Estimating the Effects of Blocking on Bluesky

Dhvani Shah*
SimPPL
Mumbai, India
Email: dhvanishah140304@gmail.com

Shristi Shetty*
SimPPL
Mumbai, India
Email: shristishetty67@gmail.com

Swapneel Mehta
SimPPL
Cambridge, United States
Email: swapneelsmehta@gmail.com

EXTENDED ABSTRACT

Safety, participation, and information sharing are foundational aspects of social media user experience and platform evolution. Platforms such as Twitter, Facebook, Reddit, and Bluesky, have implemented diverse control mechanisms—blocking, muting, reporting, feed customization to empower users to shape their online environments and moderate exposure to harassment, spam, and unwanted interactions [1], [2], [3]. Blocking is particularly notable: It is a direct, user-level intervention enabling individuals to restrict disruptive accounts, thereby altering the structure of personal feeds and the wider social graph. Research has consistently shown that such controls not only foster a sense of safety but can also influence engagement, retention, and content creation in the wake of negative experiences or increased visibility [4], [5], [6].

Despite the centrality of blocking to digital safety practices, relatively little empirical work has examined its behavioral aftermath. Most existing research focuses on why users block [4] or how moderation tools are perceived [1], but less is known about what happens to user activity after a block is issued. Does the act of blocking empower renewed participation, or does it signal disengagement? Are there measurable changes in how people communicate and share information? Addressing these questions is critical for designing moderation tools that not only mitigate harm but also promote sustained, healthy engagement.

A less-explored but equally critical element of platform dynamics is the sharing of external links—URLs leading to news, commentary, media, or resources outside the platform. Link sharing is essential for social media's role in information propagation, collective sensemaking, and networked public discourse [7], [8], [9], [10]. Understanding how safety interventions like blocking affect such link-sharing (and overall content composition) is necessary for evaluating platform health and the real-world consequences of moderation tools.

Bluesky is unique among social media platforms for making block relationships publicly observable via its API, offering unprecedented opportunities to analyze correlations between safety actions and engagement outcomes. In this study, we treat the act of blocking as a distinct, naturalistic intervention and examine how it shapes subsequent online engagement and information sharing. To ensure comparability between users who block and those who do not, we first employ Propensity Score Matching (PSM) to construct balanced groups based on observable covariates such as prior posting behavior, follower count, and engagement metrics. With these matched groups, we then apply Difference-in-Differences (DiD) and Interrupted Time Series (ITS) analyses to measure shifts in post volume, external link frequency, and thematic content before and after blocking events, controlling for likes, reposts, mentions, and other relevant factors. This combination of PSM with DiD and ITS strengthens causal inference by reducing confounding and isolating the effect of blocking itself.

A. Hypotheses

H1: Individuals will increase their posting activity after issuing a block, reflecting a regained sense of control and safety that encourages re-engagement on the platform [1], [2].

H2: Individuals will decrease the number of external links they share after issuing a block. Because link sharing often reflects persuasion, conflict, or signaling in argumentative interactions, we expect blocking to reduce the need for such outward facing behaviors [7], [3].

H3: Mentions will play a significant role in precipitating blocks. Users who experience higher rates of direct mentions, especially in contentious exchanges, are more likely to issue a block, which in turn shapes their subsequent participation [4], [6].

B. Results Overview

Our analyses reveal consistent and robust patterns. First, blocking events are followed by a significant increase in posting volume, suggesting that users regain confidence and willingness to engage after removing disruptive accounts. This supports H1 and reinforces prior findings that safety interventions restore agency [1]. Second, external link-sharing decreases after blocking, lending support to H2. This decline may reflect reduced involvement in argumentative or outward-facing interactions, and a shift toward more personal or community-focused content. Third, mentions strongly predict blocking events, validating H3. Users facing a surge in mentions particularly in heated exchanges are more likely to block, underscoring the role of interpersonal friction in triggering safety actions.

ITS analysis highlights sharp shifts immediately following blocks, while DiD estimates show statistically significant differences between blockers and non-blockers. Crucially, propensity score matching demonstrates that these results

hold even when comparing demographically and behaviorally similar users, providing stronger evidence that blocking itself, rather than pre-existing differences, drives the observed behavioral changes.

C. Detailed Results

1) Propensity Score Matching Balance: To ensure comparability between users who block and those who do not, we employed Propensity Score Matching (PSM) to create balanced groups based on observable covariates. Table I shows the covariate means for the control group (before matching), treatment group (blockers), and control group (after matching). The successful balancing indicates that our matched groups are comparable across key metrics, strengthening the causal interpretation of our findings.

TABLE I
COVARIATE MEANS FOR CONTROL (BEFORE), TREATMENT, AND CONTROL (AFTER) GROUPS

Metric	Control (Before)	Treatment	Control (After)
blocked_before_post	0.585	0.584	0.583
likes_count	119.94	747.02	498.50
repost_count	6.90	25.04	21.88
mention_count	1.64	9.85	1.97
follow_count	13.51	38.10	32.16

2) Posting Activity and Engagement Composite: Our Interrupted Time Series (ITS) analysis reveals a significant increase in posting activity following blocking events. Figure 1 illustrates this trend, showing a clear upward shift in post volume after the intervention point. This finding supports H1, suggesting that users experience a renewed sense of safety and control after blocking, which facilitates increased participation.

To provide a more comprehensive understanding of engagement changes, we developed an engagement composite variable that incorporates multiple metrics including reposts, follows, and likes. This composite measure showed similar positive trends after blocking events, indicating that the increase in activity extends beyond simple post counts to encompass broader forms of engagement. The consistency across these measures strengthens our conclusion that blocking behavior facilitates rather than inhibits platform participation.

3) Mentions and Blocking Relationship: Contrary to our initial hypothesis H3, our change point analysis revealed a complex relationship between mentions and blocking behavior. While we anticipated that mentions would directly correlate with blocking events, our analysis showed that change points in blocking time series did not consistently align with peaks in mention activity.

This finding suggests that while mentions may contribute to the decision to block, they are not the sole or primary driver. Other factors such as content toxicity, personal history with the account, or cumulative negative interactions likely play significant roles in triggering blocking behavior. The disconnect between mention spikes and blocking events indicates that users may tolerate a certain level of mention activity before resorting to blocking, or that blocking decisions are influenced by more nuanced factors beyond simple mention frequency.

4) LLM-Based Thematic Analysis of Content Sharing: We employed Large Language Model (LLM) techniques to analyze thematic shifts in content sharing following blocking events. Our analysis categorized posts into distinct themes including News, Social Media, Education, Health, Entertainment, E-Commerce, and Finance.

The results reveal significant thematic reorientation post-blocking:

News Content: Consistent decrease across all time windows, suggesting disengagement from current events and public discourse.

Social Media Content: Short-term increase (1-2 weeks) followed by normalization, indicating temporary focus on platform-related discussions.

Entertainment: Steady decline, reflecting reduced sharing of leisure-oriented content.

Education/Health: Remarkable stability, showing these topics remain independent of social conflicts triggering blocks.

Finance/E-Commerce: Mixed patterns with overall reduction in commercial content sharing.

These thematic shifts demonstrate that blocking influences not just posting volume but also content strategy. Users move away from potentially contentious topics toward community-focused or neutral content, creating curated communication patterns aligned with perceived audience safety.

5) LLM-Based Thematic Analysis of Content: We employed Large Language Model (LLM) techniques to perform thematic analysis of user content before and after blocking events. The LLM-based topic modeling categorized posts into distinct themes including News, Social Media, Education, Health, Entertainment, E-Commerce, and Finance to understand potential shifts in content preferences following moderation actions.

Our analysis revealed that blocking events did not produce significant changes in the overall thematic composition of user content. While minor fluctuations were observed across various categories, these changes were not statistically significant and did not indicate a substantial reorientation of content strategy post-blocking.

The stability in thematic patterns suggests that users maintain consistent content preferences and sharing behaviors regardless of blocking activities. This finding indicates that while blocking may affect quantitative engagement metrics, it does not substantially alter the fundamental nature or topics of content that users choose to share on the platform.

The minimal changes observed across content categories reinforce that blocking serves primarily as a relational moderation tool rather than a catalyst for content strategy transformation. Users continue to engage with their preferred topics and communities, suggesting that blocking enables them to maintain existing content patterns while managing unwanted interactions.

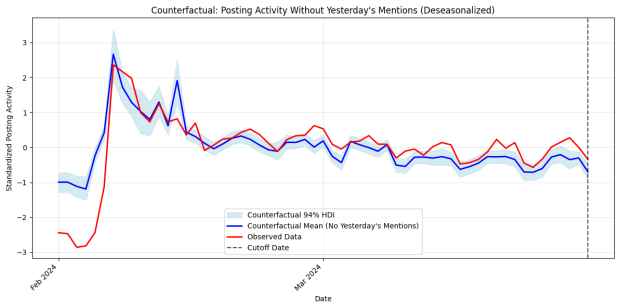

Fig. 1. Interrupted Time Series analysis of posting volume pre- and post-block. The vertical line indicates the blocking event, showing a clear increase in posting activity following the intervention.

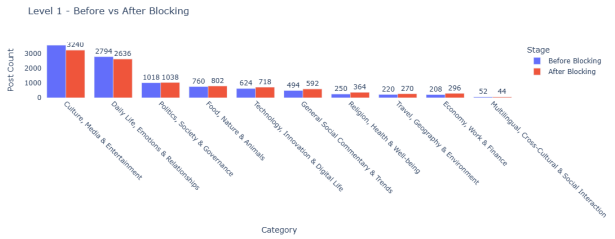

Fig. 2. Change in external link frequency before and after blocking. The plot shows a general downward trend in link sharing across most content themes following blocking events.

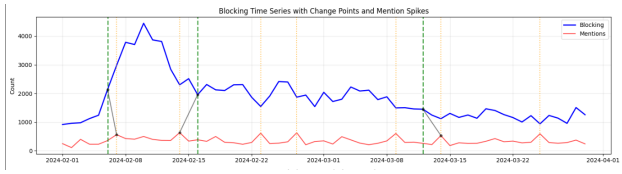

Fig. 3. Relationship between mentions and blocking events. The analysis shows that while mentions may contribute to blocking decisions, they are not the sole determinant, with change points in blocking not consistently aligning with mention peaks.

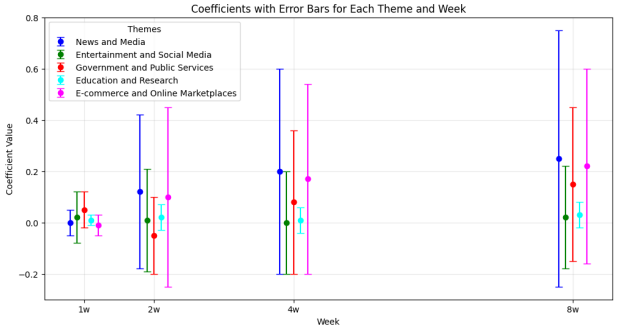

Fig. 4. Difference-in-Differences estimates of post-block engagement (with matched controls). The analysis shows significant differences in engagement patterns between users who block and matched controls who do not.

D. Figures

E. Conclusion

These findings suggest that blocking operates as both a protective and behavior-shaping mechanism: it enables users to re-engage in posting, reduces external-facing conflict behaviors, and is triggered by complex interactions rather than simple metrics like mention frequency. The theme-specific variations in response to blocking, revealed through our LLM-based topic modeling, indicate that users' content strategies evolve differently based on their topical focus and engagement patterns.

By combining ITS, DiD, propensity score matching, and advanced LLM-based thematic analysis, we provide robust evidence that blocking causally influences user behavior across multiple dimensions. The increase in posting activity coupled with decreased external-facing content suggests that blocking facilitates a shift toward more personal, community-focused engagement while reducing participation in broader, potentially contentious discourse.

The successful balancing through propensity score matching (Table I) strengthens our causal claims, demonstrating that the observed effects are not driven by pre-existing differences between users who block and those who don't. Beyond individual outcomes, these micro level adjustments may scale into macro level patterns that shape discourse and participation on decentralized platforms.

Our results underscore the importance of designing moderation tools that not only defend against harm but also promote resilience, participation, and healthier online communities. The nuanced relationship between blocking and content sharing across different themes further highlights the need for personalized moderation approaches that account for users' specific engagement patterns and content preferences.

REFERENCES

[1] S. Jhaver, I. Birman, E. Gilbert, and A. Bruckman, "Transparency in content moderation: Platforms, practices, and motivations," in *Proceedings of the ACM Conference on Computer-Supported Cooperative Work (CSCW)*. ACM, 2018, pp. 1–18.

[2] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech," in *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*. ACM, 2017, pp. 1–22.

[3] J. N. Matias, "Preventing harassment and increasing group participation through social norms in 2,190 online science discussions," in *Proceedings of the National Academy of Sciences*, vol. 116, no. 20, 2019, pp. 9785–9789.

[4] S. Jhaver, A. Bruckman, and E. Gilbert, "Didn't you know he was a troll? sharing personal stories on reddit's r/offmychest," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. ACM, 2018, pp. 1–12.

[5] G. Weld, A. X. Zhang, and M. S. Bernstein, "Predicting and mitigating user disengagement on social media after negative experiences," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. ACM, 2022, pp. 1–16.

[6] J. Majo, M. Whiting, and M. Bernstein, "Situated care: Designing online moderation as distributed work," in *Proceedings of the ACM Conference on Computer-Supported Cooperative Work (CSCW)*. ACM, 2021, pp. 1–27.

[7] H. Shwartz, N. Rosenfeld, and O. Tsur, "Knowledge and influence in social media networks: Link-sharing as information signaling," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, no. 1, 2020, pp. 623–633.

[8] Y. Tian and R. Lambiotte, "Unifying information propagation models on networks and influence maximization," *Physical Review E*, vol. 106, no. 3, p. 034316, 2022.
[9] S. S. Sundar, "Engagement with news content in online social networks," *Bellisario Media Lab, Penn State*, 2024.
[10] H. Voorveld, "Engagement with social media and social media advertising," *Journal of Advertising*, vol. 47, no. 1, pp. 38–54, 2018.