overall propensity to be blocked based on their behavioral features. However, we do not attempt to predict the exact number of blocks a user may receive in the future based on past activity; this remains an avenue for future work. Moreover, in the current formulation, features are not normalized by account age, meaning that users who joined during the observation window may have shorter behavioral histories; incorporating appropriate countermeasures could further refine future modeling efforts.

Another direction for future research involves applying link prediction techniques to understand the network dynamics associated with blocking behavior. Link prediction aims to forecast the likelihood of future connections among specific users, based on their past interactions and respective positions within the network structure. By extending the proposed approach, we could predict which users are more likely to block or be blocked by others in the future. This would provide a deeper understanding of social networks and interactions on platforms, allowing for proactive identification of potential conflicts before they escalate into blocking actions, ultimately improving user experience.

Finally, future research would benefit from developing a causal framework to understand the effects of blocking on user behavior. While our study identifies patterns associated with users who block others, it does not establish a causal relationship between blocking actions and the subsequent impact on user interactions or overall platform engagement. Investigating how blocking influences users' behavior — changes in their activity, sentiment, or social connections — might shed light on the broader consequences of blocking. This would enable platform designers to implement more effective tools and interventions to mitigate the negative effects of blocking while fostering healthier user interactions.

Ethical Statement

This study uses publicly available Bluesky data, in line with its terms of service. We ensure privacy by avoiding the identification of individuals, and we only report aggregated, non-identifying results. To support transparency and reproducibility, we will release the code necessary to replicate our analyses and experiments.

This study also recognizes that predicting blocking behavior poses ethical risks, including potential misuse for evading moderation or user targeting. We emphasize that any labeling systems must be applied cautiously to avoid user stigmatization, exclusion, or discrimination (Jhaver et al. 2018). We also acknowledge that harmful behavior is often context-dependent, emerging from situational and interpersonal dynamics rather than fixed user traits (Cheng et al. 2017), and may be misinterpreted by automated systems, which may reinforce biases. Therefore, we emphasize the need for safeguards, including user feedback mechanisms and regular algorithmic bias audits, in any real-world deployment.

We further acknowledge that the public visibility of blocking actions on Bluesky, which is not typical among other platforms, raises ethical concerns, such as potential stigmatization and surveillance. We emphasize that the findings are context-specific and should guide scientific understanding and responsible system design, with safeguards to protect users and prevent harm.

References

Ali, S.; Saeed, M. H.; Aldreabi, E.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2021. Understanding the Effect of Deplatforming on Social Networks. In *13th ACM Web Science Conference 2021*, WebSci '21, 187–195. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383301.

Ashkinaze, J.; Gilbert, E.; and Budak, C. 2024. The dynamics of (not) unfollowing misinformation spreaders. In *Proceedings of the ACM Web Conference 2024*, 1115–1125.

AT Protocol. 2024. Event Stream. https://atproto.com/specs/event-stream.

Attanasio, A.; Corso, F.; Morales, G. D. F.; and Pierri, F. 2026. Effects of Mainstream Visibility on Conspiracy Communities: Reddit after Epstein's' Suicide'. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Balduf, L.; Sokoto, S.; Ascigil, O.; Tyson, G.; Scheuermann, B.; Korczyński, M.; Castro, I.; and Król, M. 2024. Looking AT the Blue Skies of Bluesky. In *Proceedings of the 2024 ACM on Internet Measurement Conference*, IMC '24, 76–91. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705922.

Balduf, L.; Sokoto, S.; Baronchelli, A.; Castro, I.; Król, M.; Tyson, G.; Pavlou, G.; Scheuermann, B.; and Ascigil, O. 2025. Bootstrapping Social Networks: Lessons from Bluesky Starter Packs. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 178–192.

Baribi-Bartov, S.; Swire-Thompson, B.; and Grinberg, N. 2024. Supersharers of fake news on Twitter. *Science*, 384(6699): 979–982.

Baysha, O. 2020. Dividing social networks: Facebook unfriending, unfollowing, and blocking in turbulent political times. *Russian Journal of communication*, 12(2): 104–120.

Bluesky. 2024. Bluesky — Dart package. Software library. Accessed: 2024-04.

Bluesky. 2024. Firehose. https://docs.bsky.app/docs/advanced-guides/firehose. Accessed: 2024-04-29.

Bonifazi, G.; Breve, B.; Cirillo, S.; Corradini, E.; and Virgili, L. 2022. Investigating the COVID-19 vaccine discussions on Twitter through a multilayer network-based approach. *Information Processing & Management*, 59(6): 103095.

Bono, C. A.; La Cava, L.; Luceri, L.; and Pierri, F. 2024. An Exploration of Decentralized Moderation on Mastodon. In *Proceedings of the 16th ACM Web Science Conference*, WEBSCI '24, 53–58. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703348.

Burghardt, K. 2024. Data collection with Bluesky Firehose endpoint. https://github.com/KeithBurghardt/bluesky_firehose/tree/main. Accessed: 2024-04.

Chandrasekharan, E.; Jhaver, S.; Bruckman, A.; and Gilbert, E. 2022. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29(4): 1–26.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on human-computer interaction*, 1(CSCW): 1–22.

Cheng, J.; Bernstein, M.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 1217–1230.

Chuai, Y.; Tian, H.; Pröllochs, N.; and Lenzini, G. 2023. The roll-out of community notes did not reduce engagement with misinformation on Twitter. *arXiv e-prints*, arXiv–2307.

Copland, S. 2020. Reddit quarantined: Can changing platform affordances reduce hateful material online? *Internet Policy Review*, 9(4): 1–26.

Di Giovanni, M.; Pierri, F.; Torres-Lugo, C.; and Brambilla, M. 2022. VaccinEU: COVID-19 vaccine conversations on Twitter in French, German and Italian. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1236–1244.

Drolsbach, C. P.; Solovev, K.; and Pröllochs, N. 2024. Community notes increase trust in fact-checking on social media. *PNAS nexus*, 3(7): pgae217.

Erickson, N.; Mueller, J.; Shirkov, A.; Zhang, H.; Larroy, P.; Li, M.; and Smola, A. 2020. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *arXiv preprint arXiv:2003.06505*.

Failla, A.; and Rossetti, G. 2024. "I'm in the Bluesky Tonight": Insights from a year worth of social data. *PLOS One*, 19(11): e0310330.

Fox, J.; and Moreland, J. J. 2015. The dark side of social networking sites: An exploration of the relational and psychological stressors associated with Facebook use and affordances. *Computers in human behavior*, 45: 168–176.

Ghenai, A.; Noorian, Z.; Moradisani, H.; Abadeh, P.; Erentzen, C.; and Zarrinkalam, F. 2025. Exploring hate speech dynamics: The emotional, linguistic, and thematic impact on social media users. *Information Processing & Management*, 62(3): 104079.

Guidi, B. 2020. When blockchain meets online social networks. *Pervasive and Mobile Computing*, 62: 101131.

Hanu, L.; and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Horta Ribeiro, M.; Jhaver, S.; Zannettou, S.; Blackburn, J.; Stringhini, G.; De Cristofaro, E.; and West, R. 2021. Do platform migrations compromise content moderation? Evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–24.

Hunt, M. G.; Marx, R.; Lipson, C.; and Young, J. 2018. No more FOMO: Limiting social media decreases loneliness and depression. *Journal of Social and Clinical Psychology*, 37(10): 751–768.

Innes, H.; and Innes, M. 2023. De-platforming disinformation: conspiracy theories and their control. *Information, Communication & Society*, 26(6): 1262–1280.

Jhaver, S.; Boylston, C.; Yang, D.; and Bruckman, A. 2021. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on human-computer interaction*, 5(CSCW2): 1–30.

Jhaver, S.; Ghoshal, S.; Bruckman, A.; and Gilbert, E. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2): 1–33.

Kaiser, J.; Vaccari, C.; and Chadwick, A. 2022. Partisan blocking: Biased responses to shared misinformation contribute to network polarization on social media. *Journal of Communication*, 72(2): 214–240.

Kato, S. 2024. Bluesky Firehose endpoint. https://atprotodart.com/docs/lexicons/com/atproto/sync/subscriberepos/. Accessed: 2024-04.

Kleppmann, M.; Frazee, P.; Gold, J.; Graber, J.; Holmgren, D.; Ivy, D.; Johnson, J.; Newbold, B.; and Volpert, J. 2024. Bluesky and the AT protocol: Usable decentralized social media. In *Proceedings of the ACM Conext-2024 Workshop on the Decentralization of the Internet*, 1–7.

Lin, H.; Lasser, J.; Lewandowsky, S.; Cole, R.; Gully, A.; Rand, D. G.; and Pennycook, G. 2023. High level of correspondence across different news domain quality rating sets. *PNAS Nexus*, 2(9): pgad286.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4768–4777. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Martel, C.; Mosleh, M.; Yang, Q.; Zaman, T.; and Rand, D. G. 2024. Blocking of counter-partisan accounts drives political assortment on Twitter. *PNAS nexus*, 3(5): 161.

Mathew, B.; Saha, P.; Tharad, H.; Rajgaria, S.; Singhania, P.; Maity, S. K.; Goyal, P.; and Mukherjee, A. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, 369–380.

Newman, M. 2010. *Networks: An Introduction*. Oxford: Oxford University Press. ISBN 978-0-19-920665-0.

Nogara, G.; Sahneh, E. S.; DeVerna, M. R.; Liu, N.; Luceri, L.; Menczer, F.; Pierri, F.; and Giordano, S. 2026. A longitudinal analysis of misinformation, polarization and toxicity on Bluesky after its public launch. *Online Social Networks and Media*.

Pierri, F. 2024. Drivers of hate speech in political conversations on Twitter: the case of the 2022 Italian general election. *EPJ Data Science*, 13(1): 63.

Quelle, D.; and Bovet, A. 2025. Bluesky: Network topology, polarization, and algorithmic curation. *PLOS One*, 20(2): e0318034.

Raman, A.; Joglekar, S.; Cristofaro, E. D.; Sastry, N. R.; and Tyson, G. 2019. Challenges in the Decentralised Web: The Mastodon Case. *Proc. Internet Measurement Conference*.

Rathje, S.; Pretus, C.; He, J. K.; Harjani, T.; Roozenbeek, J.; Gray, K.; van der Linden, S.; and Van Bavel, J. J. 2024. Unfollowing hyperpartisan social media influencers durably reduces out-party animosity. *Preprint at https://osf. io/acbwg/download*.

Ribeiro, M. H.; Jhaver, S.; Reignier-Tayar, M.; West, R.; et al. 2024. Deplatforming norm-violating influencers on social media reduces overall online attention toward them. *arXiv preprint arXiv:2401.01253*.

Rogers, R. 2020. Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, 35(3): 213–229.

Sahneh, E. S.; Nogara, G.; DeVerna, M. R.; Liu, N.; Luceri, L.; Menczer, F.; Pierri, F.; and Giordano, S. 2024. The dawn of decentralized social media: an exploration of Bluesky's public opening. In *International Conference on Advances in Social Networks Analysis and Mining*, 422–437. Springer.

Schneider, P. J.; and Rizoiu, M.-A. 2023. The effectiveness of moderating harmful online content. *Proceedings of the National Academy of Sciences*, 120(34): e2307360120.

Seering, J. 2020. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–28.

Shen, Q.; and Rosé, C. P. 2022. A tale of two subreddits: Measuring the impacts of quarantines on political engagement on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 932–943.

Shevtsov, A.; Tzagkarakis, C.; Antonakaki, D.; and Ioannidis, S. 2022. Identification of twitter bots based on an explainable machine learning framework: The US 2020 elections case study. In *Proceedings of the international AAAI conference on web and social media*, volume 16, 956–967.

Solovev, K.; and Pröllochs, N. 2025. References to unbiased sources increase the helpfulness of community fact-checks. *arXiv preprint arXiv:2503.10560*.

Trujillo, A.; and Cresci, S. 2022. Make Reddit great again: assessing community effects of moderation interventions on r/the_donald. *Proceedings of the ACM on Human-computer Interaction*, 6(CSCW2): 1–28.

Vogels, E. A. 2021. *The state of online harassment*, volume 13. Pew Research Center Washington, DC.

Yang, K.-C.; Varol, O.; Hui, P.-M.; and Menczer, F. 2020. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 1096–1103.

Young, G. K. 2022. How much is too much: the difficulties of social media content moderation. *Information & Communications Technology Law*, 31(1): 1–16.

Zannettou, S. 2021. "I Won the Election!": an empirical analysis of soft moderation interventions on Twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 15, 865–876.

Zhang, R.; and Fu, J. S. 2020. Privacy management and self-disclosure on social network sites: The moderating effects of stress and gender. *Journal of Computer-Mediated Communication*, 25(3): 236–251.

Zhang, X.; and Gao, W. 2024. Predicting viral rumors and vulnerable users with graph-based neural multi-task learning for infodemic surveillance. *Information Processing & Management*, 61(1): 103520.

Zia, H. B.; He, J.; Raman, A.; Castro, I.; Sastry, N.; and Tyson, G. 2023. Flocking to Mastodon: Tracking the great Twitter migration. *arXiv preprint arXiv:2302.14294*.

Zignani, M.; Gaito, S.; and Rossi, G. P. 2018. Follow the "Mastodon": Structure and Evolution of a Decentralized Online Social Network. In *International Conference on Web and Social Media*.

Zuckerman, E.; and Rajendra-Nicolucci, C. 2021. Deplatforming Our Way to the Alt-Tech Ecosystem. *Knight First Amendment Institute at Columbia University, January*, 11.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, in "Data and Methods".
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we discuss them in the "Discussion".
   (e) Did you describe the limitations of your work? Yes, we discuss them in the "Discussion".
   (f) Did you discuss any potential negative societal impacts of your work? We discuss negative societal impact in "Ethical Impact".
   (g) Did you discuss any potential misuse of your work? Yes, we discuss potential misuse in "Ethical Impact".
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we provide full details on the data and models employed in the research.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA.
   (b) Did you include complete proofs of all theoretical results? NA.
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. The code to reproduce the analyses will be provided in a GitHub repository.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, we provide details in the "Dataset and Methods" section.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, we report 95% confidence intervals.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No, we did not.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, we provide details in the "Dataset and Methods" section.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, we provide details in the "Dataset and Methods" section.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes.
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, we discuss it in the "Ethical impact" section.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we discuss it in the "Ethical impact" section.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA.
   (d) Did you discuss how data is stored, shared, and deidentified? NA.

Appendix

Feature Descriptions

Table 2 provides a complete list of the 86 behavioral features used in this work, computed at the user level.

Group	#	Feature	Description
Action / Derived	1	create.app.bsky.feed.like	Number of like events created
	2	create.app.bsky.feed.post	Number of post events created
	3	create.app.bsky.feed.repost	Number of repost events created
	4	create.app.bsky.graph.follow	Number of follow events created
	5	delete.app.bsky.feed.like	Number of like events deleted
	6	delete.app.bsky.feed.post	Number of post events deleted
	7	delete.app.bsky.feed.repost	Number of repost events deleted
	8	delete.app.bsky.graph.follow	Number of follow events deleted
	9	derived.followed	Number of users following a user
	10	derived.follower	Number of users followed by a user
	11	derived.liked	Number of likes received
	12	derived.liker	Number of likes assigned
	13	derived.replied_parent	Number of replies received
	14	derived.replier_parent	Number of times replied
	15	derived.reposted	Number of reposts received
	16	derived.reposter	Number of times reposted
Post-derived	17	posts.count	Total number of posts (supported languages)
	18	posts.language_entropy	Entropy of post languages
	19	posts.num_chars_mean	Mean characters per post
	20	posts.num_chars_std	Std characters per post
	21	posts.num_digits_mean	Mean digits per post
	22	posts.num_digits_std	Std digits per post
	23	posts.num_emoji_mean	Mean emojis per post
	24	posts.num_emoji_std	Std emojis per post
	25	posts.num_lowercase_mean	Mean lowercase characters
	26	posts.num_lowercase_std	Std lowercase characters
	27	posts.num_uppercase_mean	Mean uppercase characters
	28	posts.num_uppercase_std	Std uppercase characters
	29	posts.num_spaces_mean	Mean spaces per post
	30	posts.num_spaces_std	Std spaces per post
	31	posts.identity_attack_mean	Mean “Identity attack” score (Detoxify)
	32	posts.identity_attack_std	Std “Identity attack” score (Detoxify)
	33	posts.insult_mean	Mean “Insult” score (Detoxify)
	34	posts.insult_std	Std “Insult” score (Detoxify)
	35	posts.obscene_mean	Mean “Obscene” score (Detoxify)
	36	posts.obscene_std	Std “Obscene” score (Detoxify)
	37	posts.threat_mean	Mean “Threat” score (Detoxify)
	38	posts.threat_std	Std “Threat” score (Detoxify)
	39	posts.sexual_explicit_mean	Mean “Sexual explicit” score (Detoxify)
	40	posts.sexual_explicit_std	Std “Sexual explicit” score (Detoxify)
	41	posts.severe_toxicity_mean	Mean “Severe toxicity” score (Detoxify)
	42	posts.severe_toxicity_std	Std “Severe toxicity” score (Detoxify)
	43	posts.toxicity_mean	Mean “Toxicity score” (Detoxify)
	44	posts.toxicity_std	Std “Toxicity score” (Detoxify)

Continued on next page

Table 2 – continued

Group	#	Feature	Description
Domain / URL-derived	45	domain.url_count	Average number of URLs per post
	46	domain.urls_entropy	Domain entropy
	47	domain.bias_entropy	Bias value entropy (MBFC)
	48	domain.bias_center	Fraction of “center” domains (MBFC)
	49	domain.bias_conspiracy	Fraction of “conspiracy” domains (MBFC)
	50	domain.bias_extreme-left	Fraction of “extreme-left” domains (MBFC)
	51	domain.bias_extreme-right	Fraction of “extreme-right” domains (MBFC)
	52	domain.bias_left	Fraction of “left” domains (MBFC)
	53	domain.bias_left-center	Fraction of “left-center” domains (MBFC)
	54	domain.bias_right	Fraction of “right” domains (MBFC)
	55	domain.bias_right-center	Fraction of “right-center” domains (MBFC)
	56	domain.bias_satire	Fraction of “satire” domains (MBFC)
	57	domain.bias_pro-science	Fraction of “pro-science” domains (MBFC)
	58	domain.credibility_high	Fraction of “high credibility” domains (MBFC)
	59	domain.credibility_medium	Fraction of “medium credibility” domains (MBFC)
	60	domain.credibility_low	Fraction of “low credibility” domains (MBFC)
	61	domain.factual_very_high	Fraction of “very high factual” domains (MBFC)
	62	domain.factual_high	Fraction of “high factual” domains (MBFC)
	63	domain.factual_mostly	Fraction of “mostly factual” domains (MBFC)
	64	domain.factual_mixed	Fraction of “mixed factual” domains (MBFC)
	65	domain.factual_low	Fraction of “low factual” domains (MBFC)
	66	domain.factual_very_low	Fraction of “very low factual” domains (MBFC)
	67	domain.bias_count	Count of domains with a bias value (MBFC)
	68	domain.credibility_count	Count of domains with a Credibility value (MBFC)
	69	domain.factual_count	Count of domains with a Factual value (MBFC)
	70	domain.pc1_count	Domain quality score count (Lin et al. 2023)
	71	domain.pc1_mean	Domain quality score mean (Lin et al. 2023)
Graph-based	72	graph.follows_coreness	Coreness in follows network
	73	graph.follows_degree	Degree in follows network
	74	graph.follows_pagerank	PageRank in follows network
	75	graph.likes_coreness	Coreness in likes network
	76	graph.likes_degree	Degree in likes network
	77	graph.likes_pagerank	PageRank in likes network
	78	graph.likes_strength	Weighted degree in likes network
	79	graph.replies_parent_coreness	Coreness in replies network
	80	graph.replies_parent_degree	Degree in replies network
	81	graph.replies_parent_pagerank	PageRank in replies network
	82	graph.replies_parent_strength	Weighted strength in replies network
	83	graph.reposts_coreness	Coreness in reposts network
	84	graph.reposts_degree	Degree in reposts network
	85	graph.reposts_pagerank	PageRank in reposts network
	86	graph.reposts_strength	Weighted strength in reposts network

Table 2. Description of the behavioral features, grouped by type.