Open or Blocked Skies? Community Moderation Practices in Bluesky

Saidu Sokoto
City St George's, Univ. of London
London, United Kingdom

Leonhard Balduf
TU Darmstadt
Darmstadt, Germany

Onur Ascigil
Lancaster University
Lancaster, United Kingdom

Gareth Tyson*
Hong Kong University of Science and Technology (GZ)
Guangzhou, China

Ignacio Castro
Queen Mary Univ. of London
London, United Kingdom

Björn Scheuermann
TU Darmstadt
Darmstadt, Germany

Andrea Baronchelli
City St George's, Univ. of London
London, United Kingdom

Michał Król
City St George's, Univ. of London
London, United Kingdom

Abstract

Content moderation is a major challenge for online platforms. While user-driven blocking is a common tool, its dynamics are usually hidden as moderation data is private. Bluesky makes moderation actions public-by-design, providing an unprecedented opportunity to study a community-driven moderation ecosystem at scale. We leverage this transparency to (1) map the ecosystem of moderation blocking actions across 34 million users, including both individual blocks and the through blocklists, (2) identify the signals that correlate with blocking, and (3) measure the consequences of these actions. We demonstrate that community blocking is widespread, with a volume several orders of magnitude higher than official takedowns, and affects the visibility of more than 90% of Bluesky content. The blocked accounts represent the most active, popular, toxic, and politically inclined users. However, different blocklists target different types of accounts and behaviors. Finally, blocking does not decrease the popularity and activity of the blocked users and has a limited effect on the social graph. By quantifying its dynamics and trade-offs, our study provides empirical grounding for designing future moderation systems that are transparent, pluralistic, and resistant to centralized control. Taken together, this study provides the first large-scale, quantitative analysis of a community-driven moderation ecosystem, demonstrating how individual and collective interventions influence user behavior.

CCS Concepts

• **Human-centered computing** → **Social networks**; **Empirical studies in collaborative and social computing**.

*Also affiliated with Queen Mary Univ. of London.

WWW '26, Dubai, United Arab Emirates

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792106

Keywords

Bluesky, User-driven moderation, Community moderation, Community blocklists, Decentralized social networks

ACM Reference Format:
Saidu Sokoto, Leonhard Balduf, Onur Ascigil, Gareth Tyson, Ignacio Castro, Björn Scheuermann, Andrea Baronchelli, and Michał Król. 2026. Open or Blocked Skies? Community Moderation Practices in Bluesky. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3774904.3792106

Resource Availability:
Source code of this paper has been made publicly available at https://doi.org/10.5281/zenodo.18351518.

1 Introduction

Online social platforms face persistent challenges in moderating harmful content. On centrally managed networks such as X/Twitter, moderation is typically enforced by the platform itself. However, such approaches have raised concerns about transparency, fairness, and scalability [18, 30]. Moderators must balance user safety, freedom of expression, and platform integrity, often without full visibility into user interactions or the consequences of their interventions. Recent research has highlighted that platform-driven enforcement can inadvertently suppress legitimate speech, reinforce biases, or create uneven experiences for different user communities [12, 25].

Bluesky is a recent, yet rapidly growing, addition to the social media landscape [3]. Espousing a more open and decentralized culture, it has attempted to introduce less central forms of moderation [27]. Most notably, their decentralized "labelers" have gained much attention. These allow anybody to label posts (*e.g.* as hate speech), such that others can then use such labels for content filtering [35]. Through this, Bluesky seeks to restrict the use of platform-level moderation to a minimal set of the most clear-cut or extreme cases (*e.g.* those mandated by law).

Prior works, however, have ignored an arguably more important decentralized moderation strategy supported by Bluesky. On September 12th, 2023, Bluesky introduced blocklists, supplementing the existing ability for users to block individual users. These lists

delegate the moderation tasks to the community, allowing anybody to create a blocklist and add accounts to it. Other users can subscribe to these blocklists, which results in them filtering out any content produced by members of the list (and unfollowing them). This mechanism enables users to delegate moderation decisions to others, building a foundation for community-based moderation. Although Bluesky is not the first platform to employ community blocklists, its open platform allows us to study the creation, use and impact of this moderation paradigm at scale for the first time.

Therefore, this paper presents the first longitudinal empirical study of blocking in Bluesky (§2). To achieve this, we gather a comprehensive dataset encompassing the entire Bluesky lifetime. This includes 1.2 billion posts from 34 million users (§3). Leveraging this dataset, we answer the following research questions:

- **RQ1: Mapping the Ecosystem.** What is the scale and nature of individual blocking versus blocklists on Bluesky, and what are the distinct list characteristics?
- **RQ2: Blocked User Behaviors.** What are the user's behavior and content features that correlate with users being blocked?
- **RQ3: Measuring Causal Impact.** What is the effect of being added to a blocklist on the user's subsequent activity? What effects does blocking have on reducing harmful exposure, and does it also fragment communities or reinforce echo chambers?

We identify a substantial community moderation ecosystem (§4). Overall, 12.6 % of users have blocked at least one other person, and 27.6 % of users have been blocked by at least one other individual, creating > 119 million block records. Yet, this makes up the minority of blocks. While community blocklists target only 8.78% of users and are subscribed to by 2.52% of users, they create > 16 billion unique block edges, ≈ 100× more than individual blocking, although these edges are not constrained to be unique. Yet, the blocklists are managed by only 0.043% of all the users, raising concerns about trust and re-centralization. Overall, we discover 39,589 community blocklists split between three categories: political and ideological, behavioral moderation, and content-based filtering.

This leads us to explore the characteristics of blocked users (§5). We perform this across a number of dimensions. We find that blocked users have markedly higher rates of toxic content posting. On average, non-blocked users create just 5 posts with a mean toxicity score above 0.29 (labeled using Detoxify [20]). In contrast, individually blocked posts 10× more often, with a mean toxicity score of 0.47, while users on blocklists write 72× more posts with a mean toxicity score of 0.66. Similarly, we find that the topics discussed also differ, with topics such as politics, adult content and crime being over-represented for those who go on to be blocked. Given these clear trends, we next systemize the differences between blocked vs. non-blocked users by training several models. We first develop a model to predict the likelihood of a user being blocked globally. While achieving a decent accuracy (F1 0.72), feature importance inspection shows a risk of reinforcing popularity or visibility biases instead of addressing context-dependent harms. At the same time, blocklist-specific models achieve much better accuracy and better respond to heterogeneous community norms.

Finally, we inspect the impact of the blocking (§6). We employ Generalized Propensity Score (GPS) to quantify the differences in behavior exhibited by users who are blocked. Surprisingly, community blocking can increase the popularity, activity, and toxicity of the targeted users and has a limited impact on the social graph. This suggests that community blocking can be an effective way of shaping users' social environment without silencing others.

2 Background

Bluesky is a decentralized social network built on the Authenticated Transfer Protocol (ATProto), a federated architecture designed to support a marketplace of interoperable services and clients [3, 27]. Bluesky was launched on February 1st, 2023, and opened registrations to the public on February 6th, 2024. The system's design decomposes core social media functions into distinct, community-operable components. Here, we introduce the components most relevant to our data collection.

User Repositories and Personal Data Servers (PDSes). All of a user's data — including their posts, likes, follows, and moderation actions such as blocks — is stored in a signed, version-controlled data repository. Each repository is hosted on a PDS, which a user can choose or operate themselves. A key feature of the ATProto is that for the network to function in a federated manner, these data repositories must be public. This architectural decision makes all user-generated records, including the creation of blocklists and individual block actions, publicly available for analysis.

The Firehose. To facilitate global interoperability and data access, Bluesky provides the "Firehose", an aggregated, real-time event stream. This endpoint subscribes to the commit streams from all federated PDS instances and republishes them as a single feed.

Labels & filtering. The ATProto provides a native labeling system that allows any user or service to apply metadata labels to any piece of content or user account on the network. This system is the foundation for Bluesky's community-driven moderation stack. It supports `selfLabels`, which authors apply to their own posts as content warnings, as well as labels applied by community moderators or the platform itself. These explicit, community-ascribed judgments are a core component of our feature set.

Blocking & Blocklists. To allow users to better control their feed, Bluesky introduced individual blocking on April 29th, 2023 and blocklists on September 12th, 2023. Individual blocking allows users to specifically target accounts they wish to block. In contrast, blocklists (which list multiple accounts) can be created and shared publicly. These lists, defined as a type of list record within a user's repository, contain a set of members who are to be blocked. Other users can then subscribe to these lists to adopt the same blocks. The public nature of these lists, including both their members and subscribers, allows for a detailed analysis of community-level moderation and its network effects.

3 Datasets

Bluesky Activity. We extract a complete snapshot of the network on April 14th, 2025 containing 34,251,851 Bluesky users. We download every user's repository from their respective PDS. This covers: 32,170,299 user profiles with their description and all user activity, including 6,938,743,344 likes, 1,290,686,604 posts, 888,094,540 reposts, and 2,156,778,700 follows. Furthermore, we are connected to

all known labelers to collect all labels issued by all 307 active labelers. Our dataset contains all 12,312,406 labels emitted by the official Labelers and 34,515,016 labels produced by the community. We also extract self-applied labels attached to posts by their authors.

Political Stance. We use a zero-shot classification model [11] to classify users as politically left- or right-leaning. As the model enforces a maximum input of 512 tokens, we use the profile information (display name and description) and a random sample of up to 10 posts per user as input. The model returns numerical probability scores for three hypotheses corresponding to left-leaning, right-leaning, and neutral political stances. We then assign a final categorical label by applying a confidence threshold of 0.5. Because the model works only for English, we translate all the non-English posts using the NLLB-200 model [16]. We provide additional implementation details in Appendix A.2.

Toxicity. We calculate the toxicity of every post, using the multilingual *Detoxify* model [20]. The model supports English, French, Spanish, Italian, Portuguese, Turkish, and Russian, covering 74% of all the Bluesky posts in our dataset. For each of these posts, the model outputs probability scores for seven categories of harmful content: identity attack, insult, obscene, toxicity, severe toxicity, threat, and sexually explicit. For each user, we compute the mean and standard deviation of scores in each category across all their posts to capture both typical and variable toxic behaviors. Additionally, we calculate a single toxicity score per user as the maximum toxicity observed across all their posts and categories, capturing their most extreme behavior. We skip posts in other languages, as toxicity scores of translated posts might not be meaningful.

Topics. We use the WebOrganizer classifier [41] to assign each post to one of 24 predefined topics supported by the model. For each user, we then compute a topic entropy score to quantify content diversity. A low score indicates that a user focuses on only a few topics, while a high score indicates that a user posts across a broad range of topics.

4 Mapping the Ecosystem (RQ1)

We start by mapping wider blocking patterns, focusing on the scale and uptake of both individual blocks and community blocklists. First, we look at the extent of blocking activity over time (§4.1). We then inspect each type from the perspective of both individual users (§4.2) and blocklists themselves (§4.3). Finally, we compare these mechanisms to the official Bluesky account takedowns (§4.4).

4.1 Blocking over time

Figure 1 shows the temporal evolution of moderation activity on Bluesky, including individual block events, blocklist block volume, and official takedowns (performed by Bluesky PBC), plotted on a logarithmic scale. The blocklist volume is represented as the number of created blocking-blocked relationships (*i.e.* a single subscription to a list with 100 members, adds 100 to the volume).

Blocking driven by blocklist membership consistently dominates in volume, exceeding both individual blocking and labeler takedowns by several orders of magnitude. Individual blocking remains steadier but trends upward over time, following the expanding user base. Official takedowns by Bluesky, while far fewer in absolute

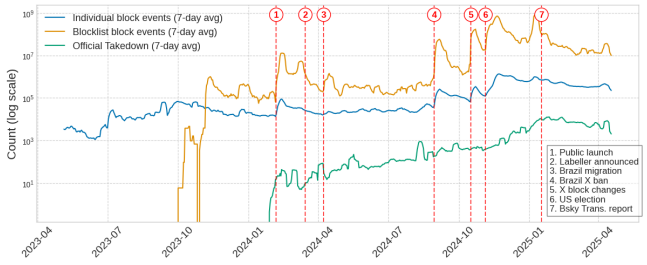


Figure 1: Blocking activity over time.

number, mirror similar temporal patterns, implying that centralized and decentralized moderation systems respond to the same stimuli. All types of blocking experience multiple spikes corresponding to major platform or geopolitical events. ① reflects early experimentation and community-led moderation efforts. Later surges align with high-profile migration waves (*e.g.* the Brazilian user influx ③ and subsequent ban of X in Brazil ④), which likely triggered the creation of new lists to manage sudden exposure to unfamiliar or contentious content. The most pronounced peaks occur in late 2024 and early 2025, coinciding with the US election ⑥ and its aftermath, highlighting how real-world polarization translates into intensified moderation behaviors on the platform. We suspect that the blocklist usage spike in January 2025 ⑦ links to a Bluesky moderation transparency report [9] that made many users aware of this feature.

4.2 Individual Blocking

We next inspect the use of blocking functionality by individual users (*i.e.* excluding the community-driven blocklists).

Blockers. Reflecting our prior observations, individual blocking is widespread among the Bluesky userbase. In total, there are 119,054,477 individual (*i.e.*, single-user) block records. Of these, the vast majority (98.4 %) are unidirectional, *i.e.* the block is not reciprocated by the other user. Indeed, 12.6 % of all users block at least one other user. Of these, the average user blocks 2 other users, with the 90th and 95th percentiles blocking 35 and 80 users individually, respectively. This suggests a clear skew, whereby the majority of blocks are issued by a small percentage of more active "blockers". The most active user blocks a remarkable 1,990,235 individual users. Surprisingly, this user has no apparent activity (*e.g.* 3 test posts) but is blocked by 23,489 other users, suggesting reciprocal activity.

Blocked Users. On the other hand, 27.6 % of users are blocked via individual blocks, twice the number of users who perform blocks themselves. This has a significant reach on the overall content ecosystem, as these individually blocked users wrote 90.2 % of all Bluesky posts. The mechanism creates 9,464,480 blocking-blocked relationships, shadowing 0.7 % of existing follow edges.

We also notice a skewed distribution, whereby certain users are regularly blocked. Whereas the median blocked user is blocked by just 1 user, the 90th percentile is blocked by 15 users, and the 95th percentile blocked by 35 users. The most-individually-blocked account is a US journalist known for controversial writings about transgender issues. This account is blocked by a remarkable 78,655 individual users (contrasting with just 9.5K followers). Many users have framed the account's arrival as a threat to the culture and

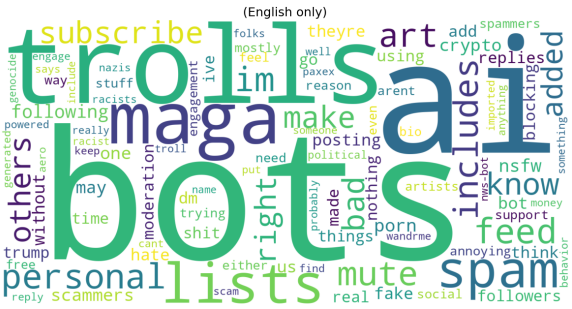


Figure 2: Word cloud of blocklist descriptions.

safety of Bluesky, particularly since many early adopters were from marginalized groups seeking refuge from harassment [31, 34].

4.3 List-based Blocking

The above considers users who individually block others. Next, we consider blocking based on the public *blocklists*.

Scale of Blocklists. We discover a total of 39,589 dedicated blocklists. Many of these seem experimental though: 25.3 % are empty (*i.e.*, without any listed users), while 56.4 % have no subscribers. The largest list includes 229,045 accounts, while the most popular one has 379,036 subscribers. We next focus on the lists with at least one subscription, which form our dataset of "active" blocklists. Collectively, those blocklists target 8.78 % of users, while only 2.52 % subscribe to them. These numbers are ≈ 3× lower than the ones of individually blocked and individually blocking, respectively. However, as flagged in §4.1, lists can block many accounts and be subscribed to by many users, amplifying their reach. In total, these blocks impact the visibility of 66.5 % posts. Moreover, blocklists are mutable: users added to a list after one subscribes are automatically blocked without oversight from subscribers. Blocklist subscriptions create 16,089,872,247 unique block edges.

Blocklist Ownership. Control over the blocklists is concentrated among a surprisingly small number of individuals. The active blocklists were created by 0.043% of all users — a number ≈ 58 times smaller than the number of subscribers. This raises concerns about re-centralization, further exacerbated by the skewed distribution of blocklist popularity. While an average list has 152 subscribers, the top-100 most popular ones cover 66.3% of subscribers and are created by only 79 individuals. An error by a single creator can thus have a significant effect, and our anecdotal evidence suggests that such errors do happen. For instance, 31.7 % of blocklisted users have never posted, with some of them included on lists dedicated to spam prevention. Reports and discussions from the community further indicate that erroneous inclusions are not uncommon, with users sharing experiences of being mistakenly blocklisted despite limited or benign activity [7, 8, 40, 42]. These findings highlight the need for greater oversight and transparency in blocklist governance.

Blocklist Themes. To gain an understanding of the themes covered by the blocklists, we investigate their self-descriptions, summarized as a word cloud in Figure 2. This reveals certain thematic foci such as MAGA and crypto, as well as more general issues such as spam and bots. We identify three recurring themes: behavioral moderation, political or ideological filtering, and content-based filtering. The most frequent terms include "trolls," "bots," and "spam," indicating that many lists are explicitly framed as defenses against inauthentic or abusive behavior. Political identifiers such as "MAGA" and "right" appear prominently, showing that blocklists are also employed to draw boundaries around ideological communities. Finally, terms like "porn/nsfw", "crypto", and "AI art" suggest that some lists function as content filters, excluding entire categories of material regardless of user behavior. Smaller but visible words such as "personal" and "mute" also reveal that a subset of lists is intended for individual curation rather than broad adoption.

We employ *ChatGPT 5* to classify each list into those three categories. We provide details on the process in Appendix A.1. The largest share of 23.4% of lists focus on behavioral moderation, followed by 8.3% for content-based filtering, and 7.7% for political or ideological filtering. The remaining unclear cases primarily arise from a lack of informative descriptions or titles rather than the need for additional categories. This suggests blocklists are not purely technical moderation tools, but encode a mix of personal preferences, political stances, and community norms.

Top Blocklists. Given the skew in popularity of the blocklists, we manually investigate the top 15 largest ones, as well as the top 15 most influential according to the number of subscribers. Collectively, these top lists cover 70% of all subscribers and 21.32% of all accounts added to blocklists. The largest lists appear to be mostly automatically generated and often lack a name and description. They focus mostly on inauthentic behavior (*e.g.* "Suspected Inauthentic Behavior") or content-based filtering (*e.g.* "nsfw"). Furthermore, their impact is limited as their total number of subscribers (49,559) is far larger than the total number of included accounts (1,605,217).

Among the most influential blocklists, the first and most prominent theme is political and ideological filtering, with 6 lists having a total of 721,792 subscribers, representing all subscriptions. These include "MAGA", "Elon Musk" or "Far right accounts".The second theme is content-based filtering with a total of 230,107 subscribers. These include 4 lists dedicated to blocking AI-generated and cryptocurrency/NFT content and Nazi symbolism. Within lists focused on combating inauthentic behavior, we notice 2 lists for content scrapers and 1 for spammers, with a total of 215,746 subscribers.

Blocklist Similarity. Finally, we hypothesize that the decentralized management of these blocklists may result in redundant efforts, where similar lists are created unnecessarily by independent parties. To quantify the relationship between blocklists, we compute the pairwise Jaccard similarity based on their membership for all 140 million unique pairs of the 16,730 blocklists with at least one subscriber and one listed user (member) in our dataset. In contrast to our hypothesis, the median Jaccard similarity is just 0.0, and 99.63 % of all pairs have a low similarity score between 0.0 and 0.01. This indicates that significant overlap between lists is a rare event and that most user-driven moderation is independent. Because most blocklists are small (relative to the global userbase), it is expected that the vast majority of pairs share no members.

Despite this overall sparsity, a small subset of blocklist pairs exhibits substantial similarity. We identify 682 pairs with ≥50% overlap, 278 pairs with ≥70% overlap, and 120 pairs with ≥90% overlap. 82 pairs overlap by 100%, *i.e.* they contain exactly the

Table 1: Average user statistics by moderation group. Values show mean counts per user and the relative increase compared to non-blocked users.

Metric	Not Blocked	Individually Blocked	On a Blocklist
Likes received	5.26	117.50 (**22.3**×)	2367.75 (**450.2**×)
Followers	7.72	64.93 (**8.4**×)	594.78 (**77.1**×)
Posts	4.78	48.63 (**10.2**×)	348.27 (**72.9**×)
Likes given	24.02	264.04 (**11.0**×)	1874.98 (**78.1**×)

same members, 16 of which have identical creators. Interestingly, among these identical lists, only 1 pair has creators that are connected via follow relationships. These clusters suggest that certain blocklists are replicated, forming pockets of concentrated membership overlap. Focusing on 178 high-impact blocklists (each with ≥1,000 subscribers), the 95th percentile of intersection size is 34 users, while the maximum observed overlap reaches 37,649 users. Although most blocklists are small and largely idiosyncratic, this subset highlights a small core of lists that target the same users, suggesting replication or convergence in moderation practices.

4.4 Comparison with Official Takedowns

Finally, we compare community-based blocking to the official Bluesky account takedown actions. Bluesky PBC has taken down 950,404 (2.95%) accounts. The reasons are unclear, yet this likely reflects the most extreme violations of community guidelines. This is ≈ 10× fewer than individually blocked users and ≈ 3× fewer than blocklist members, indicating that community moderation is operating at a far greater scale than the platform's centralized enforcement.

We find that 473,059 accounts (49.8 % of all taken-down accounts) are individually blocked by at least one other user (before they are taken down). At the same time, 359,595 (37.8 %) taken-down accounts are included in at least one blocklist with at least one subscriber. Combining these, we find that 584,747 accounts (61.5 % of all taken-down accounts) are blocked via blocklists or individually. This indicates that community actions are often aligned with official enforcement and may serve as an early warning or complementary signal for problematic accounts. At the same time, a sizable fraction of takedowns were not captured by community blocking, highlighting that official and community mechanisms operate with partly complementary scopes.

5 Studying Blocked User Behaviors (RQ2)

We explore which behaviors correlate with an increased likelihood of being blocked. We start with popularity, activity, toxicity, and discussed topics. We then train a model on these features to better distinguish the characteristics of blocked vs. non-blocked users.

5.1 Characterisation of blocked users

We start by comparing users' characteristics across three groups: *(i)* those on blocklists, *(ii)* those blocked individually by at least one user, and *(iii)* users who have never been blocked.

Activity and Popularity. Table 1 shows that users on curated blocklists are by far the most active and most popular group, followed by individually blocked users, while non-blocked users are the least active and least popular. For instance, blocklisted users receive 450× more likes and post 73× more often than non-blocked users. Individually blocked users exhibit the same pattern at a smaller scale. This indicates that blocking disproportionately affects highly visible and engaged users, highlighting potentially polarizing public opinions.

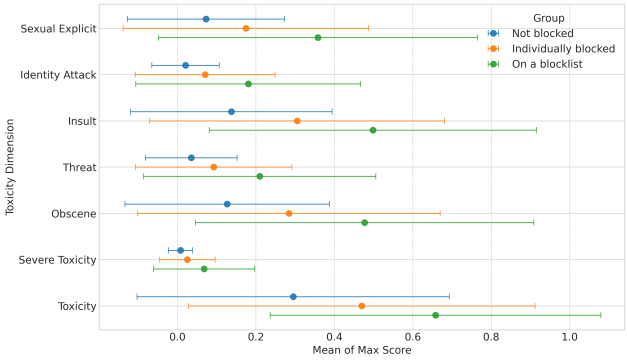


Figure 3: Mean of users' maximum toxicity scores.

This difference between the two blocking mechanisms can be partly explained by the effort involved in each type of moderation action. Adding someone to a curated blocklist requires deliberate action and often coordination among list maintainers, whereas individual blocking is quick and straightforward. As shown in §4.3, there are ≈ 3× fewer users on blocklists than those blocked individually, suggesting that blocklisted users represent the most prominent subset of accounts prone to being blocked.

Toxicity. We conjecture that many blocks may be imposed due to toxic behavior. Thus, we analyze the distribution of inferred toxicity scores across user populations using our Detoxify labels (see §3). Figure 3 compares the mean toxicity of posts across user groups for seven toxicity dimensions. The results show a clear gradient. Similar to popularity and activity, users on blocklists exhibit the highest mean toxicity across all dimensions, followed by individually blocked users, while unblocked users have the lowest scores. This pattern holds for both general measures (*e.g. toxicity* =0.66 vs. 0.47 vs. 0.29) and specific categories such as *obscene* (0.477 vs. 0.285 vs. 0.127) or *insult* (0.498 vs. 0.306 vs. 0.137). The large standard deviations for blocklisted users (e.g., 0.42 for *toxicity*) indicate substantial variability within this group, reflecting that not all blocklisted accounts are uniformly toxic. Individually blocked users, while less extreme, still show elevated toxicity compared to unblocked users across all metrics (*e.g.* 0.021 vs. 0.071 for identity attacks), suggesting that both individual and collective blocking mechanisms tend to target users who engage in more toxic or norm-violating behavior. The results imply that blocklists capture the most toxic users on the platform, but also encompass a heterogeneous mix of accounts whose inclusion extends beyond purely behavioral moderation.

Discussed topics. To explore whether the topics users post about are associated with being blocked, we analyze the topic distributions of posts from users appearing on blocklists and individually blocked by at least one user (see §3). We compare the results against the

overall platform baseline consisting of all posts classified by our topic model with a confidence score of ≥ 0.5.

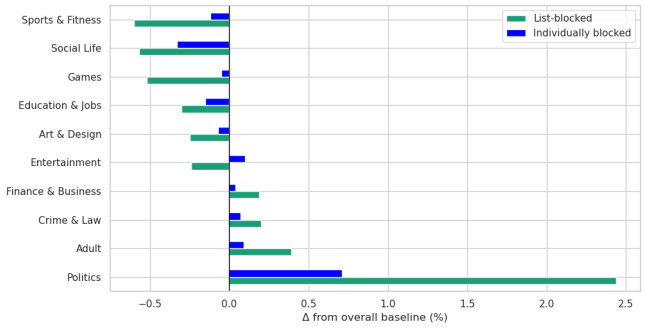


Figure 4: Topics coverage compared to non-blocked users.

Figure 4 presents the post distribution across topics for each group of users. We also present full results in Table 3 in the Appendix. We see that certain topics are over-represented in the content created by blocked users. Politics is the topic with the largest positive delta for both groups. Politics shows the largest positive deviation for both groups. For users on curated blocklists, political posts make up +2.44 percentage points (PPs) more than the baseline. Other consistently over-represented topics include Adult, Crime & Law, and Finance & Business. Conversely, topics related to general social interaction are under-represented. Posts about Sports & Fitness, Social Life, and Games are less frequent among blocked users compared to the baseline. Perhaps unsurprisingly, these findings indicate that blocking on Bluesky is correlated with the substance of the content users discuss.

Political Stance. We next examine how users' inferred political stance relates to moderation outcomes. We restrict our analysis to users with at least ten posts (4,692,323 users, corresponding to ≈ 15% of all users). Among these active users, the proportion of those who have been blocked is strikingly high: 76.7% of left-leaning and 75.6% of right-leaning users have been individually blocked, compared to 63.0% of neutral ones. Similarly, 41.3% of left-leaning and 39.5% of right-leaning users appear on curated blocklists, whereas only 26.8% of neutral users do. Expressing opinions, especially on politically charged topics, makes it likely that one will be blocked by someone. However, the rates are similar for both left- and right-leaning users, contrary to the popular perception of Bluesky as a predominantly left-leaning platform where right-wing voices are disproportionately excluded [29, 32].

Controlling for visibility. To ensure that toxicity, discussed topics, and political stance differences are not driven purely by visibility, we run 1:1 nearest-neighbor matching to never-blocked users. This yields a near-perfect balance on visibility, and the behavioral/content differences persist in the matched samples. We provide full details in Appendix B.1.

5.2 Predicting Blocking

The above highlights key differences between blocked and non-blocked users. To systematize this, we train a series of classifiers to predict blocking. We argue that if clear decision boundaries can be learned, it confirms the presence of systematic, non-random processes in moderation. Our goal is not to build a general-purpose prediction tool, but to use these classifiers as an interpretive method to decode and compare the specific "moderation philosophies" of distinct blocklists. We undertake this task via two stages. First, we train a general classifier to predict whether a user is blocked by *any* community moderation mechanism. Second, we train a series of specialized classifiers to predict membership on specific blocklists.

Methodology. We frame the task as a binary classification problem and train an XGBoost classifier [14]. XGBoost is a state-of-the-art framework widely used due to its performance, scalability, and interpretability with SHAP [28]. The input for all our models is the full set of 88 features. These include users' activity and popularity, topic profile, inferred toxicity, political stance, and labels, including self-applied as well as issued by labelers for the user account and posts. We detail the full feature list in Table 4 in the Appendix. For the general classifier, the target variable is a binary flag, indicating whether a user has been blocked by at least one other individual. Out of the 32.1M users in our dataset, 8.6 million fall into this positive class. To address class imbalance, we create an equally balanced dataset of users via random undersampling. This dataset is then split into a training set (80%) and a test set (20%), using stratification to ensure both partitions maintain a 50/50 class balance. For the per-blocklist analysis, we select the top 500 lists by subscribers and the top 500 by members. Due to some overlap, we end up with 858 distinct lists. These 858 lists cover 84.23% of active blocklist members and 94.85% of subscribers. For each of them, we train a dedicated classifier where the target is membership on that specific list, again using a balanced dataset created for each task.

General classifier. Our general classifier achieves a reasonably strong performance, with an accuracy of 73.8%, AUC of 0.8014, and F1-score of 0.7108, suggesting that user-level characteristics indeed correlate with moderation outcomes. To better understand which features drive these predictions, we perform a post-hoc SHAP analysis to interpret the model's decisions.

Figure 5 shows that the most influential features correspond to user popularity and activity, such as the number of followers, received likes, and posts. High values of these features strongly push the prediction toward the blocked class, while factors more directly related to moderation — such as average toxicity or the presence of labeler-based annotations — play a much smaller role. This imbalance reveals a critical limitation. The classifier learns to approximate blocking as a social visibility phenomenon rather than a behavioral one. In other words, it implicitly encodes the community's tendency to block prominent or active users, rather than identifying genuinely harmful content or interactions.

This highlights the difficulty of designing global moderation models that treat all users uniformly. Because blocking patterns are shaped by heterogeneous community norms, such centralized approaches risk reinforcing popularity or visibility biases instead of addressing context-dependent harms.

Per-Blocklist Classifier. While the general model reveals broad patterns, we hypothesize that individual blocklists operate on more specific criteria. Figure 6 compares the predictive performance of the general blocking classifier with the per-blocklist models. We observe that the per-blocklist classifiers perform substantially better,

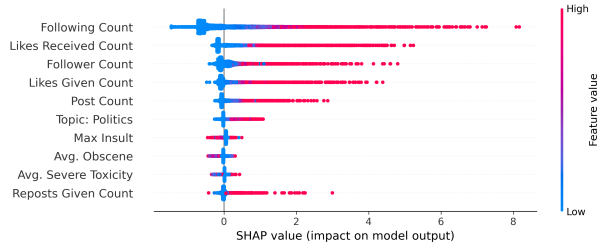


Figure 5: SHAP summary plot for the general blocking classifier, showing its 10 most influential features.

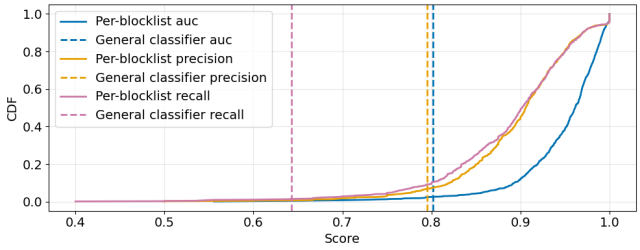


Figure 6: General vs. per blocklist classifier performance.

with > 90% outperforming the general one and many approaching perfect discrimination (*e.g.* 50% with AUC > 0.95). This indicates that lists have well-defined and distinct criteria, meaning that a one-size-fits-all moderation model would be inappropriate in practice. Importantly, these results confirm that it is possible for per-blocklist models to learn clear and reproducible decision boundaries.

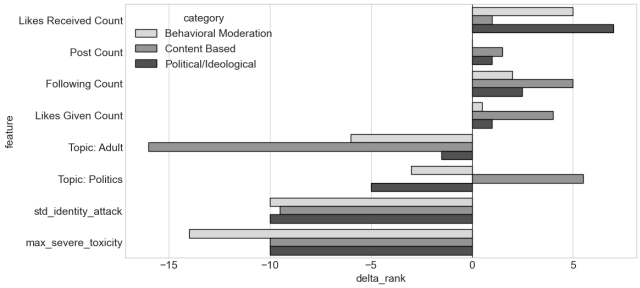


Figure 7: Feature rank difference relative to the General Classifier (negative = more important, positive = less important).

Finally, we assess the impact of individual features on the blocklist-specific classifiers. Figure 7 shows the relative rank of selected features compared to the general classifier. We group the results by blocklist categories, as defined in §4.3. Compared to the general model, blocklist-specific classifiers rely far less on popularity and activity metrics. Conversely, discussed topics (*e.g.* political content for political or ideological filtering) and toxicity dimensions (*e.g. severe toxicity* scores for behavioral moderation) play a much greater role in these bespoke per-list models. This further confirms that the criteria captured by blocklist-specific classifiers are better aligned with addressing genuinely harmful content or interactions.

6 Measuring Causal Impact (RQ3)

We assess the impact that different types of blocking.

6.1 Methodology

To estimate the causal effect of blocking intensity, we use the *Generalized Propensity Score (GPS)* [21, 23] to recover a function linking blocking "dose" to subsequent outcomes. For each user, we define a reference time t_0 as the first day they reach a given dose threshold and collect all pre-t_0 covariates. For individual blocking, dose is the number of distinct users who have blocked the user. For blocklists, dose is the total number of subscriptions across all blocklists that include the user. We use a 28-day window (two weeks before/after t_0) and define outcomes as four-week changes in activity (posts, outgoing likes, outgoing follows), popularity (incoming likes and incoming follows), and toxicity (sum of post toxicity scores): $\Delta Y = Y_{t_0+28} - Y_{t_0}$. For visualization, we stratify doses into low/medium/high bands (<1k, 1k–3k, >3k) and balance treated samples across bands. However, all estimation treats dose as continuous and fits the GPS and dose-response on raw dose values. Bluesky does not notify users of blocks or blocklist inclusion and awareness is not observable. We thus interpret effects as behavioral responses to the treatment itself (*e.g.* reduced exposure), noting that awareness may still arise via third-party tools (*e.g.* ClearSky [6]) or active checking. We provide our full methodology in Appendix A.4.

6.2 Results

Figure 8 shows the dose–response functions (DRFs) for activity (posts), toxicity and popularity indicators (received follows and likes). We show additional indicators in §C.2 in the Appendix. We truncate the blocklist (individual blocking) curves at dose ≈5,200 (≈7,800) as there are too few users at higher doses to support reliable estimates. Each curve plots the change over the following four weeks, $\Delta Y = Y_{t_0+28} - Y_{t_0}$, as a function of the dose observed at t_0. A value of $\Delta Y = 50$ on the posts means the user published 50 more posts by t_0+28 than at t_0 (*i.e.* a 28-day change, not a per-day rate). Shaded bands are 90% confidence intervals obtained by repeatedly re-running the analysis on many random samples of users.

We observe consistently higher levels of activity, toxicity and popularity among individually blocked users. At first glance, this is surprising given that, in §5.1, we found that blocklisted users generally score higher in all those metrics. However, this stems from the fact that the blocking dose is not equivalent across the two mechanisms. As discussed in §4.3, blocklists generate a large number of blocking edges, making it much easier for users to accumulate higher doses through list-based moderation than through individual blocking. Consequently, users who reach comparable doses via individual blocking represent a smaller, more active, and more visible subset of the population (those who are directly and repeatedly blocked by others).

We also observe an increase in user activity and, to a lesser extent, cumulative toxicity following blocking. The effect is initially more pronounced, but plateaus for medium and high doses. This could be explained by psychological reactance and the Streisand effect. If users perceive that they have been censored, they may respond by amplifying their engagement to reaffirm agency, seek visibility, or demonstrate resilience. Empirical studies show that

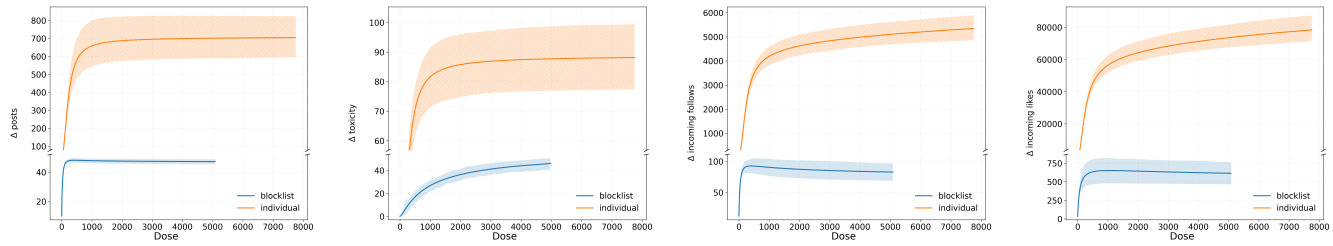


Figure 8: Dose–response functions (DRFs) for posts, toxicity, incoming follows, and incoming likes.

community-wide moderation interventions may reduce new inflows but sustain or even heighten core user activity and toxicity [13, 15, 39], and attempts to restrict expression can provoke counter-responses through reactance [24, 36].

Users' apparent popularity is affected in a similar way (more followers, more likes). This suggests that blocking itself can have a visibility-amplifying effect: once a user becomes the subject of blocks, they are likely also the focus of attention. Blocking can also activate the blocked user's community of supporters. For instance, when politically aligned or ideologically cohesive groups (*e.g.* MAGA supporters) see one of their members being blocked, they may respond by following, liking, and engaging with the affected account as an act of solidarity. At the same time, the popularity losses from blocking are minimal. Removing all edges shadowed by blocks reduces the total number of follow relationships by only 0.82 %, and the number of active users by 0.04 %, indicating that blocking, whether individual or list-based, removes relatively few existing connections. The users who block are typically not the ones who previously followed or interacted with the blocked user. For instance, just 7.3% of left-leaning and 9.1% of right-leaning users appear on blocklists curated by creators of the same political stance.

Taken together, these findings suggest that Bluesky's community-driven moderation allows users to shape their social environments without silencing others. By enabling individuals to curate their experience through blocking and blocklists, the platform preserves freedom of expression while reducing unwanted exposure.

7 Related Work

Moderation in Online Social Networks. Research on moderation largely focus on centralized social networks, analyzing platform-level strategies such as content removal, account suspension, and automated filtering [19]. These top-down approaches face well-documented challenges of scale, contextual nuance, and legitimacy, motivating growing interest in community-led forms of moderation [10]. Alongside platform governance, users engage in self-moderation through behaviors such as muting, unfollowing, and blocking. Blocking, in particular, serves as a protective mechanism against unwanted interaction [1, 17, 22]. At the same time, blocking is selective: invested users are more likely to block or unfollow politically dissimilar contacts or misinformation spreaders, which can exacerbate polarization [5, 26]. Blocking does not always reduce exposure to harmful content, since ties to misinformation spreaders often persist [2]. Our study extends this line of work by focusing on decentralized environments where blocking is not only a private action but also a mechanism of community governance. Bluesky, in particular, makes blocklists transparent, enabling collective moderation practices that are difficult to study on platforms where such actions are hidden [10].

Research on Bluesky. Bluesky has recently become the subject of studies that explore its growth, user behavior, and governance. Balduf *et al.* provide the first large-scale measurement of the network, analyzing user dynamics and moderation practices, and later show how starter packs bootstrap growth [3, 4]. Other work examines polarization, network topology, and the adoption of custom feed generators [35], as well as the dynamics of misinformation and toxicity [33]. At the level of individual interactions, Bono *et al.* analyze blocking behavior, finding correlations between being blocked and factors such as activity level, content characteristics, and network position [10]. Their study establishes blocking as a key mode of governance, but it treats blocking primarily as an individual-level action. By contrast, our work focuses on curated blocklists and community-applied labels, which represent a collective and transparent form of moderation. To our knowledge, ours is the first large-scale study to analyze these mechanisms.

8 Conclusion

This paper examined Bluesky's community moderation ecosystem. We showed that blocking is common (§4) and increasingly shapes content consumption, and that Bluesky's blocklists have enabled a small set of influential curators to exert outsized influence. While often useful, these dynamics risk outsourcing moderation to opaque community actors and may re-centralize control of the social graph, potentially conflicting with Bluesky's decentralization goals.

Our predictive models quantify regularities in who gets included on blocklists (§5), pointing to opportunities for improved support for both users and list operators (e.g., recommendations, error detection). At the same time, automation raises governance questions about oversight and accountability. This motivates more cautious human-in-the-loop tools, such as flagging unusually overlapping lists or helping users choose suitable blocklists.

Finally, our causal analysis identifies systematic behaviors before and after being blocked (§6). Being blocked may increase activity and intensify within-group engagement, suggesting that blocklist-based moderation has macro-level consequences for the wider social graph, not just individual feeds. These feedback loops can reduce cross-community exposure and strengthen cohesion—signals of segregation and echo chambers. We therefore see our results as a basis for moderation pipelines that consider downstream social effects of blocking actions, and future work will explore integrating such insights into Bluesky's community-guided tooling.

References

[1] Ahmed, S., Bee, A. W. T., Masood, M., and Tan, H. You have been blocked: exploring the psychological, personality, and cognitive traits of blocking misinformation sources on social media. *Telematics and Informatics* (2024).

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

[3] Balduf, L., Sokoto, S., Ascigil, O., Tyson, G., Scheuermann, B., Korczyński, M., Castro, I., and Król, M. Looking at the blue skies of bluesky. In *Proceedings of the 2024 ACM on Internet Measurement Conference* (2024), pp. 76–91.

[4] Balduf, L., Sokoto, S., Baronchelli, A., Castro, I., Król, M., Tyson, G., Pavlou, G., Scheuermann, B., and Ascigil, O. Bootstrapping social networks: Lessons from bluesky starter packs. In *Proceedings of the International AAAI Conference on Web and Social Media* (2025), vol. 19, pp. 178–192.

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

[6] BestskyTools. Clearsky. https://bestsky.tools/clearsky, 2026.

[7] Bluesky Community. No easy way to delete listblock records when list creator deletes/is suspended. https://github.com/bluesky-social/social-app/issues/4994, 2024.

[8] Bluesky Community. People are abusing block lists; targeted harassment. https://github.com/bluesky-social/social-app/issues/7076, 2024.

[9] Bluesky PBC. Bluesky moderation transparency report 2024. https://bsky.social/about/blog/12-jan-2025-moderation-report, 2025.

[10] Bono, C., Liu, N., Russo, G., and Pierri, F. Self-moderation in the decentralized era: decoding blocking behavior on bluesky. *arXiv preprint arXiv:2505.01174* (2025).

[11] Burnham, M., Kahn, K., Wang, R. Y., and Peng, R. X. Political debate: Efficient zero-shot and few-shot classifiers for political text. *arXiv preprint arXiv:2409.02078* (2024).

[12] Chandrasekharan, E., Gandhi, C., Mustelier, M. W., and Gilbert, E. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proc. ACM Hum.-Comput. Interact. 3*, CSCW (Nov. 2019).

[13] Chandrasekharan, E., Jhaver, S., Bruckman, A., and Gilbert, E. Quarantined! examining the effects of a community-wide moderation intervention on reddit. *ACM Transactions on Computer-Human Interaction (TOCHI) 29*, 4 (2022), 1–33.

[14] Chen, T., and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), pp. 785–794.

[15] Cima, L., Trujillo, A., Avvenuti, M., and Cresci, S. The great ban: Efficacy and unintended consequences of a massive deplatforming operation on reddit. In *Companion Publication of the 16th ACM Web Science Conference* (2024), pp. 85–93.

[16] Costa-Jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672* (2022).

[17] Fox, J., and Moreland, J. J. 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* (2015), 168–176.

[18] Gillespie, T. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[19] Gorwa, R., Binns, R., and Katzenbach, C. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society 7*, 1 (2020), 2053951719897945.

[20] Hanu, L., and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020. Accessed: October 5, 2025.

[21] Hirano, K., and Imbens, G. W. The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. 2004.

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

[23] Imai, K., and Van Dyk, D. A. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association 99*, 467 (2004), 854–866.

[24] Jansen, S. C. The streisand effect and censorship backfire. *International Journal of Communication 9* (2015), 656–671.

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

[26] Kaiser, J., Vaccari, C., and Chadwick, A. Partisan blocking: Biased responses to shared misinformation contribute to network polarization on social media. *Journal of Communication 72*, 2 (03 2022), 214–240.

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

[28] Lundberg, S. M., and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems 30* (2017).

[29] Magazine, T. Why bluesky is letting users write their own social media — bluesky still skews extremely left politically. *Time* (2025).

[30] Matias, J. N. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences 116*, 20 (2019), 9785–9789.

[31] Messman-Rucker, A. Controversy brews on bluesky after users complain about anti-trans podcaster. *The Advocate* (2024).

[32] Newsweek. Conservatives join bluesky, face abuse and censorship. *Newsweek* (2024).

[33] Nogara, G., Sahneh, E. S., DeVerna, M. R., Liu, N., Luceri, L., Menczer, F., Pierri, F., and Giordano, S. A longitudinal analysis of misinformation, polarization and toxicity on bluesky after its public launch. *arXiv preprint arXiv:2505.02317* (2025).

[34] Perez, S. Bluesky at a crossroads as users petition to ban jesse singal over anti-trans views, harassment. *TechCrunch* (2024).

[35] Quelle, D., and Bovet, A. Bluesky: Network topology, polarization, and algorithmic curation. *PloS one 20*, 2 (2025), e0318034.

[36] Rains, S. A. The nature of psychological reactance revisited: a meta-analytic review. *Human Communication Research 39*, 1 (01 2013), 47–73.

[37] Rosenbaum, P. R., and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 1 (1983), 41–55.

[38] Townsend, L., and Wallace, C. The ethics of using social media data in research: A new framework. In *The ethics of online research*. Emerald Publishing Limited, 2017, pp. 189–207.

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

[40] u/BlueskyUser. Psa: Beware of some blocking lists. https://www.reddit.com/r/BlueskySocial/comments/1gtsik9/psa_beware_of_some_blocking_lists/, 2024.

[41] Wettig, A., Lo, K., Min, S., Hajishirzi, H., Chen, D., and Soldaini, L. Organize the web: Constructing domains enhances pre-training data curation. *arXiv preprint arXiv:2502.10341* (2025).

[42] Writes, E. I've been on bluesky six months. here are my thoughts. *Medium* (2024).

A Additional Methodological Details

We provide additional details on our methodology.

A.1 Blocklist Classification Methodology

To identify the main themes of community blocklists, we used *ChatGPT 5* to classify 16,730 blocklists with at least one subscriber and one listed user (member), based on their available textual metadata into three categories: behavioral moderation (e.g., spam, bots, harassment), political or ideological filtering (e.g., MAGA, conservative, or left-wing), and content-based filtering (*e.g.* AI-generated or NSFW material). Each blocklist's name and description were pre-processed by lowercasing, removing punctuation and links, and concatenating both fields. We then applied a hybrid semantic–lexical approach that combined cosine similarity of sentence embeddings (from *paraphrase-mpnet-base-v2*) with keyword matching, assigning each list to the category with the highest combined score. Entries for which the highest combined similarity score across all categories was below 0.2 were labeled as *unclear*. A random sample of 100 results was manually inspected, achieving 85% accuracy.

A.2 Political Stance

Political stance is inferred using a zero-shot natural language inference (NLI) model [11], For each user, we construct a single document by concatenating available profile text (display name and description) with a random sample of up to 10 posts. The resulting text is truncated to a maximum length of 512 tokens, corresponding to the model's maximum supported input length.

Since the stance model operates in English, non-English posts are translated using the NLLB-200 model (facebook/nllb-200-distilled-600M) [16]. The translated document is evaluated against three hypotheses:

- "The author of this text supports politically left-leaning ideas."
- "The author of this text supports politically right-leaning ideas."
- "The author of this text expresses politically neutral ideas."

Multi-label scoring is enabled. A categorical stance label is assigned only if the highest-scoring hypothesis has probability ≥ 0.5 and is strictly greater than the other two. Users who do not meet this criterion are labeled as *Unclear*.

A.3 Validation

We manually validate the performance of the classification models used in the paper. We skip the process for toxicity and topic classification. Toxicity detection relies on the Detoxify model, which has already been evaluated on established toxicity benchmarks [20]. Topic classification follows the framework introduced by [41], which also provides large-scale evaluation and benchmarking of the underlying classifiers.

We thus focus on political stance because it relies on our custom prompt. We manually inspect the inferred stances of 100 randomly sampled users, yielding an overall classification accuracy of 76%. Looking into the misclassified cases, we observe that high-confidence predictions typically correspond to explicit political expressions in post content. When political cues are limited to profile descriptions rather than posts, the model tends to assign the *Neutral* label. As a result, the *Left* category is likely under-represented, while genuinely ambiguous cases are consistently captured by the *Unclear* class.

A.4 Measuring Causal Impact Methodology

To model the impact of blocking, we use the *Generalized Propensity Score (GPS)* [21, 23] — the continuous-treatment analog of the propensity score. This allows us to estimate the causal impact that the amount of blocking (aka the "dose") has on the blocked users' subsequent popularity and activity levels.

We define t_0 as the reference user-specific timestamp at which we *(i)* measure dose and *(ii)* take all pre-t_0 covariates; it is a common anchor for the analysis and represents the first day when the user reached a specific dose threshold. For individual blocking, the dose of a user corresponds to the number of distinct users who have blocked them. For blocklists, the dose is the total number of subscriptions across all blocklists that include the user.

We adopt a 28-day window similar to previous studies on social media [4] (2 weeks before, and 2 weeks after the block). Our causal outcomes are the four-week changes in platform activity (posts, outgoing likes, outgoing follows), popularity (incoming likes and incoming follows) and toxicity (sum of the toxicity scores of the posts): $\Delta Y = Y_{t_0+28} - Y_{t_0}$. To avoid the analysis being dominated by low-dose cases, we stratify the observed dose distribution into three bands using fixed thresholds: low (<1k), medium (1k–3k), and high (>3k). We chose 1k and 3k to guarantee a sufficient number of high-dose users and then sampled the same number of treated users from the medium and low bands as from the high band. Dose banding is used only for descriptive/visual purposes. All causal estimation treats dose as continuous. The GPS model and the dose–response function are fit using the raw numerical dose values, not the low/medium/high categories.

To account for pre-existing differences, we use pre-t_0 covariates capturing baseline popularity and activity: in/out follows, in-/out likes, posts, account age, network size, and toxicity. These variables control for differences in visibility and engagement that might otherwise confound the relationship between blocking and subsequent activity. However, drawing credible conclusions about how outcomes vary with blocking dose requires a baseline near dose ≈ 0—that is, a control group of unblocked users—whose pre-t_0 characteristics are comparable to those of users who were blocked. We use a propensity-score approach to select controls who are at-risk of being blocked. In particular, we estimate each unblocked user's probability of being blocked from pre-t_0 behavior and account features, then retain unblocked users whose probabilities are similar to those of treated users [37]. To verify that including at-risk controls yields well-balanced comparison groups, we compute standardized mean differences (SMDs) between treated and control users *before* and *after* propensity-score weighting. We estimate each user's probability of being blocked from pre-t_0 covariates and assign higher weights to controls who look likely to be blocked (*i.e.* at risk users) and downweight unlikely ones. This produces treated and control groups that are comparable on observed metrics.

We observe that after weighting, covariate balance improves substantially: for blocklists, the mean |SMD| is 0.0367 (max 0.106), and for individual blocking, the mean and max |SMD| are 0.0862 and 0.1049, respectively.

B Additional Results

We provide additional results not included in the main body of the paper.

B.1 Nearest Neighbor matching

Table 2: Differences between blocked users and non-blocked users.

Outcome (SMD, SD units)	Indv. Blocked	On Blocklist
Not matched on		
Avg. toxicity	0.108	0.155
Avg. insult	0.124	0.179
Avg. obscene	0.099	0.140
Avg. severe toxicity	0.073	0.113
Avg. identity attack	0.072	0.113
Avg. sexual explicit	0.096	0.162
Political topic share	0.163	0.224
Left stance probability	0.077	0.094
Right stance probability	0.060	0.096
Neutral stance probability	−0.055	−0.110
Matched treated users	405,569 (81.1%)	311,218 (62.2%)
Mean match distance	0.184	0.226

B.2 Full Topic Distributions by Moderation Group

Table 3 reports the full topic distributions for all posts and for blocked user groups, complementing Figure 4 in §5.1. We derive topic labels using the `WebOrganizer/TopicClassifier-NoURL` model [41

B.3 Classifiers Feature Set

Table 4 provides the full list of features used by our classifier to determine whether a specific user was blocked.

C Causal Impact

We provide additional material for our causal analysis

C.1 Covariate Balance

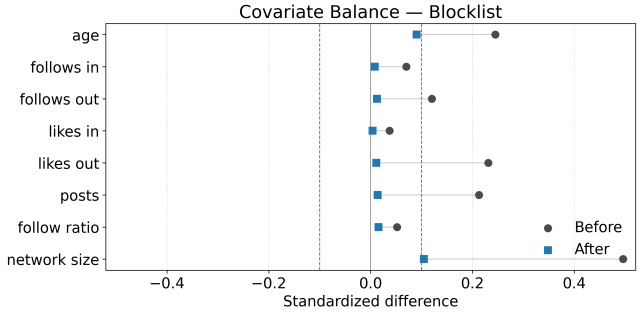


Figure 9: Covariate balance for blocklists.

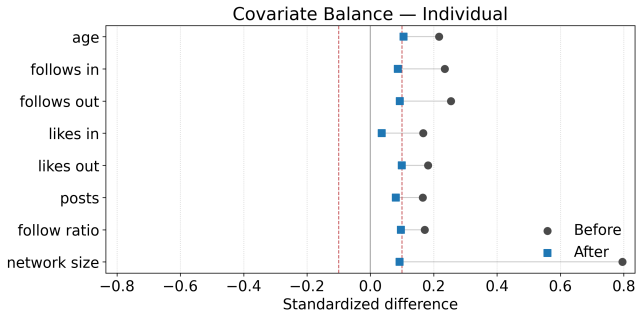
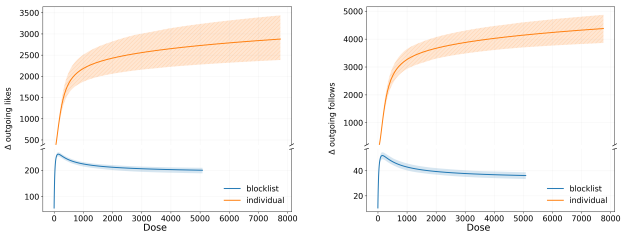


Figure 10: Covariate balance for individual blockings.

Figure 9 and Figure 10 report standardized mean differences (SMDs) for each covariate before and after matching control users with the treatment users. We include covariates measured at time $t0$, such as account age, number of posts, number of issued and received likes, number of followed and following accounts, incoming to outgoing follow ratio, and network size. The vertical dashed lines indicate the commonly used 0.1 threshold.

In both settings, the post–adjustment SMDs move substantially toward zero across all covariates, indicating strong improvement in balance relative to the pre–pre-adjustment values. Importantly, the majority of covariates fall within the ±0.10 region after adjustment, satisfying typical balance criteria used in the causal inference literature.



Outgoing likes **Outgoing follows**

Figure 11: Dose–response functions.

C.2 Remaining DRF Plots

Figure 11 provides additional DRF plots for our causal analysis. Outgoing likes and outgoing follows represent users' activity. We observe similar trends already discussed in Section 6 further confirming our conclusions.

D Ethics

Our work maps the landscape of blocklists in Bluesky, their effect on blocked users, and identifies potential signals that correlate with blocking. This work contributes to understanding the implications of opening and decentralizing social network platforms, particularly with respect to community-driven moderation. We believe that the benefits of this research significantly outweigh potential harms. We realize that our research poses limited ethical risks, which we discuss in the following.

We take multiple steps to minimise potential risks. Our data collection relies exclusively on information that is publicly available by design through the AT Protocol, including posts, moderation actions, and blocklist metadata. The data collection works as a regular client to the Bluesky Firehose, PDSes, and labelers, following the official documentation to backfill network data into a consistent snapshot. As such, our data collection likely has minimal to no impact on operators' infrastructure. All collected data are stored securely within access-controlled university infrastructure, and no external access is provided.

We follow established ethical guidelines for social data research [38] and make no attempt to infer real-world identities or to link activity across multiple accounts. We remove PII from our datasets prior to analysis, replacing the relevant fields with random integer IDs or dropping them if not required. Our analyses are conducted strictly at an aggregate level. We do not draw conclusions about individual users' intentions, awareness, or subjective experiences, nor do we make normative judgments about specific moderation decisions.

Table 3: Topic distribution for all vs. blocked users, split by positive/negative delta.

	All Posts	List-Blocked Users		Individually Blocked Users	
Topic	**% Overall**	**% of Group**	**Delta**	**% of Group**	**Delta**
Politics	11.85	14.29	+2.44	12.57	+0.71
Adult	2.42	2.81	+0.39	2.51	+0.09
Crime & Law	1.44	1.63	+0.20	1.50	+0.07
Finance & Business	2.23	2.41	+0.19	2.27	+0.04
Home & Hobbies	2.30	2.31	+0.01	2.33	+0.04
Transportation	1.40	1.41	+0.01	1.42	+0.02
History	0.49	0.52	+0.03	0.50	+0.01
Science & Tech.	1.70	1.62	-0.08	1.71	+0.01
Industrial	0.31	0.30	-0.01	0.31	-0.01
Hardware	0.86	0.83	-0.03	0.85	-0.01
Software Dev.	0.53	0.48	-0.05	0.52	-0.01
Fashion & Beauty	2.24	2.23	-0.01	2.22	-0.02
Religion	1.75	1.66	-0.09	1.72	-0.03
Games	4.66	4.14	-0.52	4.61	-0.05
Health	3.22	3.17	-0.05	3.17	-0.05
Food & Dining	3.50	3.45	-0.05	3.45	-0.05
Software	2.32	2.24	-0.08	2.26	-0.06
Travel	1.21	1.08	-0.13	1.15	-0.06
Art & Design	3.13	2.88	-0.25	3.06	-0.07
Literature	4.84	4.65	-0.19	4.77	-0.08
Sports & Fitness	6.03	5.43	-0.60	5.91	-0.12
Education & Jobs	2.52	2.22	-0.30	2.37	-0.15
Social Life	16.15	15.58	-0.57	15.82	-0.33
Entertainment	22.89	22.65	-0.24	22.99	+0.10

Table 4: Complete feature set for the blocking classifier.

Feature category	Description
	Behavioral features (6)
User activity count	Counts of posts, follows given, followers received, likes given, likes received, and reposts given.
	Content based features (62)
Topic profile	24 features representing the percentage of a user's posts in each topic (*e.g.* `topic_pct_Politics`), plus one `topic_entropy` feature.
Inferred toxicity	21 features representing the mean, standard deviation, and maximum of a user's posts across seven toxicity dimensions (*e.g.* `avg_insult`).
Self-applied labels	10 features counting the occurrences of the top 10 most common self-applied content labels (*e.g.* `selflabel_count_nudity`).
Political stance	6 features representing the probability scores for left, right, and neutral stances, and one-hot encoded categorical labels (*e.g.* `stance_LEFT`).
	Community features (20)
Community applied account labels	10 features representing one-hot encoded flags for the top 10 most common labels applied directly to user accounts (*e.g.* `has_account_label_bot`).
Community applied post labels	10 features counting the occurrences of the top 10 most common labels applied to a user's individual posts (*e.g.* `post_label_count_spam`).